Ciencia Computacional y Finanzas

José L. Fernández Pérez

Catedrático de Análisis Matemático del Departamento de Matemáticas de la Universidad Autónoma de Madrid
Director Gerente de Consultoría de Riesgos e I+D de "Tecnología, Información y Finanzas" (Grupo Analistas)

Resumen

Los modelos matemáticos pretenden captar la estructura de las relaciones causales entre las características fundamentales de cierta realidad concreta. Su utilidad es funcional. Un buen modelo es aquel que permite analizar y diagnosticar esa realidad, pronosticar su evolución (sea ésta determinista, probabilista o caótica) y tomar decisiones sobre la interacción más conveniente o el curso de acción a seguir; una suerte de test de Turing: el modelo es bueno si funciona (y si nos ayuda a comprender una realidad).

La necesidad de ser capaces de computar con el modelo escogido ha limitado tradicionalmente la ambición de su diseño. Los modelos han sido lineales, continuos, aplicables sólo a cortos espacios de tiempo, o en condiciones ideales, etc., no porque se creyera que así era la realidad, sino porque con esa primera aproximación se podía calcular, y porque, a no dudarlo, han funcionado extraordinaria, e, incluso, sorprendentemente bien. No hay mejor ejemplo de esto que las leyes de gravitación de Newton: elegantes, profundas, sencillas, y de precisión asombrosa.

Pero, hoy en día, animados por la prodigiosa potencia computacional de que ahora disponemos, nos atrevemos a modelizar realidades cada vez más complejas, donde muchas ecuaciones prescriben el comportamiento simultáneo de muchas variables, donde se incorpora la retroalimentación de causas, con efectos no-lineales que combinan ingredientes aleatorios y caóticos, y donde observamos y analizamos evoluciones temporales de largo alcance. Una complejidad que rehuye formulación cerrada, y que sólo se puede abordar mediante la simulación del modelo en el ordenador. Sistemas biológicos, económicos o financieros, el clima, o la turbulencia, enmarcan el ámbito de estas cuestiones.

El ordenador constituye un verdadero laboratorio de realidades complejas: un instrumento que permite trasladar el conocimiento organizado, a través de ese software mental que son las matemáticas, y de los modelos que concibe, en una realidad virtual sobre la que podemos actuar inocuamente. Nos permite experimentar recetas de política económica, para luego escoger la más conveniente, sin un (inadmisible) proceso de prueba y error sobre economías reales. Permite diseñar completamente un avión como el Boeing 777 pasando directamente de su concepción en el ordenador a la fase de producción, sin túneles de viento, ni prototipos. O simular explosiones termonucleares de distantes estrellas, y también de bombas atómicas, sin agredir desiertos ni atolones. O analizar los efectos de políticas alternativas de gestión ambiental sobre un ecosistema sin alterarlo irremediablemente.

Esta potencia requiere control. La simulación de un modelo no puede ser una suerte de caja negra, porque queremos entender. La complejidad de las realidades, la ambición de los modelos y la repercusión de las decisiones que emanan de su análisis generan inestabilidad. Son muchos los ejemplos de fiascos derivados de una excesiva fe en la modelización y su

simulación, al fin y a la postre (pero esto es casi una tautología) por no haber profundizado en la comprensión que la computación aporta en cuanto a diagnóstico de relaciones causales en la realidad a estudio.

En la gestión financiera, y para la toma de decisiones que conlleva, se comenzaron a desarrollar modelos científicos hace tan sólo unas decenas de años. Se trataba de modelos, ¡cómo no!, estilizados. Pero los sistemas financieros son sistemas complejos, y ya se ha generalizado el uso de ambiciosos modelos estocásticos complejos que permiten simular la evolución aleatoria integrada del negocio, de la estructura financiera, de las condiciones macroeconómicas y de los resultados de estrategias de gestión alternativas. Se trata de modelos que facilitan un proceso de decisión que tiene en cuenta no sólo un escenario medio de referencia, sino la incertidumbre inherente y la ulterior gestión activa.

Bibliografía

D.N. Arnold: *Mathematics in Industry and Government*. Presentación, NSF VIGRE meeting, Reston (Virginia), 4 de mayo de 2002.
[Disponible en http://www.ima.umn.edu/~arnold/talks/industry.pdf].

N. Barberis, R. Thaler: *A survey of behavioral finance*. National Bureau of Economic Research, Documento de trabajo 9222, septiembre de 2002.
[Disponible en http://papers.nber.org/papers/w9222].

A.J.G. Cairns: *An Introduction to Stochastic Pension Plan Modelling*. Nota técnica 94/11. Workshop on Interest Rate Risk, Vancouver, 19-20 de agosto de 1994.
[Disponible en http://www.ma.hw.ac.uk/~andrewc/papers/ajgc6.pdf].

S.P. D'Arcy: Enterprise Risk Management. *Journal of Risk Management of Korea* **12**, no. 1 (por aparecer).
[Disponible en http://www.cba.uiuc.edu/~s-darcy/papers/erm.pdf].

R. Kaufmann: *Dynamic Financial Analysis*. Presentación, 4 de julio de 2000.
[Disponible en http://www.math.ethz.ch/~kaufmann/DFA/DFA.pdf].

A.J. McNeil: *Risk Management, An Overview*. Presentación, Swiss Banking School, 17-18 de septiembre de 2001.
[Disponible en http://www.math.ethz.ch/~mcneil/ftp/bankschool.pdf].

R.J. Shiller: Human behavior and the efficiency of the financial system. En J.B. Taylor and M. Woodford (editores): *Handbook of Macroeconomics, Vol. 1* (1999), pp. 1305-1340.
[Disponible en http://cowles.econ.yale.edu/P/cd/d11b/d1172.pdf].

Enterprise Risk Management: An Analytic Approach. Monografía de la Consultora Tillinghast-Towers Perrin, 2000.
[Disponible en http://www.tillinghast.com/tillinghast/publications/reports/Enterprise_Risk_Management_An_Analytic_Approach/erm2000.pdf].

En Internet

http://www.afi.es
InfoAnalistas
Portal del Grupo Analistas Financieros Internacionales.

http://www.grupoanalistas.com
Grupo Analistas
Información corporativa del Grupo Analistas Financieros Internacionales.

http://15aniversario.afi.es
XV aniversario Grupo Analistas

http://archives.math.utk.edu/topics/computationalScience.html
Computational Science
Archivo de artículos y enlaces en Ciencia Computacional.

Ciencia Computacional, Simulación, . . . y Finanzas

José Luis Fernández Pérez

***La Laguna*, 20 de marzo de 2003**

Matemáticas

Las matemáticas $nacen$ buscando abstraer

simetrías, formas, estructuras

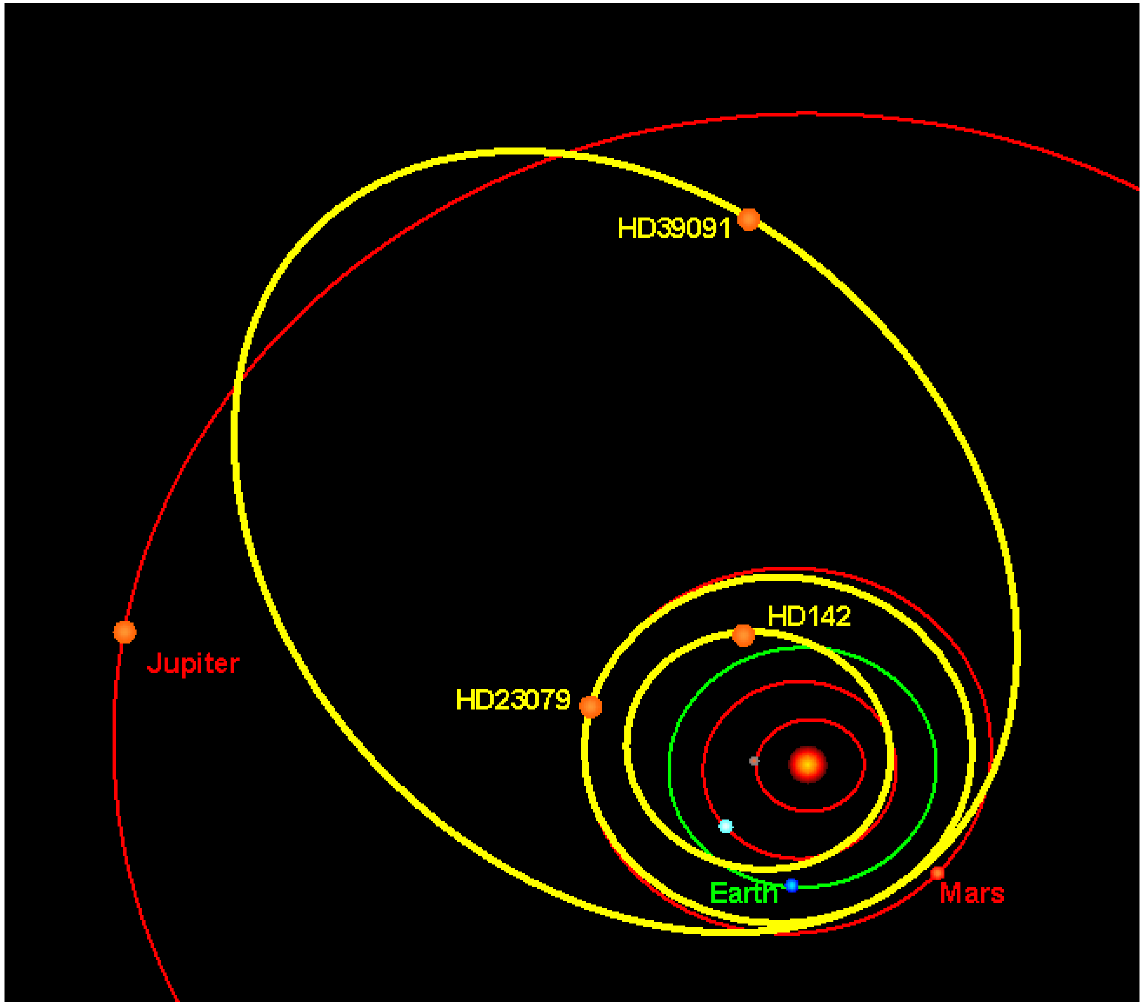
para entender el mundo.

Matemáticas: lenguaje. Matemáticas, software mental.

Simplificando . . . para entender

- Corto tiempo
- Lineal
- Pocos ingredientes
- Muy estilizados[1]
- Increíble que funcione. Un asombroso misterio.

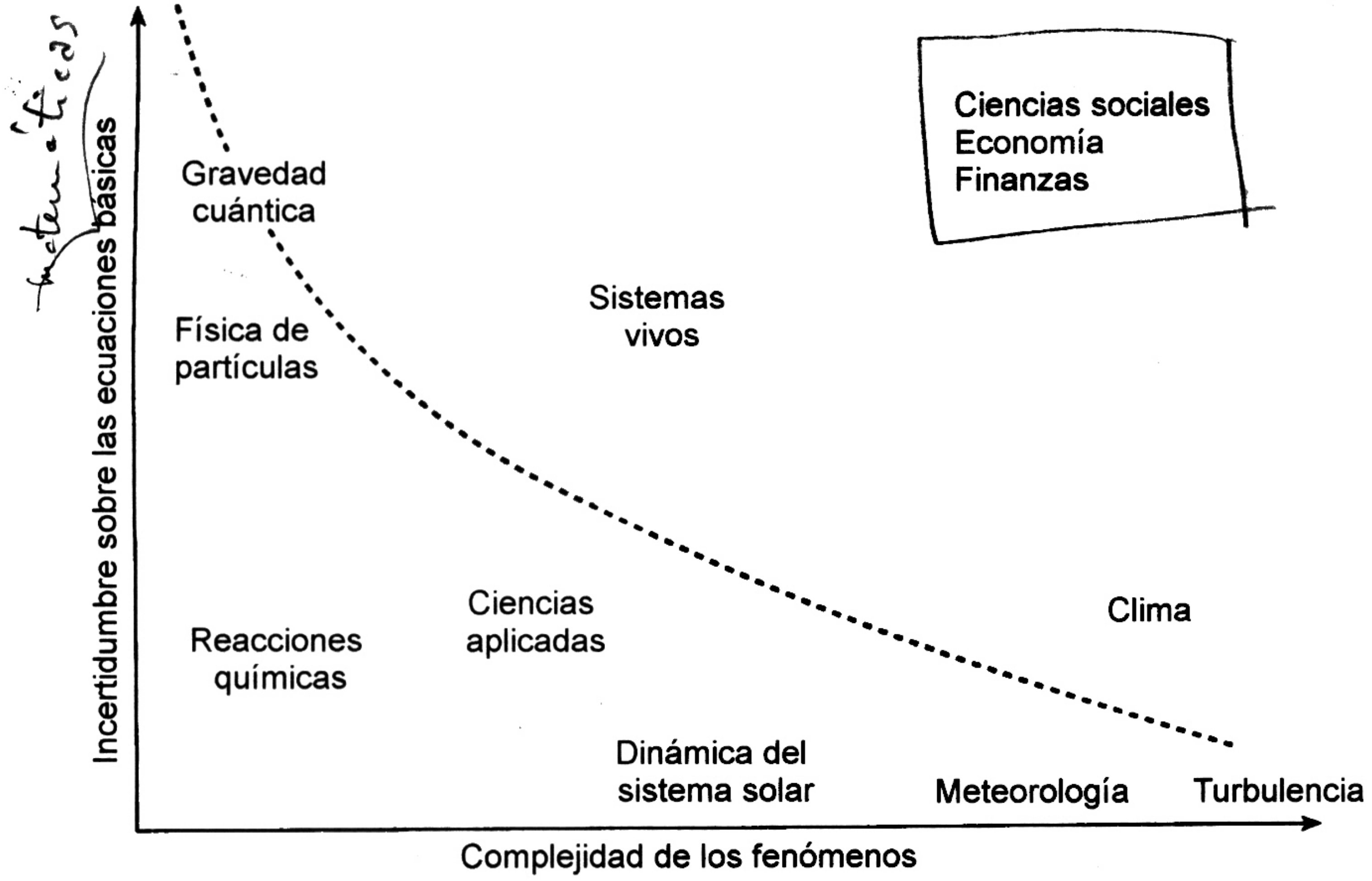
[1] Estilizar: [DRAE] Interpretar convencionalmente la forma de un objeto haciendo resaltar tan solo sus rasgos más característicos


HD39091
HD142
Jupiter
HD23079
Earth
Mars

La campiña inglesa

Sistemas Complejos

- Economía
- Biología
- Turbulencia
- Clima
- . . .


matemáticas
Incertidumbre sobre las ecuaciones básicas
Gravedad cuántica
Física de partículas
Reacciones químicas
Ciencias aplicadas
Sistemas vivos
Ciencias sociales
Economía
Finanzas
Clima
Dinámica del sistema solar
Meteorología
Turbulencia
Complejidad de los fenómenos

D. Ruelle

- Retroalimentación.
- No-linealidad y Caos.
- Dimensión, capas y cascadas.
- Aleatoriedad.
- . . .

- Regla de tres.

- Grandes Números.

- *Ceteris paribus*

Retroalimentación

> La mayor deficiencia de la raza humana es su incapacidad para comprender la función exponencial.
>
> *A. A. Bartlett*, *físico*.

- Botella de Coca-Cola y Ecología
- Interés continuo

Un niño que medía dieciocho meses en la escala de Richter.

. . . mínimos, gigantescos, qué más da:
después de todo, nadie sabe qué es lo pequeño y qué lo enorme . . .

José Hierro, *Libro de las alucinaciones*.

No-linealidad

Mecanismos perfectamente deterministas no lineales que actúan durante largo tiempo, suponen relación caótica (impredecible) ente el estado inicial del sistema y su estado futuro.

En lo caótico, lo determinista deviene en aleatorio.

Laplace. La mesa de Billar. Moneda al aire. Pascal, Cleopatra y las mariposas.

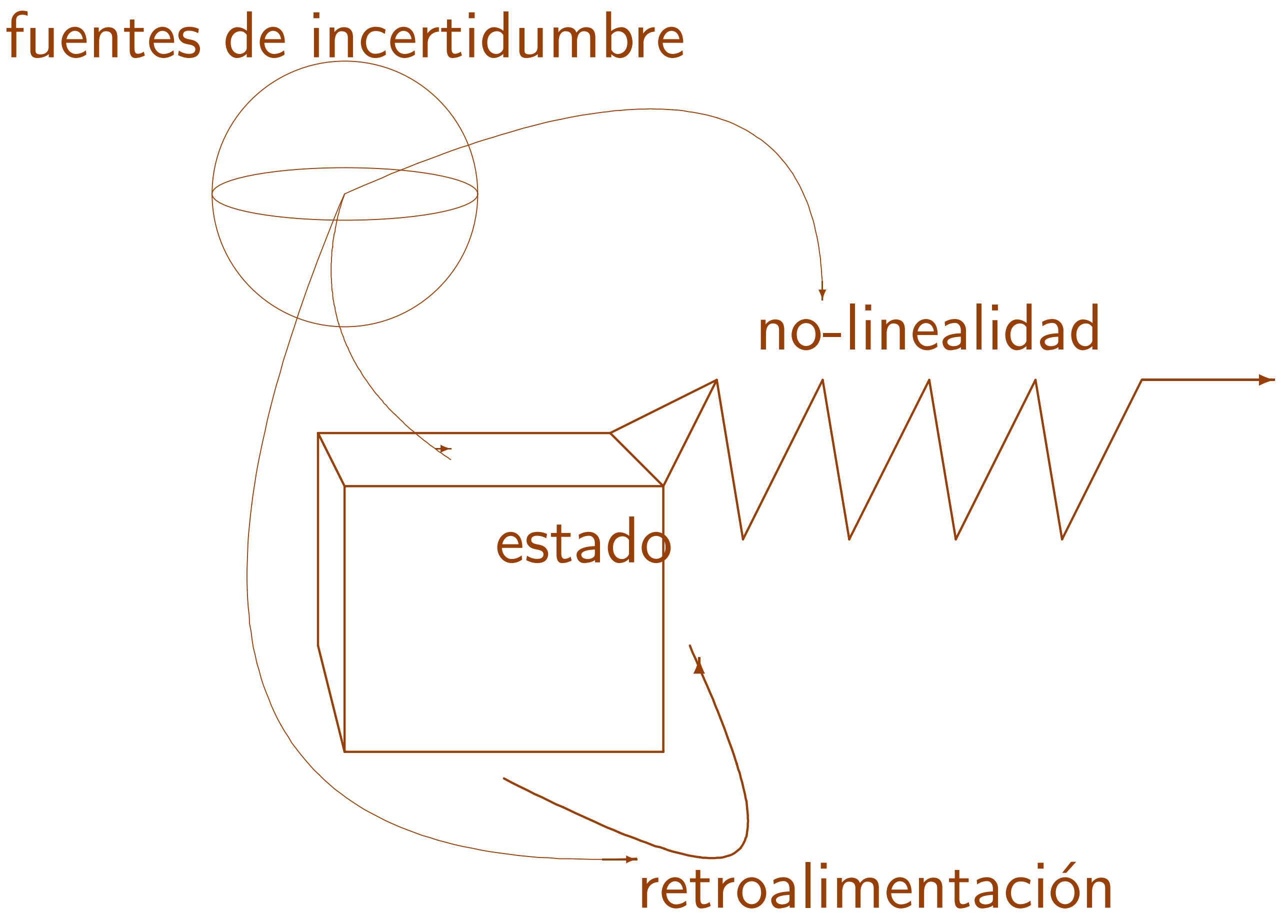
La dimensión

Sistemas grandes con un número enorme

- de variables,
- de sub-modelos,
- de capas intermedias de acción

que operan en sucesivas cascadas acumulando efectos.

Y, por supuesto: La aleatoriedad. La relación entre acción y efecto es aleatoria, intrínsecamente impredecible, a corto plazo y en acción directa.



fuentes de incertidumbre

no-linealidad

estado

retroalimentación

- En prosa: Una economía, un sistema financiero.
- En poesía: . . .

Dios, que es digital, creó la realidad analógica por la misma razón por la que nosotros hemos creado Internet: por enredar. De vez en cuando entraba en nuestro mundo como nosotros entramos ahora en la Red y disfrutaba viendo los días y las noches y el Sol y las tormentas. Y en cada una de esas incursiones, a la realidad atómica añadía alguna cosa nueva: los peces, las ranas, las serpientes, la polio, los instintos, la gripe . . . Todo ello sin calcular que la lógica de los átomos conduciría a la bomba atómica del mismo modo que la lógica digital conduce a la digitalina. Dios sólo es responsable de la puesta en marcha. Lo demás se dio por añadidura y Él fue el primero en extrañarse del modo singular que eligieron los mamíferos para reproducirse o las jirafas para llegar a la copa de los árboles. Cuando fabricas un calidoscopio, tampoco hay forma de predecir todas sus combinaciones posibles.

Con la misma extrañeza con que observaba Dios la realidad analógica, construida por Él mismo, nos asomamos ahora a la realidad virtual, hecha a nuestra imagen y semejanza. La hemos diseñado nosotros, sí, pero quién iba a imaginar que engendraría cosas tan curiosas por su cuenta. Y eso que aún estamos en el primer día de la creación como el que dice. Faltan los wap y los umts y la pantalla tridimensional, y los reptiles y las aves, y los Adanes y las Evas de ese mundo incipiente. Más que una realidad, hemos creado una lógica con capacidad para desarrollarse por sí misma, aunque la abandonáramos ahora mismo a su suerte. Dios tampoco necesitó crear los lunes ni los martes ni los miércoles... Desde el momento en que te inventas el domingo, el resto de la semana sale del huevo fecundado con cara de haberse confundido de estación.

Ahora bien, lo interesante de todo esto es el hecho de haber abierto en nuestra dimensión un agujero por el que podríamos ver el rostro de Dios, que quizá nos observa espantado por la misma abertura. No pierdan el tiempo buscándolo en dios punto com ni en satán punto es. Se trata de un hacker más experimentado que todo eso. Sepan en todo caso que, mientras navegamos, nos observa.

Juan José Millás, *Génesis*. EL PAIS

Matemáticas, *software mental*

- Simulación DRAE *Simular*. Representación de una cosa, fingiendo o imitando lo que no es. *Simulación.* Alteración aparente de la causa, la índole o el objeto verdadero de un acto o contrato.

- Modelización DRAE *Modelizar*. No'tá. *Modelo* . . . Esquema teórico, generalmente en forma matemática, de un sistema o de una realidad compleja, (por ejemplo, la evolución económica de un país), que se elabora para facilitar su comprensión y el estudio de su comportamiento.

En el *Seco*: *Simulador.* Aparato que permite reproducir artificialmente un fenómeno o un funcionamiento real. *Modelizar.* Establecer el modelo.

Los matemáticos no comprenden la realidad hasta que la encierran en una ecuación, pero los burócratas son incapaces de medir el tamaño de una catástrofe hasta que la transforman en un expediente.

Juan José Millás, *La oficina*. EL PAIS

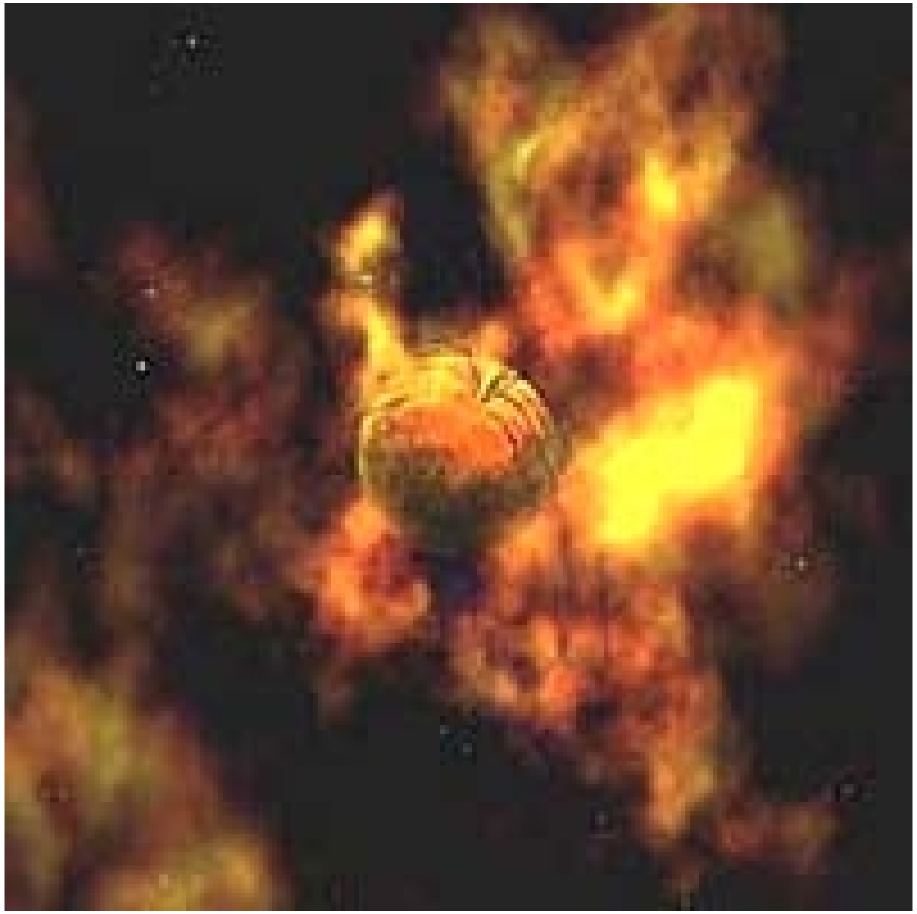
Simulación

- Laboratorio de lo complejo
  - Diagnóstico y análisis
  - Control
  - Comparación de acciones y respuestas
- Potencia computacional actual (futura!) que permite
  - modelizar ambiciosamente, sistemas complejos
  - *whatif's*, escenarios, optimización, análisis, decisión

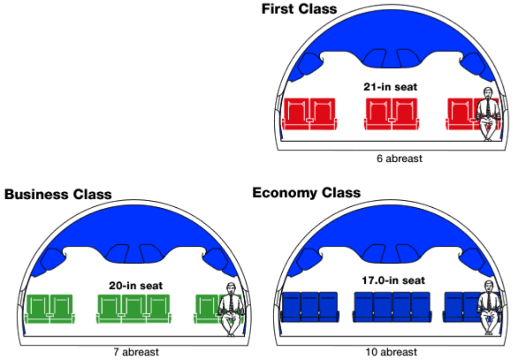
¡Un cambio de paradigma! (?)

- Arrecifes australianos

- Explosiones nucleares en estrellas

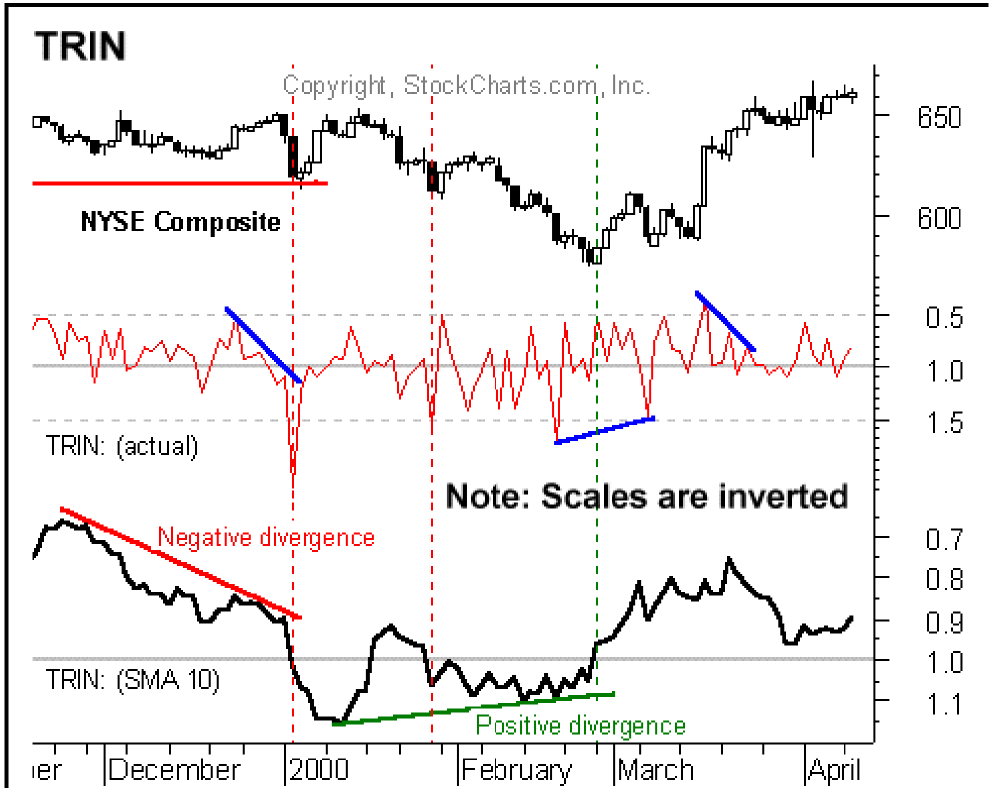


- Boeing777

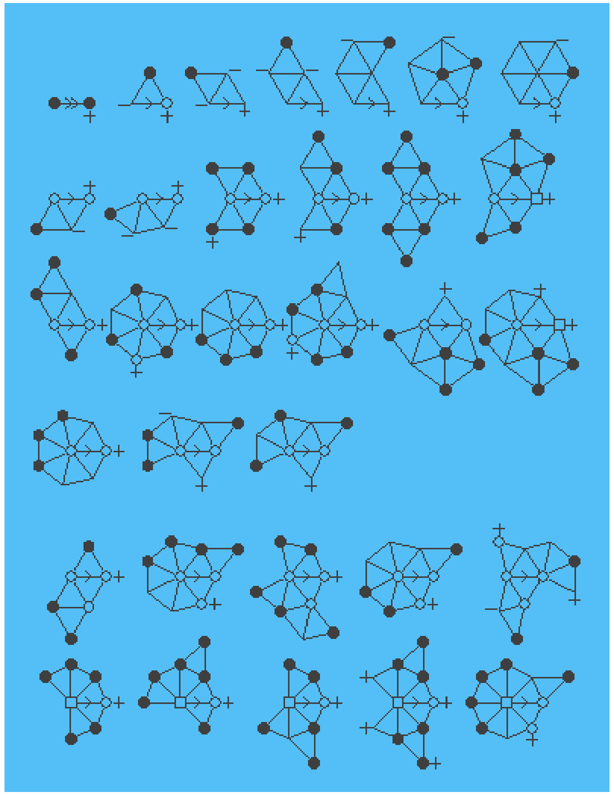

First Class
21-in seat
6 abreast
Business Class
20-in seat
7 abreast
Economy Class
17.0-in seat
10 abreast

- Plataformas petrolíferas en el Mar del Norte

- *Program trading*



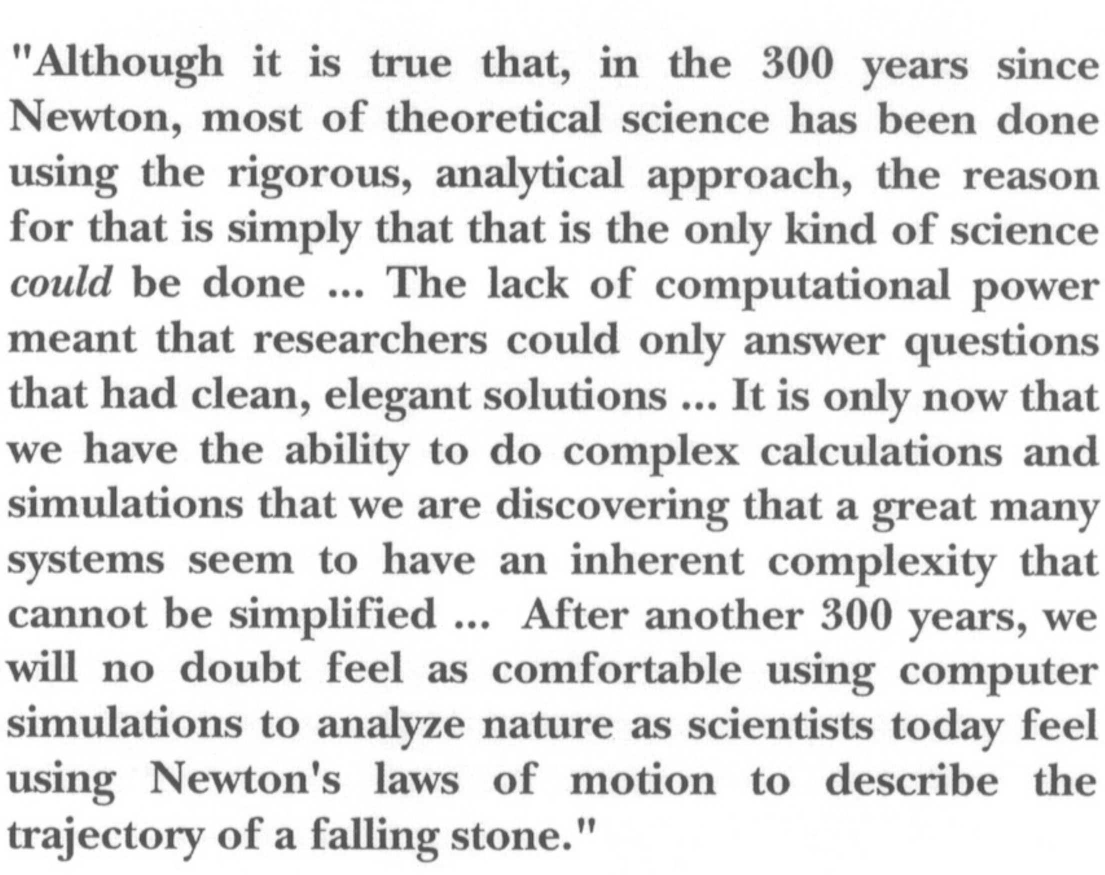
- Cuatro colores



- La paradoja de Braess

- Regulación y supervisión de mercados

Acotando con declaraciones



> "Although it is true that, in the 300 years since Newton, most of theoretical science has been done using the rigorous, analytical approach, the reason for that is simply that that is the only kind of science *could* be done ... The lack of computational power meant that researchers could only answer questions that had clean, elegant solutions ... It is only now that we have the ability to do complex calculations and simulations that we are discovering that a great many systems seem to have an inherent complexity that cannot be simplified ... After another 300 years, we will no doubt feel as comfortable using computer simulations to analyze nature as scientists today feel using Newton's laws of motion to describe the trajectory of a falling stone."

G. W. Rowe, *Theoretical models in Biology*

Most interesting problems presented by nature are likely to be formally undecidable or computationally irreducible, rendering proofs and predictions impossible. . . .

Mathematicians and scientists have managed to keep busy only by carefully choosing to work on the relatively small set of problems that have simple solutions.

S. Wolfram, *A new kind of science*.
Recensión de *J. Gray* en el *Notices* de la AMS

La urna de Polya

Una de las grandes de las ventajas de la Teoría de la Probabilidad es que nos enseña a desconfiar de nuestras primeras impresiones. *Laplace*

- VHS y Betamax; Neanderthal y Cro-Magnon
- La urna de Polya
- Democracia (?) animal y Mercados financieros

Peros . . .

Entender, entender, . . . [ciencia]

¿y, el rigor, y la certeza, . . . ?[matemáticas]

Exceso de confianza.[sociedad]

El modelo y las agentes

Efectos cuánticos cuando se modelizan comportamientos en teoría descriptiva (que no normativa).

- Encuestas / votaciones
- Evaluación científica
- *Stock options*
- Supervisión / Control de riesgos financieros: Los mercados son sistemas complejos en los que la observación altera los fundamentos. Rentabilidad, volatilidad, correlación, *todos a una*.
- Lo público y lo privado. Default cuando probabilidad de default es $\geq 10\%$.

Finanzas computacionales

- Escenario medio
- *ceteris paribus*
- Incertidumbre
- Dependencia
- . . .

Los marcos metodológicos usados en finanzas: Markowitz, Black-Scholes, CAPM, suponen modelizaciones estilizadas

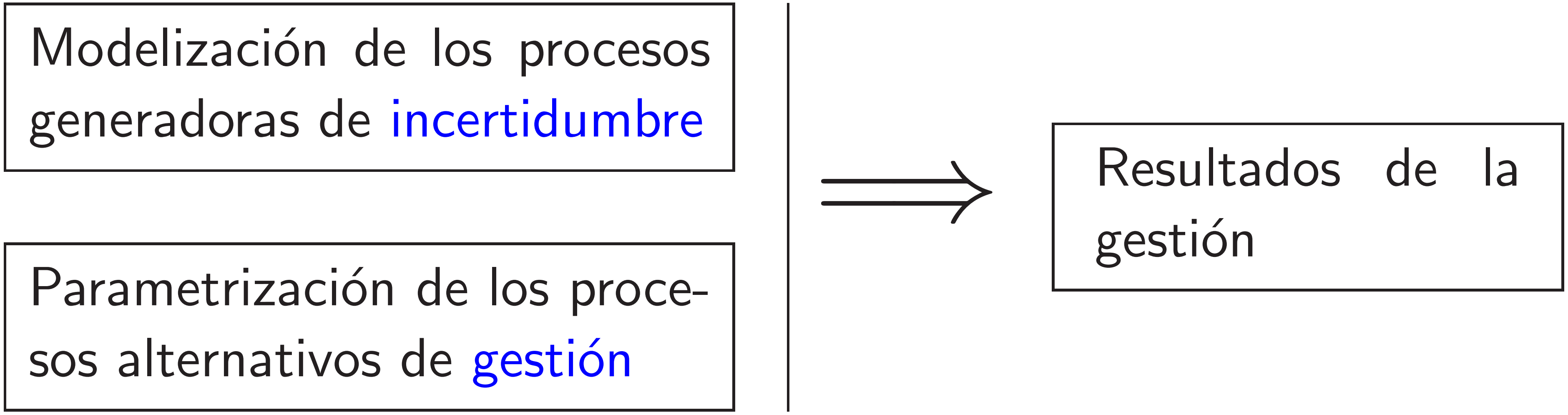
- Cálculo estocástico
  - Cálculo de Ito
  - Cálculo de Malliavin
- Hipótesis que permiten tratamiento analítico
- pero que no captan la inestabilidad e incertidumbre reales

- Premio Nobel

- El fiasco de LTCM. Finanzas forenses.

Gestión de la incertidumbre económica

Modelizar la actividad económica: del negocio y de la financiación



Modelización de los procesos generadoras de incertidumbre

Parametrización de los procesos alternativos de gestión

$\Longrightarrow$

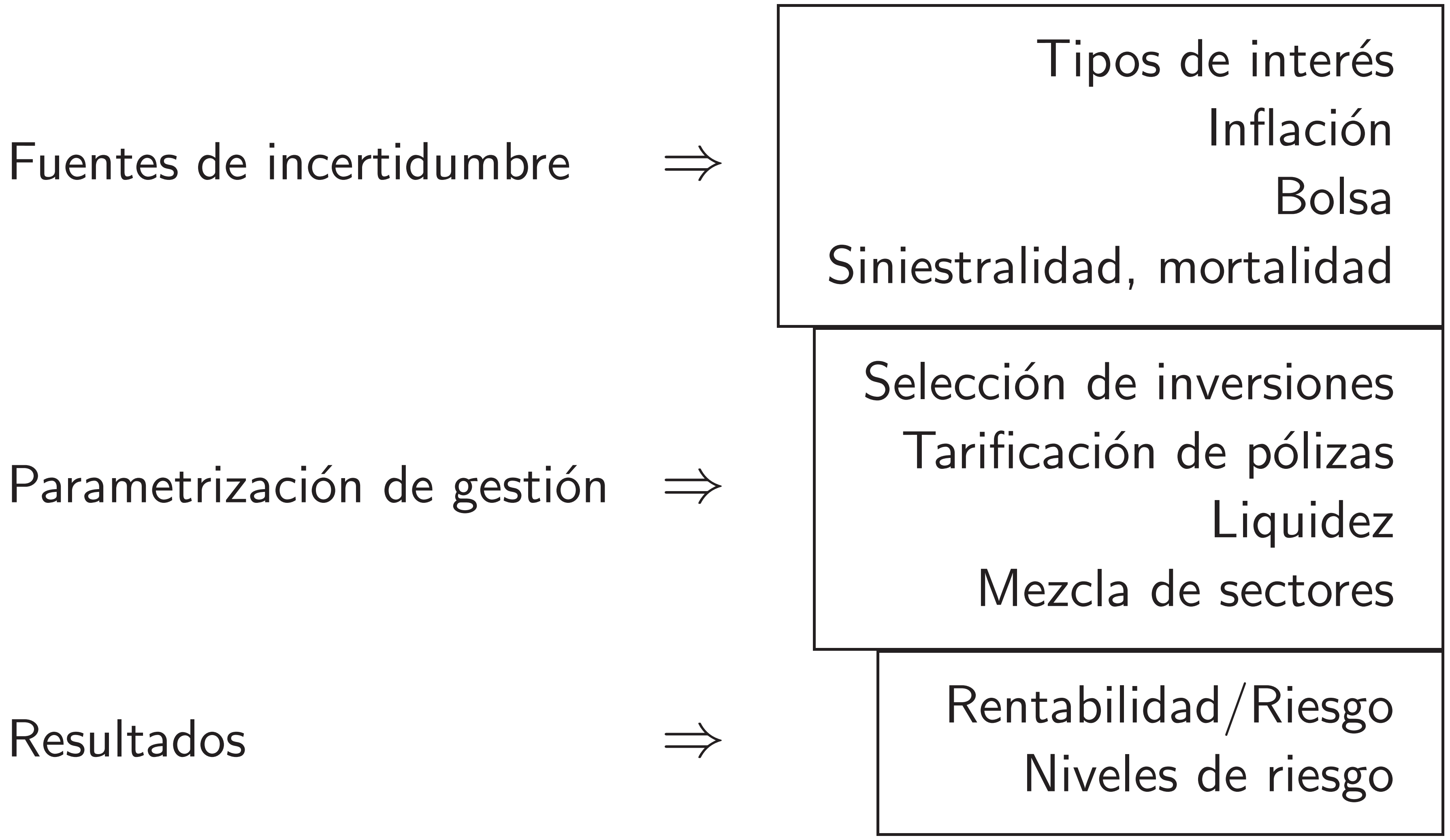
Resultados de la gestión

Énfasis en

- la modelización de los procesos exógenos,
- parametrización de las gestiones,
- determinación del criterio de optimalidad.

Optimización de la gestión.

EJEMPLO: Compañía de seguros



Fuentes de incertidumbre ⇒

Tipos de interés
Inflación
Bolsa
Siniestralidad, mortalidad

Parametrización de gestión ⇒

Selección de inversiones
Tarificación de pólizas
Liquidez
Mezcla de sectores

Resultados ⇒

Rentabilidad/Riesgo
Niveles de riesgo

- Precios de seguros.

- La situación general de las aseguradoras.

El sistema de pensiones

- *New deal*. 65 años
- Deuda nacional

- Pasivo
  - Más pensionistas
  - Más longevos
  - Mayor tasa de reposición
- Activo
  - Menor población
  - Menor carrera laboral
- El pacto de Toledo
- 3 generaciones

Fondos de pensiones

An Introduction to Stochastic Pension Plan Modelling [1] [2]

Andrew J.G. Cairns
Department of Actuarial Mathematics and Statistics,
Heriot-Watt University,
Riccarton, Edinburgh,
EH14 4AS
United Kingdom

e-mail: A.Cairns@ma.hw.ac.uk
Tel: (0)31-451-3245

Abstract

In this paper we consider models for pension plans which contain a stochastic element. The emphasis will be on the use of stochastic interest models, although we will also consider stochastic salary growth and price inflation. The paper will concentrate primarily on defined benefit pension plans. In doing so we will look at how the size of the fund and the contribution rate vary through time and examine how these are influenced by factors which are within the control of a plan's managers and advisers. These factors include the term over which surplus is amortized; the period between valuations; the delay between the valuation date and the implementation of the new contribution rate; and the asset allocation strategy.

The paper will stress the importance of having a well defined objective for a pension plan: optimal decisions and strategies can only be made when a well defined objective is in place.

The paper will also consider, briefly, defined contribution pension plans. The primary decision here relates to the construction of suitable investment strategies for individual members. Again, a well defined objective must be formulated before a sensible strategy can be designed.

Finally, computer simulation methods will be discussed.

[1] Technical Note 94/11

[2] Presented to the workshop on Interest Rate Risk, Vancouver, 19-20 August, 1994

Contents

1 Introduction

In this paper we will consider stochastic pension plans. Pension plans generally fall into one of two categories: defined benefit plans; and defined contribution plans. Both of these are common in countries such as Canada, the USA, the UK and Australia. In all of these countries defined contribution plans are growing significantly in number at the cost of defined benefit plans as employers shift the burden of investment risk over to employees.

In this work we consider how the effects of investment risk can be reduced by making effective use of factors which are within the control of the scheme. These are

- Defined benefit: the method and period of amortization; the intervaluation period; the delay in implementing a recommended contribution rate; the funding method; the valuation basis; the asset allocation strategy.
- Defined contribution: the asset allocation strategy (age dependent); the contribution rate.

In the following sections we will look at each of these factors and consider the effects which each has on levels of uncertainty. In attempting to analyse such problems, a stochastic framework is the only sensible one to use. Within a deterministic framework there is no concept of uncertainty: the very thing we are attempting to quantify and control. For some factors the effect is the intuitive one, while in others the effect may not be known until some sort of exact or numerical analysis can be carried out.

2 Defined Benefit Pension Plans

Defined benefit pension plans provide benefits to members which are defined in terms of a member's final salary (according to some definition), and the length of membership in the plan. For example,

$$\begin{aligned}
\text{Annual pension} &= \frac{N}{60} \times FPS \\
\text{where } N &= \text{number of years of plan membership} \\
FPS &= \text{final pensionable salary}
\end{aligned}$$

In defined benefit pension plans pension and other benefits do not depend on past investment performance. Instead the risk associated with future returns on the funds assets is borne by the employer. This manifests itself through the contribution rate which must vary through time as the level of the fund fluctuates above and below its target level. If these fluctuations are not dealt with (that is, if the contribution rate remains fixed) then the fund will ultimately either run out of assets from which to pay the benefits or grow exponentially out of control.

2.1 A simple model

A number of the factors which we will look at can be first investigated by looking at a very simple stochastic model. By doing so we are able to focus quite quickly on the problem and to give ourselves a good feel for what might happen when we look at more realistic and complex models. This approach follows that of Dufresne (1988, 1989 a,b, 1990), Haberman (1992, 1993 a,b, 1994), Zimbidis and Haberman (1993), Cairns (1995) and Cairns and Parker (1995).

Suppose, then, that we have a fund which has a stable membership and a stable level of benefit outgo. Assuming that all benefits and contributions are paid at the start of each year we have the following relationship:

$$AL(t+1) = (1+i'_v)\,(AL(t)+NC(t)-B(t))$$

where

$$\begin{aligned} AL(t) &= \text{actuarial liability at time } t \\ B(t) &= \text{benefit outgo at time } t \\ NC(t) &= \text{normal contribution rate at time } t \\ \text{and } \ i'_v &= \text{valuation rate of interest} \end{aligned}$$

Suppose that salary inflation is at the rate s per annum and that benefit outgo increases in line with salaries each year. Then

$$\begin{aligned} B(t) &= B.(1+s)^t \\ AL(t) &= AL.(1+s)^t \\ NC(t) &= NC.(1+s)^t \end{aligned}$$

giving

$$\begin{aligned} AL(1+s) &= (1+i'_v)(AL+NC-B) \\ \text{or } \ AL &= (1+i_v)(AL+NC-B) \end{aligned}$$

where $i_v = (1+i'_v)/(1+s)-1 = (i'_v - s)/(1+s)$ is the real valuation rate of interest. Hence

$$NC = B-(1-v_v)L$$

where $v_v = 1/(1+i_v)$.

For convenience we will work in real terms relative to salary growth. In effect this means that we may assume that $s = 0$, without losing any level of generality.

Now let $F(t)$ be the actual size of the fund at time t. Then

$$F(t+1) = (1+i(t+1))\,(F(t)+C(t)-B)$$

where $i(t+1)$ is the effective rate of interest earned on the fund during the period t up to $t+1$, and $C(t)$ is the contribution rate at time t.

$C(t)$ can be split into two parts: the normal contribution rate, NC; and an adjustment $ADJ(t)$ to allow for surplus or deficit in the fund relative to the actuarial liability. Thus

$$C(t) = NC + ADJ(t)$$

We will deal with the calculation of this adjustment in the next two sections.

The deficit or unfunded liability at time t is defined as the excess of the actuarial libility over the fund size at time t. Hence we define

$$\begin{aligned} UL(t) &= \text{unfunded liability at time } t \\ &= AL - F(t) \end{aligned}$$

In North America it is common also to look at the loss which arises over each individual year. This is defined as the difference between the expected fund size (based on the valuation assumptions) and the actual fund size at the end of the year given the history of the fund up to the start of the year. This gives us

$$\begin{aligned} L(t) &= \text{loss in year } t \\ &= E[F(t)] - F(t) \;\; \text{given the fund history up to time } t-1 \\ &= UL(t) - E[UL(t)] \;\; \text{given the fund history up to time } t-1 \end{aligned}$$

We will make use of $UL(t)$ and $L(t)$ in the next section.

No mention has been made so far of the interest rate process $i(t)$. Initially we will assume that $i(1), i(2), \ldots$ form an independent and identically distributed sequence of random variables with

$$\begin{aligned} i(t) &> -1 \;\; \text{with probability 1} \\ E[i(t)] &= i \\ Var[i(t)] &= Var[1+i(t)] = \sigma^2 \\ \Rightarrow E[(1+i(t))^2] &= (1+i)^2 + \sigma^2 \end{aligned}$$

For notational convenience we will define

$$\begin{aligned} v_1 &= \frac{1}{E[1+i(t)]} = \frac{1}{1+i} \\ v_2 &= \frac{1}{E[(1+i(t))^2]} = \frac{1}{(1+i)^2+\sigma^2} \end{aligned}$$

These will be made use of in later sections.

2.2 Two methods of amortization

The Spread Method: This is in common use in the UK. The adjustment to the contribution rate is just a fixed proportion of the unfunded liability: that is,

$$\begin{aligned} ADJ(t) &= k.UL(t) \\ \text{where } k &= \frac{1}{\ddot{a}_{\overline{m}|}} \text{ at rate } i_v \\ \text{and } m &= \text{the period of amortization.} \end{aligned}$$

The period of amortization is chosen by the actuary, and commonly ranges from 5 years to over 20 years. For accounting purposes in the UK *m* must be set equal to the average future working lifetime of the membership.

The Amortization of Losses Method: This is in common use in the USA and Canada. The adjustment is calculated as the sum of the losses in the last *m* years divided by the present value of an annuity due with a term of *m* years calculated at the valuation rate of interest: that is,

$$ADJ(t) = \frac{1}{\ddot{a}_{\overline{m}|}} \sum_{j=0}^{m-1} L(t-j)$$

The interpretation of this is that the loss made in year *s* is recovered by paying *m* equal instalments of $L(s)/\ddot{a}_{\overline{m}|}$ over the next *m* years. These *m* instalments have the same present value as the loss made in year *s*.

Dufresne (1989b) showed that the unfunded liabilities and the losses are linked in the following way:

$$\begin{aligned} UL(t) &= \sum_{j=0}^{m-1} \lambda_j L(t-j) \\ \text{where } \lambda_j &= \frac{\ddot{a}_{\overline{m-j}|}}{\ddot{a}_{\overline{m}|}} \end{aligned}$$

Intuitively this makes sense, since $\lambda_j L(t-j)$ is just the present value of the future amortization instalments in respect of the loss made at time $t-j$. Hence $UL(t)$ is equal to the present value of the outstanding instalments in respect of all losses made up until time t.

The Spread Method can also be defined in terms of the loss function. Whereas the Amortization of Losses Method recovers the loss at time *t* by taking in *m equal* instalments of $L/\ddot{a}_{\overline{m}|}$, the Spread Method recovers this by making a geometrically decreasing, infinite sequence of instalments which starts at the same level.

We are now in a position to calculate the long term mean and variance of the fund size and of the contribution rate. Details of these are provided in Dufresne (1989) (in the case when the valuation and the true mean rate of interest are equal) and Cairns (1995) (covering the case when $i \neq i_v$). For the Spread method we find that

$$
\begin{aligned}
E[F(t)] &= \frac{(1-k-v_v)AL}{(1-k-v_1)} \\
E[C(t)] &= B - \frac{(1-k-v_v)(1-v_1)AL}{(1-k-v_1)} \\
Var[F(t)] &= \frac{(1-k-v_v)^2(v_1^2-v_2)}{(1-k-v_1)^2(v_2-(1-k)^2)}AL^2 \\
Var[C(t)] &= k^2\frac{(1-k-v_v)^2(v_1^2-v_2)}{(1-k-v_1)^2(v_2-(1-k)^2)}AL^2
\end{aligned}
$$

When $i = i_v$ these simplify to

$$
\begin{aligned}
E[F(t)] &= AL \\
E[C(t)] &= B-(1-v_1)AL \\
Var[F(t)] &= \frac{(v_1^2-v_2)}{(v_2-(1-k)^2)}AL^2 \\
Var[C(t)] &= k^2\frac{(v_1^2-v_2)}{(v_2-(1-k)^2)}AL^2
\end{aligned}
$$

Now $v_1 > v_2$ and we must have $Var[F(t)]$ and $Var[C(t)]$ greater than 0. Hence we must have $(1-k)^2 < v_2 \;\Rightarrow\; k > 1-\sqrt{v_2}$. This then automatically implies that $k > 1-v_1$ and if this is combined with $k > 1-v_v$ it ensures that the mean fund size is also positive.

Looking at the Amortization of Losses Method we have, when $i = i_v$,

$$
\begin{aligned}
Var[L(t)] &= \frac{\sigma^2(1+i)^{-2}AL^2}{1-\sigma^2(1+i)^{-2}\sum_{j=1}^{m-1}\lambda_j^2} = V_\infty \;\text{ say} \\
Var[F(t)] &= V_\infty \sum_{j=0}^{m-1}\lambda_j^2 \\
Var[C(t)] &= \frac{m.V_\infty}{(\ddot{a}_{\overline{m}|})^2}
\end{aligned}
$$

2.3 The period of amortization

We now consider the first factor which we have within our control: the period of amortization, m.

For the time being, assume that $i = i_v$: we will look at the more general case in a later section. The following results can be shown to hold for the Spread Method (for example, see Dufresne, 1989b)

- $Var[F(t)]$ increases as m increases.

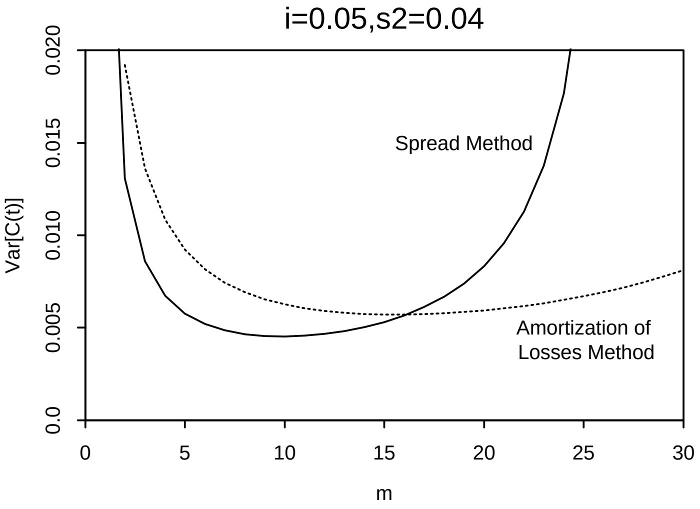


Figure 1: The effect of the period of amortization on the variance of the contribution rate with $E[i(t)] = 0.05$ and $Var[i(t)] = 0.04$.

- $Var[C(t)]$ decreases initially as m increases from 1 up to some value m^* and then increases as m increases beyond m^*. The optimal value, m^*, is such that
$k^* = 1/\ddot{a}_{\overline{m^*|}} = 1 - v_2$.

Looking at the Amortization of Losses Method no such analytical results have been proved but numerical examples show that the same qualitive behaviour holds, as illustrated in the following example.

Suppose $E[i(t)] = i = 0.05$ and $Var[i(t)] = \sigma^2 = 0.2^2$. Figure 1 illustrates how the variance of the contribution rate (with $AL = 1$) depends on m. The Spread Method has its minimum at about 10 while the Amortization of Losses Method has its minimum at about 16, and this minimum is higher.

In Figure 2 we compare the variance of the fund size against the variance of the contribution rate. We do this because we may be interested in controlling the variance of both the contribution rate *and* the fund size. As m increases each curve moves to the right, first decreasing and then increasing as m passes through m^*. Above m^* both the variance of the fund and the variance of the contribution rate are increasing. It is clear then that no value of m above m^* can be ‘optimal’ because the use of some lower value of m (say, m^*) can lower the variance of both the fund size and the contribution rate. The range $1 \leq m \leq m^*$ is the so-called *efficient* region: that is, given a value of m in this range there is no other value of m which can lower the variance of both the fund size and the contribution rate. There is therefore a trade-off between variability in the fund size and the contribution rate and settling on what we regard as an optimal spread period can only be done with reference to a more specific objective than ‘minimize variance’.

It is significant that the Amortization of Losses Method curve always lies above the Spread Method curve. This means that the Spread Method is certainly more efficient than the Amortization of Losses Method: that is, for any value of m in combination with the Amortization of Losses Method there is a (different) value m' for which the variance of both the fund size and the contribution rate can be reduced by switching to the Spread Method.

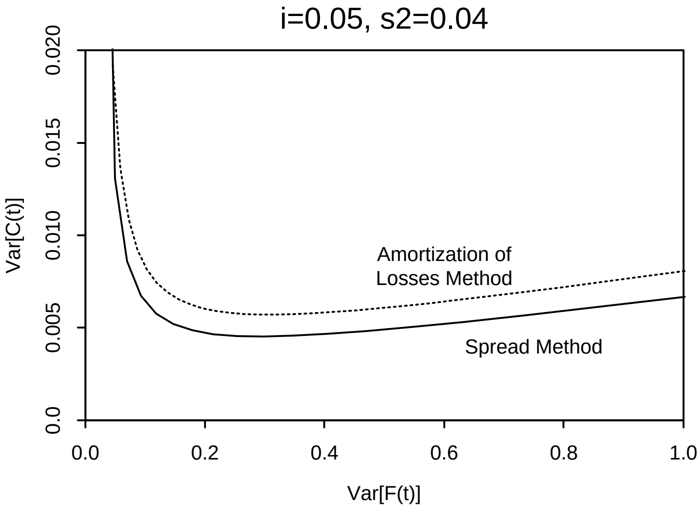


Figure 2: $E[i(t)] = 0.05$ and $Var[i(t)] = 0.04$. Comparison of $Var[F(t)]$ with $Var[C(t)]$. Notes: $Var[F(t)]$ increases as m increases; the efficient frontier for the Spread Method is always more efficient than that for the Amortization of Losses Method.

2.4 The intervaluation period

The time between valuations is nominally a factor which is within the control of the scheme. We have so far considered the case where valuations are carried out on an annual basis. Such an approach is common amongst larger funds but this is often felt to be uneconomic for smaller funds to carry out such frequent valuations. Instead smaller funds often opt for a three year period between valuations ($3\frac{1}{2}$ years being the statutory maximum in the UK).

The effects of changing from annual to triennial valuations have been considered by Haberman (1993b). He finds that under the Spread Method of amortization

- the optimal spread period for $Var[C(t)]$, m^*, increases by about 1 year;
- the variances of both $F(t)$ and $C(t)$ are increased for most values of m below about m^*.

Continuing the example of the previous section we looked at 1 and 3 year intervaluation periods. Figure 3 plots $Var[C(t)]$ against m. For low values of m lengthening the intervaluation period has the effecct of increasing the variance of $C(t)$: the intuitive effect. For higher values of m, however, the reverse is true. This perhaps reflects the fact that over each three year period $C(t)$ is being held fixed thereby reducing the overall variance.

Comparing the variances of $F(t)$ and $C(t)$ (Figure 4) we see that, in this example at least, the efficient range for annual valuations lies below that for triennial valuations. We conclude that annual valuations are preferrable, although for values of m close to m^* there is little difference in the variances, so the benefit of annual valuations is marginal.

2.5 The delay period

The original analysis asumes that the new contribution rate can be implemented at the valuation date. In reality the results of a valuation are often not known until 6 or even 12 months after the

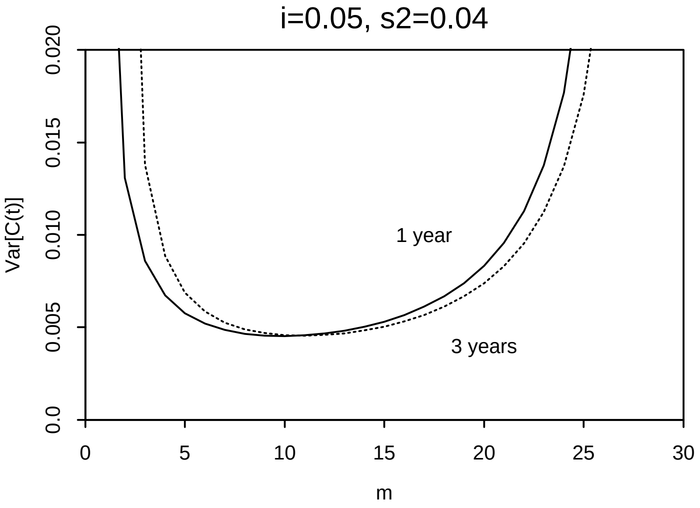


Figure 3: $E[i(t)] = 0.05$ and $Var[i(t)] = 0.04$. $Var[C(t)]$ plotted against m for annual and triennial valuations.

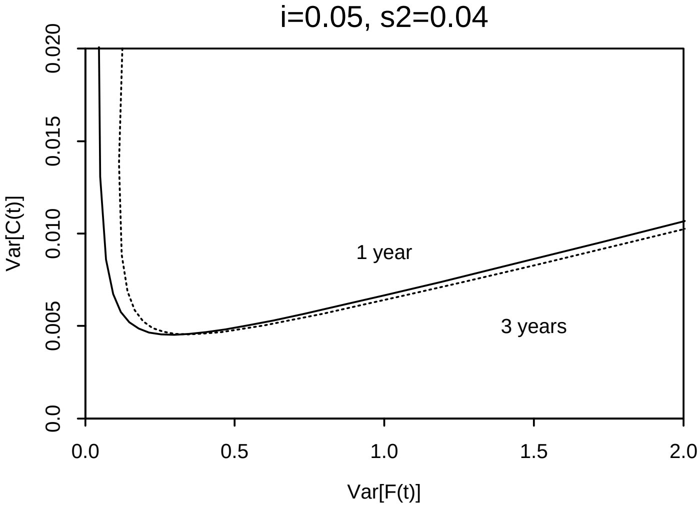


Figure 4: $E[i(t)] = 0.05$ and $Var[i(t)] = 0.04$. Comparison of $Var[F(t)]$ with $Var[C(t)]$. Note: the efficient frontier for the annual valuation case is, for most values of m less than m^*, below that for the triennial valuation case.

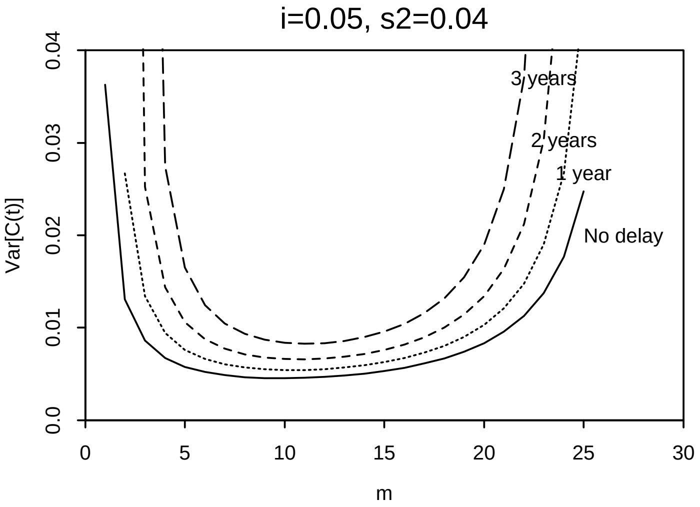


Figure 5: $E[i(t)] = 0.05$ and $Var[i(t)] = 0.04$. $Var[C(t)]$ plotted against m for delay periods of 0, 1, 2 and 3 years.

valuation date. The new recommended contribution rate is therefore typically not implemented until one year later. There is a delay period of 1 year.

This problem has been investigated by Zimbidis and Haberman (1993). In the example under consideration each extra year's delay increases the variance of $F(t)$ and $C(t)$ by at least 20% and by much more substantial amounts for small values of m. Figures 5 and 6 illustrate the results for this example. One point to note is that where there is a delay period then $Var[F(t)]$ initially decreases with m before increasing as in the no-delay case. This has the effect of reducing the efficient range for m. For example, with a delay of 3 years the efficient range is $5 \leq m \leq 11$ as compared with $1 \leq m \leq 10$ when there is no delay.

In view of the substantial increases in variance caused by a delay it is felt that the delay should be kept as short as possible and perhaps that allowance should be made in the current rate even if the final results of a valuation are not known.

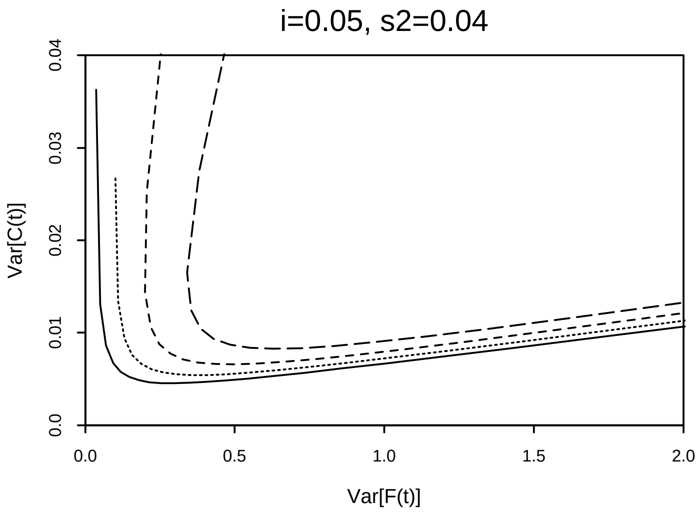


Figure 6: $E[i(t)] = 0.05$ and $Var[i(t)] = 0.04$. Comparison of $Var[F(t)]$ with $Var[C(t)]$. Increasing the delay period increases the variance of both $F(t)$ and $C(t)$.

2.6 The funding method

Recall the equilibrium equation relating *AL* to *NC*

$$AL = (1 + i_v)(AL + NC - B)$$

If we increase *AL* then *NC* balances this by falling (this is because benefits are paid from contributions plus surplus interest on the fund, which has increased). Furthermore, *AL* is determined by the funding method. The normal ordering which we find is

$$AL_{CUC} < AL_{PUC} < AL_{EAN}$$

where the subscripts represent the Current Unit Credit (CUC), Projected Unit Credit (PUC) and Entry Age Normal (EAN) methods, these being the three main funding methods appropriate for a stable membership.

The Attained Age Method has the same actuarial liability as the Projected Unit Credit Method but normally has a higher normal contribution rate which is appropriate for a closed fund, but which will give systematic rise to surplus when the fund has a stable membership. In such a case the equilibrium equation is, therefore, not satisfied. Instead the system has a higher equilibrium fund size which depends on the method and period of amortization.

The variances of $F(t)$ and $C(t)$ are both proportional to AL^2. This means that a more secure funding method (higher *AL*) gives rise to greater variability, suggesting that a method with a low actuarial liability is to be preferred. Clearly this is not a prudent strategy. It jeopardizes member's security and is more likely to violate statutory solvency requirements.

This problem can be overcome by a number of methods, including:

- the use of the normalized variances $Var[F(t)]/E[F(t)]^2$ and $Var[C(t)]/E[F(t)]^2$;
- the use of further fund objectives (for example, by conditioning on the mean fund size being at a specified level).

2.7 The strength of the valuation basis

So far we have concentrated on the case where the valuation rate of interest, i_v, is equal to the mean long term rate of interest, i. It is common, however, for valuations to be carried out on a strong (occasionally weak) basis: that is, to set $i_v < i$ (or $i_v > i$). This gives rise to a wider variety of results.

Recall that

$$\begin{aligned}
E[F(t)] &= \frac{(1-k-v_v)AL}{(1-k-v_1)} \\
E[C(t)] &= B - \frac{(1-k-v_v)(1-v_1)AL}{(1-k-v_1)} \\
Var[F(t)] &= \frac{(1-k-v_v)^2(v_1^2-v_2)}{(1-k-v_1)^2(v_2-(1-k)^2)}AL^2 \\
Var[C(t)] &= k^2\frac{(1-k-v_v)^2(v_1^2-v_2)}{(1-k-v_1)^2(v_2-(1-k)^2)}AL^2
\end{aligned}$$

We concentrate on the variance of the contribution rate and look for the existence of a minimum with respect to the period of amortization, m. There are a number of cases to consider.

1. **Strong basis:** $i_v < i$ $(v_v > v_1)$

   (these are currently observations, and not proved)

   (a) $E(C_t)$ is an increasing function of k for $k > 1-\sqrt{v_2}$.

   (b) $Var(C_t)$ has a minimum for some $1-\sqrt{v_2} < k^* < 1$.

   (c) $Var(F_t)$ is a decreasing function of k.

   From this we can see that for $k > k^*$ both the expected value and the variance of the contribution rate are increasing so that increasing k above k^* is not worthwhile. If k is decreased then we trade off a lower contribution rate for a higher variance. The optimal value therefore depends on the pension fund's utility function or objectives. This goes slightly against the conclusions of Dufresne who indicates that k^* would be the *minimum* acceptable value of k.

   For some values of k the mean contribution rate will be negative, indicating that the fund is large enough to pay for itself and at times requiring refunds to the employer. Although this seems an ideal situation, the reality is that the company must first have built up the fund to this level. It would also be likely to violate statutory surplus regulations.

   It is possible to have smaller expected fund levels and higher contribution rates, but these do not arise if the projected unit method is used in the calculation of the funding rate and using a conservative valuation rate of interest.

2. **Best estimate:** $i_v = i$ $(v_v = v_1)$

   The results of Dufresne (1989) hold.

   (a) $E(C_t)$ is a constant function of k for $k > 1-\sqrt{v_2}$.

   (b) $Var(C_t)$ has a minimum for some $1-\sqrt{v_2} < k^* < 1$.

   (c) $Var(F_t)$ is a decreasing function of k.

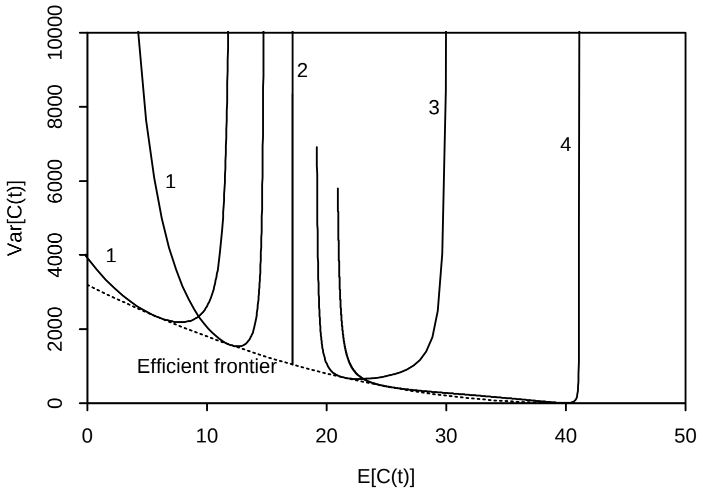


Figure 7: $E[i(t)] = 0.05$ and $Var[i(t)] = 0.04$. $Var[C(t)]$ plotted against $E[C(t)]$ for different valuation rates of interest. Moving from left to right the curves represent: $i_v = 0.03, 0.04$ (type 1, strong basis); $i_v = 0.05$ (type 2, best estimate basis); $i_v = 0.06$ (type 3, weak basis); $i_v = 0.07$ (type 4, very weak basis). The dotted line is the efficient frontier.

3. **Weak basis:** $i < i_v < \sqrt{(1+i)^2 + \sigma^2} - 1 \;\; (v_1 > v_v > \sqrt{v_2})$

   (a) $E(C_t)$ is a decreasing function of k for $k > 1 - \sqrt{v_2}$.

   (b) $Var(C_t)$ has a minimum for some $1 - \sqrt{v_2} < k^* < 1$.

   (c) $Var(F_t)$ is a decreasing function of k.

   This time we find that it may be acceptable to increase k above k^*, trading off lower contributions for higher variability.

4. **Very weak basis:** $\sqrt{(1+i)^2 + \sigma^2} - 1 < i_v \;\; (\sqrt{v_2} > v_v)$

   (a) $E(C_t)$ is a decreasing function of k for $k > 1 - v_v$ at which point it equals B and the scheme is funded on a pay as you go basis. For $1 - v_v > k > 1 - \sqrt{v_2}$ $E(C_t)$ is still a decreasing function.

   (b) $Var(C_t)$ has a minimum equal to zero at $k = 1 - v_v$. This is because the scheme is now funded on a pay as you go basis and contributions equal the constant B.

   (c) $Var(F_t)$ has a local minimum at $k = 1$, a maximum at some $1 - v_v < k^* < 1$ and a global minimum equal to zero at $k = 1 - v_v$ when the fund stays constant at zero.

The efficient frontier

Pooling these results together we can determine a curve $m(\mu_c)$ where

$$m(\mu_C) = \min\{Var(C_t) : E(C_t) = \mu_C, 1 > k > \max(1 - v_v, 1 - \sqrt{v_2}), v_v < 1\}$$

That is, $m(\mu_C)$ gives us the minimum variance attainable for a given mean contribution rate. In fact, it can be shown that $m(\mu_C)$ is convex (quadratic).

These different types of outcome are illustrated in Figure 7, with $i = 0.05$ and $\sigma^2 = 0.2^2$.

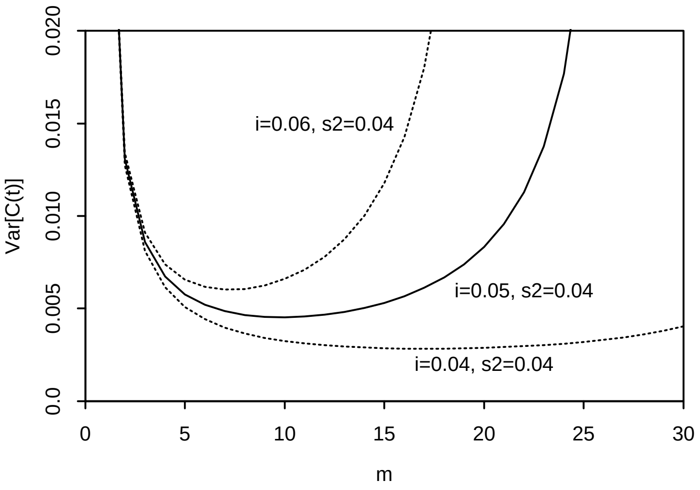


Figure 8: $E[i(t)] = 0.05$ and $Var[i(t)] = 0.04$. $Var[C(t)]$ plotted against m for different long term rates of return. The valuation rate of interest is fixed.

2.8 Sensitivity testing

In carrying out such analyses it is important to realize that the model for the rate of return including its parameter values are uncertain. First, the model we use here is only one of a range of possible models of varying complexity which all fit past data reasonably well. All of these models are, however, only an approximation to a much more complex reality. Second, the parameter values which we have used (here $i = 0.05$ and $\sigma^2 = 0.04$) are not known with certainty: for example i could equally well be 0.04 or 0.06.

In fact this can have a very significant effect on level the variability. Figures 8 and 9 illustrate this point. i is allowed to take the values 0.04, 0.05 and 0.06. In Figure 8 the effect on $Var[C(t)]$ is very significant, particularly for larger values of m. However, these results are distorted by the fact that when $i \neq i_v$ the mean fund size ($E[F(t)]$) depends on m. The normalized variance of $C(t)$ is plotted in Figure 9 and the effect can be seen to be reduced but still significant.

A change in the value of i of 1% makes a difference in m^* of about 2 years (for example, moving from $i = 0.05$ to $i = 0.06$ changes m^* from 10 to 8).

The result of these changes is not as significant as might first appear. For example, suppose we settled upon $m^* = 10$ on the basis that $i = 0.05$. If in fact the long term mean turned out to be $i = 0.06$ then amortizing over 10 years would only turn out to have been only marginally worse than if the true optimum $m^* = 8$ had been used. The fact that the actual variance of the contribution rate was perhaps 20% higher than that expected is irrelevant since the lower value would never, in fact, have been attainable.

Figure 10 shows the effects of uncertainty in σ^2 (with σ^2 taking the values 0.03, 0.04 and 0.05). The effect is again substantial, but much more uniform over the whole range of values for m. This is because σ^2 has a much more direct effect on the variance of the fund size and the contribution rate. However, as with uncertainty in i, the normalized variance is relatively stable over a range of values about the minimum, so choosing the wrong value of m will only marginally increase the long term variance.

The point to take in from this section is that we need to take care in ensuring that we look at the right quantities. We therefore need to compare the *actual* outcome based on the decision which

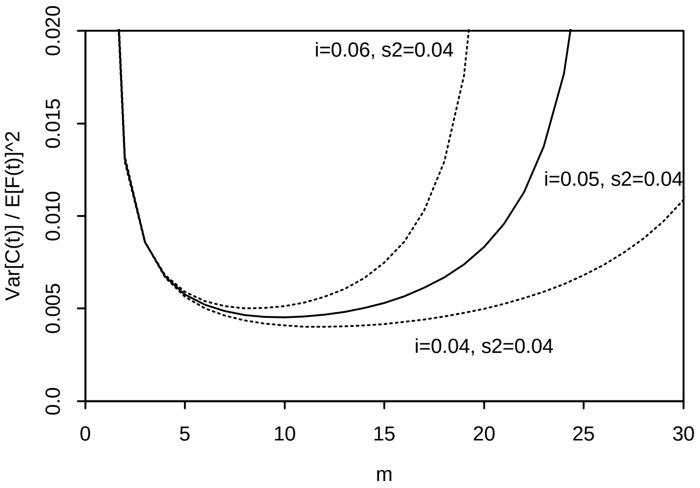


Figure 9: $E[i(t)] = 0.05$ and $Var[i(t)] = 0.04$. $Var[C(t)]/E[F(t)]^2$ plotted against m for different long term rates of return. The valuation rate of interest is fixed.

was based on incorrect assumptions with the outcome which would have *actually* happened had the decision been based on the correct assumptions. Here the differences have been shown to be minimal but if we were to find that they were significant then we may need to look carefully at our estimates to see if they can be refined and improved upon.

2.9 Objectives

We have already discussed that within the efficient region for m ($1 \leq m \leq m^*$) there is a trade off between higher variance of $F(t)$ and higher variance of $C(t)$. To settle on an optimal spread period therefore requires a specific objective or utility function. For example, we may be concerned about containing the fund size within a specified band (bounded below, say, by the minimum solvency level and above by a statutory surplus limit). We could accomodate this by specifying that $E[F(t)]$ lie in the middle of this band and that the standard deviation of $F(t)$ be no more than 10% of this mean fund size. In this case the optimum would be m^{**} which pushes the variance of $F(t)$ up to the maximum level allowable or m^* if this is lower.

If a proper optimum is to be found then the fund must have a well defined objective which will allow optimization to take place. Examples of some objectives are:

- Minimize $Var[C(t)]$ subject to $Var[F(t)] \leq V_{max}$;
- Minimize $Var[C(t)]$ subject to $E[F(t)] = \mu_F$;
- Minimize the variance of the present value of all future contributions (that is, $\sum_{t=0}^{\infty} v^t C(t)$) subject to ......;
- Maximize $E[u(F(t))]$ where $u(f)$ is utility function which depends on the fund size. For example, if $u(f) = -(f - f_0)^2$ then $E[u(F(t))] = -Var[F(t)] - (E[F(t)] - f_0)^2$, the second term being a penalty for deviation of the mean from the target of f_0.

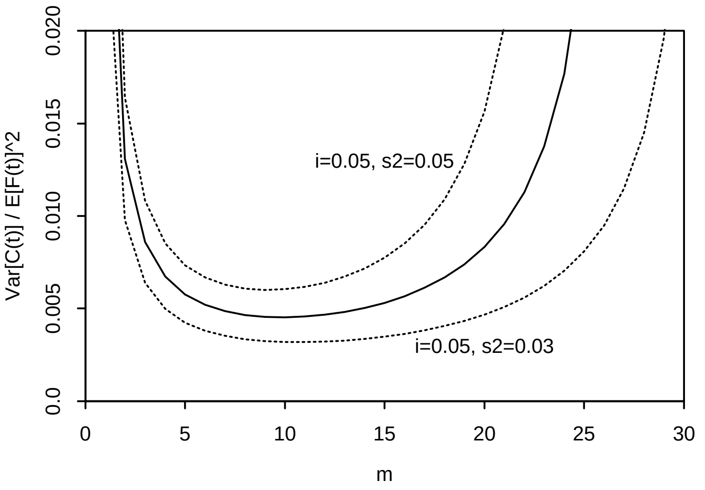


Figure 10: $E[i(t)] = 0.05$ and $Var[i(t)] = 0.04$. $Var[C(t)]/E[F(t)]^2$ plotted against m for varying levels of volatility in the rate of return. The valuation rate of interest is fixed.

Care should be taken when formulating an objective. For example, the last of these makes less sense if $E[F(t)]$ is constant for all values of m (that is if $i_v = i$); and constraints should have reasonable rather than extreme values.

2.10 Other stochastic investment models

We have used the simplest stochastic interest model here (independent and identically distributed returns) which allows us to obtain some intuitively appealing analytical results. A wide variety of more complex models are used in practice for which analytical results are not possible. However, it is expected that similar qualitative results should be available.

Autoregressive time series models: Haberman (1993a) has investigated the use of the AR(1) time series model:

$$\begin{aligned}
\delta(t) &= \delta + \alpha(\delta(t-1) - \delta) + \nu Z(t) \\
\text{where } \delta(t) &= \log(1 + i(t)) \\
Z(t) &\sim N(0,1) \\
|\alpha| &< 1 \text{ is the autoregressive parameter} \\
\delta &= \text{long term mean rate of return} \\
\nu^2 &= \text{variance parameter} \\
\text{Hence } E[\delta(t)] &= \delta \\
Var[\delta(t)] &= \sigma^2 = \frac{\nu^2}{1-\alpha^2} \\
E[1+i(t)] &= e^{\delta + \frac{1}{2}\sigma^2} \\
Var[1+i(t)] &= e^{2\delta + \sigma^2}\left(e^{\sigma^2} - 1\right)
\end{aligned}$$

It has been found that $\alpha > 0$ (positively correlated returns) decreases the value of m^* (for example, with $E[i(t)] = 0.05$ and $Var[i(t)] = 0.2^2$ m^* falls from 10 to 5 when α is changed from 0 (independent and identically distributed returns) to only 0.1). More likely is the case $\alpha < 0$ (a high return one year is followed by a low return the next year) which increases the value of m^*.

Note that such models seem more appropriate to fixed interest investments: past equity data do not show any significant signs of autocorrelation from one year to the next.

In summary the most widely used stochastic interest models are

- Independent and identically distributed returns: for example, Waters (1978), Dufresne (1990), Papachristou and Waters (1991), Parker (1993 a,b, 1994 a,b) and Aebi *et al.* (1994) give but a few examples.
- Simple autoregressive models, such as the $AR(1)$ time series model, and the Ornstein-Uhlenbeck process: for example, Dhaene (1989), Parker (1993 a,b, 1994 a,b) and Norberg and Møller (1994).
- Models for the term structure of interest rates: for example, Boyle (1978, 1980), Brennan and Schwarz (1979), Albrecht (1985), Cox, Ingersoll and Ross (1985), Beekman and Shiu (1988), Heath, Jarrow and Morton (1990, 1992), Reitano (1991), Sercu (1991) and Longstaff and Schwarz (1992).
- Models with several asset classes: for example, Wilkie (1987, 1992, 1994), and Chan (1994).

The last two of these classes are the most appropriate for the purposes of making an asset allocation decision. In an objective based setting, however, the asset allocation strategy must be considered simulataneously with other factors which are within our control (see the example in the next section).

Increasing complexity means that we need to resort to stochastic simulation in most of these cases.

2.11 Example: A two asset model

Suppose that the fund has two assets in which it can invest. The return in year t on asset j ($j = 1,2$) is $i_j(t)$ with

$$\begin{aligned} E[i_j(t)] &= i_j \text{ for } j = 1,2 \\ Cov[i_j(t), i_k(t)] &= c_{jk} = c_{kj} \quad j,k = 1,2 \end{aligned}$$

Suppose asset 1 carries a lower risk and a lower return: that is, $i_1 < i_2$ and $c_{11} < c_{22}$.

Let $i(t)$ be the overall return during year t, and suppose that a proportion p of the fund is invested in asset 1. Then

$$\begin{aligned}
E[i(t)] &= pi_1 + (1-p)i_2 = \mu(p) \ \text{ say} \\
Var[i(t)] &= Var[pi_1(t) + (1-p)i_2(t)] \\
&= Var[pi_1(t)] + Var[(1-p)i_2(t)] + 2Cov[pi_1(t), (1-p)i_2(t)] \\
&= p^2 c_{11} + (1-p)^2 c_{22} + 2p(1-p)c_{12} \\
&= \sigma^2(p) \ \text{ say}
\end{aligned}$$

(This is following the approach of Modern Portfolio Theory.)

We now put this new mean and variance into the original equations:

$$\begin{aligned}
E[F(t)] &= \frac{(1-k-v_v)AL}{(1-k-v_1)} \\
E[C(t)] &= B - \frac{(1-k-v_v)(1-v_1)AL}{(1-k-v_1)} \\
Var[F(t)] &= \frac{(1-k-v_v)^2(v_1^2-v_2)}{(1-k-v_1)^2(v_2-(1-k)^2)}AL^2 \\
Var[C(t)] &= k^2\frac{(1-k-v_v)^2(v_1^2-v_2)}{(1-k-v_1)^2(v_2-(1-k)^2)}AL^2 \\
\text{where } \ v_1 &= \frac{1}{E[1+i(t)]} = \frac{1}{1+\mu(p)} \\
v_2 &= \frac{1}{E[(1+i(t))^2]} = \frac{1}{(1+\mu(p))^2+\sigma^2(p)}
\end{aligned}$$

We now have at our disposal:

- the period of amortization;
- valuation basis;
- asset mix.

We have seen from looking at the strength of the valuation basis that a wide range of fund sizes can be attained. Optimal choices must therefore be made with reference to some specific objectives. For example,

$$\begin{aligned}
\text{minimize} \quad & Var[C(t)] \\
\text{subject to} \quad & E[F(t)] = AL' \\
& Var[F(t)] \leq (0.1AL')^2
\end{aligned}$$

where AL' is, for example, a statutory minimum plus 20%.

To find an appropriate solution one must now use numerical methods to optimize over the factors within our control. The process of optimization may proceed as follows:

1. Fix the asset proportion and the valuation rate of interest (p and i_v). Then k (therefore m) is determined by the constraint on $E[F(t)]$:

$$E[F(t)] = \frac{(1-k-v_v)}{(1-k-v_1)} AL(i_v) = AL'$$

2. Find the range of values of i_v for which $Var[F(t)] \leq (0.1AL')^2$, and within that range which i_v minimizes $Var[C(t)]$. Let this minimum be $M(p)$.

3. Minimize $M(p)$ over $0 \leq p \leq 1$.

4. Check that the optimal values are reasonable: for exampe, is i_v reasonable when compared with $E[i(t)] = \mu(p^*)$; is m^* reasonable; is p^* acceptable? If the answer to any of these questions is no then we should ask ourselves why and reformulate the objectives accordingly.

2.12 Constraints on strategies

We have already mentioned in Sections 2.6 and 2.9 that our optimal strategy may be influenced by statutory funding levels. These may be

- a minimum solvency requirement;
- a maximum surplus regulation.

Different countries have different regulations for what happens when one of these limits is breached. Typically, however, there may be a requirement to amortize the difference between the limit and the current fund size over a shorter period than normal (in the UK and Canada this is 5 years).

Another constraint may be a limit on the ability of the employer to take a refund from the fund. If no refund at all is possible then ultimately the fund will reach a stage where the fund becomes large enough to be self funding (that is, interest exceeds benefit outgo) beyond which point the fund will grow exponentially out of control. This is a certain event in a stochastic environment. More common is a (statutory) constraint that contribution refunds may only be made while the asset/liability ration remains above a specified level.

When such constraints are in place exact analyses are no longer possible. Instead numerical investigations are necessary.

2.13 Salary growth and price inflation

We have already illustrated how salary growth can be incorporated into these models. This is done by indexing the actuarial libility, the normal contribution rate and the benefit outgo in line with the total salary roll $S(t)$, and treating $i(t)$ as a real rate of return.

Salary inflation can be adequately modelled by an autoregressive process of order 1 or alternatively it can be linked to price inflation (for example, see Section 3 and Wilkie, 1994).

Problems arise when benefit outgo is not proportional to the total salary roll. For example, if pensions are paid from the fund but linked to a prices index then benefit outgo is equal to a mixture of past salary rolls increased in line with the appropriate price index.

This can be approached in two ways: by carrying out a simulation study (described in the next section); or by assuming that pensions are matched at the date of retirement by index-linked securities. In the latter case

$$\begin{aligned} B(t) &= B \times S(t) \times A(t) \\ \text{where } B &= \text{base pension benefit} \\ S(t) &= \text{salary index} \\ A(t) &= \text{real annuity rate at time } t \end{aligned}$$

The annuity rate $A(t)$ is itself governed by a random process: for example, $A(1), A(2), \ldots$ may be independent and identically distributed positive random variables.

2.14 Simulation methods

Two simulation methods are available.

Method 1: (Ergodic method)

All of the interest rate processes described are examples of *ergodic* processes (for example, see Karlin and Taylor, 1975). A consequence of this (amongst other properties) is that the fund process will satisfy

$$\begin{aligned} \bar{f}_n &= \frac{1}{n}\sum_{t=1}^{n} F(t) \rightarrow E[F(t)] \text{ almost surely as } n \rightarrow \infty \\ s_n^2 &= \frac{1}{n}\sum_{t=1}^{n} \left(F(t) - \bar{f}_n\right)^2 \rightarrow Var[F(t)] \text{ almost surely as } n \rightarrow \infty \end{aligned}$$

(If salary growth is allowed for, then $F(t)$ above should be replaced by the asset/liability ratio $F(t)/AL(t)$.)

This means that a single, long simulation run of the pension plan will give us good estimates of the means and variances of the quantities of interest. Rough calculations suggest that this simulation should be of at least 2000 years.

The simulation should be repeated for each combination of decisions being examined. For consistency and efficiency the same realization of the interest rate process should be used for each combination of decisions.

Method 2: Repeated simulation

The objective of the fund may, amongst other things, aim to minimize variance over a short period, say 10 years, rather than over the longer term. Repeated simulation is more appropriate here: that is, simulate the fund for 10 years, given appropriate initial conditions; and then repeat this, say, 200 or more times. For consistency and efficiency the same 200 scenarios of the interest rate process should be used for each combination of decisions.

3 Defined Contribution Pension Plans

Defined contribution pension plans are becoming of ever increasing importance and as such they require some long overdue investigation in order that their reliability as a pensions vehicle can be improved upon. The principal distinctions with defined benefit pension plans are that benefits are no longer based upon final salary but depend on past contribution levels and past investment returns thereby passing investment risk from the employer to the individual members.

Whereas an employer as sponsor of a defined benefit plan is able to smooth out good and bad years' investment returns, defined contribution pension plan members are rather more at the mercy of variations in returns from one year to the next. For example, Knox (1993) carried out a simulation study using a simple model which illustrated the high degree of uncertainty in the final amount of a defined contribution pension relative to final salary. This risk is well known and is a major criticism of the defined contribution set-up. Further work is therefore required to see if this risk can be reduced.

Defined contribution pension plans can be divided into two categories:

- those sponsored by an employer;
- those taken out by individuals with an insurer and with no (or only indirect) involvement on the part of an employer (Retirement Savings Plan).

From a statistical standpoint, this is an artificial distinction. Any decision which can be applied to one type should be applicable to the other: for example, the use of investment strategies which depend on the age of the individual.

3.1 Objectives

Clearly defined objectives are perhaps even more important in the decision making process associated with a defined contribution pension plan than a defined benefit pension plan. Different, member oriented objectives are required and the situation may be complicated further by the possibility that different members may have different objectives and utility functions.

An objective is most likely to be defined in terms of the the amount of pension at retirement *as a proportion of final salary* rather than as an absolute amount. Thus we define

$$\begin{aligned} P(t) &= \text{pension on retirement at time } t \\ S(t) &= \text{salary at time } t \\ \pi(t) &= P(t)/S(t) \\ &= \text{pension as a proportion of final salary} \end{aligned}$$

Now $P(t)$ depends on past contributions, past investment returns and annuity rates at retirement. If contributions are paid at the start of each year then

$$
\begin{aligned}
P(t) &= \frac{1}{A(t)} \sum_{s=0}^{t} \rho(s) S(s) \frac{F(t)}{F(s)} \\
\text{where } \rho(s) &= \text{contribution rate at time } s \\
\frac{F(t)}{F(s)} &= \text{accumulation at time } t \text{ of an investment of 1 at time } s \\
A(t) &= \text{annuity factor applied on retirement at time } t
\end{aligned}
$$

Normally it will be assumed that the contribution rate $\rho(t)$ is constant through time, although this could be used as a method of reducing uncertainty.

Each of the processes $F(t)$, $S(t)$ and $A(t)$ are random. This will exaggerate the level of uncertainty at retirement unless a suitable strategy can be found which can use one process to offset the effects of another. For example, by investing in fixed interest bonds, a fall in bond prices close to retirement will be offset by a fall in the value of $A(t)$, the cost of purchasing an annuity.

Objectives may be divided into two categories

(A) ones in which the member is told of his or her pension only at the date of retirement;

(B) ones in which the member is given advance notice of the (likely) future amount of pension and then expects the final pension to be as close to this as possible (or not too much less than).

Possible objectives of type A are:

- maximize $E[\pi(t)]$;
- maximize $E[\pi(t)]$ subject to $Var[\pi(t)] = \sigma_\pi^2$;
- maximize $Var[\pi(t)]$;
- maximize $Var[\pi(t)]$ subject to $E[\pi(t)] = \mu_\pi$;
- minimize $Pr(\pi(t) < \pi_{\min})$;
- maximize $E[u(\pi(t))]$ where $u(\cdot)$ is some utility function.

Objectives of type B include

- minimize $E[(\pi(t) - \hat{\pi}(t))^2 \mid H_s]$ where H_s gives us the history of the fund up until time t and $\hat{\pi}(t)$ is the estimated future pension based on H_s;
- maximize $E[u(\pi(t)) \mid H_s, \hat{\pi}(t)]$.

It is questionable whether some such objectives may be reasonable. For example, suppose an objective results in a strategy which locks into a given level of pension some time in advance of retirement. The problem with this is that the level which we lock into may be just as variable as the pension which could be obtained had the fund been left alone until the date of retirement. So is it really in the member's best interests to lock into a pension at too early a stage?

3.2 Investment strategies

It may be difficult to examine all possible investment strategies. However, an appropriate starting point may be to examine a small number of possibilities. For example,

- strategies which are fixed through time:
  - equities only
  - equities and matching options
  - fixed interest bonds
  - equities, fixed interest bonds and cash
  - index linked bonds
  - equities, matching options, fixed interest bonds and cash
  - etc.
- strategies which vary through time:
  - equities switching into fixed interest bonds over the last 5 years
  - fixed interest bonds
  - equities and matching options
  - equities, matching options, fixed interest bonds and cash
  - etc.
- strategies which vary through time and depend on the past history of the fund.

3.3 A simple example

Here we look at a simple example which illustrates the fallacy of an early switch into fixed interest bonds.

We simplify the situation by considering a fund which is now of size $F(0)$ and which will receive no further contributions. We are interested in the lump sum which this fund will produce at retirement as a proportion of final salary.

Three options are available:

- a zero-coupon fixed interest investment which provides a guaranteed lump sum L at retirement;
- investment in long-term index linked bonds;
- investment in equities.

The model we will use is described in the Appendix. The model and its parameters were found to fit UK experience reasonably well.

The measure of risk for each option (the variance of the logarithm of the lump sum as a proportion of final salary) is plotted in Figure 11. We can see that although the fixed pension fares

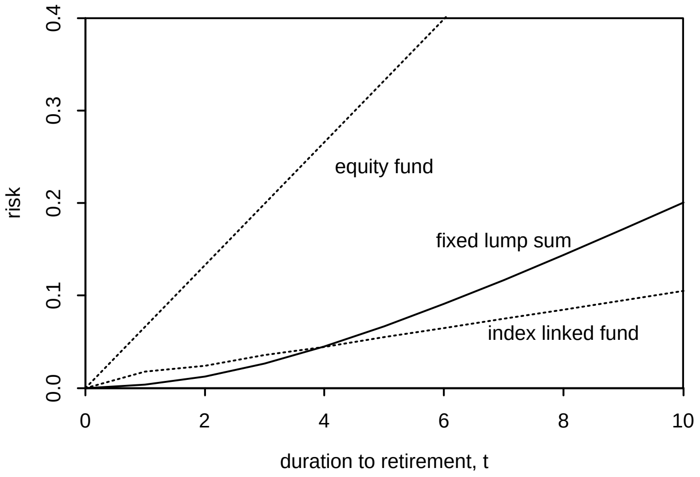


Figure 11: Risk relative to policyholder's salary for three different investment strategies. Risk is measured as $Var[L(t)/S(t)]$ where $L(t)$ is the lump sum at retirement and $S(t)$ is the final salary.

better early on the index linked option clearly becomes lower risk later on. (Note that this does not take account of uncertainty in the initial lump sum which would arise had we been considering the situation part of the way through a policy's lifetime.) The equity fund is, perhaps not surprisingly, well above the other two in terms of risk, but will also attract a reasonable risk premium. It is also likely that a fixed interest investment attracts a small risk premium over an index-linked investment so at later durations the ordering of the risks is in the order we might expect.

4 References

Aebi, M., Embrechts, P., and Mikosch, T. (1994) Stochastic discounting, aggregate claims and Mallows metric. *Advances in Applied Probability* **26** , 183-206.

Albrecht, P. (1985) A note on immunization under a general stochastic equilibrium model of the term structure. *Insurance: Mathematics and Economics* **4**, 239-245.

Beekman, J.A., and Shiu, E.S.W. (1988) Stochastic models for bond prices, function space integrals and immunization theory. *Insurance: Mathematics and Economics* **7**, 163-173.

Black, F. and Jones, R. (1987) Simplifying portfolio insurance for corporate pension plans. *Journal of Portfolio Management* **Summer**, 33-37.

Black, F. and Perold, A. (1992) Theory of constant portfolio insurance. *Journal of Economics and Control* **16**, 403-426.

Boyle, P.P. (1978) Immunization under stochastic models of the term structure. *Journal of the Institute of Actuaries* **105**, 177-188.

Boyle, P.P. (1980) Recent models of the term structure of interest rates with actuarial applications. *21st International Congress of Actuaries* , 95-104.

Brennan, M.J., and Schwarz, E.S. (1979) A continuous time approach to the pricing of bonds. *Journal of Banking and Finance* **3**, 135-155.

Cairns, A.J.G. (1995) Stochastic pension fund modelling in continuous time. *In preparation* , .

Cairns, A.J.G. and Parker, G. (1995) Stochastic pension fund modelling. *In preparation* , .

Chan, T. (1994) Some applications of Lévy processes to stochastic investment models for actuarial use. *submitted to ASTIN Bulletin* , .

Cox, J.C., Ingersoll, J.E., and Ross, S.A. (1985) A theory of the term structure of interest rates. *Econometrica* **53**, 385-407.

Dhaene, J. (1989) Stochastic interest rates and autoregressive integrated moving average processes. *ASTIN Bulletin* **19**, 131-138.

Dufresne, D. (1988) Moments of pension contributions and fund levels when rates of return are random. *Journal of the Institute of Actuaries* **115**, 535-544.

Dufresne, D. (1989a) Stability of pension systems when rates of return are random. *Insurance: Mathematics and Economics* **8**, 71-76.

Dufresne, D. (1989b) Weak convergence of random growth processes with applications to insurance. *Insurance: Mathematics and Economics* **8**, 187-201.

Dufresne, D. (1990) The distribution of a perpetuity, with applications to risk theory and pension funding. *Scandinavian Actuarial Journal* , 39-79.

Dufresne, D. (1992) On discounting when rates of return are random. *24th International*

Congress of Actuaries, Montreal **1**, 27-44.

Heath, D., Jarrow, R., and Morton, A. (1990) Bond pricing and the term structure of interest rates: a discrete time approximation. *Journal of Financial and Quantitative Analysis* **25**, 419-440.

Haberman, S. (1992) Pension funding with time delays: a stochastic approach. *Insurance: Mathematics and Economics* **11**, 179-189.

Haberman, S. (1993a) Pension funding with time delays and autoregressive rates of investment return. *Insurance: Mathematics and Economics* **13**, 45-56.

Haberman, S. (1993b) Pension funding: the effect of changing the frequency of valuations. *Insurance: Mathematics and Economics* **13**, 263-270.

Haberman, S. (1994) Autoregressive rates of return and the variability of pension fund contributions and fund levels for a defined benefit pension scheme. *Insurance: Mathematics and Economics* **14**, 219-240.

Heath, D., Jarrow, R., and Morton, A. (1992) Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation. *Econometrica* **60**, 77-105.

Karlin, S. and Taylor, H.M. (1975) *A first course in stochastic processes: 2nd Edition.* Academic Press, New York.

Knox, D. (1993) A critique of defined contribution using a simulation approach. Research paper number 7, Centre for Actuarial Studies, The University of Melbourne.

Longstaff, F.A., and Schwartz, E.S. (1992) Interest rate volatility and the term structure: a two factor general equilibrium model. *The Journal of Finance* **47**, 1259-1282.

Norberg, R., and Møller, C.M. (1994) Thiele's differential equation by stochastic interest of diffusion type. *to appear in Scandinavian Actuarial Journal* , .

Papachristou, D., and Waters, H.R. (1991) Some remarks concerning interest rates in relation to long term insurance policies. *Scandinavian Actuarial Journal* **2**, 103-117.

Parker, G. (1993a) Two stochastic approaches for discounting actuarial functions. *Proceedings of the XXIV ASTIN Colloquium* , 367-389.

Parker, G. (1993b) Distribution of the present value of future cash flows. *3rd AFIR, Rome* **2**, 831-843.

Parker, G. (1994a) Stochastic analysis of an insurance portfolio. *4th AFIR, Orlando* **1**, 49-66.

Parker, G. (1994b) Moments of the present value of a portfolio of policies. *to appear in Scandinavian Actuarial Journal* , .

Reitano, R.R. (1991a) Multivariate duration analysis. *Transactions of the Society of Actuaries* **43**, 335-391.

Sercu, P. (1991) Bond options and bond portfolio insurance. *Insurance: Mathematics and Economics* **10**, 203-230.

Waters, H.R. (1978) The moments and distributions of actuarial functions. *Journal of the Institute of Actuaries* **105**, 61-75.

Wilkie, A.D. (1987) Stochastic investment models – theory and practice. *Insurance: Mathematics and Economics* **6**, 65-83.

Wilkie, A.D. (1992) Stochastic investment models for XXIst century actuaries. *24th International Congress of Actuaries, Montreal* **5**, 119-137.

Wilkie, A.D. (1994) Stochastic modelling of long-term investment risks. *submitted to the IMA Journal of Mathematics Applied in Business and Finance* , .

Zimbidis, A. and Haberman, S. (1993) Delay, feedback and variability of pension contributions and fund levels. *Insurance: Mathematics and Economics* **13**, 271-285.

5 Appendix

$$\begin{aligned}
S(t) &= \text{salary at time } t \\
F_e(t) &= \text{equities fund at time } t \\
F_{il}(t) &= \text{index-linked fund at time } t \\
\delta_s(t) &= \log[S(t)/S(t-1)] \\
\delta_e(t) &= \log[F_e(t)/F_e(t-1)] \\
\delta_{il}(t) &= \log[F_{il}(t)/F_{il}(t-1)] \\
\\
\text{with } \delta_s(t) &= \delta_p(t) + \delta_{rs}(t) \\
\delta_e(t) &= \delta_p(t) + \delta_{rs}(t) + \delta_{re}(t) \\
\delta_{il}(t) &= \delta_p(t) + \delta_{ril}(t) \\
\\
\delta_p(t) &= \text{force of price inflation between } t-1 \text{ and } t \\
&= \delta_p + \alpha_p(\delta_p(t-1) - \delta_p) + \sigma_p Z_p(t) \\
\delta_{rs}(t) &= \text{real salary growth rate} \\
&= \delta_{rs} + \alpha_{rs}(\delta_{rs}(t-1) - \delta_{rs}) + \sigma_{rs} Z_{rs}(t) \\
\delta_{re}(t) &= \text{real equities rate of return over salaries} \\
&= \delta_{re} + \sigma_{re} Z_{re}(t) \\
\delta_{ril}(t) &= \text{real index linked return} \\
&= \delta_{ril} + \alpha_{ril}(\delta_{ril}(t-1) - \delta_{ril}) + \sigma_{ril} Z_{ril}(t)
\end{aligned}$$

where $Z_p(t)$, $Z_{rs}(t)$, $Z_{re}(t)$ and $Z_{ril}(t)$ (for $t = 0, 1, 2, \ldots$) are independent and identically distributed sequences of standard Normal random variables.

Now let

$$\begin{aligned}
y_p(t) &= \sum_{s=1}^{t} \delta_p(s) \\
y_{rs}(t) &= \sum_{s=1}^{t} \delta_{rs}(s) \\
y_{re}(t) &= \sum_{s=1}^{t} \delta_{re}(s) \\
y_{ril}(t) &= \sum_{s=1}^{t} \delta_{ril}(s) \\
\text{Then } E[y_p] &= \delta_p.t \\
Var[y_p(t)] &= \frac{\sigma_p^2}{(1-\alpha_p)^2}\left[t - \frac{2\alpha_p(1-\alpha_p^t)}{(1-\alpha_p)} + \frac{\alpha_p^2(1-\alpha_p^{2t})}{(1-\alpha_p^2)}\right] \\
E[y_{rs}] &= \delta_{rs}.t \\
Var[y_{rs}(t)] &= \frac{\sigma_{rs}^2}{(1-\alpha_{rs})^2}\left[t - \frac{2\alpha_{rs}(1-\alpha_{rs}^t)}{(1-\alpha_{rs})} + \frac{\alpha_{rs}^2(1-\alpha_{rs}^{2t})}{(1-\alpha_{rs}^2)}\right] \\
E[y_{re}(t)] &= \delta_{re}.t \\
Var[y_{re}(t)] &= \sigma_{re}^2.t \\
E[y_{ril}] &= \delta_{ril}.t \\
Var[y_{ril}(t)] &= \frac{\sigma_{ril}^2}{(1-\alpha_{ril})^2}\left[t - \frac{2\alpha_{ril}(1-\alpha_{ril}^t)}{(1-\alpha_{ril})} + \frac{\alpha_{ril}^2(1-\alpha_{ril}^{2t})}{(1-\alpha_{ril}^2)}\right]
\end{aligned}$$

We also define

$$\begin{aligned}
F_e(t) &= \exp[y_p(t) + y_{re}(t)] \\
F_{il}(t) &= \exp[y_p(t) + y_{ril}(t)] \\
S(t) &= \exp[y_p(t) + y_{rs}(t)]
\end{aligned}$$

We are interested in the three quantities

$$\begin{aligned}
L_1 &= L/S(t) \\
L_2 &= F_{il}(t)/S(t) \\
L_3 &= F_e(t)/S(t)
\end{aligned}$$

Of particular interest is the level of risk associated with each option which we measure by taking the variance of the logarithm of each quantity.

$$\begin{aligned}
Var[\log L_1] &= Var[y_p(t)] + Var[y_{rs}(t)] \\
Var[\log L_2] &= Var[y_{rs}(t)] + Var[y_{ril}(t)] \\
Var[\log L_3] &= Var[y_{re}(t)]
\end{aligned}$$

These variances are described in the main text.

Parameter values

type, θ	δ_θ	α_θ	σ_θ^2
prices, *p*	0.05	0.7	0.05^2
real salary, *rs*	0.02	0.4	0.03^2
real index-linked, *ril*	0.036	-0.5	0.13^2
real equity, *re*	0.036		0.26^2

NBER WORKING PAPER SERIES

A SURVEY OF BEHAVIORAL FINANCE

Nicholas Barberis
Richard Thaler

Working Paper 9222
http://www.nber.org/papers/w9222

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2002

Forthcoming in the *Handbook of the Economics of Finance*, edited by George Constantinides, Milt Harris and Rene Stulz. We are very grateful to Markus Brunnermeier, George Constantinides, Kent Daniel, Milt Harris, Ming Huang, Owen Lamont, Jay Ritter, Andrei Shleifer, Jeremy Stein and Tuomo Vuolteenaho for extensive comments. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

A Survey of Behavioral Finance
Nicholas Barberis and Richard Thaler
NBER Working Paper No. 9222
September 2002
JEL No. G11, G12, G30

ABSTRACT

Behavioral finance argues that some financial phenomena can plausibly be understood using models in which some agents are not fully rational. The field has two building blocks: limits to arbitrage, which argues that it can be difficult for rational traders to undo the dislocations caused by less rational traders; and psychology, which catalogues the kinds of deviations from full rationality we might expect to see. We discuss these two topics, and then present a number of behavioral finance applications: to the aggregate stock market, to the cross-section of average returns, to individual trading behavior, and to corporate finance. We close by assessing progress in the field and speculating about its future course.

Nicholas Barberis
Graduate School of Business
University of Chicago
1101 East 58th Street
Chicago, IL 60637
and NBER
nick.barberis@gsb.uchicago.edu

Richard Thaler
Graduate School of Business
University of Chicago
1101 East 58th Street
Chicago, IL 60637
and NBER
richard.thaler@gsbpop.uchicago.edu

1 Introduction

The traditional finance paradigm, which underlies many of the other articles in this handbook, seeks to understand financial markets using models in which agents are "rational". Rationality means two things. First, when they receive new information, agents update their beliefs correctly, in the manner described by Bayes' law. Second, given their beliefs, agents make choices that are normatively acceptable, in the sense that they are consistent with Savage's notion of Subjective Expected Utility (SEU).

This traditional framework is appealingly simple, and it would be very satisfying if its predictions were confirmed in the data. Unfortunately, after years of effort, it has become clear that basic facts about the aggregate stock market, the cross-section of average returns and individual trading behavior are not easily understood in this framework.

Behavioral finance is a new approach to financial markets that has emerged, at least in part, in response to the difficulties faced by the traditional paradigm. In broad terms, it argues that some financial phenomena can be better understood using models in which some agents are *not* fully rational. More specifically, it analyzes what happens when we relax one, or both, of the two tenets that underlie individual rationality. In some behavioral finance models, agents fail to update their beliefs correctly. In other models, agents apply Bayes' law properly but make choices that are normatively questionable, in that they are incompatible with SEU.[1]

This review essay evaluates recent work in this rapidly growing field. In Section 2, we consider the classic objection to behavioral finance, namely that even if some agents in the economy are less than fully rational, rational agents will prevent them from influencing security prices for very long, through a process known as arbitrage. One of the biggest successes of behavioral finance is a series of theoretical papers showing that in an economy where rational and irrational traders interact, irrationality *can* have a substantial and long-lived impact on prices. These papers, known as the literature on "limits to arbitrage," form

[1] It is important to note that most models of asset pricing use the Rational Expectations Equilibrium framework (REE), which assumes not only individual rationality but also *consistent beliefs* (Sargent, 1993). Consistent beliefs means that agents' beliefs are correct: the subjective distribution they use to forecast future realizations of unknown variables is indeed the distribution that those realizations are drawn from. This requires not only that agents process new information correctly, but that they have *enough* information about the structure of the economy to be able to figure out the correct distribution for the variables of interest.

Behavioral finance departs from REE by relaxing the assumption of individual rationality. An alternative departure is to retain individual rationality but to relax the consistent beliefs assumption: while investors apply Bayes' law correctly, they lack the information required to know the actual distribution variables are drawn from. This line of research is sometimes referred to as the literature on bounded rationality, or on structural uncertainty. For example, a model in which investors do not know the growth rate of an asset's cash flows but learn it as best as they can from available data, would fall into this class. Although the literature we discuss also uses the term bounded rationality, the approach is quite different.

one of the two buildings blocks of behavioral finance.

To make sharp predictions, behavioral models often need to specify the form of agents' irrationality. How exactly do people misapply Bayes law or deviate from SEU? For guidance on this, behavioral economists typically turn to the extensive experimental evidence compiled by cognitive psychologists on the biases that arise when people form *beliefs*, and on people's *preferences*, or on how they make decisions, given their beliefs. Psychology is therefore the second building block of behavioral finance, and we review the psychology most relevant for financial economists in Section 3.[2]

In Sections 4-8, we consider specific applications of behavioral finance: to understanding the aggregate stock market, the cross-section of average returns, and the pricing of closed-end funds in Sections 4, 5 and 6 respectively; to understanding how particular groups of investors choose their portfolios and trade over time in Section 7; and to understanding the financing and investment decisions of firms in Section 8. Section 9 takes stock and suggests directions for future research.[3]

2 Limits to Arbitrage

2.1 Market Efficiency

In the traditional framework where agents are rational and there are no frictions, a security's price equals its "fundamental value." This is the discounted sum of expected future cash flows, where in forming expectations, investors correctly process all available information, and where the discount rate is consistent with a normatively acceptable preference specification. The hypothesis that actual prices reflect fundamental values is the Efficient Markets Hypothesis (EMH). Put simply, under this hypothesis, "prices are right," in that they are set by agents who understand Bayes' law and have sensible preferences. In an efficient market, there is "no free lunch": no investment strategy can earn excess risk-adjusted average returns, or average returns greater than are warranted for its risk.

Behavioral finance argues that some features of asset prices are most plausibly interpreted as deviations from fundamental value, and that these deviations are brought about by the presence of traders who are not fully rational. A long-standing objection to this view that goes back to Friedman (1953) is that rational traders will quickly undo any dislocations

[2] The idea, now widely adopted, that behavioral finance rests on the two pillars of limits to arbitrage and investor psychology is originally due to Shleifer and Summers (1990).

[3] We draw readers' attention to two other recent surveys of behavioral finance. Shleifer (2000) provides a particularly detailed discussion of the theoretical and empirical work on limits to arbitrage, which we summarize in Section 2. Hirshleifer's (2001) survey is closer to ours in terms of material covered, although we devote less space to asset pricing, and more to corporate finance and individual investor behavior. We also organize the material somewhat differently.

caused by irrational traders. To illustrate the argument, suppose that the fundamental value of a share of Ford is $20. Imagine that a group of irrational traders becomes excessively pessimistic about Ford's future prospects and through its selling, pushes the price to $15. Defenders of the EMH argue that rational traders, sensing an attractive opportunity, will buy the security at its bargain price and at the same time, hedge their bet by shorting a "substitute" security, such as General Motors, that has similar cash flows to Ford in future states of the world. The buying pressure on Ford shares will then bring their price back to fundamental value.

Friedman's line of argument is initially compelling, but it has not survived careful theoretical scrutiny. In essence, it is based on two assertions. First, as soon as there is a deviation from fundamental value – in short, a mispricing – an attractive investment opportunity is created. Second, rational traders will immediately snap up the opportunity, thereby correcting the mispricing. Behavioral finance does not take issue with the second step in this argument: when attractive investment opportunities come to light, it is hard to believe that they are not quickly exploited. Rather, it disputes the first step. The argument, which we elaborate on in Sections 2.2 and 2.3., is that even when an asset is wildly mispriced, strategies designed to correct the mispricing can be both risky and costly, rendering them unattractive. As a result, the mispricing can remain unchallenged.

It is interesting to think about common finance terminology in this light. While irrational traders are often known as "noise traders," rational traders are typically referred to as "arbitrageurs." Strictly speaking, an arbitrage is an investment strategy that offers riskless profits at no cost. Presumably, the rational traders in Friedman's fable became known as arbitrageurs because of the belief that a mispriced asset immediately creates an opportunity for riskless profits. Behavioral finance argues that this is *not* true: the strategies that Friedman would have his rational traders adopt are not necessarily arbitrages; quite often, they are very risky.

An immediate corollary of this line of thinking is that "prices are right" and "there is no free lunch" are *not* equivalent statements. While both are true in an efficient market, "no free lunch" can also be true in an inefficient market: just because prices are away from fundamental value does not necessarily mean that there are any excess risk-adjusted average returns for the taking. In other words,

$$\boxed{\text{“prices are right”} \Rightarrow \text{“no free lunch”}}$$

but

$$\boxed{\text{“no free lunch”} \nRightarrow \text{“prices are right”}.}$$

This distinction is important for evaluating the ongoing debate on market efficiency. First, many researchers still point to the inability of professional money managers to beat the market as strong evidence of market efficiency (Rubinstein, 2000, Ross, 2001). Underlying this argument, though, is the assumption that "no free lunch" implies "prices are right." If, as we argue in Sections 2.2 and 2.3., this link is broken, the performance of money managers tells us nothing about whether prices reflect fundamental value.

Second, while some researchers accept that there is a distinction between "prices are right" and "there is no free lunch," they believe that the debate should be more about the latter statement than about the former. We disagree with this emphasis. As economists, our ultimate concern is that capital be allocated to the most promising investment opportunities. Whether this is true or not depends much more on whether prices are right than on whether there are any free lunches for the taking.

2.2 Theory

In the previous section, we emphasize the idea that when a mispricing occurs, strategies designed to correct it can be both risky and costly, thereby allowing the mispricing to survive. Here we discuss some of the risks and costs that have been identified. In our discussion, we return to the example of Ford, whose fundamental value is \$20, but which has been pushed down to \$15 by pessimistic noise traders.

Fundamental Risk

The most obvious risk an arbitrageur faces if he buys Ford's stock at \$15 is that a piece of bad news about Ford's fundamental value causes the stock to fall further, leading to losses. Of course, arbitrageurs are well aware of this risk, which is why they short a substitute security such as General Motors at the same time that they buy Ford. The problem is that substitute securities are rarely perfect, and often highly imperfect, making it impossible to remove all the fundamental risk. Shorting General Motors protects the arbitrageur somewhat from adverse news about the car industry as a whole, but still leaves him vulnerable to news that is specific to Ford – news about defective tires, say.[4]

Noise Trader Risk

Noise trader risk, an idea introduced by De Long et al. (1990a) and studied further by Shleifer and Vishny (1997), is the risk that the mispricing being exploited by the arbitrageur worsens in the short run. Even if General Motors is a *perfect* substitute security for Ford, the arbitrageur still faces the risk that the pessimistic investors causing Ford to be undervalued in the first place become even more pessimistic, lowering its price even further. Once one has granted the possibility that a price can be different from its fundamental value, then one must also grant the possibility that future price movements will increase the divergence.

[4] Another problem is that even if a substitute security exists, it may itself be mispriced. This can happen in situations involving industry-wide mispricing: in that case, the only stocks with similar future cash flows to the mispriced one are themselves mispriced.

Noise trader risk matters because it can force arbitrageurs to liquidate their positions early, bringing them potentially steep losses. To see this, note that most real-world arbitrageurs – in other words, professional portfolio managers – are not managing their own money, but rather managing money for other people. In the words of Shleifer and Vishny (1997), there is "a separation of brains and capital."

This agency feature has important consequences. Investors, lacking the specialized knowledge to evaluate the arbitrageur's strategy, may simply evaluate him based on his returns. If a mispricing that the arbitrageur is trying to exploit worsens in the short run, generating negative returns, investors may decide that he is incompetent, and withdraw their funds. If this happens, the arbitrageur will be forced to liquidate his position prematurely. Fear of such premature liquidation makes him less aggressive in combating the mispricing in the first place.

These problems can be severely exacerbated by creditors. After poor short-term returns, creditors, seeing the value of their collateral erode, will call their loans, again triggering premature liquidation.

In these scenarios, the forced liquidation is brought about by the worsening of the mispricing itself. This need not always be the case. For example, in their efforts to remove fundamental risk, many arbitrageurs sell securities short. Should the original owner of the borrowed security want it back, the arbitrageur may again be forced to close out his position if he cannot find other shares to borrow. The risk that this occurs during a temporary worsening of the mispricing makes the arbitrageur more cautious from the start.

Implementation Costs

Well-understood transaction costs such as commissions, bid-ask spreads and price impact can make it less attractive to exploit a mispricing. Since shorting is often essential to the arbitrage process, we also include short-sales constraints in the implementation costs category. These refer to anything that makes it less attractive to establish a short position than a long one. The simplest such constraint is the fee charged for borrowing a stock. In general these fees are small – D'Avolio (2002) finds that for most stocks, they range between 10 and 15 basis points – but they can be much larger; in some cases, arbitrageurs may not be able to find shares to borrow at *any* price. Other than the fees themselves, there can be legal constraints: for a large fraction of money managers – many pension fund and mutual fund managers in particular – short-selling is simply not allowed.[5]

[5] The presence of per-period transaction costs like lending fees can expose arbitrageurs to another kind of risk, *horizon risk*, which is the risk that the mispricing takes so long to close that any profits are swamped by the accumulated transaction costs. This applies even when the arbitrageur is certain that no outside party will force him to liquidate early. Abreu and Brunnermeier (2002) study a particular type of horizon risk, which they label *synchronization risk*. Suppose that the elimination of a mispricing requires the participation of a sufficiently large number of separate arbitrageurs. Then in the presence of per-period transaction costs,

We also include in this category the cost of finding and learning about a mispricing, as well as the cost of the resources needed to exploit it (Merton, 1987). Finding mispricing, in particular, can be a tricky matter. It was once thought that if noise traders influenced stock prices to any substantial degree, their actions would quickly show up in the form of predictability in returns. Shiller (1984) and Summers (1986) demonstrate that this argument is completely erroneous, with Shiller (1984) calling it "one of the most remarkable errors in the history of economic thought." They show that even if noise trader demand is so strong as to cause a large and persistent mispricing, it may generate so little predictability in returns as to be virtually undetectable.

In contrast, then, to straightforward-sounding textbook arbitrage, real world arbitrage entails both costs and risks, which under some conditions will limit arbitrage and allow deviations from fundamental value to persist. To see what these conditions are, consider two cases.

Suppose first that the mispriced security does *not* have a close substitute. By definition then, the arbitrageur is exposed to fundamental risk. In this case, sufficient conditions for arbitrage to be limited are (i) that arbitrageurs are risk averse and (ii) that the fundamental risk is systematic, in that it cannot be diversified by taking many such positions. Condition (i) ensures that the mispricing will not be wiped out by a single arbitrageur taking a large position in the mispriced security. Condition (ii) ensures that the mispricing will not be wiped out by a large number of investors each adding a *small* position in the mispriced security to their current holdings. The presence of noise trader risk or implementation costs will only limit arbitrage further.

Even if a perfect substitute does exist, arbitrage can still be limited. The existence of the substitute security immunizes the arbitrageur from fundamental risk. We can go further and assume that there are no implementation costs, so that only noise trader risk remains. De Long et al. (1990a) show that noise trader risk is powerful enough, that even with this single form of risk, arbitrage can sometimes be limited. The sufficient conditions are similar to those above, with one important difference. Here arbitrage will be limited if: (i) arbitrageurs are risk averse *and have short horizons* and (ii) the noise trader risk is systematic. As before, condition (i) ensures that the mispricing cannot be wiped out by a single, large arbitrageur, while condition (ii) prevents a large number of small investors from exploiting the mispricing. The central contribution of Shleifer and Vishny (1997) is to point out the real world relevance of condition (i): the possibility of an early, forced liquidation means that many arbitrageurs effectively have short horizons.

In the presence of certain implementation costs, conditions (ii) may not even be necessary.

arbitrageurs may hesitate to exploit the mispricing because they don't know how many *other* arbitrageurs have heard about the opportunity, and therefore how long they will have to wait before prices revert to correct values.

If it is costly to learn about a mispricing, or the resources required it to exploit it are expensive, that may be enough to explain why a large number of different individuals do not intervene in an attempt to correct the mispricing.

It is also important to note that for particular types of noise trading, arbitrageurs may prefer to trade in the *same* direction as the noise traders, thereby exacerbating the mispricing, rather than against them. For example, De Long et al. (1990b) consider an economy with positive feedback traders, who buy more of an asset this period if it performed well last period. If these noise traders push an asset's price above fundamental value, arbitrageurs do not sell or short the asset. Rather, they *buy* it, knowing that the earlier price rise will attract more feedback traders next period, leading to still higher prices, at which point the arbitrageurs can exit at a profit.

So far, we have argued that it is not easy for arbitrageurs like hedge funds to exploit market inefficiencies. However, hedge funds are not the only market participants trying to take advantage of noise traders: firm managers also play this game. If a manager believes that investors are overvaluing his firm's shares, he can benefit the firm's existing shareholders by issuing extra shares at attractive prices. The extra supply this generates could potentially push prices back to fundamental value.

Unfortunately, this game entails risks and costs for managers, just as it does for hedge funds. Issuing shares is an expensive process, both in terms of underwriting fees and time spent by company management. Moreover, the manager can rarely be *sure* that investors are overvaluing his firm's shares. If he issues shares, thinking that they are overvalued when in fact they are not, he incurs the costs of deviating from his target capital structure, without getting any benefits in return.

2.3 Evidence

From the theoretical point of view, there is reason to believe that arbitrage is a risky process and therefore that it is only of limited effectiveness. But is there any *evidence* that arbitrage is limited? In principle, any example of persistent mispricing is immediate evidence of limited arbitrage: if arbitrage were not limited, the mispricing would quickly disappear. The problem is that while many pricing phenomena can be interpreted as deviations from fundamental value, it is only in a few cases that the presence of a mispricing can be established beyond any reasonable doubt. The reason for this is what Fama (1970) dubbed the "joint hypothesis problem." In order to claim that the price of a security differs from its properly discounted future cash flows, one needs a model of "proper" discounting. Any test of mispricing is therefore inevitably a *joint* test of mispricing and of a model of discount rates, making it difficult to provide definitive evidence of inefficiency.

In spite of this difficulty, researchers have uncovered a number of financial market phenomena that are almost certainly mispricings, and persistent ones at that. These examples show that arbitrage is indeed limited, and also serve as interesting illustrations of the risks and costs described earlier.

Twin Shares

In 1907, Royal Dutch and Shell Transport, at the time completely independent companies, agreed to merge their interests on a 60:40 basis while remaining separate entities. Shares of Royal Dutch, which are primarily traded in the U.S. and in the Netherlands, are a claim to 60 percent of the total cash flow of the two companies, while Shell, which trades primarily in the U.K., is a claim to the remaining 40 percent. If prices equal fundamental value, the market value of Royal Dutch equity should always be 1.5 times the market value of Shell equity. Remarkably, it isn't.

Figure 1, taken from Froot and Dabora's (1999) analysis of this case, shows the ratio of Royal Dutch equity value to Shell equity value relative to the efficient markets benchmark of 1.5. The picture provides strong evidence of a persistent inefficiency. Moreover, the deviations are not small. Royal Dutch is sometimes 35 percent underpriced relative to parity, and sometimes 15 percent overpriced.

This evidence of mispricing is simultaneously evidence of limited arbitrage, and it is not hard to see why arbitrage might be limited in this case. If an arbitrageur wanted to exploit this phenomenon – and several hedge funds, Long Term Capital Management included, did try to – he would buy the relatively undervalued share and short the other. Table 1 summarizes the risks facing the arbitrageur. Since one share is a good substitute for the other, fundamental risk is nicely hedged: news about fundamentals should affect the two shares equally, leaving the arbitrageur immune. Nor are there any major implementation costs to speak of: shorting shares of either company is an easy matter.

The one risk that remains is noise trader risk. Whatever investor sentiment is causing one share to be undervalued relative to the other could also cause that share to become *even more* undervalued in the short term. The graph shows that this danger is very real: an arbitrageur buying a 10 percent undervalued Royal Dutch share in March 1983 would have seen it drop still further in value over the next six months. As discussed earlier, when a mispriced security has a perfect substitute, arbitrage can still be limited if (i) arbitrageurs are risk averse and have short horizons and (ii) the noise trader risk is systematic, or the arbitrage requires specialized skills, or there are costs to learning about such opportunities. It is very plausible that both (i) and (ii) are true, thereby explaining why the mispricing persisted for so long. It took until 2001 for the shares to finally sell at par.

This example also provides a nice illustration of the distinction between "prices are right" and "no free lunch" discussed in Section 2.1. While prices in this case are clearly *not* right, there are no easy profits for the taking.

Index Inclusions

Every so often, one of the companies in the S&P 500 is taken out of the index because of a merger or bankruptcy, and is replaced by another firm. Two early studies of such index inclusions, Harris and Gurel (1986) and Shleifer (1986), document a remarkable fact: when a stock is added to the index, it jumps in price by an average of 3.5 percent, and much of this jump is permanent. In one dramatic illustration of this phenomenon, when Yahoo was added to the index, its shares jumped by 24 percent in a single day.

The fact that a stock jumps in value upon inclusion is once again clear evidence of mispricing: the price of the share changes even though its fundamental value does not. Standard and Poor's emphasizes that in selecting stocks for inclusion, they are simply trying to make their index representative of the U.S. economy, not to convey any information about the level or riskiness of a firm's future cash flows.[6]

This example of a deviation from fundamental value is also evidence of limited arbitrage. When one thinks about the risks involved in trying to exploit the anomaly, its persistence becomes less surprising. An arbitrageur needs to short the included security and to buy as good a substitute security as he can. This entails considerable fundamental risk because individual stocks rarely have good substitutes. It also carries substantial noise trader risk: whatever caused the initial jump in price – in all likelihood, buying by S&P 500 index funds – may continue, and cause the price to rise still further in the short run; indeed, Yahoo went from $115 prior to its S&P inclusion announcement to $210 a month later.

Wurgler and Zhuravskaya (2002) provide additional support for the limited arbitrage view of S&P 500 inclusions. They hypothesize that the jump upon inclusion should be particularly large for those stocks with the worst substitute securities, in other words, for those stocks for which the arbitrage is riskiest. By constructing the best possible substitute portfolio for each included stock, they are able to test this, and find strong support. Their analysis also shows just how hard it is to find good substitute securities for individual stocks. For most regressions of included stock returns on the returns of the best substitute securities, the R^2 is below 25 percent.

Internet Carve-Outs

In March 2000, 3Com sold 5 percent of its wholly owned subsidiary Palm Inc. in an initial public offering, retaining ownership of the remaining 95 percent. After the IPO, a

[6] After the initial studies on index inclusions appeared, some researchers argued that the price increase might be rationally explained through information or liquidity effects. While such explanations cannot be completely ruled out, the case for mispricing was considerably strengthened by Kaul, Mehrotra and Morck (2000). They consider the case of the TS300 index of Canadian equities, which in 1996 changed the weights of some of its component stocks to meet an innocuous regulatory requirement. The reweighting was accompanied by significant price effects. Since the affected stocks were *already* in the index at the time of the event, information and liquidity explanations for the price jumps are extremely implausible.

shareholder of 3Com indirectly owned 1.5 shares of Palm. 3Com also announced its intention to spin off the remainder of Palm within 9 months, at which time they would give each 3Com shareholder 1.5 shares of Palm.

At the close of trading on the first day after the IPO, Palm shares stood at $95, putting a lower bound on the value of 3Com at $142. In fact, 3Com's price was $81, implying a market valuation of 3Com's substantial businesses outside of Palm of about -$60 per share!

This situation surely represents a severe mispricing, and it persisted for several weeks. To exploit it, an arbitrageur could buy one share of 3Com, short 1.5 shares of Palm, and wait for the spin-off, thus earning certain profits at no cost. This strategy entails no fundamental risk and no noise trader risk. Why, then, is arbitrage limited? Lamont and Thaler (2002), who analyze this case in detail, argue that implementation costs played a major role. Many investors who tried to borrow Palm shares to short were either told by their broker that no shares were available, or else were quoted a very high borrowing price. This barrier to shorting was not a legal one, but one that arose endogenously in the marketplace: such was the demand for shorting Palm, that the supply of Palm shorts was unable to meet it. Arbitrage was therefore limited, and the mispricing persisted.[7]

Some financial economists react to these examples by arguing that they are simply isolated instances with little broad relevance.[8] We think this is an overly complacent view. The "twin shares" example illustrates that in situations where arbitrageurs face only one type of risk – noise trader risk – securities can become mispriced by almost 35 percent. This suggests that if a typical stock trading on the NYSE or NASDAQ becomes subject to investor sentiment, the mispricing could be an order of magnitude larger. Not only would arbitrageurs face noise trader risk in trying to correct the mispricing, but fundamental risk as well, not to mention implementation costs.

3 Psychology

The theory of limited arbitrage shows that if irrational traders cause deviations from fundamental value, rational traders will often be powerless to do anything about it. In order to say more about the structure of these deviations, behavioral models often assume a specific form of irrationality. For guidance on this, economists turn to the extensive experimental evidence compiled by cognitive psychologists on the systematic biases that arise when people

[7] See also Mitchell, Pulvino and Stafford (2002) and Ofek and Richardson (2001) for further discussion of such "negative stub" situations, in which the market value of a company is less than the sum of its publicly traded parts.

[8] During a discussion of these issues at a University of Chicago seminar, one economist argued that these examples are "the tip of the iceberg," to which another retorted that "they *are* the iceberg."

form *beliefs*, and on people's *preferences*.[9]

In this section, we summarize the psychology that may be of particular interest to financial economists. Our discussion of each finding is necessarily brief. For a deeper understanding of the phenomena we touch on, we refer the reader to the surveys of Camerer (1995) and Rabin (1998) and to the edited volumes of Kahneman, Slovic and Tversky (1982), Kahneman and Tversky (2000) and Gilovich, Griffin and Kahneman (2002).

3.1 Beliefs

A crucial component of any model of financial markets is a specification of how agents form expectations. We now summarize what psychologists have learned about how people appear to form beliefs in practice.

Overconfidence. Extensive evidence shows that people are overconfident in their judgments. This appears in two guises. First, the confidence intervals people assign to their estimates of quantities – the level of the Dow in a year, say – are far too narrow. Their 98 percent confidence intervals, for example, include the true quantity only about 60 percent of the time (Alpert and Raiffa, 1982). Second, people are poorly calibrated when estimating probabilities: events they think are certain to occur actually occur only around 80 percent of the time, and events they deem impossible occur approximately 20 percent of the time (Fischhoff, Slovic and Lichtenstein, 1977).[10]

Optimism and Wishful Thinking. Most people display unrealistically rosy views of their abilities and prospects (Weinstein, 1980). Typically, over 90 percent of those surveyed think they are above average in such domains as driving skill, ability to get along with people and sense of humor. They also display a systematic planning fallacy: they predict that tasks (such as writing survey papers) will be completed much sooner than they actually are (Buehler, Griffin and Ross, 1994).

Representativeness. Kahneman and Tversky (1974) show that when people try to deter-

[9] We emphasize, however, that behavioral models do not *need* to make extensive psychological assumptions in order to generate testable predictions. In Section 6, we discuss Lee, Shleifer and Thaler's (1991) theory of closed-end fund pricing. That theory makes numerous crisp predictions using only the assumptions that there are noise traders with correlated sentiment in the economy, and that arbitrage is limited.

[10] Overconfidence may in part stem from two other biases, self-attribution bias and hindsight bias. Self-attribution bias refers to people's tendency to ascribe any success they have in some activity to their own talents, while blaming failure on bad luck, rather than on their ineptitude. Doing this repeatedly will lead people to the pleasing but erroneous conclusion that they are very talented. For example, investors might become overconfident after several quarters of investing success (Gervais and Odean, 2001). Hindsight bias is the tendency of people to believe, after an event has occurred, that they predicted it before it happened. If people think they predicted the past better than they actually did, they may also believe that they can predict the future better than they actually can.

mine the probability that a data set A was generated by a model B, or that an object A belongs to a class B, they often use the representativeness heuristic. This means that they evaluate the probability by the degree to which A reflects the essential characteristics of B.

Much of the time, representativeness is a helpful heuristic, but it can generate some severe biases. The first is *base rate neglect.* To illustrate, Kahneman and Tversky present this description of a person named Linda:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

When asked which of "Linda is a bank teller" (statement A) and "Linda is a bank teller and is active in the feminist movement" (statement B) is more likely, subjects typically assign greater probability to B. This is, of course, impossible. Representativeness provides a simple explanation. The description of Linda *sounds* like the description of a feminist – it is representative of a feminist – leading subjects to pick B. Put differently, while Bayes law says that

$$p(\text{statement B}|\text{description}) = \frac{p(\text{description}|\text{statement B})p(\text{statement B})}{p(\text{description})},$$

people apply the law incorrectly, putting too much weight on $p(\text{description}|\text{statement B})$, which captures representativeness, and too little weight on the base rate, $p(\text{statement B})$.

Representativeness also leads to another bias, *sample size neglect.* When judging the likelihood that a data set was generated by a particular model, people often fail to take the size of the sample into account: after all, a small sample can be just as representative as a large one. Six tosses of a coin resulting in three heads and three tails are as representative of a fair coin as 500 heads and 500 tails are in a total of 1000 tosses. Representativeness implies that people will find the two sets of tosses equally informative about the fairness of the coin, even though the second set is much more so.

Sample size neglect means that in cases where people do not initially know the data-generating process, they will tend to infer it too quickly on the basis of too few data points. For instance, they will come to believe that a financial analyst with four good stock picks is talented because four successes are not representative of a bad or mediocre analyst. It also generates a "hot hand" phenomenon, whereby sports fans become convinced that a basketball player who has made three shots in a row is on a hot streak and will score again, even though there is no evidence of a hot hand in the data (Gilovich, Vallone and Tversky, 1985). This belief that even small samples will reflect the properties of the parent population is sometimes known as the "law of small numbers" (Rabin, 2002).

In situations where people *do* know the data-generating process in advance, the law of small numbers generates a gambler's fallacy effect. If a fair coin generates five heads in a row, people will say that "tails are due". Since they believe that even a short sample should be representative of the fair coin, there have to be more tails to balance out the large number of heads.

Conservatism. While representativeness leads to an underweighting of base rates, there are situations where base rates are *over*-emphasized relative to sample evidence. In an experiment run by Edwards (1968), there are two urns, one containing 3 blue balls and 7 red ones, and the other containing 7 blue balls and 3 red ones. A random draw of 12 balls, with replacement, from one of the urns yields 8 reds and 4 blues. What is the probability the draw was made from the first urn? While the correct answer is 0.97, most people estimate a number around 0.7, apparently overweighting the base rate of 0.5.

At first sight, the evidence of conservatism appears at odds with representativeness. However, there may be a natural way in which they fit together. It appears that if a data sample is representative of an underlying model, then people overweight the data. However, if the data is not representative of any salient model, people react too little to the data and rely too much on their priors. In Edwards' experiment, the draw of 8 red and 4 blue balls is not particularly representative of either urn, possibly leading to an overreliance on prior information.

Belief Perseverance. There is much evidence that once people have formed an opinion, they cling to it too tightly and for too long (Lord, Ross and Lepper, 1979). At least two effects appear to be at work. First, people are reluctant to search for evidence that contradicts their beliefs. Second, even if they find such evidence, they treat it with excessive skepticism. Some studies have found an even stronger effect, known as confirmation bias, whereby people misinterpret evidence that goes against their hypothesis as actually being in their favor. In the context of academic finance, belief perseverance predicts that if people start out believing in the Efficient Markets Hypothesis, they may continue to believe in it long after compelling evidence to the contrary has emerged.

Anchoring. Kahneman and Tversky (1974) argue that when forming estimates, people often start with some initial, possibly arbitrary value, and then adjust away from it. Experimental evidence shows that the adjustment is often insufficient. Put differently, people "anchor" too much on the initial value.

In one experiment, subjects were asked to estimate the percentage of United Nations' countries that are African. More specifically, before giving a percentage, they were asked whether their guess was higher or lower than a randomly generated number between 0 and 100. Their subsequent estimates were significantly affected by the initial random number. Those who were asked to compare their estimate to 10, subsequently estimated 25 percent, while those who compared to 60, estimated 45 percent.

Availability Biases. When judging the probability of an event – the likelihood of getting mugged in Chicago, say – people often search their memories for relevant information. While this is a perfectly sensible procedure, it can produce biased estimates because not all memories are equally retrievable or "available", in the language of Kahneman and Tversky (1974). More recent events and more salient events – the mugging of a close friend, say – will weigh more heavily and distort the estimate.

Economists are sometimes wary of this body of experimental evidence because they believe (i) that people, through repetition, will learn their way out of biases; (ii) that experts in a field, such as traders in an investment bank, will make fewer errors; and (iii) that with more powerful incentives, the effects will disappear.

While all these factors can attenuate biases to some extent, there is little evidence that they wipe them out altogether. The effect of learning is often muted by errors of application: when the bias is explained, people often understand it, but then immediately proceed to violate it again in specific applications. Expertise, too, is often a hindrance rather than a help: experts, armed with their sophisticated models, have been found to exhibit *more* overconfidence than laymen, particularly when they receive only limited feedback about their predictions. Finally, in a review of dozens of studies on the topic, Camerer and Hogarth (1999) conclude that while incentives can sometimes reduce the biases people display, "no replicated study has made rationality violations disappear purely by raising incentives" (p.7).

3.2 Preferences

Prospect Theory

An essential ingredient of any model trying to understand asset prices or trading behavior is an assumption about investor preferences, or about how investors evaluate risky gambles. The vast majority of models assume that investors evaluate gambles according to the expected utility framework, EU henceforth. The theoretical motivation for this goes back to Von Neumann and Morgenstern (1947), VNM henceforth, who show that if preferences satisfy a number of plausible axioms – completeness, transitivity, continuity, and independence – then they can be represented by the expectation of a utility function.

Unfortunately, experimental work in the decades after VNM has shown that people systematically violate EU theory when choosing among risky gambles. In response to this, there has been an explosion of work on so-called non-EU theories, all of them trying to do a better job of matching the experimental evidence. Some of the better known models include weighted-utility theory (Chew and MacCrimmon 1979, Chew 1983), implicit EU (Chew 1989, Dekel 1986), disappointment aversion (Gul 1991), regret theory (Bell, 1982, Loomes and Sugden, 1982), rank-dependent utility theories (Quiggin 1982, Segal 1987, 1989, Yaari 1987), and prospect theory (Kahneman and Tversky 1979, Tversky and Kahneman, 1992).

Should financial economists be interested in any of these alternatives to expected utility? It may be that EU theory is a good approximation to how people evaluate a risky gamble like the stock market, even if it does not explain attitudes to the kinds of gambles studied in experimental settings. On the other hand, the difficulty the EU approach has encountered in trying to explain basic facts about the stock market suggests that it may be worth taking a closer look at the experimental evidence. Indeed, recent work in behavioral finance has argued that some of the lessons we learn from violations of EU are central to understanding a number of financial phenomena.

Of all the non-EU theories, prospect theory may be the most promising for financial applications, and we discuss it in detail. The reason we focus on this theory is, quite simply, that it is the most successful at capturing the experimental results. In a way, this is not surprising. Most of the other non-EU models are what might be called quasi-normative, in that they try to capture some of the anomalous experimental evidence by slightly weakening the VNM axioms. The difficulty with such models is that in trying to achieve two goals – normative and descriptive – they end up doing an unsatisfactory job at both. In contrast, prospect theory has no aspirations as a normative theory: it simply tries to capture people's attitudes to risky gambles as parsimoniously as possible. Indeed, Tversky and Kahneman (1986) argue convincingly that normative approaches are doomed to failure, because people routinely make choices that are simply impossible to justify on normative grounds, in that they violate dominance or invariance.

Kahneman and Tversky (1979), KT henceforth, lay out the original version of prospect theory, designed for gambles with at most two non-zero outcomes. They propose that when offered a gamble

$$(x, p; y, q),$$

to be read as "get outcome x with probability p, outcome y with probability q", where $x \leq 0 \leq y$ or $y \leq 0 \leq x$, people assign it a value of

$$\pi(p)v(x) + \pi(q)v(y), \tag{1}$$

where v and π are shown in Figure 2. When choosing between different gambles, they pick the one with the highest value.

This formulation has a number of important features. First, utility is defined over gains and losses rather than over final wealth positions, an idea first proposed by Markowitz (1952). This fits naturally with the way gambles are often presented and discussed in everyday life. More generally, it is consistent with the way people perceive attributes such as brightness, loudness, or temperature relative to earlier levels, rather than in absolute terms. Kahneman and Tversky (1979) also offer the following violation of EU as evidence that people focus on gains and losses. Subjects are asked:[11]

[11] All the experiments in Kahneman and Tversky (1979) are conducted in terms of Israeli currency. The

In addition to whatever you own, you have been given 1000. Now choose between

$$\begin{aligned} A &= (1000, 0.5) \\ B &= (500, 1). \end{aligned}$$

B was the more popular choice. The same subjects were then asked:

In addition to whatever you own, you have been given 2000. Now choose between

$$\begin{aligned} C &= (-1000, 0.5) \\ D &= (-500, 1). \end{aligned}$$

This time, C was more popular.

Note that the two problems are identical in terms of their final wealth positions and yet people choose differently. The subjects are apparently focusing only on gains and losses. Indeed, when they are not given any information about prior winnings, they choose B over A and C over D.

The second important feature is the shape of the value function v, namely its concavity in the domain of gains and convexity in the domain of losses. Put simply, people are risk averse over gains, and risk-seeking over losses. Simple evidence for this comes from the fact just mentioned, namely that in the absence of any information about prior winnings[12]

$$B \succ A, \; C \succ D.$$

The v function also has a kink at the origin, indicating a greater sensitivity to losses than to gains, a feature known as *loss aversion.* Loss aversion is introduced to capture aversion to bets of the form:

$$E = (110, \frac{1}{2}; -100, \frac{1}{2}).$$

It may seem surprising that we need to depart from the expected utility framework in order to understand attitudes to gambles as simple as E, but it is nonetheless true. In a remarkable paper, Rabin (2000) shows that if an expected utility maximizer rejects gamble E at all wealth levels, then he will also reject

$$(20000000, \frac{1}{2}; -1000, \frac{1}{2}),$$

an utterly implausible prediction. The intuition is simple: if a smooth, increasing, and concave utility function defined over final wealth has sufficient local curvature to reject E

authors note that at the time of their research, the median monthly family income was about 3000 Israeli lira.

[12] In this section $G_1 \succ G_2$ should be read as "a statistically significant fraction of Kahneman and Tversky's subjects preferred G_1 to G_2."

over a wide range of wealth levels, it must be an extraordinarily concave function, making the investor extremely risk averse over large stakes gambles.

The final piece of prospect theory is the nonlinear probability transformation. Small probabilities are overweighted, so that $\pi(p) > p$. This is deduced from KT's finding that

$$(5000, 0.001) \succ (5, 1)$$

and

$$(-5, 1) \succ (-5000, 0.001),$$

together with the earlier assumption that v is concave (convex) in the domain of gains (losses). Moreover, people are more sensitive to differences in probabilities at higher probability levels. For example, the following pair of choices,

$$(3000, 1) \succ (4000, 0.8; 0, 0.2)$$

and

$$(4000, 0.2; 0, 0.8) \succ (3000, 0.25),$$

which violate EU theory, imply

$$\frac{\pi(0.25)}{\pi(0.2)} < \frac{\pi(1)}{\pi(0.8)}.$$

The intuition is that the 20 percent jump in probability from 0.8 to 1 is more striking to people than the 20 percent jump from 0.2 to 0.25. In particular, people place much more weight on outcomes that are certain relative to outcomes that are merely probable, a feature sometimes known as the "certainty effect".

Along with capturing experimental evidence, prospect theory also simultaneously explains preferences for insurance and for buying lottery tickets. Although the concavity of v in the region of gains generally produces risk aversion, for lotteries which offer a small chance of a large gain, the overweighting of small probabilities in Figure 2 dominates, leading to risk-seeking. Along the same lines, while the convexity of v in the region of losses typically leads to risk-seeking, the same overweighting of small probabilities introduces risk aversion over gambles which have a small chance of a large loss.

Based on additional evidence, Tversky and Kahneman (1992) propose a generalization of prospect theory which can be applied to gambles with more than two outcomes. Specifically, if a gamble promises outcome x_i with probability p_i, Tversky and Kahneman (1992) propose that people assign the gamble the value

$$\sum_i \pi_i v(x_i) \tag{2}$$

where

$$v = \begin{array}{ll} x^\alpha & \text{if } x \geq 0 \\ -\lambda(-x)^\alpha & \text{if } x < 0 \end{array}$$

and

$$\begin{aligned}\pi_i &= w(P_i) - w(P_i^*) \\ w(P) &= \frac{P^\gamma}{(P^\gamma + (1-P)^\gamma)^{1/\gamma}}.\end{aligned}$$

Here, P_i (P_i^*) is the probability that the gamble will yield an outcome at least as good as (strictly better than) x_i. Tversky and Kahneman (1992) use experimental evidence to estimate $\alpha = 0.88$, $\lambda = 2.25$, and $\gamma = 0.65$. Note that λ is the coefficient of loss aversion, a measure of the relative sensitivity to gains and losses. Over a wide range of experimental contexts λ has been estimated in the neighborhood of 2.

Earlier in this section, we saw how prospect theory could explain why people made different choices in situations with identical final wealth levels. This illustrates an important feature of the theory, namely that it can accommodate the effects of problem description, or of *framing*. Such effects are powerful. There are numerous demonstrations of a 30 to 40 percent shift in preferences depending on the wording of a problem. No normative theory of choice can accommodate such behavior since a first principle of rational choice is that choices should be independent of the problem description or representation.

Framing refers to the way a problem is posed for the decision maker. In many actual choice contexts the decision maker also has flexibility in how to think about the problem. For example, suppose that a gambler goes to the race track and wins \$200 in her first bet, but then loses \$50 on her second bet. Does she code the outcome of the second bet as a loss of \$50 or as a reduction in her recently won gain of \$200? In other words, is the utility of the second loss $v(-50)$ or $v(150) - v(200)$? The process by which people formulate such problems for themselves is called *mental accounting* (Thaler, 1999). Mental accounting matters because in prospect theory, v is nonlinear.

One important feature of mental accounting is *narrow framing*, which is the tendency to treat individual gambles separately from other portions of wealth. In other words, when offered a gamble, people often evaluate it as if it is the only gamble they face in the world, rather than merging it with pre-existing bets to see if the new bet is a worthwhile addition.

Redelmeier and Tversky (1992) provide a simple illustration, based on the gamble

$$F = (2000, \frac{1}{2}; -500, \frac{1}{2}).$$

Subjects in their experiment were asked whether they were willing to take this bet; 57 percent said they would not. They were then asked whether they would prefer to play F five times or six times; 70 percent preferred the six-fold gamble. Finally they were asked:

Suppose that you have played F five times but you don't yet know your wins and losses. Would you play the gamble a sixth time?

60 percent rejected the opportunity to play a sixth time, reversing their preference from the earlier question. This suggests that some subjects are framing the sixth gamble narrowly, segregating it from the other gambles. Indeed, the 60 percent rejection level is very similar to the 57 percent rejection level for the one-off play of F.

Ambiguity Aversion

Our discussion so far has centered on understanding how people act when the outcomes of gambles have known, objective probabilities. In reality, probabilities are rarely objectively known. To handle these situations, Savage (1964) develops a counterpart to expected utility known as subjective expected utility, SEU henceforth. Under certain axioms, preferences can be represented by the expectation of a utility function, this time weighted by the individual's subjective probability assessment.

Experimental work in the last few decades has been as unkind to SEU as it was to EU. The violations this time are of a different nature, but they may be just as relevant for financial economists.

The classic experiment was described by Ellsberg (1961). Suppose that there are two urns, 1 and 2. Urn 2 contains a total of 100 balls, 50 red and 50 blue. Urn 1 also contains 100 balls, again a mix of red and blue, but the subject does not know the proportion of each.

Subjects are asked to choose one of the following two gambles, each of which involves a possible payment of \$100, depending on the color of a ball drawn at random from the relevant urn

$$\begin{aligned} a_1 &: \text{ a ball is drawn from Urn 1, \$100 if red, \$0 if blue} \\ a_2 &: \text{ a ball is drawn from Urn 2, \$100 if red, \$0 if blue.} \end{aligned}$$

Subjects are then also asked to choose between the following two gambles:

$$\begin{aligned} b_1 &: \text{ a ball is drawn from Urn 1, \$100 if blue, \$0 if red} \\ b_2 &: \text{ a ball is drawn from Urn 2, \$100 if blue, \$0 if red.} \end{aligned}$$

a_2 is typically preferred to a_1, while b_2 is chosen over b_1. These choices are inconsistent with SEU: the choice of a_2 implies a subjective probability that *fewer* than 50 percent of the balls in Urn 1 are red, while the choice of b_2 implies the opposite.

The experiment suggests that people do not like situations where they are uncertain about the probability distribution of a gamble. Such situations are known as situations of ambiguity, and the general dislike for them, as ambiguity aversion.[13] SEU does not allow

[13] An early discussion of this aversion can be found in Knight (1921), who defines risk as a gamble with

agents to express their degree of confidence about a probability distribution and therefore cannot capture such aversion.

Ambiguity aversion appears in a wide variety of contexts. For example, a researcher might ask a subject for his estimate of the probability that a certain team will win its upcoming football match, to which the subject might respond 0.4. The researcher then asks the subject to imagine a chance machine, which will display 1 with probability 0.4 and 0 otherwise, and asks whether the subject would prefer to bet on the football game – an ambiguous bet – or on the machine, which offers no ambiguity. In general, people prefer to bet on the machine, illustrating aversion to ambiguity.

Heath and Tversky (1991) argue that in the real world, ambiguity aversion has much to do with how competent an individual feels he is at assessing the relevant distribution. Ambiguity aversion over a bet can be strengthened by highlighting subjects' feelings of incompetence, either by showing them other bets in which they have more expertise, or by mentioning other people who are more qualified to evaluate the bet (Fox and Tversky, 1995).

Further evidence that supports the competence hypothesis is that in situations where people feel especially competent in evaluating a gamble, the opposite of ambiguity aversion, namely a "preference for the familiar," has been observed. In the example above, people chosen to be especially knowledgeable about football often prefer to bet on the outcome of the game than on the chance machine. Just as with ambiguity aversion, such behavior cannot be captured by SEU.

4 Application: The Aggregate Stock Market

Researchers studying the aggregate U.S. stock market have identified a number of interesting facts about its behavior. Three of the most striking are:

(i) The *Equity Premium.* The stock market has historically earned a high excess rate of return. For example, using annual data from 1871-1993, Campbell and Cochrane (1999) report that the average log return on the S&P 500 index is 3.9 percent higher than the average log return on short term commercial paper.

(ii) *Volatility.* Stock returns and price-dividend ratios are both highly variable. In the same data set, the annual standard deviation of excess log returns on the S&P 500 is 18 percent, while the annual standard deviation of the log price-dividend ratio is 0.27.

(iii) *Predictability.* Stock returns are forecastable. Using monthly, real, equal-weighted NYSE returns from 1941-1986, Fama and French (1988) show that the dividend-price ratio is able

known distribution and uncertainty as a gamble with unknown distribution, and suggests that people dislike uncertainty more than risk.

to explain 27 percent of the variation of cumulative stock returns over the subsequent four years.[14]

All three of these facts can be labelled puzzles. Fact (i) has been known as the equity premium puzzle since the work of Mehra and Prescott (1985) (see also Hansen and Singleton, 1983). Campbell (1999) calls (ii) the volatility puzzle and we refer to (iii) as the predictability puzzle. The reason they are called puzzles is that they are hard to rationalize in a simple consumption-based model.

To see this, consider the following endowment economy, which we come back to a number of times in this section. There are an infinite number of identical investors, and two assets: a risk-free asset in zero net supply, with gross return $R_{f,t}$ between time t and $t+1$, and a risky asset – the stock market – in fixed positive supply, with gross return R_{t+1} between time t and $t+1$. The stock market is a claim to a perishable stream of dividends $\{D_t\}$, where

$$\frac{D_{t+1}}{D_t} = e^{g_D + \sigma_D \varepsilon_{t+1}}, \tag{3}$$

and where each period's dividend can be thought of as one component of a consumption endowment C_t, where

$$\frac{C_{t+1}}{C_t} = e^{g_C + \sigma_C \eta_{t+1}} \tag{4}$$

and

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \omega \\ \omega & 1 \end{pmatrix} \right), \text{ i.i.d. over time.} \tag{5}$$

Investors choose consumption C_t and an allocation S_t to the risky asset to maximize

$$E_0 \sum_{t=0}^{\infty} \rho^t \frac{C_t^{1-\gamma}}{1-\gamma} \tag{6}$$

subject to the standard budget constraint.[15] Using the Euler equation of optimality,

$$1 = \rho E_t \left[\left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} R_{t+1} \right], \tag{7}$$

it is straightforward to derive expressions for stock returns and prices. The details are in the Appendix.

We can now examine the model's quantitative predictions for the parameter values in Table 2. The endowment process parameters are taken from U.S. data spanning the 20th century, and are standard in the literature. It is also standard to start out by considering *low*

[14] These three facts are widely agreed on, but they are not completely uncontroversial. A large literature has debated the statistical significance of the time series predictability, while others have argued that the equity premium is overstated due to survivorship bias (Brown, Goetzmann and Ross, 1995).

[15] For $\gamma = 1$, we replace $C_t^{1-\gamma}/1-\gamma$ with $\log(C_t)$.

values of γ. The reason is that when one computes, for various values of γ, how much wealth an individual would be prepared to give up to avoid a large-scale timeless wealth gamble, low values of γ match best with introspection as to what the answers should be (Mankiw and Zeldes, 1991). We take $\gamma = 1$, which corresponds to log utility.

In an economy with these parameter values, the average log return on the stock market would be just 0.1 percent higher than the risk-free rate, not the 3.9 percent observed historically. The standard deviation of log stock returns would be only 12 percent, not 18 percent, and the price-dividend ratio would be constant (implying, of course, that the dividend-price ratio has no forecast power for future returns).

It is useful to recall the intuition for these results. In an economy with power utility preferences, the equity premium is determined by risk aversion γ and by risk, measured as the covariance of stock returns and consumption growth. Since consumption growth is very smooth in the data, this covariance is very low, thus predicting a very low equity premium. Stocks simply do not appear risky to investors with the preferences in (6) and with low γ, and therefore do not warrant a large premium. Of course, the equity premium predicted by the model can be increased by using higher values of γ. However, other than making counterintuitive predictions about individuals' attitudes to large-scale gambles, this would also predict a counterfactually high risk-free rate, a problem known as the risk-free rate puzzle (Weil, 1989).

To understand the volatility puzzle, note that in the simple economy described above, both discount rates and expected dividend growth are constant over time. A direct application of the present value formula implies that the price-dividend ratio, P/D henceforth, is constant. Since

$$R_{t+1} = \frac{D_{t+1} + P_{t+1}}{P_t} = \frac{1 + P_{t+1}/D_{t+1}}{P_t/D_t} \frac{D_{t+1}}{D_t}, \tag{8}$$

it follows that

$$r_{t+1} = \Delta d_{t+1} + \text{const.} \equiv d_{t+1} - d_t + \text{const.}, \tag{9}$$

where lower case letters indicate log variables. The standard deviation of log returns will therefore only be as high as the standard deviation of log dividend growth, namely 12 percent.

The particular volatility puzzle seen here illustrates a more general point, first made by Shiller (1981) and LeRoy and Porter (1981), namely that it is difficult to explain the historical volatility of stock returns with *any* model in which investors are rational and discount rates are constant.

To see the intuition, consider the identity in equation (8) again. Since the volatility of log dividend growth is only 12 percent, the only way for a model to generate an 18 percent volatility of log returns is to introduce variation in the P/D ratio. But if discount rates are constant, a quick glance at a present-value formula shows that the only way to do that is to introduce variation in investors' forecasts of the dividend growth rate: a higher forecast raises the P/D ratio, a lower forecast brings it down. There is a catch here, though: if investors are rational, their expectations for dividend growth must, on average, be confirmed. In other words, times of higher (lower) P/D ratios should, on average, be followed by higher (lower) cash-flow growth. Unfortunately, price-dividend ratios are *not* reliable forecasters of dividend growth, neither in the U.S. nor in most international markets (see Campbell, 1999, for recent evidence).

Shiller and LeRoy and Porter's results shocked the profession when they first appeared. At the time, most economists felt that discount rates *were* close to constant over time, apparently implying that stock market volatility could only be fully explained by appealing to investor irrationality. Today, it is well understood that rational variation in discount rates can help explain the volatility puzzle, although we argue later that models with irrational beliefs also offer a plausible way of thinking about the data.

Both the rational and behavioral approaches to finance have made progress in understanding the three puzzles singled out at the start of this section. The advances on the rational side are well described in other articles in this handbook. Here, we discuss the behavioral approaches, starting with the equity premium puzzle and then turning to the volatility puzzle.

We do not consider the predictability puzzle separately, because in any model with a stationary P/D ratio, a resolution of the volatility puzzle is simultaneously a resolution of the predictability puzzle. To see this, recall from equation (8) that any model which captures the empirical volatility of returns must involve variation in the P/D ratio. Moreover, for a model to be a *satisfactory* resolution of the volatility puzzle, it should not make the counterfactual prediction that P/D ratios forecast subsequent dividend growth. Now suppose that the P/D ratio is higher than average. The only way it can return to its mean is if cash flows D subsequently go up, or if prices P fall. Since the P/D ratio is not allowed to forecast cash flows, it must forecast lower returns, thereby explaining the predictability puzzle.

4.1 The Equity Premium Puzzle

The core of the equity premium puzzle is that even though stocks appear to be an attractive asset – they have high average returns and a low covariance with consumption growth – investors appear very unwilling to hold them. In particular, they appear to demand a substantial risk premium in order to hold the market supply.

To date, behavioral finance has pursued two approaches to this puzzle. Both are based on preferences: one relies on prospect theory, the other on ambiguity aversion. In essence, both approaches try to understand what it is that is missing from the popular preference specification in (6) that makes investors fear stocks so much, leading them to charge a high premium in equilibrium.

Prospect Theory

One of the earliest papers to link prospect theory to the equity premium is Benartzi and Thaler (1995), BT henceforth. They study how an investor with prospect theory-type preferences allocates his financial wealth between T-Bills and the stock market. Prospect theory argues that when choosing between gambles, people compute the gains and losses for each one and select the one with the highest prospective utility. In a financial context, this suggests that people may choose a portfolio allocation by computing, for each allocation, the potential gains and losses in the value of their holdings, and then taking the allocation with the highest prospective utility. In other words, they choose ω, the fraction of financial wealth in stocks, to maximize

$$E_\pi \ v[(1-\omega)R_{f,t+1} + \omega R_{t+1} - 1], \tag{10}$$

where π and v are defined in (2). In particular, v captures loss aversion, the experimental finding that people are more sensitive to losses than to gains. $R_{f,t+1}$ and R_{t+1} are the gross returns on T-Bills and the stock market between t and $t+1$, respectively, making the argument of v the return on financial wealth. The distributions of $R_{f,t+1}$ and R_{t+1} are obtained by bootstrapping historical U.S. data.[16]

In order to implement this model, BT need to stipulate how often investors evaluate their portfolios. In other words, how long is the time interval between t and $t+1$? To see why this matters, compare two investors: energetic Nick who calculates the gains and losses in his portfolio every day, and laid-back Dick who looks at his portfolio only once per decade. Since, on a daily basis, stocks go down in value almost as often as they go up, the loss aversion built into v makes stocks appear unattractive to Nick. In contrast, loss aversion does not have much effect on Dick's perception of stocks since, at ten year horizons, stocks offer only a small risk of losing money. Rather than simply pick an evaluation interval, BT calculate how often investors would have to evaluate their portfolios to make them roughly indifferent between stocks and bonds. In other words, they compute how often investors would need to evaluate their gains and losses so that even in the face of the large historical equity premium, they would still be happy to hold the market supply of bonds and stocks.

When they solve (10) using the parametric forms for π and v estimated in experimental settings, BT find the answer to be a year, and argue that this is indeed a natural evaluation period for investors to use. The way people frame gains and losses is plausibly influenced by the way information is presented to them. Since we receive our most comprehensive mutual fund reports once a year, and do our taxes once a year, it is not unreasonable that gains and losses might be expressed as annual changes in value.

[16] In (2), π and v are defined over discrete, not continuous distributions. Benartzi and Thaler (1995) therefore summarize the historical distributions of T-Bills and stocks as discrete histograms before applying (2).

This, in turn, suggests a simple way of understanding the high historical equity premium. If investors get utility from annual changes in financial wealth and are loss averse over these changes, their fear of a major drop in financial wealth will lead them to demand a high premium as compensation. BT call the combination of loss aversion and frequent evaluations *myopic loss aversion.*

BT's result is only *suggestive* of a solution to Mehra and Prescott's equity premium puzzle. As emphasized at the start of this section, that puzzle is in large part a consumption puzzle: given the low volatility of consumption growth, why are investors so reluctant to buy a high return asset, stocks, especially when that asset's covariance with consumption growth is so low? Since BT do not consider an intertemporal model with consumption choice, they cannot address this issue directly.

To see if prospect theory can in fact help with the equity premium puzzle, Barberis, Huang and Santos (2001), BHS henceforth, make a first attempt at building it into a dynamic equilibrium model of stock returns. A simple version of their model, an extension of which we consider later, examines an economy with the same structure as the one described at the start of Section 4, but in which investors have the preferences

$$E_0 \sum_{t=0}^{\infty} \left[\rho^t \frac{C_t^{1-\gamma}}{1-\gamma} + b_0 \overline{C}_t^{-\gamma} \widehat{v}(X_{t+1}) \right]. \tag{11}$$

The investor gets utility from consumption, but over and above that, he gets utility from changes in the value of his holdings of the risky asset between t and $t+1$, denoted here by X_{t+1}. Motivated by BT's findings, BHS define the unit of time to be a year, so that gains and losses are measured annually.

The utility from these gains and losses is determined by $\widehat{v}$ where

$$\widehat{v}(X) = \left\{ \begin{array}{ll} X & \\ 2.25X & \end{array} \right. \text{for} \begin{array}{l} X \geq 0 \\ X < 0 \end{array}. \tag{12}$$

The 2.25 factor comes from Tversky and Kahneman's (1992) experimental study of attitudes to timeless gambles. This functional form is simpler than the one used by BT, v. It captures loss aversion, but ignores other elements of prospect theory, such as the concavity (convexity) over gains (losses) and the probability transformation. In part this is because it is difficult to incorporate all these features into a fully dynamic framework; but also, it is based on BT's observation that it is mainly loss aversion that drives their results.[17]

[17] The $b_0 \overline{C}_t^{-\gamma}$ coefficient on the loss aversion term is a scaling factor which ensures that risk premia in the economy remain stationary even as aggregate wealth increases over time. It involves per capita consumption $\overline{C}_t$ which is exogeneous to the investor, and so does not affect the intuition of the model. The constant b_0 controls the importance of the loss aversion term in the investor's preferences; setting $b_0 = 0$ reduces the model to the much studied case of power utility over consumption. As $b_0 \to \infty$, the investor's decisions are driven primarily by concern about gains and losses in financial wealth, as assumed by BT.

BHS show that loss aversion can indeed provide a partial explanation of the high Sharpe ratio on the aggregate stock market. However, how much of the Sharpe ratio it can explain depends heavily on the importance of the second source of utility in (11), or in short, on b_0. As a way of thinking about this parameter, BHS note that when $b_0 = 0.7$, the psychological pain of losing \$100 in the stock market, captured by the second term, is roughly equal to the consumption-related pain of having to consume \$100 less, captured by the first term. For this b_0, the Sharpe ratio of the risky asset is 0.11, about a third of its historical value.

BT and BHS are both effectively assuming that investors engage in narrow framing, both cross-sectionally and temporally. Even if they have many forms of wealth, both financial and non-financial, they still get utility from changes in the value of one specific component of their total wealth: financial wealth in the case of BT, and stock holdings in the case of BHS. And even if investors have long investment horizons, they still evaluate their portfolio returns on an annual basis.

The assumption about cross-sectional narrow framing can be motivated in a number of ways. The simplest possibility is that it captures non-consumption utility, such as regret. Regret is the pain we feel when we realize that we would be better off if we had not taken a certain action in the past. If the investor's stock holdings fall in value, he may regret the specific decision he made to invest in stocks. Such feelings are naturally captured by defining utility directly over changes in the investors' financial wealth or in the value of his stock holdings.

Another possibility is that while people actually care only about consumption-related utility, they are boundedly rational. For example, suppose that they are concerned that their consumption might fall below some habit level. They know that the right thing to do when considering a stock market investment is to merge the stock market risk with other pre-existing risks that they face – labor income risk, say – and then to compute the likelihood of consumption falling below habit. However, this calculation may be too complex. As a result, people may simply focus on gains and losses in stock market wealth alone, rather than on gains and losses in total wealth.

What about temporal narrow framing? We suggested above that the way information is presented may lead investors to care about annual changes in financial wealth even if they have longer investment horizons. To provide further evidence for this, Thaler, Tversky, Kahneman and Schwartz (1997) provide an *experimental* test of the idea that the manner in which information is presented affects the frame people adopt in their decision-making.[18]

In their experiment, subjects are asked to imagine that they are portfolio managers for a small college endowment. One group of subjects – Group I, say – is shown monthly observations on two funds, Fund A and Fund B. Returns on Fund A (B) are drawn from a normal distribution calibrated to mimic bond (stock) returns as closely as possible, although

[18] See also Gneezy and Potters (1997) for a similar experiment.

subjects are not given this information. After each monthly observation, subjects are asked to allocate their portfolio between the two funds over the next month. They are then shown the realized returns over that month, and asked to allocate once again.

A second group of investors – Group II – is shown exactly the same series of returns, except that it is aggregated at the annual level; in other words, these subjects do not see the monthly fund fluctuations, but only cumulative annual returns. After each annual observation, they are asked to allocate their portfolio between the two funds over the next year.

A final group of investors – Group III – is shown exactly the same data, this time aggregated at the five-year level, and they too are asked to allocate their portfolio after each observation.

After going through a total of 200 months worth of observations, each group is asked to make one final portfolio allocation, which is to apply over the next 400 months. Thaler et al. (1997) find that the average final allocation chosen by subjects in Group I is much lower than that chosen by people in Groups II and III. This result is consistent with the idea that people code gains and losses based on how information is presented to them. Subjects in Group I see monthly observations and hence more frequent losses. If they adopt the monthly distribution as a frame, they will be more wary of stocks and will allocate less to them.

Ambiguity Aversion

In Section 3, we presented the Ellsberg paradox as evidence that people dislike ambiguity, or situations where they are not sure what the probability distribution of a gamble is. This is potentially very relevant for finance, as investors are often uncertain about the distribution of a stock's return.

Following the work of Ellsberg, many models of how people react to ambiguity have been proposed; Camerer and Weber (1992) provide a comprehensive review. One of the more popular approaches is to suppose that when faced with ambiguity, people entertain a range of possible probability distributions and act to maximize the minimum expected utility under any candidate distribution. In effect, people behave as if playing a game against a malevolent opponent who picks the actual distribution of the gamble so as to leave them as worse off as possible. Such a decision rule was first axiomatized by Gilboa and Schmeidler (1989). Epstein and Wang (1994) showed how such an approach could be incorporated into a dynamic asset pricing model, although they did not try to assess the quantitative implications of ambiguity aversion for asset prices.

Quantitative implications *have* been derived using a closely related framework known as robust control. In this approach, the agent has a reference probability distribution in mind, but wants to ensure that his decisions are good ones even if the reference model is misspecified to some extent. Here too, the agent essentially tries to guard against a "worst-case" misspecification. Anderson, Hansen and Sargent (1998) show how such a framework can be used for portfolio choice and pricing problems, even when state equations and objective functions are nonlinear.

Maenhout (1999) applies the Anderson et al. framework to the specific issue of the equity premium. He shows that if investors are concerned that their model of stock returns is misspecified, they will charge a substantially higher equity premium as compensation for the perceived ambiguity in the probability distribution. He notes, however, that to explain the full 3.9 percent equity premium requires an unreasonably high concern about misspecification. At best then, ambiguity aversion is only a partial resolution of the equity premium puzzle.

4.2 The Volatility Puzzle

Before turning to behavioral work on the volatility puzzle, it is worth thinking about how rational approaches to this puzzle might proceed. Since, in the data, the volatility of returns is higher than the volatility of dividend growth, equation (8) makes it clear that we have to make up the gap by introducing variation in the price-dividend ratio. What are the different ways we might do this? A useful framework for thinking about this is a version of the present value formula originally derived by Campbell and Shiller (1988). Starting from

$$R_{t+1} = \frac{P_{t+1} + D_{t+1}}{P_t}, \tag{13}$$

where P_t is the value of the stock market at time t, they use a log-linear approximation to show that the log price-dividend ratio can be written

$$p_t - d_t = E_t \sum_{j=0}^{\infty} \rho^t \Delta d_{t+1+j} - E_t \sum_{j=0}^{\infty} \rho^t r_{t+1+j} + E_t \lim_{j \to \infty} \rho^j (p_{t+j} - d_{t+j}) + \text{const.}, \tag{14}$$

where lower case letters represent log variables – $p_t = \log P_t$, for example – and where $\Delta d_{t+1} = d_{t+1} - d_t$.

If the price-dividend ratio is stationary, so that the third term on the right is zero, this equation shows clearly that there are just two reasons price-dividend ratios can move around: changing expectations of future dividend growth or changing discount rates. Discount rates, in turn, can change because of changing expectations of future risk-free rates, changing forecasts of risk or changing risk aversion.

While there appear to be many ways of introducing variation in the P/D ratio, it has become clear that most of them cannot form the basis of a rational explanation of the volatility puzzle. We cannot use changing forecasts of dividend growth to drive the P/D ratio: restating the argument of Shiller (1981) and Le Roy and Porter (1981), if these forecasts are indeed rational, it must be that P/D ratios predict cash-flow growth in the time series, which they do not.[19] Nor can we use changing forecasts of future risk-free rates: again, if the forecasts are rational, P/D ratios must predict interest rates in the time series, which they do not. Even changing forecasts of risk cannot work, as there is little evidence that P/D ratios predict changes in risk in the time series. The only story that remains is therefore one about changing risk aversion, and this is the idea behind the Campbell and Cochrane (1999) model of aggregate stock market behavior. They propose a habit formation framework in which changes in consumption relative to habit lead to changes in risk aversion and hence variation in P/D ratios. This variation helps to plug the gap between the volatility of dividend growth and the volatility of returns.

Some rational approaches try to introduce variation in the P/D ratio through the third term on the right in equation (14). Since this requires investors to expect explosive growth in P/D ratios forever, they are known as models of rational bubbles. The idea is that prices are high today because they are expected to be higher next period; and they are higher next period because they are expected to be higher the period after that, and so on, forever. While such a model might initially seem appealing, a number of papers, most recently Santos and Woodford (1997), show that the conditions under which rational bubbles can survive are extremely restrictive.[20]

We now discuss some of the behavioral approaches to the volatility puzzle, grouping them by whether they focus on beliefs or on preferences.

Beliefs

One possible story is that investors believe that the mean dividend growth rate is more variable than it actually is. When they see a surge in dividends, they are too quick to believe that the mean dividend growth rate has increased. Their exuberance pushes prices up relative to dividends, adding to the volatility of returns.

A story of this kind can be derived as a direct application of representativeness and in particular, of the version of representativeness known as the law of small numbers, whereby people expect even short samples to reflect the properties of the parent population. If the investor sees many periods of good earnings, the law of small numbers leads him to believe that earnings growth has gone up, and hence that earnings will continue to be high in the

[19] There is an imporant caveat to the statement that changing cash-flow forecasts cannot be the basis of a satisfactory solution to the volatility puzzle. A large literature on structural uncertainty and learning, in which investors do not know the parameters of the cash-flow process but learn them over time, has had some success in matching the empirical volatility of returns (Brennan and Xia, 2001, Veronesi, 1999). In these models, variation in price-dividend ratios comes precisely from changing forecasts of cash-flow growth. While these forecasts are not subsequently confirmed in the data, investors are not considered irrational – they simply don't have enough data to infer the correct model. In related work, Barksy and De Long (1993) generate return volatility in an economy where investors forecast cash flows using a model that is wrong, but not easily rejected with available data.

[20] Brunnermeier (2001) provides a comprehensive review of this literature.

future. After all, the earnings growth rate cannot be "average". If it were, then according to the law of small numbers, earnings should *appear* average, even in short samples: some good earnings news, some bad earnings news, but not several good pieces of news in a row.

Another belief-based story relies more on private, rather than public information, and in particular, on overconfidence about private information. Suppose that an investor has seen public information about the economy, and has formed a prior opinion about future cash-flow growth. He then does some research on his own and becomes overconfident about the information he gathers: he overestimates its accuracy and puts too much weight on it relative to his prior. If the private information is positive, he will push prices up too high relative to current dividends, again adding to return volatility.[21]

Price-dividend ratios and returns might also be excessively volatile because investors extrapolate *past returns* too far into the future when forming expectations of future returns. Such a story might again be based on representativeness and the law of small numbers. The same argument for why investors might extrapolate past cash flows too far into the future can be applied here to explain why they might do the same thing with past returns.

The reader will have noticed that we do not cite any specific papers in connection with these behavioral stories. This is because these ideas were originally put forward in papers whose primary focus is explaining *cross-sectional* anomalies such as the value premium, even though they also apply here in a natural way. In brief, many of those papers – which we discuss in detail in Section 5 – generate certain cross-sectional anomalies by building excessive time series variation into the price-earnings ratios of individual stocks. It is therefore not surprising that the mechanisms proposed there might also explain the substantial time series variation in *aggregate*-level price-earnings ratios. In fact, it is perhaps satisfying that these behavioral theories simultaneously address both aggregate and firm-level evidence.

We close this section with a brief mention of "money illusion", the confusion between real and nominal values first discussed by Fisher (1928), and more recently investigated by Shafir et al. (1997). In financial markets, Modigliani and Cohn (1979) and more recently, Ritter and Warr (2002), have argued that part of the variation in P/D ratios and returns may be due to investors mixing real and nominal quantities when forecasting future cash flows. The value of the stock market can be determined by discounted real cash flows at real rates, or nominal cash flows at nominal rates. At times of especially high or especially low inflation though, it is possible that some investors mistakenly discount *real* cash flows

[21] Campbell (2000), among others, notes that behavioral models based on cash-flow forecasts often ignore potentially important interest rate effects. If investors are forecasting excessively high cash-flow growth, pushing up prices, interest rates should also rise, thereby dampening the price rise. One response is that interest rates are governed by expectations about *consumption* growth, and in the short run, consumption and dividends can be somewhat delinked: even if dividend growth is expected to be high, this need not necessarily trigger an immediate interest rate response. Alternatively, one can try to specify investors' expectations in such a way that interest rate effects become less important. Cecchetti, Lam and Mark (2000) take a step in this direction.

at *nominal* rates. If inflation increases, so will the nominal discount rate. If investors then discount the *same* set of cash flows at this higher rate, they will push the value of the stock market down. Of course, this calculation is incorrect: the same inflation which pushes up the discount rate should also push up future cash flows. On net, inflation should have little effect on market value. Such real vs. nominal confusion may therefore cause excessive variation in P/D ratios and returns and seems particularly relevant to understanding the low market valuations during the high inflation years of the 1970s, as well as the high market valuations during the low inflation 1990s.

Preferences

Barberis, Huang and Santos (2001) show that a straightforward extension of the version of their model discussed in Section 4.1 can explain both the equity premium and volatility puzzles. To do this, they appeal to experimental evidence about dynamic aspects of loss aversion. This evidence suggests that the degree of loss aversion is not the same in all circumstances but depends on prior gains and losses. In particular, Thaler and Johnson (1990) find that after prior gains, subjects take on gambles they normally do not, and that after prior losses, they refuse gambles that they normally accept. The first finding is sometimes known as the "house money effect", reflecting gamblers' increasing willingness to bet when ahead. One interpretation of this evidence is that losses are less painful after prior gains because they are cushioned by those gains. However, after being burned by a painful loss, people may become more wary of additional setbacks.[22]

To capture these ideas, Barberis, Huang and Santos (2001) modify the utility function in (11) to

$$E_0 \sum_{t=0}^{\infty} \left[\rho^t \frac{C_t^{1-\gamma}}{1-\gamma} + b_0 \overline{C}_t^{-\gamma} \widetilde{v}(X_{t+1}, z_t) \right]. \tag{15}$$

Here, z_t is a state variable that tracks past gains and losses on the stock market. For any fixed z_t, the function $\widetilde{v}$ is a piecewise linear function similar in form to $\widehat{v}$, defined in (12). However, the investors' sensitivity to losses is no longer constant at 2.25, but is determined by z_t, in a way that reflects the experimental evidence described above.

A model of this kind can help explain the volatility puzzle. Suppose that there is some good cash-flow news. This pushes the stock market up, generating prior gains for investors, who are now less scared of stocks: any losses will be cushioned by the accumulated gains. They therefore discount future cash flows at a lower rate, pushing prices up still further relative to current dividends and adding to return volatility.

[22] It is important to distinguish Thaler and Johnson's (1990) evidence from other evidence presented by Kahneman and Tversky (1979) and discussed in Section 3, showing that people are risk averse over gains and risk seeking over losses. One set of evidence pertains to one-shot gambles, the other to sequences of gambles. Kahneman and Tversky's (1979) evidence suggests that people are willing to take risks in order to avoid a loss; Thaler and Johnson's (1990) evidence suggests that if these efforts are unsuccessful and the investor suffers an unpleasant loss, he will *subsequently* act in a more risk averse manner.

5 Application: The Cross-section of Average Returns

While the behavior of the aggregate stock market is not easy to understand from the rational point of view, promising rational models have nonetheless been developed and can be tested against behavioral alternatives. Empirical studies of the behavior of *individual* stocks have unearthed a set of facts which is altogether more frustrating for the rational paradigm. Many of these facts are about the *cross-section* of average returns: they document that one group of stocks earns higher average returns than another. These facts have come to be known as "anomalies" because they cannot be explained by the simplest and most intuitive model of risk and return in the financial economist's toolkit, the Capital Asset Pricing Model, or CAPM.

We now outline some of the more salient findings in this literature and then consider some of the rational and behavioral approaches in more detail.

The Size Premium

This anomaly was first documented by Banz (1981). We report the more recent findings of Fama and French (1992). Every year from 1963 to 1990, Fama and French group all stocks traded on the NYSE, Amex, and Nasdaq into deciles based on their market capitalization, and then measure the average return of each decile over the next year. They find that for this sample period, the average return of the smallest stock decile is 0.74 percent per month higher than the average return of the largest stock decile. This is certainly an anomaly relative to the CAPM: while stocks in the smallest decile do have higher betas, the difference in risk is not enough to explain the difference in average returns.[23]

Long-term Reversals

Every three years from 1926 to 1982, De Bondt and Thaler (1985) rank all stocks traded on the NYSE by their prior three year cumulative return and form two portfolios: a "winner" portfolio of the 35 stocks with the best prior record and a "loser" portfolio of the 35 worst performers. They then measure the average return of these two portfolios over the three years subsequent to their formation. They find that over the whole sample period, the average annual return of the loser portfolio is higher than the average return of the winner portfolio by about 8 percent per year.

The Predictive Power of Scaled-price Ratios

These anomalies, which are about the cross-sectional predictive power of variables like the book-to-market (B/M) and earnings-to-price (E/P) ratios, where some measure of fun-

[23] The last decade of data has served to reduce the size premium considerably. Gompers and Metrick (2001) argue that this is due to demand pressure for large stocks resulting from the growth of institutional investors, who prefer such stocks.

damentals is scaled by price, have a long history in finance going back at least to Graham (1949), and more recently Dreman (1977), Basu (1983) and Rosenberg, Reid, and Lanstein (1985). We concentrate on Fama and French's (1992) more recent evidence.

Every year, from 1963 to 1990, Fama and French group all stocks traded on the NYSE, AMEX, and Nasdaq into deciles based on their book-to-market ratio, and measure the average return of each decile over the next year. They find that the average return of the highest-B/M-ratio decile, containing so called "value" stocks, is 1.53 percent per month higher than the average return on the lowest-B/M-ratio decile, "growth" or "glamour" stocks, a difference much higher than can be explained through differences in beta between the two portfolios. Repeating the calculations with the earnings-price ratio as the ranking measure produces a difference of 0.68 percent per month between the two extreme decile portfolios, again an anomalous result.[24]

Momentum

Every month from January 1963 to December 1989, Jegadeesh and Titman (1993) group all stocks traded on the NYSE into deciles based on their prior six month return and compute average returns of each decile over the six months after portfolio formation. They find that the decile of biggest prior winners outperforms the decile of biggest prior losers by an average of 10 percent on an annual basis.

Comparing this result to De Bondt and Thaler's (1985) study of prior winners and losers illustrates the crucial role played by the length of the prior ranking period. In one case, prior winners continue to win; in the other, they perform poorly.[25] A challenge to both behavioral and rational approaches is to explain why extending the formation period switches the result in this way.

There is some evidence that tax-loss selling creates seasonal variation in the momentum effect. Stocks with poor performance during the year may later be subject to selling by investors keen to realize losses that can offset capital gains elsewhere. This selling pressure means that prior losers continue to lose, enhancing the momentum effect. At the turn of the year, though, the selling pressure eases off, allowing prior losers to rebound and weakening the momentum effect. A careful analysis by Grinblatt and Moskowitz (1999) finds that on net, tax-loss selling may explain part of the momentum effect, but by no means all of it. In any case, while selling a stock for tax purposes is rational, a model of predictable price movements based on such behavior is not. Roll (1983) calls such explanations "stupid" since

[24] Ball (1978) and Berk (1995) point out that the size premium and the scaled-price ratio effects emerge naturally in any model where investors apply different discount rates to different stocks: if investors discount a stock's cash flows at a higher rate, that stock will typically have a lower market capitalization and a lower price-earnings ratio, but also higher returns. Note, however, that this view does not shed any light on whether the variation in discount rates is rationally justifiable or not.

[25] In fact, De Bondt and Thaler (1985) also report that one-year big winners outperform one-year big losers over the following year, but do not make much of this finding.

investors would have to be stupid not to buy in December if prices were going to increase in January.

A number of studies have examined stock returns following important corporate announcements, a type of analysis known as an event study. Jay Ritter's chapter in this volume discusses many of these studies in detail; here, we summarize them briefly.

Event Studies of Earnings Announcements

Every quarter from 1974 to 1986, Bernard and Thomas (1989) group all stocks traded on the NYSE and AMEX into deciles based on the size of the surprise in their most recent earnings announcement. "Surprise" is measured relative to a simple random walk model of earnings. They find that on average, over the 60 days after the earnings announcement, the decile of stocks with surprisingly good news outperforms the decile with surprisingly bad news by an average of about 4 percent, a phenomenon known as post-earnings announcement drift. Once again, this difference in returns is not explained by differences in beta between the two portfolios. A later study by Chan, Jegadeesh and Lakonishok (1996) measures surprise in other ways – relative to analyst expectations, and by the stock price reaction to the news – and obtains similar results.[26]

Event Studies of Dividend Initiations and Omissions

Michaely, Thaler and Womack (1995) study firms which announced initiation or omission of a dividend payment between 1964 and 1988. They find that on average, the shares of firms initiating (omitting) dividends significantly outperform (underperform) the market portfolio over the year after the announcement.

Event Studies of Stock Repurchases

Ikenberry, Lakonishok and Vermaelen (1995) look at firms which announced a share repurchase between 1980 and 1990, while Mitchell and Stafford (2001) study firms which did either self-tenders or share repurchases between 1960 and 1993. The latter study finds that on average, the shares of these firms outperform a control group matched on size and book-to-market by a substantial margin over the four year period following the event.

Event Studies of Primary and Secondary Offerings

Loughran and Ritter (1995) study firms which undertook primary or secondary equity offerings between 1970 and 1990. They find that the average return of shares of these firms

[26] Vuolteenaho (2002) combines a clean-surplus accounting version of the present value formula with Campbell's (1991) log-linear decomposion of returns to estimate a measure of cash-flow news that is potentially more accurate than earnings announcements. Analogous to the post-earnings announcement studies, he finds that stocks with good cash-flows news subsequently have higher average returns than stocks with disappointing cash-flow news.

over the five year period after the issuance is markedly below the average return of shares of non-issuing firms matched to the issuing firms on size. Brav and Gompers (1997) and Brav, Geczy and Gompers (2001) argue that this anomaly may not be distinct from the scaled-price anomaly listed above: when the returns of event firms are compared to the returns of firms matched on both size and book-to-market, there is very little difference.

Long-term event studies like the last three analyses summarized above raise some thorny statistical problems. In particular, conducting statistical inference with long-term buy-and-hold post-event returns is a treacherous business. Barber and Lyon (1997), Lyon, Barber and Tsai (1999), Brav (2000), Fama (1998), Loughran and Ritter (2000) and Mitchell and Stafford (2001) are just a few of the papers that discuss this topic. Cross-sectional correlation is one important issue: if a certain firm announces a share repurchase shortly after another firm does, their four-year post event returns will overlap and cannot be considered independent. Although the problem is an obvious one, it is not easy to deal with effectively. Some recent attempts to do so, such as Brav (2000), suggest that the anomalous evidence in the event studies on dividend announcements, repurchase announcements, and equity offerings is statistically weaker than initially thought, although how much weaker remains controversial.

A more general concern with *all* the above empirical evidence is data-mining. After all, if we sort and rank stocks in enough different ways, we are bound to discover striking – but completely spurious – cross-sectional differences in average returns.

A first response to the data-mining critique is to note that the above studies do not use the kind of obscure firm characteristics or marginal corporate announcements that would suggest data-mining. Indeed, it is hard to think of an important class of corporate announcements that has *not* been associated with a claim about anomalous post-event returns. A more direct check is to perform out-of-sample tests. Interestingly, a good deal of the above evidence *has* been replicated in other data sets. Fama, French and Davis (2000) show that there is a value premium in the subsample of U.S. data that precedes the data set used in Fama and French (1992), while Fama and French (1998) document a value premium in international stock markets. Rouwenhourst (1998) shows that the momentum effect is alive and well in international stock market data.

If the empirical results are taken at face value, then the challenge to the rational paradigm is to show that the above cross-sectional evidence emerges naturally from a model with fully rational investors. In special cases, models of this form reduce to the CAPM, and we know that this does not explain the evidence. More generally, rational models predict a multifactor pricing structure,

$$\overline{r}_i - r_f = \beta_{i,1}(\overline{F}_1 - r_f) + \ldots + \beta_{i,K}(\overline{F}_K - r_f), \tag{16}$$

where the factors proxy for marginal utility growth and where the loadings $\beta_{i,k}$ come from a time series regression of excess stock returns on excess factor returns,

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_{i,1}(F_{1,t} - r_{f,t}) + \ldots + \beta_{i,K}(F_{K,t} - r_{f,t}) + \varepsilon_{i,t}. \tag{17}$$

To date, it has proved difficult to derive a multi-factor model which explains the cross-sectional evidence, although this remains a major research direction.

Alternatively, one can skip the step of *deriving* a factor model, and simply try a specific model to see how it does. This is the approach of Fama and French (1993, 1996). They show that a certain three factor model does a good job explaining the average returns of portfolios formed on size and book-to-market rankings. Put differently, the α_i intercepts in regression (17) are typically close to zero for their choice of factors. The specific factors they use are the return on the market portfolio, the return on a portfolio of small stocks minus the return on a portfolio of large stocks – the "size" factor – and the return on a portfolio of value stocks minus the return on a portfolio of growth stocks – the "book-to-market" factor. By constructing these last two factors, Fama and French are isolating common factors in the returns of small stocks and value stocks, and their three factor model can be loosely motivated by the idea that this comovement is a systematic risk that is priced in equilibrium.

The low α_i intercepts obtained by Fama and French (1993, 1996) are not necessarily cause for celebration. After all, these authors began their investigation only after it was already known that small stocks and value stocks earn high average returns. Moreover, as Roll (1977) emphasizes, in any specific sample, it is always possible to mechanically construct a one factor model that prices average returns *exactly*.[27] This sounds a cautionary note: just because a factor model happens to work well does not necessarily mean that we are learning anything about the economic drivers of average returns. To be fair, Fama and French (1996) themselves admit that their results can only have their full impact once it is explained what it is about investor preferences and the structure of the economy that leads people to price assets according to their model.

One general feature of the rational approach is that it is loadings or betas, and not firm characteristics that determine average returns. For example, a risk-based approach would argue that value stocks earn high returns not because they have high book-to-market ratios, but because such stocks happen to have a high loading on the book-to-market factor. Daniel and Titman (1997) cast doubt on this specific prediction by performing double sorts of stocks on both book-to-market ratios and loadings on book-to-market factors, and showing that stocks with different loadings but the same book-to-market ratio do *not* differ in their average returns. These results appear quite damaging to rational approach. However, using a longer data set and a different methodology, Fama, French and Davis (2000) claim to reverse Daniel and Titman's findings. We expect further developments on this controversial front.

[27] For any sample of observations on individual returns, choose any one of the ex-post mean-variance efficient portfolios. Roll (1977) shows that there is an exact linear relationship between the sample mean returns of the individual assets and their betas, computed with respect to the mean-variance efficient portfolio.

More generally, rational approaches to the cross-sectional evidence face a number of other obstacles. First, rational models typically measure risk as the covariance of returns with marginal utility of consumption. Stocks are risky if they fail to pay out at times of high marginal utility – in "bad" times – and instead pay out when marginal utility is low – in "good" times. The problem is that for many of the above findings, there is little evidence that the portfolios with anomalously *high* average returns do poorly in bad times, whatever plausible measure of bad times is used. For example, Lakonishok, Shleifer and Vishny (1994) show that in their 1968 to 1989 sample period, value stocks do well relative to growth stocks even when the economy is in recession. Similarly, De Bondt and Thaler (1987) find that their loser stocks have higher betas than winners in up markets and lower betas in down markets – an attractive combination that no one would label "risky".

Second, some of the portfolios in the above studies – the decile of stocks with the lowest book-to-market ratios for example – earn average returns below the risk-free rate. It is not easy to explain why a rational investor would willingly accept a lower return than the T-Bill rate on a volatile portfolio.

Finally, in some of the examples given above, it is not just that one portfolio outperforms another on average. In some cases, the outperformance is present in almost every period of the sample. For example, in Bernard and Thomas' (1989) study, firms with surprisingly good earnings outperform those with surprisingly poor earnings in 46 out of the 50 quarters studied. It is not easy to see any risk here than might justify the outperformance.

There are a number of behavioral models which try to explain some of the above phenomena. We classify them based on whether their mechanism centers on beliefs or on preferences.

5.1 Belief-based Models

Barberis, Shleifer and Vishny (1998), BSV henceforth, argue that much of the above evidence is the result of systematic errors that investors make when they use public information to form expectations of future cash flows. They build a model that incorporates two of the updating biases from Section 3: conservatism, the tendency to underweight new information relative to priors; and representativeness, and in particular the version of representativeness known as the law of small numbers, whereby people expect even short samples to reflect the properties of the parent population.

When a company announces surprisingly good earnings, conservatism means that investors react insufficiently, pushing the price up too little. Since the price is too low, subsequent returns will be higher on average, thereby generating both post-earnings announcement drift and momentum. After a *series* of good earnings announcements, though, representativeness causes people to overreact and push the price up too high. The reason is that after many periods of good earnings, the law of small numbers leads investors to believe that this is a firm with particularly high earnings growth, and hence to forecast high earnings in the future. After all, the firm cannot be "average". If it were, then according the to law of small numbers, its earnings should *appear* average, even in short samples. Since the price is now too high, subsequent returns are too low on average, thereby generating long-term reversals and a scaled-price ratio effect.

To capture these ideas mathematically, BSV consider a model with a representative risk neutral investor in which the true earnings process for all assets is a random walk. Investors, however, do not use the random walk model to forecast future earnings. They think that at any time, earnings are being generated by one of two regimes: a "mean-reverting" regime, in which earnings are more mean-reverting than in reality, and a "trending" regime in which earnings trend more than in reality. The investor believes that the regime generating earnings changes exogenously over time and sees his task as trying to figure out which of the two regimes is currently generating earnings.

This framework offers one way of modelling the updating biases described above. Including a "trending" regime in the model captures the effect of representativeness by allowing investors to put more weight on trends than they should. Conservatism suggests that people may put too little weight on the latest piece of earnings news relative to their prior beliefs. In other words, when they get a good piece of earnings news, they effectively act as if part of the shock will be reversed in the next period, in other words, as if they believe in a "mean-reverting" regime. BSV confirm that for a wide range of parameter values, this model does indeed generate post-earnings announcement drift, momentum, long-term reversals and cross-sectional forecasting power for scaled-price ratios.[28]

Daniel, Hirshleifer and Subrahmanyam (1998, 2001), DHS henceforth, stress biases in the interpretation of *private*, rather than public information. Imagine that the investor does some research on his own to try to determine a firm's future cash flows. DHS assume that he is overconfident about this information; in particular, they argue that investors are more likely to be overconfident about private information they have worked hard to generate than about public information. If the private information is positive, overconfidence means that investors will push prices up too far relative to fundamentals. Future public information will slowly pull prices back to their correct value, thus generating long-term reversals and a scaled-price effect. To get momentum and a post-earnings announcement effect, DHS assume that the public information alters the investor's confidence in his original private information in an asymmetric fashion, a phenomenon known as self-attribution bias: public news which confirms the investor's research strongly increases the confidence he has in that research. Disconfirming public news, though, is given less attention, and the investor's confidence in the private information remains unchanged. This asymmetric response means that initial

[28] Poteshman (2001) finds evidence of a BSV-type expectations formation process in the options market. He shows that when pricing options, traders appear to underreact to individual daily changes in instantaneous variance, while overreacting to longer sequences of increasing or decreasing changes in instantaneous variance.

overconfidence is on average followed by even greater overconfidence, generating momentum.

Chopra, Lakonishok and Ritter (1992) and La Porta et al. (1997) provide compelling evidence that supports the idea that investors make irrational forecasts of future cash flows. If, as BSV and DHS argue, long-term reversals and the predictive power of scaled-price ratios are driven by excessive optimism or pessimism about future cash flows followed by a correction, then most of the correction should occur at those times when investors find out that their initial beliefs were too extreme, in other words, at earnings announcement dates. The data strongly confirms this prediction. Chopra et al. show that after portfolio formation, De Bondt and Thaler's (1985) "winner" portfolio performs particularly poorly in the few days around earnings' announcements. La Porta et al. show that the same is true for a portfolio of growth stocks. It is very hard to give a rational reason for why these portfolios earn such low average returns over just a few days of the year.

Perhaps the simplest way of capturing much of the cross-sectional evidence is positive feedback trading, where investors buy more of an asset which has recently gone up in value (De Long et al., 1990b, Barberis and Shleifer, 2003). If a company's stock price goes up this period on good earnings, positive feedback traders buy the stock in the following period, causing a further price rise. On the one hand, this generates momentum and post-earnings announcement drift. On the other hand, since the price has now risen above what is justified by fundamentals, subsequent returns will on average be too low, generating long-term reversals and a scaled-price ratio effect.

The simplest way of motivating positive feedback trading is extrapolative expectations, where investors' expectations of future returns are based on past returns. This in turn, may be due to representativeness and to the law of small numbers in particular. The same argument made by BSV as to why investors might extrapolate past cash flows too far into the future can be applied here to explain why they might extrapolate past *returns* too far into the future. De Long et al. (1990b) note that institutional features such as portfolio insurance or margin calls can also generate positive feedback trading.

Positive feedback trading also plays a central role in the model of Hong and Stein (1999), although in this case it emerges endogenously from more primitive assumptions. In this model, two boundedly rational groups of investors interact, where bounded rationality means that investors are only able to process a subset of available information. "Newswatchers" make forecasts based on private information, but do not condition on past prices. "Momentum traders" condition only on the most recent price change.

Hong and Stein also assume that private information diffuses slowly through the population of newswatchers. Since these investors are unable to extract each others' private information from prices, the slow diffusion generates momentum. Momentum traders are then added to the mix. Given what they are allowed to condition on, their optimal strategy is to engage in positive feedback trading: a price increase last period is a sign that good private information is diffusing through the economy. By buying, momentum traders hope to profit from the continued diffusion of information. This behavior preserves momentum, but also generates price reversals: since momentum traders cannot observe the extent of news diffusion, they keep buying even after price has reached fundamental value, generating an overreaction that is only later reversed.

These four models differ most in their explanation of momentum. In two of the models – BSV and Hong and Stein (1999) – momentum is due to an initial underreaction followed by a correction. In De Long et al. (1990b) and DHS, it is due to an initial overreaction followed by even more overreaction. Within each pair, the stories are different again.[29]

Hong, Lim and Stein (2000) present supportive evidence for the view of HS that momentum is due simply to slow diffusion of private information through the economy. They argue that the diffusion of information will be particularly slow among small firms and among firms with low analyst coverage, and that the momentum effect should therefore be more prominent there, a prediction they confirm in the data. They also find that among firms with low analyst coverage, momentum is almost entirely driven by prior losers continuing to lose. They argue that this too, is consistent with a diffusion story. If a firm not covered by analysts is sitting on good news, it will do its best to convey the news to as many people as possible, and as quickly as possible; bad news, however, will be swept under the carpet, making its diffusion much slower.

5.2 Belief-based Models with Institutional Frictions

Some authors have argued that models which combine mild assumptions about investor irrationality with institutional frictions may offer a fruitful way of thinking about some of the anomalous cross-sectional evidence.

The institutional friction that has attracted the most attention is short-sale constraints. As mentioned in Section 2.2., these can be thought of as anything which makes investors less willing to establish a short position than a long one. They include the direct cost of shorting, namely the lending fee; the risk that the loan is recalled by the lender at an inopportune moment; as well as legal restrictions: a large fraction of mutual funds are not allowed to short stocks.

Several papers argue that when investors differ in their beliefs, the existence of short-sale constraints can generate deviations from fundamental value and in particular, explain why stocks with high price-earnings ratios earn lower average returns in the cross-section. The simplest way of motivating the assumption of heterogeneous beliefs is overconfidence, which

[29] In particular, the models make different predictions about how individual investors would trade following certain sequences of past returns. Armed with transaction-level data, Hvidkjaer (2001) exploits this to provide initial evidence that may distinguish the theories.

is why that assumption is often thought of as capturing a mild form of irrationality. In the absence of overconfidence, investors' beliefs converge rapidly as they hear each other's opinions and hence deduce each other's private information.

There are at least two mechanisms through which differences of opinion and short-sale constraints can generate price-earnings ratios that are too high, and thereby explain why price-earnings ratios predict returns in the cross-section.

Miller (1977) notes that when investors hold different views about a stock, those with bullish opinions will, of course, take long positions. Bearish investors, on the other hand, want to short the stock, but being unable to do so, they sit out of the market. Stock prices therefore reflect only the opinions of the most optimistic investors which, in turn, means that they are too high and that they will be followed by lower returns.

Harrison and Kreps (1978) and Scheinkman and Xiong (2001) argue that in a dynamic setting, a second, speculation-based mechanism arises. They show that when there are differences in beliefs, investors will be happy to buy a stock for more than its fundamental value in anticipation of being able to sell it later to other investors even more optimistic than themselves. Note that short-sale constraints are essential to this story: in their absence, an investor can profit from another's greater optimism by simply shorting the stock. With short-sale constraints, the only way to do so is to buy the stock first, and then sell it on later.

Both types of models make the intriguing prediction that stocks which investors disagree about more will have higher price-earnings ratios and lower subsequent returns. Three recent papers test this prediction, each using a different measure of differences of opinion.

Dieter, Malloy and Scherbina (2003) use IBES data on analyst forecasts to obtain a direct measure of heterogeneity of opinion. They group stocks into quintiles based on the level of dispersion in analysts' forecasts of current year earnings and confirms that the highest dispersion portfolio earns lower average returns than the lowest dispersion portfolio.

Chen, Hong and Stein (2002) use "breadth of ownership" – defined roughly as the fraction of mutual funds that hold a particular stock – as a proxy for divergence of opinion about the stock. The more dispersion in opinions there is, the more mutual funds will need to sit out the market due to short sales constraints, leading to lower breadth. Chen et al. predict, and confirm in the data, that stocks experiencing a decrease in breadth subsequently have lower average returns compared to stocks whose breadth increases.

Jones and Lamont (2002) use the cost of short-selling a stock – in other words, the lending fee – to measure differences of opinion about that stock. The idea is that if there is a lot of disagreement about a stock's prospects, many investors will want to short the stock, thereby pushing up the cost of doing so. Jones and Lamont confirm that stocks with higher lending fees have higher price-earnings ratios and earn lower subsequent returns. It is interesting to note that their data set spans the years from 1926 to 1933. At that time, there existed a centralized market for borrowing stocks and lending fees were published daily in the Wall Street Journal. Today, by contrast, stock lending is an over-the-counter market, and data on lending fees is harder to come by.

In other related work, Hong and Stein (2003) analyze the implications of short sales constraints and differences of opinion for higher order moments, and show that they lead to skewness. The intuition is that when a stock's price goes down, more information is revealed: by seeing at what point they enter into the market, we learn the valuations of those investors whose pessimistic views could not initially be reflected in the stock price, because of short sales constraints. When the stock market goes up, the sidelined investors stay out of the market and there is less information revelation. This increase in volatility after a downturn is the source of the skewness.

One prediction of this idea is that stocks which investors disagree about more should exhibit greater skewness. Chen, Hong and Stein (2001) test this idea using increases in turnover as a sign of investor disagreement. They show that stocks whose turnover increases subsequently display greater skewness.

5.3 Preferences

Earlier, we discussed Barberis, Huang and Santos (2001) which tries to explain *aggregate* stock market behavior by combining loss aversion and narrow framing with an assumption about how the degree of loss aversion changes over time. Barberis and Huang (2001) show that applying the same ideas to individual stocks can generate the evidence on long-term reversals and on scaled-price ratios. The key idea is that when investors hold a number of different stocks, narrow framing may induce them to derive utility from gains and losses in the value of *individual* stocks. The specification of this additional source of utility is exactly the same as in BHS, except that it is now applied at the individual stock level instead of at the portfolio level: the investor is loss averse over individual stock fluctuations and the pain of a loss on a specific stock depends on that stock's past performance.

To see how this model generates a value premium, consider a stock which has had poor returns several periods in a row. Precisely because the investor focuses on individual stock gains and losses, he finds this painful and becomes especially sensitive to the possibility of further losses on the stock. In effect, he perceives the stock as riskier, and discounts its future cash flows at a higher rate: this lowers its price-earnings ratio and leads to higher subsequent returns, generating a value premium. In one sense, this model is narrower than those in the "beliefs" section, Section 5.1., as it does not claim to address momentum. In another sense, it is broader, in that it simultaneously explains the equity premium and derives the risk-free rate endogenously.

The models we describe in Sections 5.1., 5.2., and 5.3 focus primarily on momentum, long-term reversals, the predictive power of scaled-price ratios and post-earnings announcement drift. What about the other examples of anomalous evidence with which we began Section 5? In Section 7, we argue that the long-run return patterns following equity issuance and repurchases may be the result of rational managers responding to the kinds of noise traders analyzed in the preceding behavioral models. In short, if investors cause prices to swing away from fundamental value, managers may try to time these cycles, issuing equity when it is overpriced, and repurchasing it when it is cheap. In such a world, equity issues will indeed be followed by low returns, and repurchases by high returns. The models we have discussed so far do not, however, shed light on the size anomaly, nor on the dividend announcement event study.

6 Application: Closed-end Funds and Comovement

6.1 Closed-end Funds

Closed-end funds differ from more familiar open-end funds in that they only issue a fixed number of shares. These shares are then traded on exchanges: an investor who wants to buy a share of a closed-end fund must go to the exchange and buy it from another investor at the prevailing price. By contrast, should he want to buy a share of an open-end fund, the fund would create a new share and sell it to him at its net asset value, or NAV, the per share market value of its asset holdings.

The central puzzle about closed-end funds is that fund share prices differ from NAV. The typical fund trades at a discount to NAV of about 10 percent on average, although the difference between price and NAV varies substantially over time. When closed-end funds are created, the share price is typically above NAV; when they are terminated, either through liquidation or open-ending, the gap between price and NAV closes.

A number of rational explanations for the average closed-end fund discount have been proposed. These include expenses, expectations about future fund manager performance, and tax liabilities. These factors can go some way to explaining certain aspects of the closed-end fund puzzle. However, none of them can satisfactorily explain *all* aspects of the evidence. For example, agency costs such as management fees can explain why funds usually sell at discounts, but not why they typically initially sell at a premium, nor why discounts tend to vary from week to week.

Lee, Shleifer and Thaler (1991), LST henceforth, propose a simple behavioral view of these closed-end fund puzzles. They argue that some of the individual investors who are the primary owners of closed-end funds are noise traders, exhibiting irrational swings in their expectations about future fund returns. Sometimes they are too optimistic, while at other times, they are too pessimistic. Changes in their sentiment affect fund share prices and hence also the difference between prices and net asset values.[30]

This view provides a clean explanation of all aspects of the closed-end fund puzzle. Owners of closed-end funds have to contend with two sources of risk: fluctuations in the value of the funds' assets, and fluctuations in noise trader sentiment. If this second risk is systematic – we return to this issue shortly – rational investors will demand compensation for it. In other words, they will require that the fund's shares trade at a discount to NAV.

This also explains why new closed-end funds are often sold at a premium. Entrepreneurs will choose to create closed-end funds at times of investor exuberance, when they know that they can sell fund shares for more than they are worth. On the other hand, when a closed-end fund is liquidated, rational investors no longer have to worry about changes in noise trader sentiment because they know that at liquidation, the fund price will equal NAV. They therefore no longer demand compensation for this risk, and the fund price rises towards NAV.

An immediate prediction of the LST view is that prices of closed-end funds should comove strongly, even if the cash-flow fundamentals of the assets held by the funds do not: if noise traders become irrationally pessimistic, they will sell closed-end funds across the board, depressing their prices regardless of cash-flow news. LST confirm in the data that closed-end fund discounts are highly correlated.

The LST story depends on noise trader risk being systematic. There is good reason to think that it is. If the noise traders who hold closed-end funds also hold other assets, then negative changes in sentiment, say, will drive down the prices of closed-end funds *and* of their other holdings, making the noise trader risk systematic. To check this, LST compute the correlation of closed-end fund discounts with another group of assets primarily owned by individuals, small stocks. Consistent with the noise trader risk being systematic, they find a significant positive correlation.

6.2 Comovement

The LST model illustrates that behavioral models can make interesting predictions not only about the *average* level of returns, but also about patterns of comovement. In particular, it explains why the prices of closed-end funds comove so strongly, and also why closed-end funds as a class comove with small stocks. This raises the hope that behavioral models might be able to explain other puzzling instances of comovement as well.

[30] For the noise traders to affect the *difference* between price and NAV rather than just price, it must be that they are more active traders of closed-end fund shares than they are of assets owned by the funds. As evidence for this, LST point out that while funds are primarily owned by individual investors, the funds' assets are not.

Before studying this in more detail, it is worth setting out the traditional view of return comovement. The simplest rational explanation of return comovement is that it is due to cash-flow comovement: there will be a common factor in the returns of a group of assets if there is a common factor in news about their future earnings. There is little doubt that many instances of return comovement can be explained by cash flows: stocks in the automotive industry move together primarily because their earnings are correlated.

The closed-end fund evidence shows that cash-flow view of comovement is at best, incomplete: in that case, the prices of closed-end funds comove even though their fundamentals do not.[31] Other evidence is just as puzzling. Froot and Dabora (1999) study "twin stocks", which are claims to the same cash-flow stream, but are traded in different locations. The Royal Dutch/Shell pair, discussed in Section 2, is perhaps the best known example. If return comovement is simply a reflection of cash-flow comovement, these two stocks should be perfectly correlated. In fact, as Froot and Dabora show, Royal Dutch comoves strongly with the S&P 500 index of U.S. stocks, while Shell comoves with the FTSE index of U.K. stocks.

Fama and French (1993) uncover salient common factors in the returns of small stocks, as well as in the returns on value stocks. In order to test the rational view of comovement, Fama and French (1995) investigate whether these strong common factors can be traced to common factors in the earnings of these stocks. While they do uncover a common factor in the earnings of small stocks, as well as in the earnings of value stocks, these cash-flow factors are weaker than the factors in returns and there is little evidence that the return factors are driven by the cash-flow factors. Once again, there appears to be comovement in returns that has little to do with cash-flow comovement.[32]

In response to this evidence, researchers have begun to posit behavioral theories of comovement. LST is one such theory. To state their argument more generally, they start by observing that many investors choose to trade only a subset of all available securities. As these investors' risk aversion or sentiment changes, they alter their exposure to the particular securities they hold, thereby inducing a common factor in the returns of these securities. Put differently, this "habitat" view of comovement predicts that there will be a common factor in the returns of securities that are the primary holdings of a specific subset of investors,

[31] Bodurtha et al. (1993) and Hardouvelis et al. (1994) provide further interesting examples of a delinking between cash-flow comovement and return comovement in the closed-end fund market. They study closed-end *country* funds, whose assets trade in a different location from the funds themselves and find that the funds comove as much with the national stock market in the country where they are traded as with the national stock market in the country where their *assets* are traded. For example, a closed-end fund invested in German equities but traded in the U.S. typically comoves as much with the U.S. stock market as with the German stock market.

[32] In principle, comovement can also be rationally generated through changes in discount rates. However, changes in interest rates or risk aversion induce a common factor in the returns on *all* stocks, and do not explain why a particular group of stocks comoves. A common factor in news about the risk of certain assets may also be a source of comovement for those assets, but there is little direct evidence to support such a mechanism in the case of small stocks or value stocks.

such as individual investors. This story seems particularly appropriate for thinking about closed-end funds, and also for Froot and Dabora's evidence.

A second behavioral view of comovement was recently proposed by Barberis and Shleifer (2003). They argue that to simplify the portfolio allocation process, many investors first group stocks into categories such as small-cap stocks or automotive industry stocks, and then allocate funds across these various categories. If these categories are also adopted by noise traders, then as these traders move funds from one category to another, the price pressure from their coordinated demand will induce common factors in the returns of stocks that happen to be classified into the same category, even if those stocks' cash flows are largely uncorrelated. In particular, this view predicts that when an asset is added to a category, it should begin to comove more with that category than before.

Barberis, Shleifer and Wurgler (2001) test this "category" view of comovement by taking a sample of stocks that has been added to the S&P 500, and computing the betas of these stocks with the S&P 500 both before and after they are included. Based on both univariate and multivariate regressions, they show that upon inclusion, a stock's beta with the S&P 500 rises significantly, as does the fraction of its variance that is explained by the S&P 500, while its beta with stocks outside the index falls.[33] This result does not sit well with the cash-flow view of comovement – addition to the S&P 500 carries no information about the covariance of a stock's cash flows with other stocks' cash flows – but emerges naturally from a model where prices are affected by category-level demand shocks. Little is known, at this point, about how investors form categories in the first place, but an intriguing start on this problem is provided by Mullainathan (2000).

7 Application: Investor Behavior

Behavioral finance has also had some success in explaining how certain groups of investors behave, and in particular, what kinds of portfolios they choose to hold and how they trade over time. The goal here is less controversial than in the previous three sections: it is simply to explain the actions of certain investors, and not necessarily to claim that these actions also affect prices. Two factors make this type of research increasingly important. First, now that the costs of entering the stock market have fallen, more and more individuals are investing in equities. Second, the world-wide trend toward defined contribution retirement savings plans, and the possibility of individual accounts in social security systems mean that individuals are more responsible for their own financial well-being in retirement. It is therefore natural to ask how well they are handling these tasks.

We now describe some of the evidence on the actions of investors and the behavioral ideas that have been used to explain it.

[33] Similar results from univariate regressions can also be found in earlier work by Vijh (1994).

Insufficient Diversification

A large body of evidence suggests that investors diversify their portfolio holdings much less than is recommended by normative models of portfolio choice.

First, investors exhibit a pronounced "home bias". French and Poterba (1991) report that investors in the U.S., Japan and the U.K. allocate 94 percent, 98 percent, and 82 percent of their overall equity investment, respectively, to *domestic* equities. It has not been easy to explain this fact on rational grounds (Lewis, 1999). Indeed, normative portfolio choice models that take human capital into account typically advise investors to *short* their national stock market, because of its high correlation with their human capital (Baxter and Jermann, 1997).

Some studies have found an analog to home bias *within* countries. Using an especially detailed data set from Finland, Grinblatt and Keloharju (2001) find that investors in that country are much more likely to hold and trade stocks of Finnish firms which are located close to them geographically, which use their native tongue in company reports, and whose chief executive shares their cultural background. Huberman (2001) studies the geographic distribution of shareholders of U.S. Regional Bell Operating Companies (RBOCs) and finds that investors are much more likely to hold shares in their local RBOC than in out-of-state RBOCs. Finally, studies of allocation decisions in 401(k) plans find a strong bias towards holding own company stock: over 30 percent of defined contribution plan assets in large U.S. companies are invested in employer stock, much of this representing voluntary contributions by employees (Benartzi, 2001).

In Section 3, we discussed evidence showing that people dislike ambiguous situations, where they feel unable to specify a gamble's probability distribution. Often, these are situations where they feel that they have little competence in evaluating a certain gamble. On the other hand, people show an excessive liking for familiar situations, where they feel they are in a better position than others to evaluate a gamble.

Ambiguity and familiarity offer a simple way of understanding the different examples of insufficient diversification. Investors may find their national stock markets more familiar – or less ambiguous – than foreign stock indices; they may find firms situated close to them geographically more familiar than those located further away; and they may find their employer's stock more familiar than other stocks.[34] Since familiar assets are attractive, people invest heavily in those, and invest little or nothing at all in ambiguous assets. Their portfolios therefore appear undiversified relative to the predictions of standard models that ignore the investor's degree of confidence in the probability distribution of a gamble.

Not all evidence of home bias should be interpreted as a preference for the familiar.

[34] Particularly relevant to this last point is survey data showing that people consider their own company stock less risky than a diversified index (Driscoll et al., 1995).

Coval and Moskowitz (1999) show that U.S. mutual fund managers tend to hold stocks whose company headquarters are located close to their funds' headquarters. However, Coval and Moskowitz's (2001) finding that these local holdings subsequently perform well suggests that an information story is at work here, not a preference for the familiar. It is simply less costly to research local firms and so fund managers do indeed focus on those firms, picking out the stocks with higher expected returns. There is no obvious information-based explanation for the results of French and Poterba (1991), Huberman (2001) or Benartzi (2001), while Grinblatt and Keloharju (2001) argue against such an interpretation of their findings.

Naive Diversification

Benartzi and Thaler (2001) find that when people *do* diversify, they do so in a naive fashion. In particular, they provide evidence that in 401(k) plans, many people seem to use strategies as simple as allocating $1/n$ of their savings to each of the n available investment options, whatever those options are. Some evidence that people think in this way comes from the laboratory. Benartzi and Thaler ask subjects to make an allocation decision in each of the following three conditions: first, between a stock fund and a bond fund; next, between a stock fund and a balanced fund, which invests 50 percent in stocks and 50 percent in bonds; and finally, between a bond fund and a balanced fund. They find that in all three cases, a 50:50 split across the two funds is a popular choice, although of course this leads to very different effective choices between stocks and bonds: the average allocation to stocks in the three conditions was 54 percent, 73 percent and 35 percent respectively.

The $1/n$ diversification heuristic or other naive diversification strategies predicts that in 401(k) plans which offer predominantly stock funds, investors will allocate more to stocks. Benartzi and Thaler test this in a sample of 170 large retirement savings plans. They divide the plans into three groups based on the fraction of funds – low, medium, or high – they offer that are stock funds. The allocation to stocks increases across the three groups, from 49 percent to 60 percent to 64 percent, confirming the initial prediction.

Excessive Trading

One of the clearest predictions of rational models of investing is that there should be very little trading. In a world where rationality is common knowledge, I am reluctant to buy if you are ready to sell. In contrast to this prediction, the volume of trading on the world's stock exchanges is very high. Furthermore, studies of individuals and institutions suggest that both groups trade more than can be justified on rational grounds.

Barber and Odean (2000) examine the trading activity from 1991 to 1996 in a large sample of accounts at a national discount brokerage firm. They find that after taking trading costs into account, the average return of investors in their sample is well below the return of standard benchmarks. Put simply, these investors would do a lot better if they traded less. The underperformance in this sample is largely due to transaction costs. However, there is also some evidence of poor security selection: in a similar data set covering the 1987 to 1993 time period, Odean (1999) finds that the average gross return of stocks that investors buy, over the year after they buy them, is lower than the average gross return of stocks that they sell, over the year after they sell them.

The most prominent behavioral explanation of such excessive trading is overconfidence: people believe that they have information strong enough to justify a trade, whereas in fact the information is too weak to warrant any action. This hypothesis immediately predicts that people who are more overconfident will trade more and, because of transaction costs, earn lower returns. Consistent with this, Barber and Odean (2000) show that the investors in their sample who trade the most earn by far the lowest average returns. Building on evidence that men are more overconfident than women, and using the same data as in their earlier study, Barber and Odean (2001) predict and confirm that men trade more and earn lower returns on average.

Working with the same data again, Barber and Odean (2002a) study the subsample of individual investors who switch from phone-based to online trading. They argue that for a number of reasons, the switch should be accompanied by an increase in overconfidence. First, better access to information and a greater degree of control – both features of an online trading environment – have been shown to increase overconfidence. Moreover, the investors who switch have often earned high returns prior to switching, which may only increase their overconfidence further. If this is indeed the case, they should trade more actively after switching and perform worse. Barber and Odean confirm these predictions.

The Selling Decision

Several studies find that investors are reluctant to sell assets trading at a loss relative to the price at which they were purchased, a phenomenon labelled the "disposition effect" by Shefrin and Statman (1985). Working with the same discount brokerage data used in the Odean (1999) study from above, Odean (1998) finds that the individual investors in his sample are more likely to sell stocks which have gone up in value relative to their purchase price, rather than stocks which have gone down.

It is hard to explain this behavior on rational grounds. Tax considerations point to the selling of losers, not winners.[35] Nor can one argue that investors rationally sell the winners because of information that their future performance will be poor. Odean reports that the average performance of stocks that people sell is better than that of stocks they hold on to.

Two behavioral explanations of these findings have been suggested. First, investors may have an irrational belief in mean-reversion. A second possibility relies on prospect theory and narrow framing. We have used these ingredients before, but this time it is not loss

[35] Odean (1998) does find that in December, investors prefer to sell past losers rather than past winners, but overall, this effect is swamped by a strong preference for selling past winners in the remaining 11 months.

aversion that is central, but rather the concavity (convexity) of the value function in the region of gains (losses).

To see the argument, suppose that a stock that was originally bought at \$50 now sells for \$55. Should the investor sell it at this point? Suppose that the gains and losses of prospect theory refer to the sale price minus the purchase price. In that case, the utility from selling the stock now is $v(5)$. Alternatively, the investor can wait another period, whereupon we suppose that the stock could go to \$50 or \$60 with equal probability; in other words, we abstract from belief-based trading motives by saying that the investor expects the stock price to stay flat. The expected value of waiting and selling next period is then $\frac{1}{2}v(0) + \frac{1}{2}v(10)$. Since the value function v is concave in the region of gains, the investor sells now. In a different scenario, the stock may currently be trading at \$45. This time, the comparison is between $v(-5)$ and $\frac{1}{2}v(-10) + \frac{1}{2}v(0)$, assuming a second period distribution of \$40 and \$50 with equal probability. Convexity of v pushes the investor to wait. Intuitively, by not selling, he is gambling that the stock will eventually break even, saving him from having to experience a painful loss.

The disposition effect is not confined to individual stocks. In an innovative study, Genesove and Mayer (2001) find evidence of a reluctance to sell at a loss in the housing market. They show that sellers whose expected selling price is below their original purchase price, set an asking price that exceeds the asking price of sellers with comparable houses. Moreover, this is not simply wishful thinking on the sellers' part that is later corrected by the market: sellers facing a possible loss do actually transact at considerably higher prices than other sellers.

Coval and Shumway (2000) study the behavior of professional traders in the Treasury Bond futures pit at the CBOT. If the gains and losses of prospect theory are taken to be daily profits and losses, the curvature of the value function implies that traders with profits (losses) by the middle of the trading day will take less (more) risk in their afternoon trading. This prediction is borne out in the data.

Grinblatt and Han (2001) argue that the investor behavior inherent in the disposition effect may be behind a puzzling feature of the cross-section of average returns, namely momentum in stock returns. Due to the concavity of the value function in the region of gains, investors will be keen to sell a stock which has earned them capital gains on paper. The selling pressure that results may initially depress the stock price, generating higher returns later. On the other hand, if the holders of a stock are facing capital losses, convexity in the region of losses means that they will only sell if offered a price premium; the price is therefore initially inflated, generating lower returns later. Grinblatt and Han provide supportive evidence for their story by regressing, in the cross-section, a stock's return on its past 12-month return as well as on a measure of the capital gain or loss faced by its holders. This last variable is computed as the current stock price minus investors' average cost basis, itself inferred from past volume. They find that the capital gain or loss variable steals a substantial amount of explanatory power from the past return.

The Buying Decision

Odean (1999) presents useful information about the stocks the individual investors in his sample choose to buy. Unlike "sells", which are mainly prior winners, "buys" are evenly split between prior winners and losers. Conditioning on the stock being a prior winner (loser) though, the stock is a big prior winner (loser). In other words, a good deal of the action is in the extremes.

Odean argues that the results for stock purchases are in part due to an attention effect. When buying a stock, people do not tend to systematically sift through the thousands of listed shares until they find a good "buy." They typically buy a stock that has caught their attention and perhaps the best attention draw is extreme past performance, whether good or bad.

Among individual investors, attention is less likely to matter for stock sales because of a fundamental way in which the selling decision differs from the buying decision. Due to short sales constraints, when individuals are looking for a stock to sell, they limit their search to those stocks that they currently own. When buying stocks, though, people have a much wider range of possibilities to choose from, and factors more related to attention may enter the decision.

Using the same discount brokerage data as in their earlier papers, Barber and Odean (2002b) test the idea that for individual investors, buying decisions are more driven by attention than are selling decisions. On any particular day, they create portfolios of "attention-getting" stocks using a number of different criteria: stocks with abnormally high trading volume, stocks with abnormally high or low returns, and stocks with news announcements. They find that whichever criterion is used, the individual investors in their sample are more likely to be purchasers of these high-attention stocks than sellers.

8 Application: Corporate Finance

8.1 Security Issuance, Capital Structure and Investment

An important strand of research in behavioral finance asks whether irrational investors such as those discussed in earlier sections affect the financing and investment decisions of firms.

We first address this question theoretically, and ask how a rational manager interested in maximizing true firm value – in other words, the stock price that will prevail once any mispricing has worked its way out of valuations – should act in the face of irrational investors. Stein (1996) provides a useful framework for thinking about this, as well as about other issues that arise in this section. He shows that when a firm's stock price is too high, the rational manager should issue more shares so as to take advantage of investor exuberance. Conversely, when the price is too low, he should repurchase shares. We refer to this model of security issuance as the "market timing" view.

What evidence there is to date on security issuance appears remarkably consistent with this framework. First, at the aggregate level, the share of new equity issues among total new issues – the "equity share" – is higher when the overall stock market is more highly valued. In fact, Baker and Wurgler (2000) show that the equity share is a reliable predictor of future stock returns: a high share predicts low, and sometimes negative stock returns. This is consistent with managers timing the market, issuing more equity at its peaks, just before it sinks back to more realistic valuation levels.

At the individual firm level, a number of papers have shown that the book-to-market ratio of a firm is a good cross-sectional predictor of new equity issuance (see Koracjzyk, Lucas and Macdonald 1991, Jung, Kim and Stulz 1996, Loughran, Ritter and Rydqvist 1994, Pagano, Panneta and Zingales 1998, Baker and Wurgler 2002a). Firms with high valuations issue more equity while those with low valuations repurchase their shares. Moreover, long-term stock returns after an IPO or SEO are low (Loughran and Ritter, 1995), while long term returns after the announcement of a repurchase are high (Ikenberry, Lakonishok and Vermaelen, 1995). Once again, this evidence is consistent with managers timing the market in their own securities.

More support for the market timing view comes from survey evidence. Graham and Harvey (2001) report that 67 percent of surveyed CFOs said that "the amount by which our stock is undervalued or overvalued" was an important consideration when issuing common stock.

The success of the market timing framework in predicting patterns of equity issuance offers the hope that it might also be the basis of a successful theory of capital structure. After all, a firm's capital structure simply represents its cumulative financing decisions over time. Consider, for example, two firms which are similar in terms of characteristics like firm size, profitability, fraction of tangible assets, and current market-to-book ratio, which have traditionally been thought to affect capital structure. Suppose, however, that in the past, the market-to-book ratio of firm A has reached much higher levels than that of firm B. Since, under the market timing theory, managers of firm A may have issued more shares at that time to take advantage of possible overvaluation, firm A may have more equity in its capital structure today.

In an intriguing recent paper, Baker and Wurgler (2002a) confirm this prediction. They show that all else equal, a firm's weighted-average historical market-to-book ratio, where more weight is placed on years in which the firm made an issuance of some kind, whether debt or equity, is a good cross-sectional predictor of the fraction of equity in the firm's capital structure today.

There is some evidence, then, that irrational investor sentiment affects financing decisions. We now turn to the more critical question of whether this sentiment affects actual investment decisions. Once again, we consider the benchmark case in Stein's (1996) model, in which the manager is both rational and interested in maximizing the firm's true value.

Suppose that a firm's stock price is too high. As discussed above, the manager should issue more equity at this point. More subtly, though, Stein shows that he should *not* channel the fresh capital into any actual new investment, but instead keep it in cash or in another fairly priced capital market security. While investors' exuberance means that, in *their* view, the firm has many positive net present value (NPV) projects it could undertake, the rational manager knows that these projects are not, in fact, positive NPV and that in the interest of true firm value, they should be avoided. Conversely, if the manager thinks that his firm's stock price is irrationally low, he should repurchase shares at the advantageously low price but not scale back actual investment. In short, irrational investors may affect the timing of security issuance, but they should not affect the firm's investment plans.

Once we move beyond this simple benchmark case, though, there emerge several channels through which sentiment might affect investment after all. First, the above argument properly applies only to *non-equity dependent* firms; in other words, to firms which because of their ample internal funds and borrowing capacity do not need the equity markets to finance their marginal investments.

For equity-dependent firms, however, investor sentiment and, in particular, excessive investor pessimism, may distort investment: when investors are excessively pessimistic, such firms may have to forgo attractive investment opportunities because it is too costly to finance them with undervalued equity. This thinking leads to a cross-sectional prediction, namely that the investment of equity-dependent firms should be more sensitive to gyrations in stock price than the investment of non-equity dependent firms.

Other than this equity-dependence mechanism, there are other channels through which investor sentiment might distort investment. Consider the case where investors are excessively optimistic about a firm's prospects. Even if a manager is in principle interested in maximizing true value, he faces the danger that if he refuses to undertake projects investors perceive as profitable, they may depress stock prices, exposing him to the risk of a takeover, or more simply, try to have him fired.[36]

Even if the manager is rational, this does not mean he will choose to maximize the firm's

[36] Shleifer and Vishny (2001) argue that in a situation such as this, where the manager feels forced to undertake some kind of investment, the best investment of all may be an acquisition of a less overvalued firm, in other words, one more likely to retain its value in the long run. This observation leads to a parsimonious theory of takeover waves, which predicts, among other things, an increase in stock-financed acquisitions at times of high dispersion in valuations.

true value. The agency literature has argued that some managers may maximize other objectives – the size of their firm, say – as a way of enhancing their prestige. This suggests another channel for investment distortion: managers might use investor exuberance as a cover for doing negative NPV "empire building" projects.

Finally, investor sentiment can also affect investment if managers put some weight on investors' opinions, perhaps because they think investors know something they don't. Managers may then mistake excessive optimism for well-founded optimism and get drawn into making negative NPV investments.

An important goal of empirical research, then, is to try to understand whether sentiment does affect investment, and if so, through which channel. Early studies produced little evidence of investment distortion. In aggregate data, Blanchard, Rhee and Summers (1993) find that movements in price apparently unrelated to movements in fundamentals have only weak forecasting power for future investment: the effects are marginally statistically significant and weak in economic terms. To pick out two particular historical episodes: the rise in stock prices through the 1920s did not lead to a commensurate rise in investment, nor did the crash of 1987 slow investment down appreciably. Morck, Shleifer and Vishny (1993) reach similar conclusions using firm level data, as do Baker and Wurgler (2002a): in their work on capital structure, they show that not only do firms with higher market-to-book ratios in their past have more equity in their capital structure today, but also that the equity funds raised are typically used to increase cash balances and *not* to finance new investment.

More recently though, Polk and Sapienza (2001) report stronger evidence of investment distortion. They identify overvalued firms as firms with high accruals, defined as earnings minus actual cash flow, and as firms with high net issuance of equity. Firms with high accruals may become overvalued if investors fail to understand that earnings are overstating actual cash flows, and Chan et al. (2001) confirm that such firms indeed earn low returns. Overvalued firms may also be identified through their opportunistic issuance of equity, and we have already discussed the evidence that such firms earn low long-run returns. Controlling for actual investment opportunities as accurately as possible, Polk and Sapienza find that the firms they identify as overvalued appear to invest more than other firms, suggesting that sentiment does influence investment.

Further evidence of distortion comes from Baker, Stein and Wurgler's (2001) test of the cross-sectional prediction that equity-dependent firms will be more sensitive to stock price gyrations than will non-equity dependent firms. They identify equity-dependent firms on the basis of their low cash balances, among other measures, and find that these firms have an investment sensitivity to stock prices about three times as high as that of non-equity dependent firms. This study therefore provides initial evidence that for some firms at least, sentiment may distort investment, and that it does so through the equity-dependence channel.

8.2 Dividends

A major open question in corporate finance asks why firms pay dividends. Historically, dividends have been taxed at a higher rate than capital gains. This means that stockholders who pay taxes would always prefer that the firm repurchase shares rather than pay a dividend. Since the tax exempt shareholders would be indifferent between the dividend payment and the share repurchase, the share repurchase is a Pareto improving action. Why then, do investors seem perfectly happy to accept a substantial part of their return in the form of dividends? Or, using behavioral language, why do firms choose to frame part of their return as an explicit payment to stockholders, and in so doing, apparently make some of their shareholders worse off?

Shefrin and Statman (1984) propose a number of behavioral explanations for why investors exhibit a preference for dividends. Their first idea relies on the notion of self-control. Many people exhibit self-control problems. On the one hand, we want to deny ourselves an indulgence, but on the other hand, we quickly give in to temptation: today, we tell ourselves that tomorrow we will not overeat, and yet, when tomorrow arrives, we again eat too much. To deal with self-control problems, people often set rules, such as "bank the wife's salary, and only spend from the husband's paycheck". Another very natural rule people might create to prevent themselves from overconsuming their wealth is "only consume the dividend, but don't touch the portfolio capital". In other words, people may like dividends because dividends help them surmount self-control problems through the creation of simple rules.

A second rationale for dividends is based on mental accounting: by designating an explicit dividend payment, firms make it easier for investors to segregate gains from losses and hence to increase their utility. To see this, consider the following example. Over the course of a year, the value of a firm has increased by \$10 per share. The firm could choose *not* to pay a dividend and return this increase in value to investors as a \$10 capital gain. Alternatively, it could pay a \$2 dividend, leaving an \$8 capital gain. In the language of prospect theory, investors will code the first option as $v(10)$. They may also code the second option as $v(10)$, but the explicit segregation performed by the firm may encourage them to code it as $v(2) + v(8)$. This will, of course, result in a higher perceived utility, due to the concavity of v in the domain of gains.

This manipulation is equally useful in the case of losses. A firm whose value has declined by \$10 per share over the year can offer investors a \$10 capital loss or a \$12 capital loss combined with a \$2 dividend gain. While the first option will be coded as $v(-10)$, the second is more likely to be coded as $v(2) + v(-12)$, again resulting in a higher perceived utility, this time because of the convexity of v in the domain of losses.

The utility enhancing trick in these examples depends on investors segregating the overall gain or loss into different components. The key insight of Shefrin and Statman is that by paying dividends, firms make it easier for investors to perform this segregation.

Finally, Shefrin and Statman argue that by paying dividends, firms help investors avoid regret. Regret is a frustration that people feel when they imagine having taken an action that would have led to a more desirable outcome. It is stronger for errors of commission – cases where people suffer because of an action they took – than for errors of omission – where people suffer because of an action they *failed* to take.

Consider a company which does not pay a dividend. In order to finance consumption, an investor has to sell stock. If the stock subsequently goes up in value, the investor feels substantial regret because the error is one of commission: he can readily imagine how not selling the stock would have left him better off. If the firm had paid a dividend and the investor was able to finance his consumption out of it, a rise in the stock price would not have caused so much regret. This time, the error would have been one of omission: to be better off, the investor would have had to reinvest the dividend.

Shefrin and Statman try to explain why firms pay dividends at all. Another question asks how dividend paying firms decide on the size of their dividend. The classic paper on this subject is Lintner (1956). His treatment is based on extensive interviews with executives of large American companies in which he asked the respondent, often the CFO, how the firm set dividend policy. Based on these interviews Lintner proposed what we would now call a behavioral model. In his model, firms first establish a target dividend payout rate based on notions of fairness, in other words, on what portion of the earnings it is fair to return to shareholders. Then, as earnings increase and the dividend payout ratio falls below the target level, firms increase dividends only when they are confident that they will not have to reduce them in the future.

There are several behavioral aspects to this model. First, the firm is not setting the dividend to maximize firm value or shareholder after-tax wealth. Second, perceptions of fairness are used to set the target payout rate. Third, the asymmetry between an increase in dividends and a decrease is explicitly considered. Although fewer firms now decide to start paying dividends, for those that do Lintner's model appears to be valid to this day (Benartzi, Michaely and Thaler 1997, Fama and French 2001).

Baker and Wurgler (2002b) argue that changes in dividend policy may also reflect changing investor sentiment about dividend-paying firms relative to their sentiment about non-paying firms. They argue that for some investors, dividend-paying firms and non-paying firms represent salient categories and that these investors exhibit changing sentiment about the categories. For instance, when investors become more risk averse, they may prefer dividend-paying stocks because of a confused notion that these firms are less risky (the well-known "bird in the hand" fallacy). If managers are interested in maximizing short-run value, perhaps because it is linked to their compensation, they may be tempted to change their dividend policy in the direction favored by investors.

Baker and Wurgler find some supportive evidence for their theory. They measure relative investor sentiment about dividend-paying firms as the log market-to-book ratio of paying firms minus the log market-to-book ratio of non-paying firms, and find that in the time series, a high value of this measure one year predicts that in the following year, a higher fraction of non-paying firms initiate a dividend and a larger fraction of newly-listed firms choose to pay one. Similar results obtain for other measures of sentiment about dividend-paying firms.

8.3 Models of Managerial Irrationality

The theories we have discussed so far interpret the data as reflecting actions taken by rational managers in response to irrationality on the part of investors. Other papers have argued that some aspects of managerial behavior are the result of irrationality on the part of managers themselves.

Much of Section 2 was devoted to thinking about whether rational agents might be able to correct dislocations caused by irrational traders. Analogously, before we consider models of irrational managers, we should ask to what extent rational agents can undo their effects.

On reflection, it doesn't seem any easier to deal with irrational managers than irrational investors. It is true that many firms have mechanisms in place designed to solve agency problems and to keep the manager's mind focused on maximizing firm value: giving him stock options for example, or saddling him with debt. The problem is that these mechanisms are unlikely to have much of an effect on irrational managers. These managers *think* that they are maximizing firm value, even if in reality, they are not. Since they think that they are already doing the right thing, stock options or debt are unlikely to change their behavior.

In the best known paper on managerial irrationality, Roll (1986) argues that much of the evidence on takeover activity is consistent with an economy in which there are *no* overall gains to takeovers, but in which managers are overconfident, a theory he terms the "hubris hypothesis". When managers think about taking over another firm, they conduct a valuation analysis of that firm, taking synergies into account. If managers are overconfident about the accuracy of their analysis, they will be too quick to launch a bid when their valuation exceeds the market price of the target. Just as overconfidence among individual investors may lead to excessive trading, so overconfidence among managers may lead to excessive takeover activity.

The main predictions of the hubris hypothesis are that there will be a large amount of takeover activity, but that the total combined gain to bidder and target will be zero; that on the announcement of a bid, the price of the target will rise and the value of the bidder will fall by a similar amount. Roll examines the available evidence and concludes that it is impossible to reject any of these predictions.

Heaton (2002) analyses the consequences of managerial optimism whereby managers overestimate the probability that the future performance of their firm will be good. He shows that it can explain pecking order rules for capital structure: since managers are optimistic relative to the capital markets, they believe their equity is undervalued, and are therefore reluctant to issue it unless they have exhausted internally generated funds or the debt market. Managerial optimism can also explain the puzzlingly high correlation of investment and cash flow: when cash flow is low, managers' reluctance to use external markets for financing means that they forgo an unusually large number of projects, lowering investment at the same time.

Malmendier and Tate (2001) test Heaton's model by investigating whether firms with excessively optimistic CEOs display a greater sensitivity of investment to cash flow. They detect excessive optimism among CEOs by examining at what point they exercise their stock options: CEOs who hold on to their options longer than recommended by normative models of optimal exercise are deemed to be have an overly optimistic forecast of their stock's future price. Malmendier and Tate find that the investment of these CEOs' firms is indeed more sensitive to cash flow than the investment of other firms.[37]

9 Conclusion

Behavioral finance is a young field, with its formal beginnings in the 1980s. Much of the research we have discussed was completed in the past five years. Where do we stand? Substantial progress has been made on numerous fronts.

Empirical investigation of apparently anomalous facts. When De Bondt and Thaler's (1985) paper was published, many scholars thought that the best explanation for their findings was a programming error. Since then their results have been replicated numerous times by authors both sympathetic to their view and by those with alternative views. At this stage, we think that most of the empirical facts are agreed upon by most of the profession, although the interpretation of those facts is still in dispute. This is progress. If we all agree that the planets do orbit the sun, we can focus on understanding why.

Limits of Arbitrage. Twenty years ago, many financial economists thought that the Efficient Markets Hypothesis had to be true because of the forces of arbitrage. We now understand that this was a naive view, and that the limits of arbitrage can permit substantial mispricing. It is now also understood by most that the absence of a profitable investment strategy, because of risks and costs, does not imply the absence of mispricing. Prices can be

[37] Another paper which can be included in the managerial irrationality category is Loughran and Ritter's (2002) explanation for why managers issuing shares appear to leave significant amounts of money "on the table," as evidenced by the high average return of IPOs on their first day of trading. The authors note that the IPOs with good first day performance are often those IPOs in which the price has risen far above its filing range, giving the managers a sizeable wealth gain. One explanation is therefore that since managers are already enjoying a major windfall, they do not care too much about the fact that they could have been even wealthier.

very wrong without creating profit opportunities.

Understanding Bounded Rationality. Thanks largely to the work of cognitive psychologists such as Daniel Kahneman and Amos Tversky, we now have a long list of robust empirical findings that catalogue some of the ways in which actual humans form expectations and make choices. There has also been progress in writing down formal models of these processes, with prospect theory being the most notable. Economists once thought that behavior was either rational or impossible to formalize. We now know that models of bounded rationality are both possible and also much more accurate descriptions of behavior than purely rational models.

Behavioral Finance Theory Building. In the past few years there has been a burst of theoretical work modelling financial markets with less than fully rational agents. These papers relax the assumption of complete rationality either through the belief formation process or through the decision-making process. Like the work of psychologists discussed above, these papers are important existence proofs, showing that it is possible to think coherently about asset pricing while incorporating salient aspects of human behavior.

Investor Behavior. We have now begun the important job of trying to document and understand how investors, both amateurs and professionals, make their portfolio choices. Until recently such research was notably absent from the repertoire of financial economists, perhaps because of the mistaken belief that asset pricing can be modeled without knowing anything about the behavior of the agents in the economy.

This is a lot of accomplishment in a short period of time, but we are still much closer to the beginning of the research agenda than we are to the end. We know enough about the perils of forecasting to realize that most of the future progress of the field is unpredictable. Still, we cannot resist venturing a few observations on what may be coming next.

First, much of the work we have summarized is narrow. Models typically capture something about investors' beliefs, or their preferences, or the limits of arbitrage, but not all three. This comment applies to most research in economics, and is a natural implication of the fact that researchers are boundedly rational too. Still, as progress is made we expect theorists to begin to incorporate more than one strand into their models.

An example can, perhaps, illustrate the point. The empirical literature repeatedly finds that the asset pricing anomalies are more pronounced in small and mid-cap stocks than in the large cap sector. It seems likely that this finding reflects limits of arbitrage: the costs of trading smaller stocks are higher, and the low liquidity keeps many potential arbitrageurs uninterested. While this observation may be an obvious one, it has not found its way into formal models. We expect investigation of the interplay between limits of arbitrage and cognitive biases to be an important research area in the coming years.

Second, there are obviously competing behavioral explanations for some of the empirical facts. Some critics view this as a weakness of the field. It is sometimes said that the long list of cognitive biases summarized in Section 3 offer behavioral modelers so many degrees of freedom that anything can be explained. We concede that there are numerous degrees of freedom, but note that rational modelers have just as many options to choose from. As Arrow (1986) has forcefully argued, rationality per se does not yield many predictions. The predictions come from auxiliary assumptions.

There is really only one scientific way to compare alternative theories, behavioral or rational, and that is with empirical tests. One kind of test looks for novel predictions the theory makes. For example, Lee, Shleifer and Thaler (1991) test their model's prediction that small firm returns will be correlated with closed-end fund discounts, while Hong, Lim and Stein (2000) test the implication of the Hong and Stein (1999) model that momentum will be stronger among stocks with thinner analyst coverage.

Another sort of test is to look for evidence that agents actually behave the way a model claims they do. The Odean (1998) and Genesove and Mayer (2001) investigations of the disposition effect using actual market behavior fall into this category. Bloomfield et al. (2002) offers an experimental test of the behavior theorized by Barberis, Shleifer and Vishny (1998). Of course, such tests are never airtight, but we should be skeptical of theories based on behavior that is undocumented empirically. Since behavioral theories claim to be grounded in realistic assumptions about behavior, we hope behavioral finance researchers will continue to give their assumptions empirical scrutiny. We would urge the same upon authors of rational theories.[38]

We have two predictions about the outcome of the exercise of direct tests of the assumptions of economic models. First, we will find out that most of our current theories, both rational and behavioral, are wrong. Second, substantially better theories will emerge.

[38] Directly testing the validity of a model's assumptions is not common practice in economics, perhaps because of Milton Friedman's influential argument that one should evaluate theories based on the validity of their predictions rather than the validity of their assumptions. Whether or not this is sound scientific practice, we note that much of the debate over the past 20 years has occured precisely because the evidence has not been consistent with the theories, so it may be a good time to start worrying about the assumptions. If a theorist wants to claim that fact X can be explained by behavior Y, it seems prudent to check whether people actually do Y.

10 Appendix

We show that for the economy laid out in (3)-(6), there is an equilibrium in which the risk-free rate is constant and given by

$$R_f = \frac{1}{\rho} e^{\gamma g_C - \frac{1}{2}\gamma^2 \sigma_C^2} \tag{18}$$

and in which the price-dividend ratio is a constant f, and satisfies

$$1 = \rho \frac{1+f}{f} e^{g_D - \gamma g_C + \frac{1}{2}(\sigma_D^2 + \gamma^2 \sigma_C^2 - 2\gamma \sigma_C \sigma_D \omega)}. \tag{19}$$

In this equilibrium, returns are therefore given by

$$R_{t+1} = \frac{D_{t+1} + P_{t+1}}{P_t} = \frac{1 + P_{t+1}/D_{t+1}}{P_t/D_t} . \frac{D_{t+1}}{D_t} = \frac{1+f}{f} e^{g_D + \sigma_D \varepsilon_{t+1}}. \tag{20}$$

To see this, start from the Euler equations of optimality, obtained through the usual perturbation arguments,

$$\begin{aligned} 1 &= \rho R_f E_t \left[\left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} \right] && (21) \\ 1 &= \rho E_t \left[R_{t+1} \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} \right]. && (22) \end{aligned}$$

Computing the expectation in (21) gives (18). We conjecture that in this economy, there is an equilibrium in which the price-dividend ratio is a constant f, so that returns are given by (20). Substituting this into (22) and computing the expectation gives (19), as required. For given parameter values, the quantitative implications for P/D ratios and returns are now easily computed.

11 References

Abreu, D., and M. Brunnermeier (2002), "Synchronization Risk and Delayed Arbitrage," forthcoming, *Journal of Financial Economics.*

Alpert M., and H. Raiffa (1982), "A Progress Report on the Training of Probability Assessors," in D. Kahneman, P. Slovic and A. Tversky, eds., *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.

Anderson E., Hansen L., and T. Sargent (1998), "Risk and Robustness in Equilibrium," Working paper, University of Chicago.

Arrow, K. (1986), "Rationality of Self and Others," in R. Hogarth and M. Reder, eds., *Rational Choice*, Chicago: University of Chicago Press.

Baker, M., Stein, J., and J. Wurgler (2001), "When Does the Market Matter? Stock Prices and the Investment of Equity Dependent Firms," Working paper, Harvard University.

Baker, M., and J. Wurgler (2000), "The Equity Share in New Issues and Aggregate Stock Returns," *Journal of Finance* 55, 2219-2257.

Baker, M., and J. Wurgler (2002a), "Market Timing and Capital Structure," *Journal of Finance* 57, 1-32.

Baker, M., and J. Wurgler (2002b), "A Catering Theory of Dividends," Working paper, Harvard University.

Ball, R. (1978), "Anomalies in Relations between Securities' Yields and Yield Surrogates," *Journal of Financial Economics* 6, 103-126.

Banz, R. (1981), "The Relation between Return and Market Value of Common Stocks," *Journal of Financial Economics* 9, 3-18.

Barber, B., and J. Lyon (1997), "Detecting Long-run Abnormal Stock Returns: the Empirical Power and Specification of Test Statistics," *Journal of Financial Economics* 43, 341-372.

Barber, B., and T. Odean (2000), "Trading is Hazardous to Your Wealth: The Common Stock Performance of Individual Investors," *Journal of Finance* 55, 773-806.

Barber, B., and T. Odean (2001), "Boys will be Boys: Gender, Overconfidence, and Common Stock Investment," *Quarterly Journal of Economics* 141, 261-292.

Barber, B., and T. Odean (2002a), "Online Investors: Do the Slow Die First?" *Review of Financial Studies* 15, 455-487.

Barber, B., and T. Odean (2002b), "All that Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors," Working paper, UC Berkeley.

Barberis, N., and M. Huang (2001), "Mental Accounting, Loss Aversion and Individual Stock Returns," *Journal of Finance* 56, 1247-1292.

Barberis, N., Huang M., and T. Santos (2001), "Prospect Theory and Asset Prices," *Quarterly Journal of Economics* 116, 1-53.

Barberis, N., and A. Shleifer (2003), "Style Investing," forthcoming, *Journal of Financial Economics.*

Barberis, N., Shleifer A., and R. Vishny (1998), "A Model of Investor Sentiment," *Journal of Financial Economics* 49, 307-345.

Barberis, N., Shleifer A., and J. Wurgler (2001), "Comovement," Working paper, University of Chicago.

Barsky, R., and B. De Long (1993), "Why Does the Stock Market Fluctuate?" *Quarterly Journal of Economics* 107, 291-311.

Basu, S. (1983), "The Relationship Between Earnings Yield, Market Value and Return for NYSE Common Stocks: Further Evidence," *Journal of Financial Economics* 12, 129-156.

Baxter, M., and U. Jermann (1997), "The International Diversification Puzzle is Worse than You Think," *American Economic Review* 87, 170-180.

Bell, D. (1982), "Regret in Decision Making Under Uncertainty," *Operations Research* 30, 961-981.

Benartzi, S. (2001), "Excessive Extrapolation and the Allocation of 401(k) Accounts to Company Stock," *Journal of Finance* 56, 1747-1764.

Benartzi, S., Michaely R., and R. Thaler (1997), "Do Changes in Dividends Signal the Future or the Past?" *Journal of Finance* 52, 1007-34.

Benartzi, S., and R. Thaler (1995), "Myopic Loss Aversion and the Equity Premium Puzzle," *Quarterly Journal of Economics* 110, 75-92.

Benartzi, S., and R. Thaler (2001), "Naïve Diversification Strategies in Defined Contribution Savings Plans," *American Economic Review* 91, 79-98.

Berk, J. (1995), "A Critique of Size Related Anomalies," *Review of Financial Studies* 8, 275-286.

Bernard, V. and J. Thomas (1989), "Post-Earnings Announcement Drift: Delayed Price Response or Risk Premium?" *Journal of Accounting Research* (Supplement), 1-36.

Blanchard, O., Rhee C., and L. Summers (1993), "The Stock Market, Profit, and Investment," *Quarterly Journal of Economics* 108, 115-136.

Bloomfield, R., and J. Hales (2002), "Predicting the Next Step of a Random Walk: Experimental Evidence of Regime-Shifting Beliefs," *Journal of Financial Economics* 65, 397-414.

Bodurtha J., Kim, D., and C.M. Lee (1993), "Closed-end Country Funds and U.S. Market Sentiment," *Review of Financial Studies* 8, 879-918.

Brav, A. (2000), "Inference in Long-horizon Event Studies," *Journal of Finance* 55, 1979-2016.

Brav A. and P. Gompers (1997), "Myth or Reality? The Long-run Underperformance of Initial Public Offerings: Evidence from Venture and Non-venture-backed Companies," *Journal of Finance* 52, 1791-1821.

Brav, A., Geczy C. and P. Gompers (2000), "Is the Abnormal Return Following Equity Issuances Anomalous?" *Journal of Financial Economics* 56, 209-249.

Brennan, M., and Y. Xia (2001), "Stock Return Volatility and the Equity Premium," *Journal of Monetary Economics* 47, 249-283.

Brown, S., Goetzmann, W. and S. Ross (1995), "Survival," *Journal of Finance* 50, 853-873.

Brunnermeier, M. (2001), *Asset Pricing under Asymmetric Information – Bubbles, Crashes, Technical Analysis, and Herding*, Oxford: Oxford University Press.

Buehler R., Griffin D., and M. Ross (1994), "Exploring the Planning Fallacy: Why People Underestimate their Task Completion Times," *Journality of Personality and Social Psychology* 67, 366-381.

Camerer, C. (1995), "Individual Decision Making," in J. Kagel and A. Roth, eds., *Handbook of Experimental Economics*, Princeton: Princeton University Press.

Camerer, C., and R. Hogarth (1999), "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor Production Framework," *Journal of Risk and Uncertainty* 19, 7-42.

Camerer, C., and M. Weber (1992), "Recent Developments in Modeling Preferences: Uncertainty and Ambiguity," *Journal of Risk and Uncertainty* 5, 325-70.

Campbell J.Y. (1991), "A Variance Decomposition for Stock Returns," *Economic Journal* 101, 157-179.

Campbell, J.Y. (1999), "Asset Prices, Consumption and the Business Cycle," in J. Taylor and M. Woodford, eds., *Handbook of Macroeconomics*, Amsterdam: North-Holland.

Campbell, J.Y. (2000), "Asset Pricing at the Millenium," *Journal of Finance* 55, 1515-1567.

Campbell, J.Y., and J. Cochrane (1999), "By Force of Habit: A Consumption-based Explanation of Aggregate Stock Market Behavior," *Journal of Political Economy* 107, 205-51.

Campbell, J.Y. and R. Shiller (1988), "Stock Prices, Earnings and Expected Dividends," *Journal of Finance* 43, 661-676.

Cecchetti, S., Lam P., and N. Mark (2000), "Asset Pricing with Distorted Beliefs: Are Equity Returns Too Good to Be True?" *American Economic Review* 90, 787-805.

Chan, K., Chan L., Jegadeesh N., and J. Lakonishok (2001), "Earnings Quality and Stock Returns," Working paper, University of Illinois.

Chan, L., Jegadeesh N., and J. Lakonishok (1996), "Momentum Strategies," *Journal of Finance* 51, 1681-1713.

Chen, J., Hong H. and J. Stein (2001), "Forecasting Crashes: Trading Volume, Past Returns and Conditional Skewness in Stock Prices," *Journal of Financial Economics* 61, 345-381.

Chen, J., Hong H., and J. Stein (2002), "Breadth of Ownership and Stock Returns", forthcoming, *Journal of Financial Economics.*

Chew, S. (1983), "A Generalization of the Quasilinear Mean with Applications to the Measurement of Income Inequality and Decision Theory Resolving the Allais Paradox," *Econometrica* 51, 1065-1092.

Chew, S. (1989), "Axiomatic Utility Theories with the Betweenness Property," *Annals of Operations Research* 19, 273-98.

Chew, S., and K. MacCrimmon (1979), "Alpha-nu Choice Theory: An Axiomatization of Expected Utility," Working paper, University of British Columbia.

Chopra, N., Lakonishok J., and J. Ritter (1992), "Measuring Abnormal Performance: Do Stocks Overreact?" *Journal of Financial Economics* 31, 235-268.

Coval, J., and T. Moskowitz (1999), "Home Bias at Home: Local Equity Preference in Domestic Portfolios," *Journal of Finance* 54, 2045-73.

Coval, J., and T. Moskowitz (2001), "The Geography of Investment: Informed Trading and Asset Prices," *Journal of Political Economy* 109, 811-841.

Coval, J. and T. Shumway (2000), "Do Behavioral Biases Affect Prices?" Working paper, University of Michigan.

Daniel, K., Hirshleifer D., and A. Subrahmanyam (1998), "Investor Psychology and Security Market Under- and Overreactions," *Journal of Finance* 53, 1839-1885.

Daniel, K., Hirshleifer D., and A. Subrahmanyam (2001), "Overconfidence, Arbitrage and Equilbrium Asset Pricing," *Journal of Finance* 56, 921-965.

Daniel, K. and S. Titman (1997), "Evidence on the Characteristics of Cross-Sectional Variation in Stock Returns," *Journal of Finance* 52, 1-33.

D'Avolio, G. (2002), "The Market for Borrowing Stock," forthcoming, *Journal of Financial Economics.*

De Bondt, W., and R. Thaler (1985), "Does the Stock Market Overreact?" *Journal of Finance* 40, 793-808.

De Bondt, W., and R. Thaler (1987), "Further Evidence on Investor Overreaction and Stock Market Seasonality," *Journal of Finance* 42, 557-581.

De Long, J.B., Shleifer A., Summers L., and R. Waldmann (1990a), "Noise Trader Risk in Financial Markets," *Journal of Political Economy* 98, 703-738.

De Long, J.B., Shleifer A., Summers L., and R. Waldmann (1990b), "Positive Feedback Investment Strategies and Destabilizing Rational Speculation," *Journal of Finance* 45, 375-395.

Dekel, E. (1986), "An Axiomatic Characterization of Preferences Under Uncertainty: Weakening the Independence Axiom," *Journal of Economic Theory* 40, 304-18.

Diether, K., Malloy C., and A. Scherbina (2003), "Stock Prices and Differences of Opinion: Empirical Evidence that Stock Prices Reflect Optimism," forthcoming, *Journal of Finance.*

Dreman, D. (1977), *Psychology and the Stock Market: Investment Strategy Beyond Random Walk*, New York: Warner Books.

Driscoll, K., Malcolm J., Sirul M., and P. Slotter (1995), "1995 Gallup Survey of Defined Contribution Plan Participants," John Hancock Financial Services.

Edwards, W. (1986), "Conservatism in Human Information Processing," in B. Kleinmutz, ed., *Formal Representation of Human Judgment*, New York: John Wiley and Sons.

Ellsberg, D. (1961), "Risk, Ambiguity, and the Savage Axioms," *Quarterly Journal of Economics* 75, 643-69.

Epstein, L. and T. Wang (1994), "Intertemporal Asset Pricing under Knightian Uncertainty," *Econometrica* 62, 283-322.

Fama, E. (1970), "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance* 25, 383-417.

Fama, E. (1998), “Market Efficiency, Long-Term Returns and Behavioral Finance,” *Journal of Financial Economics* 49, 283-307.

Fama, E. and K. French (1988), “Dividend Yields and Expected Stock Returns,” *Journal of Financial Economics* 22, 3-25.

Fama, E. and K. French (1992), “The Cross-Section of Expected Stock Returns,” *Journal of Finance* 47, 427-465.

Fama, E. and K. French (1993), “Common Risk Factors in the Returns of Bonds and Stocks,” *Journal of Financial Economics* 33, 3-56.

Fama, E., and K. French (1995), “Size and Book-to-Market Factors in Earnings and Returns,” *Journal of Finance* 50, 131-155.

Fama, E. and K. French (1996), “Multifactor Explanations of Asset Pricing Anomalies,” *Journal of Finance* 51, 55-84.

Fama, E. and K. French (1998), “Value vs. Growth: The International Evidence,” *Journal of Finance* 53, 1975-1999.

Fama, E., and K. French (2001), “Disappearing Dividends: Changing Firm Characteristics or Lower Propensity to Pay?” *Journal of Financial Economics* 60, 3-43.

Fama, E., French K., and J. Davis (2000), “Characteristics, Covariances and Average Returns 1929-1997,” *Journal of Finance* 55, 389-406.

Fischhoff B., Slovic, P., and S. Lichtenstein (1977), “Knowing With Certainty: The Appropriateness of Extreme Confidence,” *Journal of Experimental Pyschology: Human Perception and Performance* 3, 552-564.

Fisher, I. (1928), *Money Illusion*, New York: Adelphi.

Fox, C., and A. Tversky (1995), “Ambiguity Aversion and Comparative Ignorance,” *Quarterly Journal of Economics* 110, 585-603.

French, K., and J. Poterba (1991), “Investor Diversification and International Equity Markets,” *American Economic Review* 81, 222-226.

Friedman, M. (1953), “The Case for Flexible Exchange Rates,” in *Essays in Positive Economics*, Chicago: University of Chicago Press.

Froot, K. and E. Dabora (1999), “How Are Stock Prices Affected by the Location of Trade?” *Journal of Financial Economics* 53, 189-216.

Genesove, D., and C. Mayer (2001), “Loss Aversion and Seller Behavior: Evidence from the Housing Market,” *Quarterly Journal of Economics* 116, 1233-1260.

Gervais, S., and T. Odean (2001), "Learning to be Overconfident," *Review of Financial Studies* 14, 1-27.

Gilboa I., and D. Schmeidler (1989), "Maxmin Expected Utility with a Non-unique Prior," *Journal of Mathematical Economics* 18, 141-53.

Gilovich, T., Griffin D., and D. Kahneman eds. (2002), *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge: Cambridge University Press.

Gilovich, T., Vallone, R., and A. Tversky (1985), "The Hot Hand in Basketball: On the Misperception of Random Sequences," *Cognitive Psychology* 17, 295-314.

Gneezy, U., and J. Potters (1997), "An Experiment on Risk Taking and Evaluation Periods," *Quarterly Journal of Economics* 112, 631-645.

Gompers, P. and A. Metrick (2001), "Institutional Investors and Equity Prices," *Quarterly Journal of Economics* 116, 229-259.

Graham, B. (1949), *The Intelligent Investor: A Book of Practical Counsel*, Harper and Row: New York.

Graham, J., and C. Harvey (2001), "The Theory and Practice of Corporate Finance: Evidence from the Field," *Journal of Financial Economics* 60, 187-243.

Grinblatt, M., and B. Han (2001), "The Disposition Effect and Momentum," Working paper, UCLA.

Grinblatt, M., and M. Keloharju (2001), "How Distance, Language, and Culture Influence Stockholdings and Trades," *Journal of Finance* 56, 1053-73.

Grinblatt, M., and T. Moskowitz (1999), "The Cross-section of Expected Returns and its Relation to Past Returns," Working paper, University of Chicago.

Gul, F. (1991), "A Theory of Disappointment in Decision Making under Uncertainty," *Econometrica* 59, 667-86.

Hansen, L., and K. Singleton (1983), "Stochastic Consumption, Risk Aversion and the Temporal Behavior of Asset Returns," *Journal of Political Economy* 91, 249-268.

Hardouvelis, G., La Porta R., and T. Wizman (1994), "What Moves the Discount on Country Equity Funds?" in J. Frankel, ed., *The Internationalization of Equity Markets*, Chicago: The University of Chicago Press.

Harris, L., and E. Gurel (1986), "Price and Volume Effects Associated with Changes in the S&P 500: New Evidence for the Existence of Price Pressure," *Journal of Finance* 41, 851-60.

Harrison, J.M., and D. Kreps (1978), "Speculative Investor Behavior in a Stock Market with Heterogeneous Expectations," *Quarterly Journal of Economics* 92, 323-336.

Heath, C., and A. Tversky (1991), "Preference and Belief: Ambiguity and Competence in Choice under Uncertainty," *Journal of Risk and Uncertainty* 4, 5-28.

Heaton, J.B. (2002), "Managerial Optimism and Corporate Finance," forthcoming, *Financial Managment.*

Hirshleifer, D. (2001), "Investor Psychology and Asset Pricing," *Journal of Finance* 56, 1533-1597.

Hong, H., Lim T., and J. Stein (2000), "Bad News Travels Slowly: Size, Analyst Coverage, and the Profitability of Momentum Strategies," *Journal of Finance* 55, 265-295.

Hong, H., and J. Stein (1999), "A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets," *Journal of Finance* 54, 2143-2184.

Hong, H., and J. Stein (2003), "Differences of Opinion, Short-sale Constraints and Market Crashes," forthcoming, *Review of Financial Studies.*

Huberman, G. (2001), "Familiarity Breeds Investment," *Review of Financial Studies* 14, 659-680.

Hvidkjaer, S. (2001), "A Trade-based Analysis of Momentum," Working paper, University of Maryland.

Ikenberry, D., Lakonishok J., and T. Vermaelen (1995), "Market Underreaction to Open Market Share Repurchases," *Journal of Financial Economics* 39, 181-208.

Jegadeesh, N. and S. Titman (1993), "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance* 48, 65-91.

Jones, C., and O. Lamont (2002), "Short-sale Constraints and Stock Returns," forthcoming, *Journal of Financial Economics.*

Jung, K., Kim Y., and R. Stulz (1996), "Timing, Investment Opportunities, Managerial Discretion, and the Security Issue Decision," *Journal of Financial Economics* 42, 159-185.

Kahneman, D., Slovic P., and A. Tversky eds. (1982), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.

Kahneman, D., and A. Tversky (1974), "Judgment Under Uncertainty: Heuristics and Biases," *Science* 185, 1124-31.

Kahneman, D., and A. Tversky (1979), "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica* 47, 263-91.

Kahneman, D., and A. Tversky eds. (2000), *Choices, Values and Frames*, Cambridge: Cambridge University Press.

Kaul, A., Mehrotra V., and R. Morck (2000), "Demand Curves for Stocks Do Slope Down: New Evidence from an Index Weights Adjustment," *Journal of Finance* 55, 893-912.

Knight, F. (1921), *Risk, Uncertainty and Profit*, Boston, New York: Houghton Mifflin.

Korajczyk, R., Lucas D., and R. MacDonald (1991), "The Effects of Information Releases on the Pricing and Timing of Equity Issues," *Review of Financial Studies* 4, 685-708.

La Porta, R., Lakonishok J., Shleifer A., and R. Vishny (1997), "Good News for Value Stocks: Further Evidence on Market Efficiency," *Journal of Finance* 49, 1541-1578.

Lakonishok, J., Shleifer A., and R. Vishny (1994), "Contrarian Investment, Extrapolation and Risk," *Journal of Finance* 49, 1541-1578.

Lamont, O., and R. Thaler (2002), "Can the Market Add and Subtract? Mispricing in Tech Stock Carve-Outs," forthcoming, *Journal of Political Economy.*

Lee, C., Shleifer, A., and R. Thaler (1991), "Investor Sentiment and the Closed-end Fund Puzzle," *Journal of Finance* 46, 75-110.

LeRoy S. and R. Porter (1981), "The Present-value Relation: Tests Based on Implied Variance Bounds," *Econometrica* 49, 97-113.

Lewis, K. (1999), "Trying to Explain Home Bias in Equities and Consumption," *Journal of Economic Literature* 37, 571-608.

Lintner, J. (1956), "Distribution of Incomes of Corporations among Dividends, Retained Earnings and Taxes," *American Economic Review* 46, 97-113.

Loomes G., and R. Sugden (1982), "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty," *The Economic Journal* 92, 805-824.

Lord, C., Ross L., and M. Lepper (1979), "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence," *Journal of Personality and Social Psychology* 37, 2098-2109.

Loughran, T., and J. Ritter (1995), "The New Issues Puzzle," *Journal of Finance* 50, 23-50.

Loughran, T., and J. Ritter (2000), "Uniformly Least Powerful Tests of Market Efficiency," *Journal of Financial Economics* 55, 361-389.

Loughran, T., and J. Ritter (2002), "Why Don't Issuers Get Upset About Leaving Money on the Table?" *Review of Financial Studies* 15, 413-443.

Loughran, T., Ritter J., and K. Rydqvist (1994), "Initial Public Offerings: International Insights," *Pacific Basin Finance Journal* 2, 165-199.

Lyon, J., Barber, B. and C. Tsai (1999), "Improved Methods for Tests of Long-run Abnormal Stock Returns," *Journal of Finance* 54, 165-201.

Maenhout, P. (1999), "Robust Portfolio Rules and Asset Pricing," Working paper, Insead.

Malmendier, U., and G. Tate (2001), "CEO Overconfidence and Corporate Investment," Working paper, Harvard University.

Mankiw, N.G., and S. Zeldes (1991), "The Consumption of Stockholders and Non-stockholders," *Journal of Financial Economics* 29, 97-112.

Markowitz, H. (1952), "The Utility of Wealth," *Journal of Political Economy* 60, 151-158.

Mehra, R. and E. Prescott (1985), "The Equity Premium: A Puzzle," *Journal of Monetary Economics* 15, 145-161.

Merton, R. (1987), "A Simple Model of Capital Market Equilibrium with Incomplete Information," *Journal of Finance* 42 483-510.

Michaely, R., Thaler R., and K. Womack (1995), "Price Reactions to Dividend Initiations and Omissions," *Journal of Finance* 50, 573-608.

Miller, E. (1977), "Risk, Uncertainty and Divergence of Opinion," *Journal of Finance* 32, 1151-1168.

Mitchell, M., Pulvino T., and E. Stafford (2000), "Limited Arbitrage in Equity Markets," *Journal of Finance* 57, 551-584.

Mitchell, M., and E. Stafford (2001), "Managerial Decisions and Long-term Stock Price Performance," *Journal of Business* 73, 287-329.

Modigliani, F. and R. Cohn (1979), "Inflation and the Stock Market," *Financial Analysts Journal* 35, 24-44.

Morck, R., Shleifer A., and R. Vishny (1993), "The Stock Market and Investment: Is the Market a Sideshow?" Brookings Papers on Economic Activity.

Mullainathan (2000), "Thinking Through Categories," Working paper, MIT.

Odean, T. (1998), "Are Investors Reluctant to Realize their Losses?" *Journal of Finance* 53, 1775-1798.

Odean, T. (1999), "Do Investors Trade Too Much?" *American Economic Review* 89, 1279-1298.

Ofek, E., and M. Richardson (2001), "Dot-com Mania: Market Inefficiency in the Internet Sector," Working paper, New York University.

Pagano, M., Panetta F., and L. Zingales (1998), "Why do Companies Go Public? An Empirical Analysis," *Journal of Finance* 53, 27-64.

Polk, C., and P. Sapienza (2001), "The Real Effects of Investor Sentiment," Working paper, Northwestern University.

Poteshman, A. (2001), "Underreaction, Overreaction and Increasing Misreaction to Information in the Options Market," *Journal of Finance* 56, 851-876.

Quiggin, J. (1982), "A Theory of Anticipated Utility," *Journal of Economic Behavior and Organization* 3, 323-43.

Rabin, M., (1998), "Psychology and Economics," *Journal of Economic Literature* 36, 11-46.

Rabin, M. (2000), "Risk Aversion and Expected Utility Theory: A Calibration Theorem," *Econometrica* 68, 1281-1292.

Rabin, M., (2002), "Inference By Believers in the Law of Small Numbers," *Quarterly Journal of Economics* 117, 775-816.

Redelmeier D. and A. Tversky (1992), "On the Framing of Multiple Prospects," *Psychological Science* 3, 191-193.

Ritter, J., and R. Warr (2002), "The Decline of Inflation and the Bull Market of 1982 to 1997," *Journal of Financial and Quantitative Analysis* 37, 29-61.

Roll, R. (1977), "A Critique of the Asset Pricing Theory's Tests: Part I," *Journal of Financial Economics* 4, 129-174.

Roll, R. (1983), "Vas ist Das?" *Journal of Portfolio Management* 9, 18-28.

Roll, R. (1986), "The Hubris Hypothesis of Corporate Takeovers," *Journal of Business* 59, 197-216.

Rosenberg, B., Reid, K., and R. Lanstein (1985), "Persuasive Evidence of Market Inefficiency," *Journal of Portfolio Management* 11, 9-17.

Ross, S. (2001), *Lectures Notes on Market Efficiency*, Sloan School of Management.

Rouwenhorst, G. (1998), "International Momentum Strategies," *Journal of Finance* 53, 267-284.

Rubinstein, M. (2001), "Rational Markets: Yes or No? The Affirmative Case," *Financial Analysts Journal* (May-June), 15-29.

Santos, M. and M. Woodford (1997), "Rational Asset Pricing Bubbles," *Econometrica* 65, 19-58.

Sargent T. (1993), *Bounded Rationality in Macroeconomics*, Oxford: Oxford University Press.

Savage, L. (1964), *The Foundations of Statistics*, New York: Wiley.

Scheinkman J. and W. Xiong (2001), "Overconfidence and Speculative Bubbles," Working paper, Princeton University.

Segal, U. (1987), "Some Remarks on Quiggin's Anticipated Utility," *Journal of Economic Behavior and Organization* 8, 145-154.

Segal, U. (1989), "Anticipated Utility: A Measure Representation Approach," *Annals of Operations Research* 19, 359-73.

Shafir, E., Diamond, P. and A. Tversky (1997), "Money Illusion," *Quarterly Journal of Economics* 112, 341-374.

Shefrin, H. and M. Statman (1984), "Explaining Investor Preference for Cash Dividends," *Journal of Financial Economics* 13, 253-282.

Shefin, H., and M. Statman (1985), "The Disposition to Sell Winners too Early and Ride Losers too Long," *Journal of Finance* 40, 777-790.

Shiller, R. (1981), "Do Stock Prices Move too Much to be Justified by Subsequent Changes in Dividends?" *American Economic Review* 71, 421-436.

Shiller, R. (1984), "Stock Prices and Social Dynamics," *Brookings Papers on Economic Activity* 2, 457-498.

Shleifer, A. (1986), "Do Demand Curves for Stocks Slope Down?" *Journal of Finance* 41, 579-90.

Shleifer, A. (2000), *Inefficient Markets: An Introduction to Behavioral Finance*, Oxford: Oxford University Press.

Shleifer, A., and L. Summers (1990), "The Noise Trader Approach to Finance," *Journal of Economic Perspectives* 4, 19-33.

Shleifer, A., and R. Vishny (1997), "The Limits of Arbitrage," *Journal of Finance* 52, 35-55.

Shleifer, A., and R. Vishny (2001), "Stock Market Driven Acquisitions," Working paper, Harvard University.

Stein, J. (1996), "Rational Capital Budgeting in an Irrational World," *Journal of Business*

69, 429-55.

Summers, L. (1986), “Does the Stock Market Rationally Reflect Fundamental Values?” *Journal of Finance* 41, 591-601.

Thaler, R. (1999), “Mental Accounting Matters,” in D. Kahneman and A. Tversky, eds., *Choice, Values and Frames*, Cambridge: Russell Sage Foundation.

Thaler, R., and E. Johnson (1990), “Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice,” *Management Science* 36, 643-660.

Thaler, R., Tversky A., Kahneman D., and A. Schwartz (1997), “The Effect of Myopia and Loss Aversion on Risk-Taking: An Experimental Test,” *Quarterly Journal of Economics* 112, 647-661.

Tversky, A., and D. Kahneman (1986), “Rational Choice and the Framing of Decisions,” *Journal of Business* 59, 251-78.

Tversky, A., and D. Kahneman (1992), “Advances in Prospect Theory: Cumulative Representation of Uncertainty,” *Journal of Risk and Uncertainty* 5, 297-323.

Veronesi, P. (1999), “Stock Market Overreaction to Bad News in Good Times: A Rational Expectations Equilibrium Model,” *Review of Financial Studies* 12, 975-1007.

Vijh, A. (1994), “S&P 500 Trading Strategies and Stock Betas,” *Review of Financial Studies* 7, 215-251.

Von Neumann, J., and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton: Princeton University Press.

Vuolteenaho T. (2002), “What Drives Firm-level Stock Returns?” *Journal of Finance* 57, 233-264.

Weil, P. (1989), “The Equity Premium Puzzle and the Risk-free Rate Puzzle,” *Journal of Monetary Economics* 24, 401-421.

Weinstein, N. (1980), “Unrealistic Optimism about Future Life Events” *Journal of Personality and Social Psychology* 39, 806-820.

Wurgler, J., and K. Zhuravskaya (2002), “Does Arbitrage Flatten Demand Curves for Stocks?”, forthcoming, *Journal of Business.*

Yaari, M. (1987), “The Dual Theory of Choice Under Risk,” *Econometrica* 55, 95-115.

Table 1: Arbitrage costs and risks that arise in exploiting mispricing: fundamental risk (FR), noise trader risk (NTR) and implementation costs (IC).

	FR	NTR	IC
Royal Dutch/Shell	×	$\surd$	×
Index Inclusions	$\surd$	$\surd$	×
Palm/3-Com	×	×	$\surd$

Table 2: Parameter values for a simple consumption-based model.

Parameter	
g_C	1.84%
σ_C	3.79%
g_D	1.5%
σ_D	12.0%
ω	0.15
γ	1.0
ρ	0.98

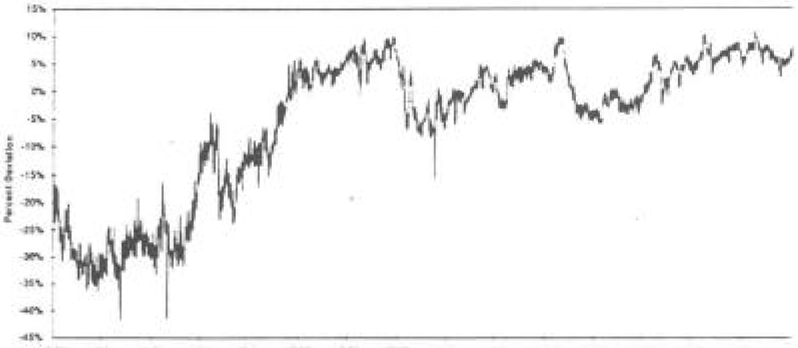


Figure 1. Log deviations from Royal Dutch/Shell parity. Source: Froot and Dabora (1999).

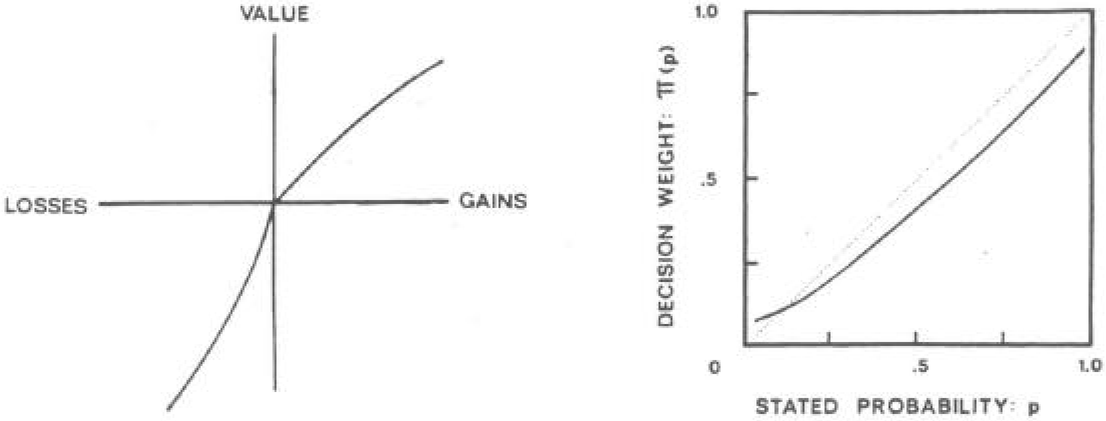


Figure 2. The two panels show Kahneman and Tversky's (1979) proposed value function v and probability weighting function π.

Human Behavior and the Efficiency of the Financial System

by

Robert J. Shiller[*]

Abstract

Recent literature in empirical finance is surveyed in its relation to underlying behavioral principles, principles which come primarily from psychology, sociology and anthropology. The behavioral principles discussed are: prospect theory, regret and cognitive dissonance, anchoring, mental compartments, overconfidence, over- and underreaction, representativeness heuristic, the disjunction effect, gambling behavior and speculation, perceived irrelevance of history, magical thinking, quasi-magical thinking, attention anomalies, the availability heuristic, culture and social contagion, and global culture.

Theories of human behavior from psychology, sociology, and anthropology have helped motivate much recent empirical research on the behavior of financial markets. In this paper I will survey both some of the most significant theories (for empirical finance) in these other social sciences and the empirical finance literature itself.

Particular attention will be paid to the implications of these theories for the efficient markets hypothesis in finance. This is the hypothesis that financial prices efficiently incorporate all public information and that prices can be regarded as optimal estimates of true investment value at all times. The efficient markets hypothesis in turn is based on more primitive notions that people behave rationally, or accurately maximize expected utility, and are able to process all available information. The idea behind the term "efficient markets hypothesis," a term coined by Harry Roberts (1967),[1] has a long history in financial research, a far longer history than the term itself has. The hypothesis (without the words

[*] This paper was prepared for John B. Taylor and Michael Woodford, Editors, *Handbook of Macroeconomics*. An earlier version was presented at a conference "Recent Developments in Macroeconomics" at the Federal Reserve Bank of New York, February 27–8, 1997. The author is indebted to Ricky Lam for research assistance, and to Michael Krause, Virginia Shiller, Andrei Shleifer, David Wilcox, and the editors for helpful comments. This research was supported by the National Science Foundation.

[1] The Roberts (1967) paper has never been published; the fame of his paper apparently owes to the discussion of it in Fama (1970).

efficient markets) was given a clear statement in Gibson (1889), and has apparently been widely known at least since then, if not long before. All this time there has also been tension over the hypothesis, a feeling among many that there is something egregiously wrong with it; for an early example, see MacKay (1841). In the past couple of decades the finance literature, has amassed a substantial number of observations of apparent anomalies (from the standpoint of the efficient markets hypothesis) in financial markets. These anomalies suggest that the underlying principles of rational behavior underlying the efficient markets hypothesis are not entirely correct and that we need to look as well at other models of human behavior, as have been studied in the other social sciences.

The organization of this paper is different from that of other accounts of the literature on behavioral finance (for example, De Bondt and Thaler, 1996 or Fama, 1997): this paper is organized around a list of theories from the other social sciences that are used by researchers in finance, rather than around a list of anomalies. I organized the paper this way because, in reality, most of the fundamental principles that we want to stress here really do seem to be imported from the other social sciences. No surprise here: researchers in these other social sciences have done most of the work over the last century on understanding less-than-perfectly-rational human behavior. Moreover, each anomaly in finance typically has more than one possible explanation in terms of these theories from the other social sciences. The anomalies are observed in complex real world settings, where many possible factors are at work, not in the experimental psychologist's laboratory. Each of their theories contributes a little to our understanding of the anomalies, and there is typically no way to quantify or prove the relevance of any one theory. It is better to set forth the theories from the other social sciences themselves, describing when possible the controlled experiments that demonstrate their validity, and give for each a few illustrations of applications in finance.

Before beginning, it should be noted that theories of human behavior from these other social sciences often have underlying motivation that is different from that of economic theories. Their theories are often intended to be robust to application in a variety of everyday, unstructured experiences, while the economic theories are often intended to be robust in the different sense that, even if the problems the economic agents face become very clearly defined, their behavior will not change after they learn how to solve the problems. Many of the underlying behavioral principles from psychology and other social sciences that are discussed below are unstable and the hypothesized behavioral phenomena may disappear when the situation becomes better structured and people have had a lot of opportunity to learn about it. Indeed, there are papers in the psychology literature claiming that many of the cognitive biases in human judgment under uncertainty uncovered by experimental psychologists will disappear when the experiment is changed so that the probabilities and issues that the experiment raises are explained clearly enough to subjects (see, for example, Gigerenzer, 1991). Experimental subjects can in many cases be convinced, if given proper instruction, that their initial behavior in the experimental situation was irrational, and they will then correct their ways.

To economists, such evidence is taken to be more damning to the theories than it would be by the social scientists in these other disciplines. Apparently economists at large have not fully appreciated the extent to which enduring patterns can be found in this 'unstable' human behavior. The examples below of application of theories from other social sciences to understanding anomalies in financial markets will illustrate.

Each section below, until the conclusion, refers to a theory taken from the literature in psychology, sociology or anthropology. The only order of these sections is that I have placed first theories that seem to have the more concrete applications in finance, leaving some more impressionistic applications to the end. In the conclusion I attempt to put these theories into perspective, and to recall that there are also important strengths in conventional economic theory and in the efficient markets hypothesis itself.

Prospect Theory

Prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992) has probably had more impact than any other behavioral theory on economic research. Prospect theory is very influential despite the fact that it is still viewed by much of the economics profession at large as of far less importance than expected utility theory. Among economists, prospect theory has a distinct, though still prominent, second place to expected utility theory for most research.

I should say something first about the expected utility theory that still retains the position of highest honor in the pantheon of economic tools. It has dominated much economic theory so long because the theory offers a parsimonious representation of truly rational behavior under uncertainty. The axioms (Savage, 1954) from which expected utility theory is derived are undeniably sensible representations of basic requirements of rationality. For many purposes, it serves well to base an economic theory on such assumptions of strictly rational behavior, especially if the assumptions of the model are based on simple, robust realities, if the model concerns well-considered decisions of informed people, and if the phenomenon to be explained is one of stable behavior over many repetitions, where learning about subtle issues has a good chance of occurring.

Still, despite the obvious attractiveness of expected utility theory, it has long been known that the theory has systematically mispredicted human behavior, at least in certain circumstances. Allais (1953) reported examples showing that in choosing between certain lotteries, people systematically violate the theory. Kahneman and Tversky (1979) give the following experimental evidence to illustrate one of Allais' examples. When their subjects were asked to choose between a lottery offering a 25% chance of winning 3,000 and a lottery offering a 20% chance of winning 4,000, 65% of their subjects chose the latter, while when subjects were asked to choose between a 100% chance of winning 3,000 and an 80% chance of winning 4,000, 80% chose the former. Expected utility theory predicts that they should not choose differently in these two cases, since the second choice is the same as the first except that all probabilities are multiplied by the same constant. Their preference for the first choice in the lottery when it is certain in this example illustrates what is called the "certainty effect," a preference for certain outcomes.

Prospect theory is a mathematically-formulated alternative to the theory of expected utility maximization, an alternative that is supposed to capture the results of such experimental research. (A prospect is the Kahneman–Tversky name for a lottery as in the

Allais example above.) Prospect theory actually resembles expected utility theory in that individuals are represented as maximizing a weighted sum of "utilities," although the weights are not the same as probabilities and the "utilities" are determined by what they call a "value function" rather than a utility function.

The weights are, according to Kahneman and Tversky (1979) determined by a function of true probabilities which gives zero weight to extremely low probabilities and a weight of one to extremely high probabilities. That is, people behave as if they regard extremely improbable events as impossible and extremely probable events as certain. However, events that are just very improbable (not extremely improbable) are given too much weight; people behave as if they exaggerate the probability. Events that are very probable (not extremely probable) are given too little weight; people behave as if they underestimate the probability. What constitutes an extremely low (rather than very low) probability or an extremely high (rather than very high) probability is determined by individuals' subjective impression and prospect theory is not precise about this. Between the very low and very high probabilities, the weighting function (weights as a function of true probabilities) has a slope of less than one.

This shape for the weighting function allows prospect theory to explain the Allais certainty effect noted just above. Since the 20% and 25% probabilities are in the range of the weighting function where its slope is less than one, the weights people attach to the two outcomes are more nearly equal than are the probabilities, and people tend just to choose the lottery that pays more if it wins. In contrast, in the second lottery choice the 80% probability is reduced by the weighting function while the 100% probability is not; the weights people attach to the two outcomes are more unequal than are the probabilities, and people tend just to choose the outcome that is certain.

If we modify expected utility function only by substituting the Kahneman and Tversky weights for the probabilities in expected utility theory, we might help explain a number of puzzling phenomena in observed human behavior toward risk. For a familiar example, such a modification could explain the apparent public enthusiasm for high-prize lotteries, even though the probability of winning is so low that expected payout of the lottery is not high. It could also explain such phenomenon as the observed tendency for overpaying for airline flight insurance (life insurance policies that one purchases before an airline flight, that has coverage only during that flight), Eisner and Strotz (1961).

The Kahneman–Tversky weighting function may explain observed overpricing of out-of-the-money and in-the-money options. Much empirical work on stock options pricing has uncovered a phenomenon called the "options smile" (see Mayhew, 1995, for a review.). This means that both deep out-of-the-money and deep in-the-money options have relatively high prices, when compared with their theoretical prices using Black–Scholes formulae, while near-the-money options are more nearly correctly priced. Options theorists, accustomed to describing the implied volatility of the stock implicit in options prices, like to state this phenomenon not in terms of option prices but in terms of these implied volatilities. When the implied volatility for options of various strike prices at a point in time derived using the Black–Scholes (1973) formula are plotted, on the vertical axis, against the strike price on the horizontal axis, the curve often resembles a smile. The curve is higher both for low strike price (out-of-the-money) options and for high strike price (in-the-money) options than it is for middle-range strike prices. This options smile might possibly be explained in terms of the distortion in probabilities represented by the Kahneman–Tversky weighting function, since the theory would suggest that people act as if they overestimate the small probability that the price of the underlying crosses the strike price and underestimate the high probability that the price remains on the same side of the strike price. The Kahneman–Tversky weighting function might even explain the down-turned corners of the mouth that some smiles exhibit (see Fortune, 1996) if at these extremes the discontinuities at the extremes of the weighting function become relevant.[2]

We now turn to the other foundation of prospect theory, the Kahneman and Tversky (1979) value function. The value function differs from the utility function in expected utility theory in a very critical respect: the function (of wealth or payout) has a kink in it at a point, the "reference point," the location of which is determined by the subjective impressions of the individual. The reference point is the individual's point of comparison, the "status quo" against which alternative scenarios are contrasted. Taking value as a function of wealth, the Kahneman–Tversky (1979) value function is upward sloping everywhere, but with an abrupt decline in slope at the reference point (today's wealth or whatever measure of wealth that is psychologically important to the subject). For wealth levels above the reference point, the value function is concave downward, just as are conventional utility functions. At the reference point, the value function may be regarded, from the fact that its slope changes abruptly there, as infinitely concave downward. For wealth levels below the reference point, Kahneman and Tversky found evidence that the value function is concave upward, not downward. People are risk lovers for losses, they asserted.

Perhaps the most significant thing to notice about the Kahneman–Tversky value function is just the discontinuity in slope at the reference value, the abrupt downward change in slope as one moves upward past the reference value. Prospect theory does not nail down accurately what determines the location of the reference point, just as it does not nail down accurately, for the weighting function, what is the difference between very high probabilities and extremely high probabilities. The theory does not specify these matters because experimental evidence has not produced any systematic patterns of behavior that can be codified in a general theory. However, the reference point is thought to be determined by some point of comparison that the subject finds convenient, something readily visible or suggested by the wording of a question.

This discontinuity means that, in making choices between risky outcomes, people will behave in a risk averse manner, no matter how small the amounts at stake are. This is a contrast to the prediction of expected utility theory with a utility function of wealth without

[2] There are other potential explanations of the options smile in terms of nonnormality or jump processes for returns, and these have received the attention in the options literature. Such explanations might even provide a complete rational basis for the smile, though it is hard to know for sure. Since the 1987 stock market crash, the options smile has usually appeared distorted into an options "leer," with the left side of the mouth higher (e.g., the deep out-of-the-money puts are especially overpriced), see Bates (1995), Jackwerth and Rubinstein (1995) and Bates (1991). Public memories of the 1987 crash are apparently at work in producing this "leer."

kinks, for which, since the utility function is approximately linear for small wealth changes, people should behave as if they are risk neutral for small bets. That people would usually be risk neutral for small bets would be the prediction of expected utility theory even if the utility function has such a slope discontinuity, since the probability that wealth is currently at the kink is generally zero. With prospect theory, in contrast, the kink always moves with wealth to stay at the perceived current level of wealth (or the current point of reference); the kink is always relevant.

Samuelson (1963) told a story which he perceived as demonstrating a violation of expected utility theory, and, although it came before Kahneman and Tversky's prospect theory, it illustrates the importance of the kink in the value function. Samuelson reported that he asked a lunch colleague whether he would accept a bet that paid him $200 with a probability of .5 and lost him $100 with a probability of .5. The colleague said he would not take the bet, but that he would take a hundred of them. With 100 such bets, his expected total winnings are $5,000 and he has virtually no chance of losing any money. It seems intuitively compelling to many people that one would readily take the complete set of bets, even if any element of the set is unattractive. Samuelson proved that if his colleague would answer the same way at any wealth level, then he necessarily violates expected utility theory.

Samuelson's colleague is not, however, in violation of prospect theory. When viewing a single bet, the kink in the value function is the dominant consideration. If he were to judge 100 bets sequentially, the kink would always be relevant (the reference point would move with each successive bet) and he would reject all of them. But if he were to judge 100 bets together, the collective outcomes would be far above today's value function kink, and the bet is, by prospect theory, clearly desirable.

The failures to accept many such bets when one considers them individually has been called "myopic loss aversion" by Benartzi and Thaler (1995). They argue that, under estimated values for the magnitude of the kink in the Kahneman–Tversky value function, the "equity premium puzzle" of Mehra and Prescott (1985) can be resolved; see also Siegel and Thaler (1997).

Today, the term "equity premium puzzle," coined by Mehra and Prescott (1985), is widely used to refer to the puzzlingly high historical average returns of stocks relative to bonds.[3] The equity premium is the difference between the historical average return in the stock market and the historical average return on investments in bonds or treasury bills. According to Siegel (1994), the equity premium of U.S. stocks over short-term government bonds has averaged 6.1% a year for the United States for 1926 to 1992, and so one naturally

[3] Mehra and Prescott did not discover the equity premium. Perhaps that honor should go to Smith (1925), although there must be even earlier antecedents in some forms. Mehra and Prescott's original contribution seems to have been, in the context of present-value investor intertemporal optimizing models, to stress that the amount of risk aversion that would justify the equity premium, given the observed correlation of stocks with consumption, would imply much higher riskless interest rates than we in fact see.

wonders why people invest at all in debt if it is so outperformed by stocks.[4] Those who have tried to reconcile the equity premium with rational investor behavior commonly point out the higher risk that short-run stock market returns show: investors presumably are not fully enticed by the higher average returns of stocks since stocks carry higher risk. But, such riskiness of stocks is not a justification of the equity premium, at least assuming that investors are mostly long term. Most investors ought to be investing over decades, since most of us expect to live for many decades, and to spend the twilight of their lives living off savings. Over long periods of times, it has actually been long-term bonds (whose payout is fixed in nominal terms), not the stocks, that have been more risky in real terms, since the consumer price index has been, despite its low variability from month to month, very variable over long intervals of time, see Siegel (1994). Moreover, stocks appear strictly to dominate bonds: there is no thirty-year period since 1871 in which a broad portfolio of stocks was outperformed either by bonds or treasury bills.[5]

Benartzi and Thaler show (1995) that if people use a one-year horizon to evaluate investments in the stock market, then the high equity premium is explained by myopic loss aversion. Moreover, prospect theory does not suggest that in this case riskless real interest rates need be particularly high. Thus, if we accept prospect theory and that people frame stock market returns as short-term, the equity premium puzzle is solved.

Benartzi and Thaler (1996) demonstrated experimentally that when subjects are asked to allocate their defined contribution pension plans between stocks and fixed incomes, their responses differed sharply depending on how historical returns were presented to them. If they were shown 30 one-year returns, their median allocation to stocks was 40%, but if they were shown 30-year returns their median allocation to stocks was 90%. Thaler, Tversky, Kahneman and Schwartz (1997) shows further experiments confirming this response.

Loss aversion has also been used to explain other macroeconomic phenomena, savings behavior (Bowman, Minehart and Rabin, 1993) and job search behavior (Bryant, 1990).

Regret and Cognitive Dissonance

There is a human tendency to feel the pain of regret at having made errors, even small errors, not putting such errors into a larger perspective. One "kicks oneself" at having done something foolish. If one wishes to avoid the pain of regret, one may alter one's behavior in ways that would in some cases be irrational unless account is taken of the pain of regret.

The pain of regret at having made errors is in some senses embodied in the Kahneman–

[4] Siegel (1994, p. 20). However, Siegel notes that the U.S. equity premium was only 1.9% per year 1816–70 and 2.8% per year 1871–1925.

[5] Siegel (1994, p. 31). It should be noted that one must push the investor horizon up to a fairly high number, around 30 years, before one finds that historically stocks have always outperformed bonds since 1871; for ten year periods of time one finds that bonds often outperform stocks. There are not many thirty-year periods in stock market history, so this information might be judged as insubstantial. Moreover, Siegel notes that even with a thirty-year period stocks did not always outperform bonds in the U.S. before 1871.

Tversky notion of a kink in the value function at the reference point. There are also other ways of representing how people behave who feel pain of regret. Loomes and Sugden (1982) have suggested that people maximize the expected value of a “modified utility function” which is a function of the utility they achieve from a choice as well as the utility they would have achieved from another choice that was considered. Bell (1982) proposed a similar analysis.

Regret theory may apparently help explain the fact that investors defer selling stocks that have gone down in value and accelerate the selling of stocks that have gone up in value, Shefrin and Statman (1985). Regret theory may be interpreted as implying that investors avoid selling stocks that have gone down in order not to finalize the error they make and not to feel the regret. They sell stocks that have gone up in order that they cannot regret failing to do so before the stock later fell, should it do so. That such behavior exists has been documented using volume of trade data by Ferris, Haugen and Makhija (1988) and Odean (1996b).

Cognitive dissonance is the mental conflict that people experience when they are presented with evidence that their beliefs or assumptions are wrong; as such, cognitive dissonance might be classified as a sort of pain of regret, regret over mistaken beliefs. As with regret theory, the theory of cognitive dissonance (Festinger, 1957) asserts that there is a tendency for people to take actions to reduce cognitive dissonance that would not normally be considered fully rational: the person may avoid the new information or develop contorted arguments to maintain the beliefs or assumptions. There is empirical support that people often make the errors represented by the theory of cognitive dissonance. For example, in a classic study, Erlich, Guttman, Schopenback and Mills (1957) showed that new car purchasers selectively avoid reading, after the purchase is completed, advertisements for car models that they did not choose, and are attracted to advertisements for the car they chose.

McFadden (1974) modelled the effect of cognitive dissonance in terms of a probability of forgetting contrary evidence and showed how this probability will ultimately distort subjective probabilities. Goetzmann and Peles (1993) have argued that the same theory of cognitive dissonance could explain the observed phenomenon that money flows in more rapidly to mutual funds that have performed extremely well than flows out from mutual funds that have performed extremely poorly: investors in losing funds are unwilling to confront the evidence that they made a bad investment by selling their investments.

Anchoring

It is well-known that when people are asked to make quantitative assessments their assessments are influenced by suggestions. An example of this is found in the results survey researchers obtain. These researchers often ask people about their incomes using questionnaires in which respondents are instructed to indicate which of a number of income brackets, shown as choices on the questionnaire, their incomes fall into. It has been shown that the answers people give are influenced by the brackets shown on the questionnaire. The tendency to be influenced by such suggestions is called “anchoring” by psychologists.

In some cases, at least, anchoring may be rational behavior for respondents. They may rationally assume that the deviser of the questionnaire uses some information (in this case, about typical people's incomes) when devising the questionnaire. Not fully remembering their own income, they may rely on the information in the brackets to help them answer better. If the brackets do contain information, then it is rational for subjects to allow themselves to be influenced by the brackets.

While anchoring undoubtedly has an information-response component in many circumstances, it has also been shown that anchoring behavior persists even when information is absent. In one experiment Tversky and Kahneman (1974), subjects were given simple questions whose answers were in percentages, e.g., the percentage of African nations in the United Nations. A wheel of fortune with numbers from 1 to 100 was spun before the subjects. Obviously, the number at which the wheel of fortune stopped had no relevance to the question just asked. Subjects were asked whether their answer was higher or lower than the wheel of fortune number, and then to give their own answer. Respondents' answers were strongly influenced by the "wheel of fortune." For example, the median estimates of the percentage of African countries in the United Nations were 25 and 45 for groups that received 10 and 65, respectively, as starting points (p. 184).

Values in speculative markets, like the stock market, are inherently ambiguous. Who would know what the value of the Dow Jones Industrial Average should be? Is it really "worth" 6,000 today? Or 5,000 or 7,000? or 2,000 or 10,000? There is no agreed-upon economic theory that would answer these questions. In the absence of any better information, past prices (or asking prices or prices of similar objects or other simple comparisons) are likely to be important determinants of prices today.

That anchoring affects valuations, even by experts, was demonstrated by Northcraft and Neale (1987) in the context of real estate valuation. All subjects were taken to a house for sale, asked to inspect the house for up to 20 minutes, and were given a ten-page packet of information about the house and about other houses in the area, giving square footage and characteristics of the properties, and prices of the other properties. The same packet was given to all subjects except that the asking price of the property under consideration and its implied price per square foot were changed between subjects. Subjects were asked for their own opinions of its appraisal value, appropriate listing price, purchase price, and the lowest offer the subject would accept for the house if the subject were the seller. The real estate agents who were given an asking price of $119,900 had a mean predicted appraisal value of $114,204, listing price of $117,745, purchase price of $111,454 and a lowest acceptable offer of $111,136, while the real estate agents who were given an asking price of $149,900 had a mean appraisal value of $128,754, listing price of $130,981, predicted purchase price of $127,318, and a lowest offer of $123,818. The changed asking prices thus swayed their valuations by 11% to 14% of the value of the house. Similar results were found with amateur subjects. While this experiment does not rule out that the effect of the asking price was due to a rational response to the assumed information in the asking price, the effects of asking price are remarkably large, given that so much other information on the house was also given. Moreover, when subjects were asked afterwards to list the items of information that weighed most heavily in their valuations, only 8% of the expert subjects and only 9% of the amateur subjects listed asking price of the property under consideration among the top three items. Note that the valuation problem presented to these subjects is far less difficult or ambiguous than the problem of determining the “correct” value for the stock market, since here they are implicitly being asked to assume that the comparable properties are correctly valued. (See also McFadden, 1974 and Silberman and Klock, 1989.)

One might object that the notion that anchoring on past prices helps determine present price in the stock market might be inconsistent with the low serial correlation of stock price changes, that is with the roughly random-walk behavior of daily or monthly stock prices that has been widely noted.[6] This conclusion is not warranted however. Models of “smart money” (i.e., people who are unusually alert to profit opportunities in financial markets) seeking to exploit serial correlation in price, models which also include ordinary investors, are consistent with the implications that serial correlation is low and yet the anchoring remains important for the level of stock prices (see Shiller, 1984, 1990).

By extension from these experimental results, it is to be presumed that very many economic phenomena are influenced by anchoring. Gruen and Gizycki (1993) used it to explain the widely observed anomaly[7] that forward discounts to not properly explain subsequent exchange rate movements. The anchoring phenomenon would appear relevant to the “sticky prices” that are so talked about by macroeconomists. So long as past prices are taken as suggestions of new prices, the new prices will tend to be close to the past prices. The more ambiguous the value of a commodity, the more important a suggestion is likely to be, and the more important anchoring is likely to be for price determination.

The anchoring phenomenon may help to explain certain international puzzles observed in financial markets. U.S. investors who thought in the late 1980s that Japanese stock price–earnings ratios were outrageously high then may have been influenced by the readily-available anchor of (much lower) U.S. price–earnings ratios. By the mid 1990s, many U.S. investors feel that the Tokyo market is no longer overpriced (see Shiller, Kon-Ya and Tsutsui, 1996), even though price–earnings ratios remain much higher than in the U.S. perhaps because the anchor of the widely-publicized high Tokyo price–earnings ratios of the late 1980s appears to be another anchor.

Anchoring may also be behind certain forms of money illusion. The term money illusion, introduced by Fisher (1928), refers to a human tendency to make inadequate allowance, in economic decisions, for the rate of inflation, and to confuse real and nominal quantities. Shafir, Diamond and Tversky (1997) have shown experimentally that people tend to give different answers to the same hypothetical decision problem depending on whether the problem was presented in a way that stressed nominal quantities or in a way that

[6] The notion that speculative prices approximately describe "random walks" was first proposed by Bachelier (1900, 1964). It became widely associated with the efficient markets hypothesis, the hypothesis that market prices efficiently incorporate all available information, with the work of Fama (1970). For further information on the literature on the random walk and efficient markets theory see also Cootner (1964), Malkiel (1981), and Fama (1991).

[7] For a discussion of the anomaly, see Backus, Foresi and Telmer (1995) and Froot and Thaler (1990).

stressed real quantities. The quantities that were shown in the question (whether nominal or real) may have functioned as anchors.[8]

Mental Compartments

Related to the anchoring and framing phenomena is a human tendency to place particular events into mental compartments based on superficial attributes. Instead of looking at the big picture, as would be implied by expected utility theory, they look at individual small decisions separately.

People may tend to place their investments into arbitrarily separate mental compartments, and react separately to the investments based on which compartment they are in. Shefrin and Statman (1994) have argued that individual investors think naturally in terms of having a "safe" part of their portfolio that is protected from downside risk and a risky part that is designed for a chance of getting rich. Shefrin and Thaler (1988) have argued that people put their sources of income into three categories, current wage and salary income, asset income, and future income, and spend differently out of the present values of these different incomes. For example, people are reluctant to spend out of future income even if it is certain to arrive.

The tendency for people to allow themselves to be influenced by their own mental compartments might explain the observed tendency for stock prices to jump up when the stock is added to the Standard and Poor Stock Index (see Shleifer, 1986). It might also help explain the widely noted "January effect" anomaly. This anomaly, that stock prices tend to go up in January, has been observed in as many as 15 different countries (Gultekin and Gultekin, 1983). The anomaly cannot be explained in terms of effects related to the tax year, since it persists also in Great Britain (whose tax year begins in April) and Australia (whose tax year begins in July), see Thaler (1987). If people view the year end as a time of reckoning and a new year as a new beginning, they may be inclined them to behave differently at the turn of the year, and this may explain the January effect.

A tendency to separate out decisions into separate mental compartments may also be behind the observed tendency for hedgers to tend to hedge specific trades, rather than their overall profit situation. René Stulz (1996, p. 8), in summarizing the results of his research and that of others on the practice of risk management by firms, concludes that:

> It immediately follows from the modern theory of risk management that one should be concerned about factors that affect the present value of future cash flows. This is quite different from much of the current practice of risk management where one is concerned about hedging transaction risk or the risk of transactions expected to occur in the short run.

[8] There appears to be much more to money illusion than just anchoring; people associate nominal quantities with opinions about the economy, anticipated behavior of the government, fairness, and prestige, opinions that are not generally shared by economists, see Shiller (1997a,b).

The Wharton/CIBC Wood Gundy 1995 Survey of Derivatives Usage by U.S. Non-Financial Firms (Bodnar and Marston, 1996) studied 350 firms: 176 firms in the manufacturing sector, 77 firms in the primary products sector, and 97 firms in the service sector. When asked by the Wharton surveyors what was the most important objective of hedging strategy, 49% answered managing "volatility in cashflows," 42% answered managing "volatility in accounting earnings," and only 8% answered managing "the market value of the firm" (1% answered "managing balance sheet accounts and ratios"). Fifty percent of the respondents in the survey reported frequently hedging contractual commitments, but only 8% reported frequently hedging competitive/economic exposure.

It is striking that only 8% reported that their most important objective is the market value of the firm, since maximizing the market value of the firm is, by much financial theory, the ultimate objective of the management of the firm. It is of course hard to know just what people meant by their choices of answers, but there is indeed evidence that firms are driven in their hedging by the objective of hedging specific near-term transactions, and neglect consideration of future transactions or other potential factors that might also pose longer run risks to the firm. In the Wharton study, among respondents hedging foreign currency risks, 50% reported hedging anticipated transactions less than one year off, but only 11% report frequently hedging transactions more than one year off. This discrepancy is striking, since most of the value of the firm (and most of the concerns it has about its market value) must come in future years, not the present year.[9]

Overconfidence, Over- and Under-Reaction and the Representativeness Heuristic

People often tend to show, in experimental settings, excessive confidence about their own judgments. Lichtenstein, Fischhoff and Philips (1977) asked subjects to answer simple factual questions (e.g., "Is Quito the capital of Ecuador?") and then asked them to give the probability that their answer was right: subjects tended to overestimate the probability that they were right, in response to a wide variety of questions.

Such studies have been criticized (see Gigerenzer, 1991) as merely reflecting nothing more than a difference between subjective and frequentist definitions of probability, i.e., critics claimed that individuals were simply reporting a subjective degree of certainty, not the fraction times they are right in such circumstances. However, in reaction to such criticism, Fischhoff, Slovic and Lichtenstein (1977) repeated the experiments asking the

[9] Recent surveys of hedging behavior of firms indicates that despite extensive development of derivative products, actual use of these products for hedging is far from optimal. Of the firms cited in the Wharton/study, only 40.5% reported using derivatives at all. On the other hand, Dolde (1993) surveyed 244 Fortune 500 companies and concluded that over 85% used swaps, forwards, futures or options in managing financial risk. Nance, Smith and Smithson (1993) in a survey of 194 firms reported that 62% used hedging instruments in 1986. These studies concentrated on rather larger companies than did the Wharton study. Overall, these studies may be interpreted as revealing a surprisingly low fraction of respondents who do any hedging, given that firms are composed of many people, any one of whom might be expected to initiate the use of derivatives.

subjects for probability odds that they are right and very clearly explaining what such odds mean, and even asking them to stake money on their answer. The overconfidence phenomenon persisted. Moreover, in cases where the subjects said they were certain they were right, they were in fact right only about 80% of the time: there is no interpretation of subjective probability that could reconcile this result with correct judgments.

A tendency towards overconfidence among ordinary investors seems apparent when one interviews them. One quickly hears what seem to be overconfident statements. But how can it be that people systematically are so overconfident? Why wouldn't people learn from life's experiences to correct their overconfidence?

Obviously, people do learn substantially in circumstances when the consequences of their errors are repeatedly presented to them, and sometimes they even overreact and show too little confidence. But still there seems to be a common bias towards overconfidence. Overconfidence is apparently related to some deep-set psychological phenomena: Ross (1987) argues that much overconfidence is related to a broader difficulty with "situational construal," a difficulty in making adequate allowance for the uncertainty in one's own view of the broad situation, a more global difficulty tied up with multiple mental processes. Overconfidence may also be traced to the "representativeness heuristic," Tversky and Kahneman (1974), a tendency for people to try to categorize events as typical or representative of a well-known class, and then, in making probability estimates, to overstress the importance of such a categorization, disregarding evidence about the underlying probabilities.[10] One consequence of this heuristic is a tendency for people to see patterns in data that is truly random, to feel confident, for example, that a series which is in fact a random walk is not a random walk.[11]

Overconfidence itself does not imply that people overreact (or underreact) to all news. In fact, evidence on the extent of overreaction or underreaction of speculative asset prices to news has been mixed.

There has indeed been evidence of overreaction. The first substantial statistical evidence for what might be called a general market overreaction can be found in the literature on excess volatility of speculative asset prices, Shiller (1979, 1981a,b) and LeRoy and Porter (1981). We showed statistical evidence that speculative asset prices show persistent deviations from the long-term trend implied by the present-value efficient markets model, and then, over horizons of many years, to return to this trend. This pattern of price behavior, it was argued, made aggregate stock prices much more volatile than would be implied by the efficient markets model. It appears as if stock prices overreact to some news, or to their own past values, before investors come to their senses and correct the prices. Our arguments led to a spirited debate about the validity of the efficient markets model in the

[10] People tend to neglect "base rates," the unconditional probabilities or frequencies of events, see Meehl and Rosen (1955).

[11] Rabin (1996) characterizes this judgment error as a tendency to over-infer the probability distribution from short sequences. Part of overconfidence may be nothing more than simple forgetting of contrary evidence; a tendency to forget is by its very nature not something that one can learn to prevent.

finance literature, a literature that has too many facets to summarize here, except to say that it confirms there are many potential interpretations of any statistical results based on limited data.[12] My own view of the outcome of this debate is that it is quite likely that speculative asset prices tend to be excessively volatile. Certainly, at the very least, one can say that no one has been able to put forth any evidence that there is not excess volatility in speculative asset prices. For an evaluation of this literature, see Shiller (1989), Campbell and Shiller (1988, 1989), West (1988), and Campbell, Lo and MacKinlay (1997, Ch. 7).

Since then, papers by De Bondt and Thaler (1985), Fama and French (1988), Poterba and Summers (1988), and Cutler, Poterba and Summers (1991) have confirmed the excess volatility claims by showing that returns tend to be negatively autocorrelated over horizons of three to five years, that an initial overreaction is gradually corrected. Moreover, Campbell and Shiller (1988, 1989) show that aggregate stock market dividend yields or earnings yields are positively correlated with subsequently observed returns over similar intervals; see also Dreman and Berry (1995).[13] Campbell and Shiller (1998) connect this predictive power to the observed stationarity of these ratios. Since the ratios have no substantial trend over a century and appear mean reverting over much shorter time intervals, the ratio must predict future changes in either the numerator (the dividend or earnings) or the denominator (the price); we showed that it has been unequivocally the denominator, the price, that has restored the ratios to their mean after they depart from it, and not the numerator. La Porta (1996) found that stocks for which analysts projected low earnings growth tended to show upward price jumps on earnings announcement dates, and stocks for which analysts projected high earnings growth tended to show downward price jumps on earnings announcement dates. He interprets this as consistent with a hypothesis that analysts (and the market) excessively extrapolated past earnings movements and only gradually correct their errors as earnings news comes in. The behavior of initial public offerings around announcement dates appears also to indicate some overreaction and later rebound, see Ibbotson and Ritter (1988) and Ritter (1991).

On the other hand, there has also been evidence of what might be called underreaction. Most days when big news breaks have been days of only modest stock market price movements, the big movements tending to come on days when there is little news, see Cutler, Poterba and Summers (1989). Cutler, Poterba and Summers (1991) also found that

[12] There has been some confusion about the sense in which the present-value efficient markets model puts restrictions on the short-run (or high frequency) movements in speculative asset prices. The issues are laid out in Shiller (1979), (appendix). Kleidon (1986) rediscovered the same ideas again, but gave a markedly different interpretation of the implications for tests of market efficiency.

[13] An extensive summary of the literature on serial correlation of US stock index returns is in Campbell, Lo and MacKinlay (1997). Chapter 2 documents the positive serial correlation of returns over short horizons, but concludes that the evidence for negative serial correlation of returns over long horizons is weak. Chapter 7, however, shows evidence that long-horizon returns are negatively correlated with the price-earnings ratio and price-dividend ratio. Recent critics of claims that long-horizon returns can be forecasted include Goetzmann and Jorion (1992), Nelson and Kim (1993) and Kirby (1997). In my view, they succeed in reducing the force of the evidence, but not the conclusion that long-horizon returns are quite probably forecastable.

for a number of indices of returns on major categories of speculative assets there has been a tendency for positive autocorrelation of short-run returns over short horizons, less than a year; see also Jegadeesh and Titman (1993) and Chan, Jegadeesh and Lakonishok (1996).[14] This positive serial correlation in return indices has been interpreted as implying an initial underreaction of prices to news, to be made up gradually later. Bernard and Thomas (1992) found evidence of underreaction of stock prices to changes, from the previous year, in company earnings: prices react with a lag to earnings news; see also Ball and Brown (1968).[15] Irving Fisher (1930, Ch. XXI, pp. 493–94) thought that, because of human error, nominal interest rates tend to underreact to inflation, so that there is a tendency for low real interest rates in periods of high inflation, and high real rates in periods of low inflation. More recent data appear to confirm this behavior of real interest rates, and data on inflationary expectations also bear out Fisher's interpretation that the phenomenon has to do with human error; see De Bondt and Bange (1992) and Shefrin (1997).[16]

Does the fact that securities prices sometimes underreact pose any problems for the psychological theory that people tend to be overconfident? Some observers seem to think that it does. In fact, however, overconfidence and overreaction are quite different phenomena. People simply cannot overreact to everything: if they are overconfident they will make errors, but not in any specified direction in all circumstances. The concepts of overreaction or underreaction, while they may be useful in certain contexts, are not likely to be good psychological foundations on which to organize a general theory of economic behavior.

The fact that both overreaction and underreaction are observed in financial markets has been interpreted by Fama (1997) as evidence that the anomalies from the standpoint of efficient markets theory are just "chance results," and that therefore the theory of market efficiency survives the challenge of its critics. He is right, of course, that both overreaction and underreaction together may sometimes seem a little puzzling. But one is not likely to want to dismiss these as "chance results" if one has an appreciation for the psychological theory that might well bear on these phenomena. In his survey of behavioral finance Fama

[14] Lo and MacKinlay (1988) and Lehmann (1990), however, find evidence of *negative* serial correlation of individual weekly stock returns between successive weeks. As explained by Lo and MacKinlay (1990), weekly returns on portfolios of these same stocks still exhibit positive serial correlation from week to week because the cross-covariances between returns of individual stocks are positive. They conclude that this pattern of cross-covariances is not what one would expect to find based on theories of investor inertia. Lehmann, however, has a different interpretation of the negative week-to-week serial correlation of individual weekly stock returns, that the negative serial correlation reflects nothing more than the behavior of market makers facing order imbalances and asymmetric information.

[15] Firms' management appear acutely aware that earnings growth has a psychological impact on prices, and so attempt to manage earnings accounting to provide a steady growth path. Impressive evidence that they do so is found in Degeorge, Patel and Zeckhauser (1997).

[16] Modigliani and Cohn (1979) argue that public failure to understand the relation of interest rates to inflation has caused the stock market to overreact to nominal interest rate changes.

(1997) makes no more than a couple of oblique references to any literature from the other social sciences. In fact, Fama states that the literature on testing market efficiency has no clearly stated alternative, “the alternative hypothesis is vague, market inefficiency” (p. 1). Of course, if one has little appreciation of these alternative theories then one might well conclude that the efficient markets theory, for all its weaknesses, is the best theory we have. Fama appears to believe that the principal alternative theory is just one of consistent overreaction or underreaction, and says that “since the anomalies literature has not settled on a testable alternative to market efficiency, to get the ball rolling, I assume that reasonable alternatives must predict either over-reaction or under-reaction” (p. 2). The psychological theories reviewed here cannot be reduced to such simple terms, contrary to Fama’s expectations.

Barberis, Shleifer and Vishny (1997) provide a psychological model, involving the representativeness heuristic as well as a principle of conservatism (Edwards, 1968), that offers a reconciliation of the overreaction and underreaction evidence from financial markets; see also Daniel, Hirshleifer and Subrahmanyam (1997) and Wang (1997). More work could be done in understanding when it is that people overreact in financial markets and when it is that they underreact. Understanding these overreaction and underreaction phenomena together appears to be a fertile field for research at the present time. There is neither reason to think that it is easy obtain such an understanding, nor reason to despair that it can ever be done.

Overconfidence may have more clear implications for the volume of trade in financial markets than for any tendency to overreact. If we connect the phenomenon of overconfidence with the phenomenon of anchoring, we see the origins of differences of opinion among investors, and some of the source of the high volume of trade among investors. People may fail to appreciate the extent to which their own opinions are affected by anchoring to cues that randomly influenced them, and take action when there is little reason to do so.

The extent of the volume of trade in financial markets has long appeared to be a puzzle. The annual turnover rate (shares sold divided by all shares outstanding) for New York Stock Exchange Stocks has averaged 18% a year from the 1950s through the 1970s, and has been much higher in certain years. The turnover rate was 73% in 1987 and 67% in 1930. It does not appear to be possible to justify the number of trades in stocks and other speculative assets in terms of the normal life-cycle ins and outs of the market. Theorists have established a “nonspeculation theorem” that states that rational agents who differ from each other only in terms of information and who have no reason to trade in the absence of information will not trade (Milgrom and Stokey, 1982l; Geanakoplos, 1992).

Apparently, many investors do feel that they do have speculative reasons to trade often, and apparently this must have to do with some tendency for each individual to have beliefs that he or she perceives as better than others’ beliefs. It is as if most people think they are above average.

Odean (1996a), in analyzing individual customer accounts at a nationwide discount brokerage house, examined the profits that customers made on trades that were apparently not motivated by liquidity demands, tax loss selling, portfolio rebalancing, or a move to lower-risk securities. On the remaining trades, the returns on the stocks purchased was on average lower, not higher, than on those sold. This appears to be evidence of overconfidence among these investors.

Within the week of the stock market crash of October 19, 1987 I sent out questionnaires to 2,000 wealthy individual investors and 1,000 institutional investors, asking them to recall their thoughts and reasons for action on that day; see Shiller (1987b). There were 605 completed responses from individuals and 284 responses from institutions. One of the questions I asked was: "Did you think at any point on October 19, 1987 that you had a pretty good idea when a rebound was to occur?" Of individual investors, 29.2% said yes, of institutional investors, 28.0% said yes. These numbers seem to be surprisingly high: one wonders why people thought they knew what was going to happen in such an unusual situation. Among those who bought on that day, the numbers were even higher, 47.1% and 47.9% respectively. The next question on the questionnaire was "If yes, what made you think you knew when a rebound was to occur?" Here, there was a conspicuous absence of sensible answers; often the answers referred to "intuition" or "gut feeling." It would appear that the high volume of trade on the day of the stock market crash, as well as the occurrence, duration, and reversal of the crash was in part determined by overconfidence in such intuitive feelings.[17]

If people are not independent of each other in forming overconfident judgments about investments, and if these judgments change collectively through time, then these "noisy" judgments will tend to cause prices of speculative assets to deviate from their true investment value. Then a "contrarian" investment strategy, advocated by Graham and Dodd (1934) and Dreman (1977) among many others, a strategy of investing in assets that are currently out of favor by most investors, ought to be advantageous. Indeed, there is much evidence that such contrarian investment strategy does pay off, see for example, De Bondt and Thaler (1985), Fama and French (1988, 1992), Fama (1991), and Lakonishok, Shleifer and Vishny (1994). That a simple contrarian strategy may be profitable may appear to some to be surprising: one might think that "smart money," by competing with each other to benefit from the profit opportunities, would ultimately have the effect of eliminating any such profit opportunities. But, there are reasons to doubt that such smart money will indeed have this effect; see Shiller (1984), De Long et al. (1990a,b), and Shleifer and Vishny (1996).[18]

[17] See also Case and Shiller (1988) for a similar analysis of recent real estate booms and busts. On the other hand, Garber (1990) analyzes some famous speculative bubbles, including the tulipomania in the 17th century, and concludes that they may have been rational.

[18] Even public expectations of a stock market crash does not prevent the stock market from rising; there is evidence from options prices that the stock market crash of 1987 was in some sense expected before it happened; see Bates (1991, 1995). Lee, Shleifer and Thaler (1991) argue that investor expectations, or rather "sentiment" can be measured by closed-end mutual fund discounts, which vary through time.

The Disjunction Effect

The disjunction effect is a tendency for people to want to wait to make decisions until information is revealed, even if the information is not really important for the decision, and even if they would make the same decision regardless of the information. The disjunction effect is a contradiction to the "sure-thing principle" of rational behavior (Savage, 1954).

Experiments showing the disjunction effect were performed by Tversky and Shafir (1992). They asked their subjects whether they would take one of the bets that Samuelson's lunch colleague, discussed above, had refused a coin toss in which one has equal chances to win $200 or lose $100. Those who took the one bet were then asked whether they then wanted to take another such bet. If they were asked after the outcome of the first bet was known, then it was found that a majority of respondents took the second bet whether or not they had won the first. However, a majority would not take the bet if they had to make the decision before the outcome of the bet was known. This is a puzzling result: if one's decision is the same regardless of the outcome of the first bet, then it would seem that one would make the same decision before knowing the outcome. Tversky and Shafir gave their sense of the possible thought patterns that accompany such behavior: if the outcome of the first bet is known and is good, then subjects think that they have nothing to lose in taking the second, and if the outcome is bad they want to try to recoup their losses. But if the outcome is not known, then they have no clear reason to accept the second bet.

The disjunction effect might help explain changes in the volatility of speculative asset prices or changes in the volume of trade of speculative asset prices at times when information is revealed. Thus, for example, the disjunction effect can in principle explain why there is sometimes low volatility and low volume of trade just before an important announcement is made, and higher volatility or volume of trade after the announcement is made. Shafir and Tversky (1992) give the example of presidential elections, which sometimes induce stock market volatility when the election outcome is known even though many skeptics may doubt that the election outcome has any clear implications for market value.

Gambling Behavior and Speculation

A tendency to gamble, to play games that bring on unnecessary risks, has been found to pervade widely divergent human cultures around the world and appears to be indicative of a basic human trait, Bolen and Boyd (1968). Kallick et al. (1975) estimated that 61% of the adult population in the United States participated in some form of gambling or betting in 1974. They also estimated that 1.1% of men and 0.5% of women are "probably compulsive gamblers," while an additional 2.7% of men and 1% of women are "potential compulsive gamblers." These figures are not trivial, and it is important to keep in mind that compulsive gambling represents only an extreme form of the behavior that is more common.

The tendency for people to gamble has provided a puzzle for the theory of human behavior under uncertainty, since it means that we must accommodate both risk-avoiding behavior (as evidenced by people's willingness to purchase insurance) with an apparent riskloving behavior. Friedman and Savage (1948) proposed that the co-existence of these behaviors might be explained by utility functions that become concave upward in extremely high range, but such an explanation has many problems. For one thing, people who gamble do not appear to be systematically risk seekers in any general sense, instead they are seeking specific forms of entertainment or arousal.[19] Moreover, the gambling urge is compartmentalized in people's lives, it tends to take for each individual only certain forms: people specialize in certain games. The favored forms of gambling tend to be associated with a sort of ego involvement: people may feel that they are especially good at the games they favor or that they are especially lucky with these.

The complexity of human behavior exemplified by the gambling phenomenon has to be taken into account in understanding the etiology of bubbles in speculative markets. Gamblers may have very rational expectations, at some level, for the likely outcome of their gambling, and yet have other feelings that drive their actual behavior. Economists tend to speak of quantitative "expectations" as if these were the only characterization of people's outlooks that mattered. It is my impression, from interviews and survey results, that the same people who are highly emotionally involved with the notion that the stock market will go up may give very sensible, unexciting, forecasts of the market if asked to make quantitative forecasts.

The Irrelevance of History

One particular kind of overconfidence that appears to be common is a tendency to believe that history is irrelevant, not a guide to the future, and that the future must be judged afresh now using intuitive weighing only of the special factors we see now. This kind of overconfidence discourages taking lessons from past statistics; indeed most financial market participants virtually never study historical data for correlations or other such statistics; they take their anchors instead from casual recent observations. Until academic researchers started collecting financial data, most was just thrown away as irrelevant.

One reason that people may think that history is irrelevant is a human tendency toward historical determinism, a tendency to think that historical events should have been known in advance. According to historian Florovsky (1969, p. 364):

> In retrospect we seem to perceive the *logic* of events, which unfold themselves in a regular order, according to a recognizable pattern, with an alleged inner necessity, so that we get the impression that it really could not have happened otherwise.

Fischhoff (1975) attempted to demonstrate this tendency towards historical determinism

[19] According to the American Psychiatric Association's DSM–IV (1994), "Most individuals with Pathological Gambling say that they are seeking 'action' (an aroused, euphoric state) even more than money. Increasingly larger bets, or greater risks, may be needed to continue to produce the desired level of excitement" (p. 616).

by presenting experimental subjects with incomplete historical stories, stories that are missing the final outcome of the event. The stories were from historical periods remote enough in time that the subjects would almost certainly not know the actual outcome. Subjects were asked to assign probabilities to each of four different possible conclusions to the story (only one of which was the true outcome). There were two groups of subjects, one of which was told that one of the four outcomes had in fact happened. The probability given to the outcomes was on average 10% higher when people were told it was the actual outcome.

Fischhoff's demonstration of a behavior consistent with belief in historical determinism may not demonstrate the full magnitude of such behavior, because it does not capture the effects of social cognition of past events, a cognition that may tend to remember historical facts that are viewed as causing subsequent historical events, or are connected to them, and to forget historical facts that seem not to fit in with subsequent events. It will generally be impossible to demonstrate such phenomena of social cognition in short laboratory experiments.

A human tendency to believe in historical determinism would tend to encourage people to assume that past exigencies (the stock market crash of 1929, the great depression, the world wars, and so on) were probably somewhat known in advance, or, at least, that before these events people had substantial reason to worry that they might happen. There may tend to be a feeling that there is nothing definite on the horizon now, as there presumably was before these past events.[20] It is in this human tendency toward believing history is irrelevant that the equity premium puzzle, discussed above, may have its most important explanation. People may tend just not to think that the past stock market return history itself gives any indication of the future, at least not until they perceive that authorities are in agreement that it does.

According to the representativeness heuristic, discussed above, people may see past return history as relevant to the future only if they see the present circumstances as representative in some details of widely remembered past periods. Thus, for example, the public appears to have made much, just before the stock market crash of 1987, of similarities in that period to the period just before the crash of 1929. Newspapers, including the *Wall Street Journal* on the morning of the stock market crash of October 19, 1987, showed plots of stock prices before October 1929 superimposed on a plot of stock prices before October 1987, suggesting comparisons. In this way, historical events can be remembered and viewed as relevant, but this is not any systematic analysis of past data.

Lack of learning from historical lessons regarding financial and economic uncertainties may explain why many investors show little real interest in diversification around the world and why most investors appear totally uninterested in the correlation of their investments with their labor income, violating with their behavior one of the most fundamental premises of financial theory. Most people do not make true diversification around the world a high priority, and virtually no one is short the company that he or she works for, or is short the

[20] This feeling can of course be disrupted, if a sudden event calls to mind parallels to a past event, or if the social cognition memorializes and interprets a past event as likely to be repeated.

stock market in one's own country, as would be suggested by economic theory.[21]

A prominent reason that most people appear apathetic about schemes to protect them from price level uncertainty in nominal contracts is that they just do not seem to think that past actual price level movements are any indicator of future uncertainty. In a questionnaire I distributed (1997a) to a random sample from phone books in the U.S.A. and Turkey, the following question was posed:

> We want to know how accurately you think that financial experts in America (Turkey) can predict the price level in 2006, ten years from now. Can you tell us, if these experts think that a "market basket" of goods and services that the typical person buys will cost $1,000 (100 million TL) in 2006, then you think it will probably actually cost:
>
> (Please fill in your lower and upper bounds on the price:)
>
> Between $__________ (TL) and $__________ (TL)

The median ratio between high and low was 4/3 for U.S. respondents and 3/2 for Turkish respondents. Only a few respondents wrote numbers implying double- or triple-digit ratios, even in Turkey. The ratios not far from one that most respondents revealed would seem to suggest excessive confidence in the predictability of price levels. Note that in Turkey the CPI increased three-fold between 1964 and 1974, 31-fold between 1974 and 1984, and 128-fold between 1984 and 1994. But, Turkish respondents appear to connect the price level movements with prior political and social events that may be perceived as having largely predicted the price movements, events that are themselves not likely to be repeated in the same way. While these people have apparently learned to take certain steps to protect themselves from price level uncertainty (such as not investing in long-term nominal bonds), they do not appear to have a well-developed understanding of the potential uncertainty of the Turkish Lira that would allow them to deal systematically with such uncertainty. For example, they have shown relatively little interest in government indexed bonds.

Magical Thinking

B. F. Skinner (1948) in what is now regarded as a classic experiment fed starved experimental pigeons small quantities of food at regular fifteen-second intervals with no dependence whatsoever on the bird's behavior. Even though the feeding was unaffected by their behavior, the birds began to behave as if they had a "superstition" that something in their behavior caused the feeding (see also McFadden, 1974). Each pigeon apparently conditioned itself to exhibit a specific behavior to get the food, and because each bird

[21] Kusko, Poterba and Wilcox (1997) showed, using data on 10,000 401k plan participants in a manufacturing firm, found that barely 20% of participants directed *any* of their own balances into an S&P index fund, while nearly 25% of participants directed *all* of their discretionary balances into a fund invested completely in the own company stock.

exhibited its characteristic behavior so reliably, it was never deconditioned:

> One bird was conditioned to turn counter-clockwise in the cage, making two or three turns between reinforcements. Another repeatedly thrust its head into one of the upper corners of the cage. A third developed a " tossing" response, as if placing its head beneath an invisible bar and lifting it repeatedly.... (1948, p. 168)

Arbitrary behaviors that are so generated are referred to with the term "magical thinking" by psychologists.

A wide variety of economic behaviors are likely to be generated in exactly the same way that the arbitrary behaviors of the pigeons are generated. Thus, for example, firms' investment or management decisions that happened to precede increases in sales or profits may tend to be repeated, and if this happens in a period of rising profits (as when the economy is recovering from a recession) the notion that these decisions were the cause of the sales or profit increase will be reinforced. Because firms are similar to each other and observe each other, the magical thinking may be social, rather than individual, and hence may have aggregate effects.

Roll (1986), with his hubris hypothesis concerning corporate takeovers, argued that managers of bidder firms may become overconfident of their own abilities to judge firms, because of their luck in their first takeovers. This overconfidence can cause them to overbid in subsequent takeover attempts.

The tendency for speculative markets to respond to certain news variables may be generated analogously. The U.S. stock market used often to be buoyed by positive news about the economy, but in recent years it appears to tend to be moved in the opposite direction by such news. This new "perverse" movement pattern for the stock market is sometimes justified in the media by a theory that the good news will cause the Federal Reserve to tighten monetary policy and that then the higher interest rates will lower the stock market. But the whole belief could be the result of a chain of events that was set off by some initial chance movements of the stock market. Because people believe these theories they may then behave so that the stock price does indeed behave as hypothesized, the initial correlations will persist later, and thereby reinforce the belief.

Quasi-Magical Thinking

The term quasi-magical thinking, as defined by Shafir and Tversky (1992), is used to describe situations in which people act as if they erroneously believe that their actions can influence an outcome (as with magical thinking) but in which they in fact do not believe this. It includes acting as if one thinks that one can take actions that will, in effect, undo what is obviously predetermined, or that one can change history.

For example, Quattrone and Tversky (1984) divided subjects into a control and experimental group and then asked people in both groups to see how long they could bear to hold their hands in some ice water. In the experimental group subjects were told that people with strong hearts were better able to endure the ice water. They found that those in the experimental group in fact held their hands in the ice water longer. If indeed, as appears to be the case, those in the experimental group held their hands in the ice water longer to prove that they had strong hearts, then this would be quasi-magical, since no notion was involved that there was any causal link from holding hands in ice water to strengthening the heart.

While this particular experimental outcome might also be explained as the result of a desire for self deception, Shafir and Tversky report as well as other experiments that suggest that people do behave as if they think they can change predetermined conditions. Shafir and Tversky (1992) show, with an experimental variant of Newcomb's Paradox, that people behave as if they can influence the amount of money already placed in a box.

Quasi-magical thinking appears to operate more strongly when outcomes of future events, rather than historical events, are involved. Langer (1975) showed that people place larger bets if invited to bet before a coin is tossed than after (where the outcome has been concealed), as if they think that they can better influence a coin not yet tossed.

It appears likely that such quasi-magical thinking explains certain economic phenomena that would be difficult to explain the basis of strictly rational behavior. Such thinking may explain why people vote, and why shareholders exercise their proxies. In most elections, people must know that the probability that they will decide the election must be astronomically small, and they would thus rationally decide not to vote. Quasi-magical thinking, thinking that in good societies people vote and so if I vote I can increase the likelihood that we have a good society or a good company, might explain such voting. The ability of labor union members or oligopolists to act in concert with their counterparts, despite an incentive to free-ride, or defect, may also be explained by quasi-magical thinking.

The disposition effect (Shefrin and Statman, 1985) referred to above, the tendency for individuals to want to hold losers and sell winners might also be related to quasi-magical thinking, if people feel at some level that holding on to losers can reverse the fact that they have already lost. Public demand for stocks at a time when they are apparently overvalued may be influenced by quasi-magical thinking, a notion that if I hold, then the stocks will continue to rise.

Attention Anomalies and the Availability Heuristic

William James (1890, p. 402) criticized earlier psychologists, who in their theories effectively assumed that the human mind takes account of all sensory input, for taking no note of the phenomenon of selective attention:

> But the moment one thinks of the matter, one sees how false a notion of experience that is which would make it tantamount to the mere presence to the senses of an outward order. Millions of items of the outward order are present to my senses which never properly enter into my experience. Why? Because they have no *interest* for me. *My experience is what I agree to*

> *attend to*. Only those items which I *notice* shape my mind — without selective interest, experience is utter chaos.

The same criticism might equally well be applied to expected utility maximization models in economics, for assuming that people attend to all facts that are necessary for maximization of the assumed objective function (Berger, 1994, elaborates on this point).

Attention is associated with language; the structure of our language invites attention to categories that are represented in the language. Taylor (1989) showed, for example, that certain concepts of "the self" were apparently absent from languages in the time of Augustine. The language shapes our attention to even the most inward of phenomena.

In economics, certain terms were apparently virtually absent from popular discourse fifty or more years ago: gross national product, the money supply, the consumer price index. Now, many economists are wont to model individual attention to these concepts as if they were part of the external reality that is manifest to all normal minds.

Attention may be capricious because it is affected by the "salience" of the object; whether it is easily discerned or not (Taylor and Thompson, 1982) or by the "vividness" of the presentation, whether the presentation has colorful details. Judgments may be affected, according to the "availability heuristic," that is, by the "ease with which instances or associations come to mind" (Tversky and Kahneman, 1974).

Investment fashions and fads, and the resulting volatility of speculative asset prices, appear to be related to the capriciousness of public attention (Shiller, 1984, 1987). Investor attention to categories of investments (stocks versus bonds or real estate, investing abroad versus investing at home) seems to be affected by alternating waves of public attention or inattention. Investor attention to the market at all seems to vary through time, and major crashes in financial markets appear to be phenomena of attention, in which an inordinate amount of public attention is suddenly focussed on the markets.[22]

Economic theories that are most successful are those that take proper account of the limitations and capriciousness of attention. One reason that the hypothesis of no unexploited arbitrage opportunities (a hypothesis that has led to the Black–Scholes (1973) option pricing theory, the Ross (1976) arbitrage pricing theory, and other constructs of finance) has been so successful is that it does not rely on pervasive public attention. The essence of the no-arbitrage assumption, when it is used successfully to produce theories in finance, is that the arbitrage opportunities, were they to ever exist, would be exploited and eliminated even if only a tiny fraction of investors were paying attention to the opportunity.

Culture and Social Contagion

The concept of culture, central to sociology and cultural anthropology ever since the work of Tylor (1871), Durkheim (1893) and Weber (1947), is related to the selective attention that the human mind exhibits. There is a social cognition, reenforced by conversation, ritual and

[22] There is evidence that the stock market crash of 1987 can be viewed in these terms, see Shiller (1989).

symbols, that is unique to each interconnected group of people; to each nation, tribe, or social group. People tend not to remember well facts or ideas that are not given attention in the social cognition, even though a few people may be aware of such facts. If one speaks to groups of people about ideas that are foreign to their culture, one may find that someone in the group will know of the ideas, and yet the ideas have no currency in the group and hence have no influence on their behavior at large.

The array of facts, suppositions, symbols, categories of thought that represent a culture have subtle and far-reaching affects on human behavior. For a classic example, Durkheim (1897), in a careful study of differing suicide rates across countries, found that there was no apparent explanation for these differing rates other than cultural differences.

Cultural anthropologists have used methods of inferring elements of primitive culture by immersing themselves in the society, observing their everyday life, and talking and listening to them nonjudgmentally, letting them direct the conversation. From such learning, for example, Lévy–Strauss (1966, pp. 9–10) wrote persuasively that the customs of primitive people that we may tend to view as inexplicably savage actually arise as a logical consequence of a belief system common to all who belong to the society, a belief system which we can grow to understand only with great difficulty:

> The real question is not whether the touch of a woodpecker's beak does in fact cure toothache. It is rather whether there is a point of view from which a woodpecker's beak and a man's tooth can be seen as 'going together' (the use of this congruity for therapeutic purposes being only one of its possible uses) and whether some initial order can be introduced into the universe by means of these groupings.... The thought we call primitive is founded on this demand for order.

The same methods that cultural anthropologists use to study primitive peoples can also be used to study modern cultures. O'Barr and Conley (1992) studied pension fund managers using personal interviews and cultural anthropological methods. They concluded that each pension fund has its own culture, associated often with a colorful story of the origin of their own organization, akin to the creation myths of primitive peoples. The culture of the pension fund is a belief system about investing strategy and that culture actually drives investment decisions. Cultural factors were found to have great influence because of a widespread desire to displace responsibility for decisions onto the organization, and because of a desire to maintain personal relationships within the organization.[23]

Psychological research that delineates the factors that go into the formation of culture has been undertaken under the rubric of social psychology and attitude change, or under social cognition. There is indeed an enormous volume of research in these areas. For surveys, one may refer to McGuire (1985) for attitude change or Levine and Resnick (1993)

[23] The psychologist Janis (1972) has documented with case studies how social patterns ("groupthink") within decision making groups can cause even highly intelligent people to make disastrously wrong decisions.

for social cognition.

One difficulty that these researchers have encountered with experimental work is that of disentangling the "rational" reasons for the imitation of others with the purely psychological. Some recent economic literature has indeed shown the subtlety of the informational influences on people's behavior (learning from each other), see Bannerjee (1992), Bikhchandani et al. (1992), Leahy (1994), and Shiller (1995).

A Global Culture

We see many examples of imitation across countries apparently widely separated by both physical and language barriers. Fashions of dress, music, and youthful rebellion, are obvious examples. The convergence of seemingly arbitrary fashions across nations is evidence that something more is at work in producing internationally-similar human behavior than just rational reactions to common information sets relevant to economic fundamentals, see Featherstone (1990).

And yet it will not be an easy matter for us to decide in what avenues global culture exerts its influence (Hannerz, 1990, p. 237):

> There is now a world culture, but we had better make sure that we understand what this means. It is marked by an organization of diversity rather than by a replication of uniformity. No total homogenization of systems of meaning and expression has occurred, nor does it appear likely that there will be one any time soon. But the world has become one network of social relationships, and between its different regions there is a flow of meanings as well as of people and goods.

Sociologists have made it their business to study patterns of influence within cultures, and we ought to be able to learn something about the nature of global culture from their endeavors. For example, one study of patterns of influence regarded as a classic among sociologists is the in-depth study of the town of Rovere by sociologist Robert Merton (1957). After extensive study of the nature of interpersonal influence, he sought meaningful ways to categorize people. He found that it was meaningful to divide people into two broad categories: locals (who follow local news and derive status by their connectedness with others) and cosmopolitans (who orient themselves instead to world news and derive status from without the community). He found that the influence of cosmopolitans on locals transcended both their numbers and their stock of useful information. We must bear this conclusion in mind when deciding how likely it is that incipient cultural trends are pervasive across many different nations.

Reading such sociological studies inclines us to rather different interpretations of globally similar behaviors than might occur naturally to many traditional economists. Why did the real estate markets in many cities around the world rise together into the late 1980s and fall in the early 1990s? (See Goetzmann and Wachter, 1996 and Hendershott, 1997.) Why have the stock markets of the world moved somewhat together? Why did the stock markets of the world show greater tendency to move together after the stock market crash of 1987? (See von Furstenberg and Jeon, 1989 and King, Sentana and Wadhwani, 1994.) If we recognize the global nature of culture, there is no reason to assume that these events have anything to do with genuine information about economic fundamentals.

Concluding Remarks

Since this paper was written in response to an invitation to summarize literature on behavioral theory in finance, it has focussed exclusively on this topic, neglecting the bulk of finance literature. Because of its focus on anomalies and departures from conventional notions of rationality, I worry that the reader of this paper can get a mistaken impression about the place of behavioral theory in finance, and of the importance of conventional theory.

The lesson from the literature surveyed here, and the list of varied behavioral phenomena, is not that "anything can happen" in financial markets. Indeed, while the behavioral theories have much latitude for interpretation, when they are combined with observations about behavior in financial markets, they allow us to develop theories that do have some restrictive implications. Moreover, conventional efficient markets theory is not completely out the window. I could have, had that been the goal of this paper, found very many papers that suggest that markets are impressively efficient in certain respects.

Financial anomalies that intuitive assessments of human nature might lead one to expect to find, or anomalies one hears casually about, often turn out to be tiny, ephemeral, or nonexistent. There is, for example, virtually no Friday the thirteenth effect (Chamberlain et al., 1991; Dyl and Maberly, 1988). Investors apparently aren't that foolish.

Heeding the lessons of the behavioral research surveyed here is not going to be simple and easy for financial researchers. Doing research that is sensitive to lessons from behavioral research does not mean entirely abandoning research in the conventional expected utility framework. The expected utility framework can be a workhorse for some sensible research, if it is used appropriately. It can also be a starting point, a point of comparison from which to frame other theories.

It is critically important for research to maintain an appropriate perspective about human behavior and an awareness of its complexity. When one does produce a model, in whatever tradition, one should do so with a sense of the limits of the model, the reasonableness of its approximations, and the sensibility of its proposed applications.

References

Allais, M. (1953). "Le Comportement de l'Homme Rationnel devant le Risque, Critique des Postulats et Axiomes de l'Ecole Americaine," *Econometrica*, 21:503–546.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM–IV).* Washington, DC: American Psychiatric Association.

Bachelier, L.(1964). "Theory of Speculation." In Paul Cootner (ed.), *The Random Character of Stock Market Prices*. Cambridge, MA: MIT Press, pp. 17–75. (Originally submitted as doctoral dissertation to the Faculty of Sciences of the Academy of Paris, 1900.)

Backus, D. K., S. Foresi and C. I. Telmer (1995). "Interpreting the Forward Premium Anomaly," *Canadian Journal of Economics*, 28: S108–119.

Ball, R. and P. Brown (1968). "AnEmpirical Examination of Accounting Income Numbers," *Journal of Accounting Research*, 6: 159–178.

Bannerjee, A. V. (1992). "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, 107(3): 797–817.

Barberis, N., A. Shleifer and R. Vishny (1997). "A Model of Investor Sentiment," reproduced, University of Chicago, presented at the NBER–Sage workshop on Behavioral Economics, Cambridge, MA.

Bates, D. S. (1991). "The Crash of '87: Was It Expected? The Evidence from Options Markets," *Journal of Finance*, 46(3): 1009–1044.

Bates, D. S. (1995). "Post-`87 Crash Fears in S&P Futures Options," reproduced, Wharton School, University of Pennsylvania.

Bell, D. E. (1982). "Regret in Decision Making Under Uncertainty," *Operations Research*, 30(5): 961–981.

Benartzi, S. and R. H. Thaler (1995). "Myopic Loss Aversion and the Equity Premium Puzzle." *Quarterly Journal of Economics*, 110(1): 73–92.

Benartzi, S. and R. H. Thaler (1996). "Risk Aversion or Myopia: The Fallacy of Small Numbers and its Implications for Retirement Saving," reproduced.

Berger, L. A. (1994). "Mutual Understanding, The State of Attention, and the Ground for Interaction in Economic Systems," *Business and Ethics Quarterly*.

Bernard, V. L. and J. K. Thomas (1992). "Evidence that Stock Prices Do Not Fully Reflect the Implications of Current Earnings for Future Earnings," *Journal of Accounting Economics*, 13: 305–340.

Bernard, V. (1992). "Stock Price Reactions to Earnings Announcements." In R. Thaler (ed.), *Advances in Behavioral Finance*. New York: Russell Sage Foundation.

Bikhchandani, S., D. Hirshleifer and I. Welch (1992). "A Theory of Fashion, Social Custom, and Cultural Change," *Journal of Political Economy*, 100(5): 992–1026.

Black, F. and M. Scholes (1973). "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81: 637–654.

Bodnar, G. and R. Marston (1996). "1995 Survey of Derivatives Usage by US Non-Financial Firms," reproduced, Wharton School, University of Pennsylvania.

Bolen, D. W. and W. H. Boyd (1968). "Gambling and the Gambler: A Review and Preliminary Findings," *Archives of General Psychiatry*, 18(5): 617–29.

Bowman, D., D. Minehart and M. Rabin (1993). "Loss Aversion in a Savings Model," University of California Working Paper in Economics 93–12.

Bryant, R. R. (1990). "Job Search and Information Processing in the Presence of Nonrational Behavior," *Journal of Economic Behavior and Organization*, 14(2): 249–260.

Campbell, J. Y., A. W. Lo and A. C. MacKinlay (1997). *The Econometrics of Financial Markets*. Princeton, NJ: Princeton University Press.

Campbell, J. Y. and R. J. Shiller (1988). "Stock Prices, Earnings, and Expected Dividends," *Journal of Finance*, 43: 661–676.

Campbell, J. Y. and R. J. Shiller (1989). "The Dividend Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, 1: 195–228.

Campbell, J. Y. and R. J. Shiller (1998). "Valuation Ratios and the Long-Term Stock Market Outlook," forthcoming, *Journal of Portfolio Management*.

Case, K. E. and R. J. Shiller (1988). "The Behavior of Home Buyers in Boom and Post-Boom Markets," *New England Economic Review* (Nov./Dec.): 29–46.

Chamberlain, T. W., C. S. Cheung, and C. C. Y. Chang (1991). "The Friday the Thirteenth Effect," *Quarterly Journal of Business and Economics*, 30(2): 111–117.

Chan, L., N. Jegadeesh, and J. Lakonishok (1996). "Momentum Strategies," NBER Working Paper 5375, forthcoming, *Journal of Finance*.

Cootner, P. H. (1964). *The Random Character of Stock Market Prices*. Cambridge, MA: MIT Press.

Cutler, D. M., J. M.Poterba, and L. H. Summers (1989). "What Moves Stock Prices?" *Journal of Portfolio Management*, 15(3): 4–12.

Cutler, D. M., J. M.Poterba, and L. H. Summers (1991). "Speculative Dynamics," *Review of Economic Studies*, 58(3): 529–546.

Daniel, K., D. Hirshleifer, and A. Subrahmanyam (1997). "A Theory of Overconfidence, Self-Attribution, and Security Market Over- and Underreaction," unpublished paper, presented at the NBER–Sage workshop on Behavioral Economics, Cambridge, MA.

De Bondt, W. F. and M. M. Bange (1992). "Inflation Forecast Errors and Time Variation in Term Premia," *Journal of Finance and Quantitative Analysis*, 24: 479–496.

De Bondt, W. F. and R. H. Thaler (1985). "Does the Stock Market Overreact?" *Journal of Finance*, 40: 793–805.

De Bondt, W. and R. H. Thaler (1996). "Financial Decision-Making in Markets and Firms: A Behavioral Perspective." In *Handbook in Operations Research and Management Science*, Vol. 9 North–Holland.

De Long, J. B., A. Shleifer, L. Summers, and R. J. Waldman (1990a). "Noise Trader Risk in Financial Markets," *Journal of Political Economy*, 98: 703–38.

De Long, J. B., A. Shleifer, L. Summers, and R. J. Waldman (1990b). "Positive Feedback Investment Strategies and Destabilizing Rational Speculation," *Journal of Finance*, 45(2): 379–395.

Degeorge, F., J. Patel, and R. Zeckhauser (1997). "Earnings Manipulations to Exceed Thresholds," unpublished paper, HEC School of Management, presented at the NBER–Sage Workshop on Behavioral Finance, Cambridge, MA.

Dolde, W. (1993). "The Trajectory of Corporate Financial Risk Management," *Journal of Applied Corporate Finance*, 6:33–41.

Dreman, D. (1977). *Psychology and the Stock Market: Why the Pros Go Wrong and How to Profit*. New York: Warner Books.

Dreman, D. and M. Berry (1977). "Investor Overreaction and the Low P/E Effect," unpublished paper, Dreman Foundation, presented at the NBER–Sage Workshop in Behavioral Finance, Cambridge, MA.

Durkheim, É. (1893). "Représentations individuelles et représentations collectives," *Revue de Métaphysique et de Morale*, 6: 273–302.

Durkheim, É. (1897). *Le Suicide*.

Dyl, E. A. and E. D. Maberly (1988). "The Anomaly That Isn't There: A Comment on Friday the Thirteenth," *Journal of Finance*, 43(5): 1285–1286.

Edwards, W. (1968). "Conservatism in Human Information Processing." In B. Kleinmutz (ed.), *Formal Representation of Human Judgment*. New York: John Wiley & Sons.

Eisner, R. and R. H. Strotz (1961). "Flight Insurance and the Theory of Choice," *Journal of Political Economy*, 69: 355–368.

Erlich, D., P. Guttman, P. Schopenbach, and J. Mills (1957). "Postdecision Exposure to Relevant Information," *Journal of Abnormal and Social Psychology*, 54: 98–102.

Fama, E. F. (1970). "Efficient Capital Markets: A Review of Empirical Work," *Journal of Finance*, 25: 383–417.

Fama, E. F. (1991). "Efficient Capital Markets II," *Journal of Finance*, 46(5): 1575–1617.

Fama, E.F. (1992). "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47: 427–465.

Fama, E. F. (1997). "Market Efficiency, Long-Term Returns, and Behavioral Finance," Center for Research in Security Prices Working Paper 448, University of Chicago.

Fama, E. F. and K. R. French (1988). "Permanent and Temporary Components of Stock Returns," *Journal of Political Economy*, 96: 246–273.

Featherstone, M. (1990). *Global Culture: Nationalism, Globalization and Modernity*. London: Sage Publications.

Ferris, S. P., R. A. Haugen and A. K. Makhija (1988). "Predicting Contemporary Volume with Historic Volume at Differential Price Levels: Evidence Supporting the Disposition Effect," *Journal of Finance*, 43(3): 677–697.

Festinger, L. (1957). A Theory of Cognitive Dissonance., Stanford, CA: Stanford University Press.

Fisher, I. (1928). *The Money Illusion*. New York: Adelphi.

Fisher, I. (1930). *The Theory of Interest*. New York: MacMillan.

Fischhoff, B. (1975). "Hindsight is not Foresight: The Effect of Outcome Knowledge on Judgment Under Uncertainty," *Journal of Experimental Psychology: Human Perception and Performance*, 1: 288–299.

Fischhoff, B., P. Slovic and S. Lichtenstein (1977). "Knowing With Uncertainty: The Appropriateness of Extreme Confidence," *Journal of Experimental Psychology: Human Perception and Performance*, 3: 552–564.

Florovsky, G. (1969). "The Study of the Past." In R. H. Nash (ed.), *Ideas of History*, Vol. 2. New York: Dutton.

Fortune, P. (1996). "Anomalies in Option Pricing: The Black–Scholes Model Revisited," *New England Economic Review*, March/April, 17–40.

Friedman, M. and L. J. Savage (1948). "The Utility Analysis of Choices Involving Risk," *Journal of Political Economy*, 56: 279–304.

Froot, K. and R. Thaler (1990). "Anomalies: Foreign Exchange," *Journal of Economic Perspectives*, 4(3): 179–192.

Garber, P. (1990). "Famous First Bubbles," *Journal of Economic Perspectives*, 42(2): 35–54.

Geanakoplos, J. (1992). "Common Knowledge," *Journal of Economic Perspectives*, 6(4): 53–82.

Gertler, M. (1994). "Monetary Policy, Business Cycles, and the Behavior of Small Manufacturing Firms," *Quarterly Journal of Economics*, 109(2): 309–340.

Gibson, G. R. (1889). *The Stock Markets of London, Paris and New York*. New York: G. P. Putnam's Sons.

Gigerenzer, G. (1991). "How to Make Cognitive Illusion Disappear: Beyond 'Heuristics and Biases'," *European Review of Social Psychology*, 2: 83–115.

Goetzmann, W. N. and P. Jorion (1993). "Testing the Predictive Power of Dividend Yields," *Journal of Finance*,48: 63.

Goetzmann, W. N. and N. Peles (1993). "Cognitive Dissonance and Mutual Fund Investors," reproduced, Yale School of Management.

Goetzmann, W. N. and S. M. Wachter (1996). "The Global Real Estate Crash: Evidence from an International Database," reproduced, Yale University.

Graham, B. and D. L. Dodd (1934). *Security Analysis*. New York: McGraw Hill.

Gruen, D. K. and M. C. Gizycki (1993). "Explaining Forward Discount Bias: Is It Anchoring?" Princeton University Woodrow Wilson School Discussion Paper in Economics 164.

Gultekin, M. and N. B. Gultekin (1983). "Stock Market Seasonality: International Evidence," *Journal of Financial Economics*, 12: 469–481.

Hannerz, U. (1990). "Cosmopolitans and Locals in World Culture," *Theory, Culture and Society*, 7: 237–51.

Harawini, G., and D. B. Keim (1995). "On the Predictability of Common Stock Returns: Worldwide Evidence," In R. Jarrow, M. Maksiomovic and W. T. Ziemba (eds.), *Finance*, Handbooks in Operations Research and Management Science. Amsterdam: North–Holland, Volume 9.

Hendershott, P. H. (1997). "Systematic Valuation Errors and Property Cycles: A Clinical Study of the Sydney Office Market," reproduced, Ohio State University.

Ibbotson, R. and J. R. Ritter (1988). "Initial Public Offerings," *Journal of Applied Corporate Finance*, 1: 37–45.

Jackwerth, J. C. and M. Rubinstein (1995). "Recovering Probability Distributions from Contemporaneous Security Prices," reproduced, Haas School of Business, University of California, Berkeley.

James, W. (1890). *Principles of Psychology*. New York: Dover Publications (reprinted 1950).

Janis, I. (1972). *Victims of Groupthink*. Houston: Boston.

Jegadeesh, N. and S. Titman (1993). "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance*, 48(1): 65–91.

Kahneman, D. and A. Tversky (1979). "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, 47: 263–291.

Kahneman, D., P. Slovic, and A. Tversky (1974). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge, England: Cambridge University Press.

Kallick, M., D. Suits, T. Dielman, and J. Hybels (1975). *A Survey of American Gambling Attitudes and Behavior*. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.

King, M., E. Sentana, and S. Wadhwani (1994). "Volatility and Links Between National Stock Markets," *Econometrica*, 62(4): 901–934.

Kirby, C. (1997). "Measuring the Predictable Variation in Stock and Bond Returns," *Review of Financial Studies*, 10: 579–630.

Kusko, A. L., J. M. Poterba, and D. Wilcox (1997). "Employee Decision with Respect to 401(k) Plans: Evidence from Individual-Level Data." Forthcoming in O. S. Mitchell and S. J. Schieber (eds.), *Living with Defined Contribution Pension Plans: Remaking Responsibility for Retirement*. Philadelphia: University of Pennsylvania Press.

Lakonishok, J., A. Shleifer, and R. W. Vishny (1994). "Contrarian Investment, Extrapolation and Risk," *The Journal of Finance*, 49(5): 1541–1578.

Langer, E. J. (1975). "The Illusion of Control," *Journal of Personality and Social Psychology*, 32: 311–328.

La Porta, R. (1996). "Expectations and the Cross-Section of Stock Returns," *Journal of Finance*, 51: 1715–1742.

Leahy, J. (1994). "Miracle on Sixth Avenue," reproduced, Harvard University.

Lee, C., A. Shleifer, and R. Thaler (1991). "Investor Sentiment and the Closed-End Fund Puzzle," *Journal of Finance*, 46: 75–109.

Lehmann, B. N. (1990). "Fads, Martingales and Market Efficiency," *Quarterly Journal of Economics*, 105(1): 1–28.

Lehmann, B. N. (1991). "Asset Pricing and Intrinsic Values: A Review Essay," *Journal of Monetary Economics*, 28: 485–500.

LeRoy, S. F., and R. D. Porter (1981). "Stock Price Volatility: A Test Based on Implied Variance Bounds," *Econometrica*, 49: 97–113.

Levine, J. and L. B. Resnick (1993). "Social Foundations of Cognition," *Annual Review of Psychology*, 44: 585–612.

Lévy–Strauss, C. (1966). *The Savage Mind*. Chicago, IL: The University of Chicago Press.

Lo, A. W. and A. C. MacKinlay (1988). "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test," *Review of Financial Studies*, 1: 41–66.

Lo, A. W. and A. C. MacKinlay (1990). "When Are Contrarian Profits Due to Stock Market Overreaction?" *Review of Financial Studies*, 3: 175–208.

Loomes, G. and R. Sugden (1982). "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty," *The Economic Journal*, 92: 805–824.

Mackay, D. (1841). *Memoirs of Extraordinary Popular Delusions*. London: Bentley.

Malkiel, B. (1981). *A Random Walk Down Wall Street*, 2nd ed. New York: Norton.

Mayhew, S. (1995). "Implied Volatility," *Financial Analysts' Journal*, 51(4): 8–20.

McFadden, D. (1974). "On Some Facets of Betting." In M. Balch et al. (eds.), *Essays on Economic Behavior Under Uncertainty*. North–Holland, pp. 99–122.

McGuire, W. J. (1985). "Attitudes and Attitude Change." In G. Lindzey and E. Aronson (eds.), *Handbook of Social Psychology*. Reading, MA: Addison Wesley, pp.233–346.

Meehl, P. and A. Rosen (1955). "Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, and Cutting Scores," *Psychological Bulletin*, 52: 194–215.

Mehra, R. and E. C. Prescott (1985). "The Equity Premium: A Puzzle," *Journal of Monetary Economics*, 15: 145–162.

Merton, R. K. (1957). *Social Theory and Social Structure*. Glencoe, IL: Free Press.

Milgrom, P. and N. Stokey (1982). "Information, Trade, and Common Knowledge," *Econometrica*, 49: 219–222.

Modigliani, F. and R. Cohn (1979). "Inflation, Rational Valuation, and the Market," *Financial Analysts Journal*, 35: 24–44.

Nance, D. R., C. W. Smith, and C. W. Smithson (1993). "On the Determinants of Corporate Hedging," *Journal of Finance*, 48: 267–284.

Nelson, C. and M. Kim (1993). "Predictable Stock Returns: The Role of Small Sample Bias," *Journal of Finance*, 48(2): 641–661.

Northcraft, G. B. and M. A. Neale (1987). "Experts, Amateurs, and Real Estate: An Anchoring-and-Adjustment Perspective on Property Pricing Decisions," *Organizational Behavior and Human Decision Processes*, 39: 84–97.

O'Barr, W. M. and J. M. Conley (1992). *Fortune and Folly: The Wealth and Power of Institutional Investing*. Homewood, IL: Irwin.

Odean, T. (1996a). "Are Investors Reluctant to Realize Their Losses?" unpublished paper, University of California, Berkeley.

Odean, T. (1996b). "Why Do Investors Trade Too Much?" unpublished paper, University of California, Berkeley.

Poterba, J. M. and L. H. Summers (1988). "Mean Reversion in Stock Prices: Evidence and Implications," *Journal of Financial Economics*, 22(1): 27–59.

Quattrone, G. A. and A. Tversky (1984). "Causal versus Diagnostic Contingencies: On Self Deception and on the Voter's Illusion," *Journal of Personality and Social Psychology*, 46(2): 237–248.

Rabin, M. (1996). "Psychology and Economics," *Journal of Economic Literature*.

Ritter, Jay R. (1991). "The Long-Run Performance of Initial Public Offerings," *Journal of Finance*, 46: 3–28.

Roberts, H. V. (1967). "Statistical versus Clinical Prediction of the Stock Market," unpublished paper presented to the Seminar on the Analysis of Security Prices, University of Chicago.

Roll, R. (1986). "The Hubris Hypothesis of Corporate Takeovers," *Journal of Business*, 59(2): 541–566.

Ross, L. (1987). "The Problem of Construal in Social Inference and Social Psychology." In N. Grunberg, R. E. Nisbett and J. Singer (eds.), *A Distinctive Approach to Psychological Research: The Influence of Stanley Schachter*. Hillsdale, NJ: Erlbaum.

Ross, S. A. (1976). "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13: 341–360.

Samuelson, P. A. (1963). "Risk and Uncertainty: A Fallacy of Large Numbers," *Scientia*, 98(4–5): 108–113.

Savage, L. J. (1954). "The Sure-Thing Principle." In Leonard J. Savage, *The Foundations of Statistics*. New York: John Wiley, pp. 21–26.

Shafir, E., P. Diamond and A. Tversky (1997). "On Money Illusion," *Quarterly Journal of Economics*, 92: 341–374.

Shafir, E. and A. Tversky (1992). "Thinking Through Uncertainty: Nonconsequential Reasoning and Choice," *Cognitive Psychology*, 24: 449–474.

Shefrin, H. (1997). "Erroneous Investor Beliefs: Implications for the Term Structure of Interest Rates,' unpublished paper, Santa Clara University.

Shefrin, H. and M. Statman (1985). "The Disposition to Sell Winners Too Early and Ride Losers Too Long," *Journal of Finance*, 40: 777–790.

Shefrin, H. and M. Statman (1994). "Behavioral Portfolio Theory," unpublished paper, Santa Clara University.

Shefrin, H. and R. H. Thaler (1988). "The Behavioral Life-Cycle Hypothesis," *Economic Inquiry*. Reprinted in R. H. Thaler, *Quasi Rational Economics*.

Shiller, R. J. (1979). "The Volatility of Long Term Interest Rates and Expectations Models of the Term Structure," *Journal of Political Economy*, 87: 1190–1219.

Shiller, R. J. (1981a). "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?" *American Economic Review*, 71(3): 421–436.

Shiller, R. J. (1981b). "The Use of Volatility Measures in Assessing Market Efficiency," *Journal of Finance*, 36: 291-304.

Shiller, R. J. (1984). "Stock Prices and Social Dynamics," *Brookings Papers on Economic Activity*, II, pp. 457–98.

Shiller, R. J. (1987a). "Fashions, Fads and Bubbles in Financial Markets." In Jack Coffee (ed.), *Knights, Raiders and Targets: The Impact of the Hostile Takeover*. Oxford, England: Oxford University Press.

Shiller, R. J. (1987b). "Investor Behavior in the October 1987 Stock Market Crash: Survey Evidence," National Bureau of Economic Research Working Paper 2446. (Reprinted in Robert Shiller, *Market Volatility*, 1989.)

Shiller, R. J. (1989). *Market Volatility*. Cambridge, MA: MIT Press.

Shiller, R. J. (1990). "Market Volatility and Investor Behavior," *American Economic Review*, 80(2): 58–62.

Shiller, R. J. (1995). "Conversation, Information, and Herd Behavior," *American Economic Review*, 85(2): 181–185.

Shiller, R. J. (1997a). "Public Resistance to Indexation: A Puzzle," *Brookings Papers on Economic Activity*, I, pp. 159–228.

Shiller, R. J. (1997b). "Why Do People Dislike Inflation?" In C. Romer and D. Romer (eds.), *Reducing Inflation: Motivation and Strategy*. National Bureau of Economic Research and University of Chicago Press, pp. 13–65.

Shiller, R. J., F. Kon-Ya, Y. Tsutsui (1996). "Why Did the Nikkei Crash? Expanding the Scope of Expectations Data Collection," *Review of Economics and Statistics*, 78: 156–164.

Shleifer, A. (1986). "Do Demand Curves for Stocks Slope Down?" *Journal of Finance*, 41: 579–589.

Shleifer, A. and L. Summers (1990). "The Noise Trader Approach to Finance," *Journal of Econometrics*, 4(2): 19–23.

Shleifer, A. and R. Vishny (1996). "The Limits of Arbitrage," NBER Working Paper 5167, forthcoming, *Journal of Finance*.

Siegel, J. J. (1994). *Stocks for the Long Run*. New York: Irwin.

Siegel, J. J. and R. H. Thaler (1997). "Anomalies: The Equity Premium Puzzle," *Journal of Economic Perspectives*, 11(1): 191–200.

Silberman, J. and M. Klock (1989). "The Behavior of Respondents in Contingent Valuation: Evidence on Starting Bids," *Journal of Behavioral Economics*, 18: 51–60.

Skinner, B. F. (1948). "Superstition in the Pigeon," *Journal of Experimental Psychology*, 38:168–72. (Reprinted in same journal (1992), 121(3):273–274.)

Slovic, P. (1972). "Psychological Study of Human Judgment: Implications for Investment Decision Making Under Uncertainty," *Journal of Finance*, 27(4): 779–799.

Smith, E. L. (1925). *Common Stocks as Long-Term Investments*. New York: MacMillan.

Stulz, R.M. (1996). "Rethinking Risk Management," *Journal of Applied Corporate Finance*, 9(5): 8–24.

Taylor, C. (1989). *Sources of the Self: The Making of Modern Identity*. Cambridge: University Press.

Taylor, S. E. and S. C. Thompson (1982). "Stalking the Elusive Vividness Effect," *Psychological Review*, 89: 155–181.

Thaler, R. H. (1987). "The January Effect," *Journal of Economic Perspectives*, 1(1): 197–201.

Thaler, R. H. (1987). "Seasonal Movements in Security Prices II: Weekend, Holiday, Turn of the Month and Intraday Effects," *Journal of Economic Perspectives*, 1(1): 169–177.

Thaler, R. H., A. Tversky, D. Kahneman and A. Schwartz (1997). "The Effect of Myopia and Loss Aversion on Risk Taking: An Experimental Test," *Quarterly Journal of Economics*, 112(2): 647–661.

Turner, J.C. (1991). *Social Influence*. Buckingham, England: Open University Press.

Tversky, A. and D. Kahneman (1974). "Judgment Under Uncertainty: Heuristics and Biases," *Science*, 185: 1124–1131.

Tversky, A. and D. Kahneman (1992). "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty*, 5: 297–323.

Tversky, A. and E. Shafir (1992). "The Disjunction Effect in Choice Under Uncertainty," *Psychological Science*, 3(5): 305–309.

Tylor, E. B. (1871). *Primitive Culture*.

von Furstenberg, G. M. and B. N. Jeon (1989). "Relations Among Stock Markets Around the World," *Brookings Papers on Economic Activity*, I, 125–180.

Wang, F. A. (1997). "Overconfidence, Delegated Fund Management, and Survival," unpublished paper, Columbia University, presented at the NBER–Sage Workshop on Behavioral Finance, Cambridge, MA.

Weber, M. (1947). *The Theory of Social and Economic Organization*, edited by T. Parsons. Glencoe, IL: The Free Press.

West, K. D. (1988). "Bubbles, Fads, and Stock Price Volatility: A Partial Evaluation," *Journal of Finance*, 43: 639–655.

Zimmerman, J. L. (1983). "Taxes and Firm Size," *Journal of Accounting and Economics*, 5: 119–149.

Dynamic Financial Analysis

Roger Kaufmann

RiskLab

ETH Zürich

July 4, 2000

Home page: http://www.math.ethz.ch/~kaufmann

E-mail: kaufmann@math.ethz.ch

RiskLab: http://www.risklab.ch

Dynamic Financial Analysis

- General ideas
- DFA and solvency testing
- Identifying sources of stochastic behaviour
- Strengths, weaknesses and limitations of DFA
- DFA in action

Idea

For analyzing the financial effects of different strategies for insurance companies over a given time horizon there are two primary techniques in use today:

- *Scenario testing* projects results under specific scenarios in the future. The disadvantage of this deterministic approach is the fact that only a few arbitrary scenarios are tested in order to decide how good a strategy is.
- Stochastic simulation, better known as *Dynamic Financial Analysis (DFA)*.
  Here many different scenarios are generated stochastically with the aim of giving information about the distribution of some important variables, like surplus or loss ratio.

Fixing the Time Period

- We would like to model over as long a time period as possible in order to see the long-term effects of a chosen strategy.
- Simulated values get more and more unreliable the longer this time period is.
- A compromise must be made in order to fix the length of the simulated time period.

What Does DFA Stand for?

- *Dynamic* means stochastic or variable, as opposed to static or fixed.
- *Financial* reflects the fact that not only the underwriting business is simulated but rather the total of all assets and liabilities.
- *Analysis* is defined as an examination of the whole complex, its elements and their interrelationships.

Which Risks Should be Modelled?

- Asset risk:
  - How will assets develop?
- Liability risk:
  - Which liabilities will be incurred?
  - When will they be incurred?
  - How big are they?
- Interrelation between both sides:
  - How do these risks depend on each other?
- It is neither possible nor appropriate to model all sources of risk: It can be dangerous to place confidence in a detailed, but perhaps inappropriate model. It is often better to use a simple model that captures only the key features.

Aim of DFA

DFA gives the opportunity to compare the effects of different strategies before applying them to reality.

It does not necessarily give an optimal solution but leaves the decision of selecting a strategy to management.

So DFA serves as a decision tool that requires a good understanding of insurance business and some analytical/actuarial skills to be successfully implemented.

Applications of DFA Models

Before using a DFA model, management has to choose a financial or economic measure which should be analyzed.

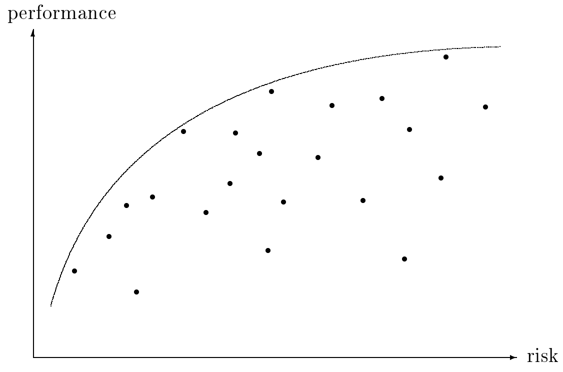
The most common concept is the *efficient frontier concept*:

1. Choose a measure for performance,

   e.g. expected surplus.

2. Choose a measure for risk, e.g.
   - ruin probability,
   - quantiles (VaR) of distribution of surplus,
   - conditional expected loss.
3. Compare different strategies by plotting the measured risk and the measured performance.

Comparing Strategies with Respect to Performance and Risk



Efficient Frontier

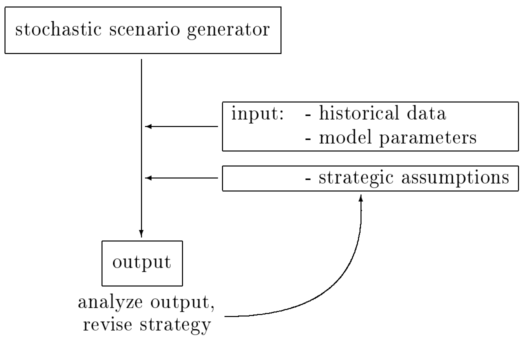
Link Between DFA and Solvency Testing

A better known concept than DFA is *solvency testing*, which deals with one central question:

Does the company have enough capital compared to the level of risk to which it is exposed, i.e. does the company have enough capital to keep the probability of ruin below a given level for the risks taken?

DFA gives us an estimate for the distribution of the surplus. A negative surplus is equivalent to the company becoming insolvent. Therefore DFA can also help answer the question of survival/ruin that is asked in solvency testing.

Main Structure of a DFA Model



stochastic scenario generator

input: - historical data
- model parameters

- strategic assumptions

output

analyze output,
revise strategy

Which Variables are Generated Stochastically?

An important step in the process of building an appropriate model is to identify the most important variables, and the sources of stochastic behaviour.

There are many possible ways of doing this.

A reasonable approach is the one implemented in Dynamo: Several different risk categories are selected and each is modelled with the help of a stochastic generator.

- Non-catastrophe losses
- Catastrophes
- Interest rates
- Stock returns
- Business cycles
- Payment patterns

Non-Catastrophe Losses for Each LOB

Aging phenomenon: The loss ratio – i.e. the ratio of losses divided by earned premiums – decreases when the age of policy increases. Therefore it might prove useful to divide insurance business into three classes, as done in Dynamo:

- New business (superscript 0)
- Renewal business – first annual (superscript 1)
- Renewal business – second annual and subsequent (superscript 2)

For every class we can simulate

- Number of losses ($j = 0, 1, 2$)

  $$N_t^j \sim \text{NB, Pois, Bin}, \ldots$$

- Mean severity $X_t^j = \frac{\sum_{i=1}^{N_t^j} X_t^j(i)}{N_t^j}$

  $$X_t^j \sim \text{Gamma, GPD}, \ldots$$

- Losses in year t $$\sum_{j=0}^{2} N_t^j X_t^j$$

Catastrophes

- Number of catastrophes

  $$N_t \sim \text{NB, Pois, Bin}, \ldots$$
  $$N_1, N_2, \ldots \text{ i.i.d.}$$

- Severity of an individual catastrophe $i = 1, \ldots, N_t$.

  $$X_t(i) \sim \text{lognormal, Pareto, GPD}, \ldots$$
  $$X_t(1), \ldots, X_t(N_t) \text{ i.i.d.}$$

- Total severity is divided up among LOBs affected by event.

  $$X_{t,k}(i) = a_{t,k}(i)\, X_t(i), \quad \begin{aligned} k &= 1, \ldots, l, \\ l &= \#\ \text{LOBs}, \\ \textstyle\sum_{k=1}^{l} a_{t,k}(i) &= 1. \end{aligned}$$

- Catastrophe losses in year t

  $$\sum_{k=1}^{l} b_{t,k} \left(\sum_{i=1}^{N_t} X_{t,k}(i) \right),$$
  $$b_{t,k} = \text{market share of the company.}$$

Interest Rate Generator

- Interest rates r_t (financial assets)

  CIR: $dr_t = a\,(b - r_t)\,dt + s\,\sqrt{r_t}\,dZ_t$,

  Z_t = a standard Brownian motion.

  Yearly discretization:

  - $r_t = r_{t-1} + a\,(b - r_{t-1}) + s\,\sqrt{|r_{t-1}|}\,Z_t$,
  - $r_t = r_{t-1} + a\,(b - r_{t-1}) + s\,\sqrt{{r_{t-1}}^+}\,Z_t$,
  - $r_t = \left(r_{t-1} + a\,(b - r_{t-1}) + s\,\sqrt{r_{t-1}}\,Z_t\right)^+$,

  $$Z_t \sim \mathcal{N}(0,1), \ Z_1, Z_2, \ldots \text{ i.i.d.}$$

- Long term interest rates $R_{t,T}$

  $$R_{t,T} = \frac{r_t\, B_T - \ln A_T}{T},$$

  where

  $$A_T = \left(\frac{2G\, e^{(a+G)\,T/2}}{(a+G)\,(e^{GT} - 1) + 2G} \right)^{2ab/s^2},$$
  $$B_T = \frac{2(e^{GT} - 1)}{(a+G)\,(e^{GT} - 1) + 2G},$$
  $$G = \sqrt{a^2 + 2s^2}.$$

- Return on stock portfolio r_t^S

  CAPM:

  $$\mathbb{E}[r_t^S | R_{t,1}] = (e^{R_{t,1}} - 1) + \beta_t \Big(\mathbb{E}[r_t^M | R_{t,1}] - (e^{R_{t,1}} - 1) \Big),$$

  where

  $$\mathbb{E}[r_t^M | R_{t,1}] = a^M + b^M\,(e^{R_{t,1}} - 1),$$
  $$\beta_t = \frac{\text{Cov}(r_t^S,\, r_t^M)}{\text{var}(r_t^M)},$$
  $$e^{R_{t,1}} - 1 = \text{risk-free return.}$$

  Assuming a lognormal distribution for $1 + r_t^S$ leads to

  $$1 + r_t^S \sim \text{lognormal}\,(\mu_t, \sigma^2),$$

  with μ_t chosen to yield

  $$m_t = e^{\mu_t + \frac{\sigma^2}{2}},$$

  where

  $m_t = 1 + \mathbb{E}[r_t^S | R_{t,1}]$,

  σ^2 = estimated variance of logarithmic historical values.

- Inflation i_t (loss payments)

$$i_t = a^I + b^I\, r_t + \sigma^I\, \epsilon^I_t,$$
$$\epsilon^I_t \sim \mathcal{N}(0,1),\ \epsilon^I_1, \epsilon^I_2, \dots \text{ i.i.d.}$$

- Impact of i_t on each LOB
  - Impact on mean number of losses:

    A reasonable model is
$$\mathbb{E}[N^j_t] = (1+\delta^N_t)\, \mathbb{E}[N^j_{t-1}],$$
$$\text{var}[N^j_t] = (1+\delta^N_t)^2\, \text{var}[N^j_{t-1}],$$
    where
$$\delta^N_t = \max(a^N + b^N\, i_t + \sigma^N\, \epsilon^N_t, -1),$$
$$\epsilon^N_t \sim \mathcal{N}(0,1),\ \epsilon^N_1, \epsilon^N_2, \dots \text{ i.i.d.}$$
  - Impact on mean loss severity:

    A reasonable model is
$$\mathbb{E}[X^j_t] = (1+\delta^X_t)\, \mathbb{E}[X^j_{t-1}],$$
$$\text{var}[X^j_t] = \frac{(1+\delta^X_t)^2}{1+\delta^N_t}\, \text{var}[X^j_{t-1}],$$
$$\mathbb{E}[X_t(i)] = (1+\delta^X_t)\, \mathbb{E}[X_{t-1}(i)],$$
$$\text{var}[X_t(i)] = (1+\delta^X_t)^2\, \text{var}[X_{t-1}(i)],$$
    where
$$\delta^X_t = \max(a^X + b^X\, i_t + \sigma^X\, \epsilon^X_t, -1),$$
$$\epsilon^X_t \sim \mathcal{N}(0,1),\ \epsilon^X_1, \epsilon^X_2, \dots \text{ i.i.d.}$$

Business Cycles by LOB

Is there strong competition among insurance companies in this LOB? Is there a general recession?

We can use a homogeneous Markov chain model where we classify each LOB for every year into one of the following states

1 Weak competition

2 Average competition

3 Strong competition

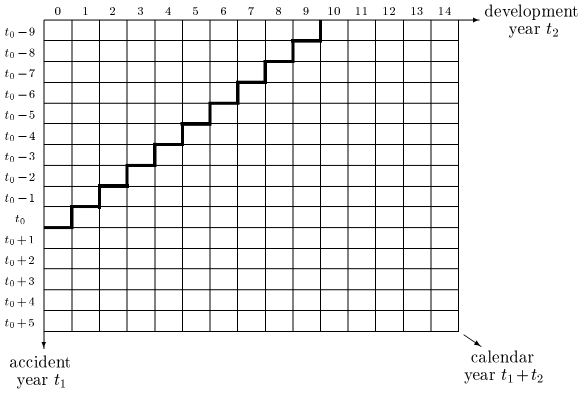
When the company writes l LOBs, there are 3^l states of the world. Because business cycles of different LOBs are strongly correlated, only few of the 3^l states are attainable. So we have to model $L \ll 3^l$ states.

Transition probabilities p_{ij}, $i, j \in \{1, \dots, L\}$ from one year to the next are equal for every year. (Markov chain is homogeneous.)

Main effect of business cycles: The weaker the competition, the higher the premiums.

Payment patterns

When are losses paid?



Paid losses in the upper triangle bounded by the solid line are known, while those in the lower triangle must be simulated.

To model percentages of paid losses we can use for example beta distributions.

Critical Appraisal of DFA:

Strengths of DFA

Compared to scenario testing where only a few arbitrary and possibly unrepresentative scenarios are considered, DFA gives better information on the effects of chosen strategies, because DFA simulates dynamically many different scenarios.

Because of the large number of simulations a DFA model can run, it gives us information not only on behaviour under ordinary circumstances, but also when extremal events occur. Of course the stochastic generators must be sufficiently flexible to generate occasional extreme values.

Weaknesses of DFA

Because reality is complex, it's not possible to model all sources of risk. We have to restrict attention to some key risk factors. So in a DFA model there is not only the randomness by reason of the inherent variability, but also the uncertainty caused by incomplete knowledge.

Generally DFA overestimates probability of ruin since it does not take into consideration that an insurance company has the opportunity to make additional capital available – e.g. by issuing stocks – when it runs the risk of ruin.

Limitations of DFA

DFA does not provide an optimal strategy. It serves as a decision tool that helps management compare different strategies. When a DFA model is used without enough actuarial knowledge, it is only a black box of limited utility.

Because reality can never be represented perfectly, we should of course always be cautious, and never rely completely upon the output produced by a DFA model.

DFA in Action

Model assumptions are:

- Time horizon: 10 years.
- Performance measure: expected surplus.
- Risk measure: ruin probability.
- Only 1 LOB.
- New business and renewal business are not modelled separately.
- Number of non-catastrophe losses $\sim$ NB (154, 0.025).
- Mean severity of non-catastrophe losses $\sim$ Gamma (9.091, 242), inflation-adjusted.
- Number of catastrophes $\sim$ Pois (18).
- Severity of individual catastrophes $\sim$ lognormal (13, 1.5^2), inflation-adjusted.
- Market share: 5%.
- Written premiums in the last year: 20 million.
- Expenses: 28.5% of written premiums.
- Optional excess of loss reinsurance with deductible 500 000 (inflation-adjusted), and cover ∞.
- Premiums for reinsurance: 175 000 p.a. (inflation-adjusted).

- For interest rates we use the discretization $r_t = r_{t-1} + a\,(b - r_{t-1}) + s\,\sqrt{|r_{t-1}|}\,Z_t$.
- Parameters for interest rate generator: $a = 0.25,\ b = 5\%,\ s = 0.1,\ r_1 = 2\%$.
- Parameters for generating return on stock portfolio: $a^M = 4\%,\ b^M = 0.5,\ \beta_t \equiv 0.5,$ $\sigma = 0.15$.
- Parameters for modelling inflation: $a^I = 0\%,\ b^I = 0.75,\ \sigma^I = 0.025$.
- No impact of inflation on the number of claims for the modelled LOB.
- Parameters for modelling the impact of inflation on the severity of claims for the modelled LOB: $a^X = 3.5\%,\ b^X = 0.5,\ \sigma^X = 0.02$.
- Business cycles: 1 = weak, 2 = average, 3 = strong. State in year 0: 1 (weak). Transition probabilities: $p_{11} = 60\%,\ p_{12} = 25\%,\ p_{13} = 15\%,$ $p_{21} = 25\%,\ p_{22} = 55\%,\ p_{23} = 20\%,$ $p_{31} = 10\%,\ p_{32} = 25\%,\ p_{33} = 65\%.$
- Payment patterns are deterministic.

- All liquidity is reinvested. There are only two investment possibilities:
  1) buy a risk-free bond with maturity one year,
  2) buy an equity portfolio with a fixed beta.
- Market valuation: assets and liabilities are stated at market value, i.e. assets are stated at their current market values, liabilities are discounted at the appropriate term spot rate determined by the model.
- No transaction costs.
- No taxes.
- No dividends paid.
- Initial surplus: 12 million.

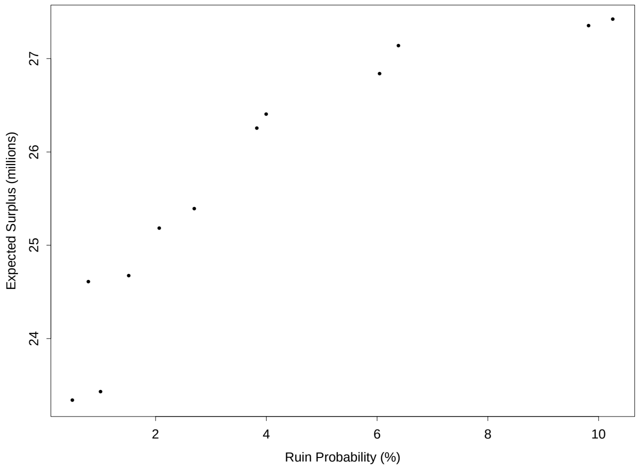
In this model one can choose:

- How many simulations should be run.
- Whether reinsurance should be purchased or not.
- How the liquidity is divided between bond and portfolio.

Example with 10 000 Runs

Expected surplus & ruin probabilities for twelve different strategies:

	with reinsurance	without reinsurance
100 % bonds 0 % stocks	23.33 mio. 0.50 %	23.42 mio. 1.01 %
50 % bonds 50 % stocks	25.17 mio. 2.07 %	25.38 mio. 2.70 %
0 % bonds 100 % stocks	27.34 mio. 9.82 %	27.41 mio. 10.26 %
≤ 5 mio. bonds rest stocks	26.83 mio. 6.05 %	27.13 mio. 6.39 %
≤10 mio. bonds rest stocks	26.25 mio. 3.83 %	26.40 mio. 4.00 %
≤20 mio. bonds rest stocks	24.60 mio. 0.79 %	24.66 mio. 1.52 %



Enterprise Risk Management

Forthcoming in the
Journal of Risk Management of Korea
Volume 12, Number 1

Stephen P. D'Arcy
Fellow of the Casualty Actuarial Society
John C. Brogan Faculty Scholar in Risk Management and Insurance
and Professor of Finance

University of Illinois at Urbana-Champaign

May 30, 2001

Contact Information:

Address:	Department of Finance 340 Wohlers Hall 1206 S. Sixth Street Champaign, IL 61820 U.S.A.
Telephone:	217-333-0772
FAX:	217-244-3102
E-mail:	s-darcy@uiuc.edu

Introduction

Enterprise Risk Management is a relatively new term that is quickly becoming viewed as the ultimate approach to risk management. Consultants are advertising their ability to perform enterprise risk management. Auditors are examining how to incorporate enterprise risk management approaches into company audits.[1] Presentations are being made on this topic at many actuarial, risk management and other insurance meetings.[2] Seminars devoted to this topic are being conducted to explain the process, provide examples of applications and discuss advances in the field. Papers on enterprise risk management are beginning to appear in journals and books on the topic are starting to be published.[3] Some universities are even starting to offer courses titled enterprise risk management. It appears that a new field of risk management is opening up, one requiring new and specialized expertise, one that will make other forms of risk management incomplete and less attractive. This paper will explain what enterprise risk management is, why it has developed so quickly, how it differs from traditional risk management, what new skills are involved in this process and what advantages and opportunities this approach offers compared to prior techniques.

[1] See the Institute of Internal Auditors website for an extensive list of references and discussion of enterprise risk management.

[2] See the CAS website, and particularly the presentations by Friedel, Kawamoto, Miccolis, and Miccolis and Shah.

[3] See Davenport and Bradley (2000), Deloach and Temple (2000), Doherty (2000), Guthrie, et al (1999), Lam (2000) and Shimpi (1999).

Definition of Enterprise Risk Management

Enterprise risk management is, in essence, the latest name for an overall risk management approach to business risks. Precursors to this term include corporate risk management, business risk management, holistic risk management, strategic risk management and integrated risk management. Although each of these terms has a slightly different focus, in part fostered by the risk elements that were of primary concern to organizations when each term first emerged, the general concepts are quite similar.

According to the Casualty Actuarial Society (CAS), enterprise risk management is defined as:

> "The process by which organizations in all industries assess, control, exploit, finance and monitor risks from all sources for the purpose of increasing the organization's short and long term value to its stakeholders."

The CAS then proceeds to enumerate the types of risk subject to enterprise risk management as hazard, financial, operational and strategic. Hazard risks are those risks that have traditionally been addressed by insurers, including fire, theft, windstorm, liability, business interruption, pollution, health and pensions. Financial risks cover potential losses due to changes in financial markets, including interest rates, foreign exchange rates, commodity prices, liquidity risks and credit risk. Operational risks cover a wide variety of situations, including customer satisfaction, product development, product failure, trademark protection, corporate leadership, information technology, management fraud and information risk. Strategic risks include such factors as completion, customer preferences, technological innovation and regulatory or political impediments. Although there can be disagreement over which category would apply to a specific instance, the primary point is that enterprise risk management considers all types of risk an organization faces.

A common thread of enterprise risk management is that the overall risks of the organization are managed in aggregate, rather than independently. Risk is also viewed as a potential profit opportunity, rather than as something simply to be minimized or eliminated. The level of decision making under enterprise risk management is also shifted, from the insurance risk manager, who would generally seek to control risk, to the chief executive officer, or board of directors, who would be willing to embrace profitable risk opportunities (Kawamoto, 2001).

Basically, though, enterprise risk management simply represents a return to the original roots of risk management, a field that was first developed in the 1950s by a group of innovative insurance professors. The first risk management text, presciently titled *Risk Management and the Business Enterprise*, was published in 1963, after six years of development, by Robert I. Mehr and Bob Hedges. As initially introduced in this text, the objective of risk management is, "to maximize the productive efficiency of the enterprise." The basic premise of this text was that risks should be managed in a comprehensive manner, and not simply insured.

The initial focus of risk management was on what is now termed hazard risk. This specialty area developed its own terminology and techniques for addressing risk. Financial risks began to be addressed much later, and by a separate business segment of most organizations. This field also developed its own terminology and techniques for addressing risk, independently of those used in traditional risk management. Each specialty area also developed different methods for reporting the risks the organization faced within each area. Since the hazard risk manager and the financial risk manager both generally reported to a common position, frequently the treasurer or chief financial officer of the firm, the different, and separate, approaches to dealing with risk created a problem. Potentially, each area could be expending resources to deal with a risk that, in aggregate, would cancel out within the firm. Also, the tolerance for risk applied in each area could be vastly different between hazard risks and financial risks. These discrepancies provided the impetus for developing a common terminology and common techniques for dealing with risk. In addition, this common approach could then be applied to other risks, such as operational and strategic risks, that could adversely affect the organization. This common approach to dealing with all risks that a firm faces is the heart of enterprise risk management, and represents an encompassing application of Mehr and Hedges objective," to maximize the productive efficiency of the enterprise."

Historical Development

Risk management has been practiced for thousands of years.[4] One can imagine a proto-risk manager burning a fire at night to keep wild animals away. Early lenders must have quickly learned to reduce the risk of loan defaults by limiting the amount loaned to any one individual and by restricting loans to those considered most likely to repay them. Individuals and firms could manage the risk of fire through the choice of building materials and safety practices, or after the introduction of fire insurance in 1667, by shifting it to an insurer. However, it wasn't until the 1960s that the field was formally named, principles developed and guidelines established.   Robert Mehr and

[4] For an excellent overview of the treatment of risk through the ages, see Bernstein (1996).

Bob Hedges, widely acclaimed as the fathers of risk management, enumerated the following steps for the risk management process:

1. Identifying loss exposures
2. Measuring loss exposures
3. Evaluating the different methods for handling risk
   - Risk assumption
   - Risk transfer
   - Risk reduction
4. Selecting a method
5. Monitoring results

Initially, the risk management process focused on what has been termed "pure risks." Pure risks are those in which there is either a loss or no loss. Either something bad happens, or it doesn't. The states of possible outcomes in a pure risk situation do not allow for any outcome more favorable than the current position.

A typical example of a pure risk is owning a house. Your house may burn down, be hit by an earthquake or be infested by insects. If none of these, or other, unfavorable developments occur, then you are in the no loss position. This is no better than where you started, but no worse either.

The other classification of risk is "speculative risk." In a speculative risk, there is the possibility of a gain. For example, investing in the stock market generates the possibility of a loss (the stock could go down in value), the possibility that the value would not change (the stock price remains where you bought it), and the possibility of a gain (the stock price could increase).

Traditional risk management has focused on pure risks for several reasons. First, the field of risk management was developed by individuals who taught or worked in the insurance field, so the focus was on risks that insurers would be willing to write. In fact, some risk managers job duties are limited to buying insurance, an unfortunate limitation since many other options are readily available and should be explored. Another reason for the focus on pure risks is that in many cases these represented the most serious short term threats to the financial position of an organization at the time this field was founded. A fire could quickly put a firm out of business. Efforts to reduce the likelihood of a fire occurring, or to minimize the damage a fire would cause, or to establish a contingency plan to keep the business going in the event of a fire, or to purchase an insurance policy to compensate the owners for the damages caused by a fire, were easily seen to be beneficial to the firm. Finally, there were simply not a lot of reasons or options for dealing with financial risks such as interest rate changes, foreign exchange rate movements or equity market fluctuations, when this field was first developing.

At the time the field of risk management first emerged, interest rates were stable, foreign exchange rates were intentionally maintained within narrow bands and inflation was not yet a concern to most corporations. Thus, financial risks were not a major issue for most businesses. Indeed, the field of finance was primarily institutional at the time. Although Markowitz had proposed portfolio theory (Markowitz, 1952), the Capital Asset Pricing Model had not yet been developed. The mathematics for quantifying financial risk were not sufficient to put these risks in the same framework as most pure risks. The primary risks of the time were hazard risks: the risk of fire, windstorm or other property damage, or liability. Environmental risks had not yet developed into significant losses. Pensions were, at this point, neither guaranteed nor regulated.

Given the primary risks facing businesses were hazard risks, the initial focus of risk management was on these types of risks. Risks were quantified, the evaluation of different methods of dealing with risk was advanced and standardized, and an extensive terminology for managing risk was developed. Such terms as maximum possible loss (the largest loss that could occur) and maximum probable loss (the largest loss that is likely to occur) were introduced to help define risk exposure. Probability and statistical analysis were used to estimate the range of likely losses and the effect of adopting steps to mitigate these risks.

Risk managers did their job quite effectively. Firms almost universally handled their hazard risk in an appropriate manner. When they didn't, such as the MGM Grand Hotel that found it was not adequately insured for liability coverage after a major fire, new methods of handling risk, in this case retroactive insurance, were developed (Smith and Witt, 1985). Rarely did companies face financial ruin as a result of failure to manage their hazard risks effectively.

Beginning in the 1970s, financial risk became an important source of uncertainty for firms and, shortly thereafter, tools for handling financial risk were developed. These new tools allowed financial risks to be managed in a similar fashion to the ways that pure risks had been managed for decades. In 1972 the major developed countries ended the Bretton Woods agreement which had kept exchange rates stable for three decades. The result of ending the Bretton Woods agreement was to introduce instability in exchange rates. As foreign exchange rates varied, the balance sheets and operating results of corporations engaging in international trade began to fluctuate. This instability affected the performance of many firms. Also during the 1970s, oil prices began to rise as the Organization of Petroleum Exporting Countries (OPEC) developed agreements to reduce production to raise prices. Later in the same decade, a policy shift by the U. S. Federal Reserve to focus on fighting inflation (a result of oil price increases) instead of stabilizing interest rates led to a rapid rise, and increasing volatility, of interest rates in the United States, and had a spillover effect in other nations as well. Thus, volatility in foreign exchange rates, prices and interest rates caused financial risk to become an important concern for institutions.

Although financial risk had become a major concern for institutions by the early 1980s, organizations did not begin to apply the standard risk management tools and techniques to this area. The reasons for this failure were based on the artificial categorization of risk into pure risk and speculative risk (D'Arcy, 1999). Since fixed income assets, investments denominated in foreign currency and operating results that were affected by inflation or foreign exchange rates all had the possibility of a gain, they represented speculative risk. Risk managers had built a wall around their specialty, called pure risk, within which they operated. When a new risk area emerged, they did not expand to incorporate it into their domain. To do so would have required learning about financial instruments and moving away from the type of risks commonly covered by insurance. This would have been a bold move, but one that the innovative thinkers who developed risk management would have espoused. This failure was costly to organizations, and to the risk management field. With the emergence of enterprise risk management, traditional risk managers will be pushed into a wider arena of risk analysis, one that incorporates financial risk management and other forms of risk analysis. Thus, the refusal to expand into financial risks did not prevent risk managers from having to learn about financial risk management, it simply delayed it by a few decades.

A Primer in Financial Risk Management

The basic tools of financial risk management are forwards, futures, swaps and options (Smithson, 1998). These contracts are all termed derivatives, since their values are derived from some other instrument's value. Forwards are contracts entered into today in which the exchange will take place at some future date. The terms of the contract, the price, the date and the specific characteristics of the underlying asset, are all determined when the contract is established, but no money changes hands when the contract is initiated. At the specified date, each party is obligated to consummate the transaction. Since each forward contract is individually negotiated between the two parties, there is considerable flexibility regarding the terms of the contract. However, since forwards are contracts between the two parties, the risk of failure to perform exists, in the same manner that credit risk is a factor in any loan. In financial markets, this risk is termed counterparty risk. Also, since the contracts are specialized agreements between two parties, the contract is not liquid and can be very hard to terminate prior to the specified date if conditions were to change for one or both of the parties.

Futures contracts were developed to address the credit risk and liquidity concerns of forward contracts. Similar to forwards, futures are entered into today for an exchange that will take place at some future date. The terms of the contract are determined when the contract is entered into and no money changes hands when the contract is initiated. However, there are several significant differences between forward and futures. First, a clearinghouse (a firm that guarantees the performance of the parties in an exchange-traded derivatives transaction - Hull, 2000) serves as an intermediary to the contract. Each party is contracting with the clearinghouse, not with the other party. Thus, the risk of nonperformance is significantly reduced. Next, in order to reduce the risk of default, several financial requirements are introduced. Each party must post collateral, termed margin, with its broker. The amount of the margin that must be posted initially is determined for each futures contract (initial margin). Also, each day futures contracts are "marked-to-market" with cash payments flowing from one party to the other based on changes in the value of the futures contract. Thus, if the price of a futures contract increases by $500, then the party that is short the contract (has sold the asset) pays $500 to the party that is long the contract (has bought the asset). These funds come out of, and flow into, the respective margin accounts. If the margin account, falls below a predetermined value (maintenance margin), then a deposit must be made into the margin account to restore it to the initial margin level.

Swaps are agreements between two parties to exchange a series of cash flows based on a predetermined arrangement. Early swaps were based on exchanging a series of payments based on different currencies. For example, one company would pay a predetermined sum in Korean won and the other party would pay in US dollars each quarter for several years. Often the value of the exchanges would be netted (the respective values of each payment would be determined, and one party would pay the counterparty the difference in values). The most common swap today is an interest rate swap in which one party pays a fixed interest rate and the other pays a floating interest rate based on a set index such as the London Interbank Offer Rate (LIBOR). However, swaps can also be based on commodity prices or equity values. Similar to forwards and futures, swaps do not involve a payment by either party went the transaction is initiated.

The final basic tool of financial risk management is an option. An option provides the right, but not the obligation, to engage in a financial transaction at a predetermined price in the future. The owner of the option has the choice about consummating the transaction. The seller of the option is required to fulfill the contract if the buyer chooses. Since an option represents one sided risk, there is an initial cost to purchasing an option, which is termed the option premium. Options can be based on equities, bonds, interest rates, commodities, foreign exchange rates, or any other financial variable. A call option provides the right to buy the underlying asset at the predetermined price; a put option provides the right to sell the underlying asset. Although all options have these general characteristics, many specialized forms of options have been generated to produce a wide variety of different payoffs.

Introduction of Financial Risk Management

Forwards, futures and options had all been traded based on non-financial assets long before they were adapted to deal with financial risk. Swaps were not introduced until 1981, when the first currency swap was announced (Smithson, 1998). However, it did not take long after financial risk began to affect institutions for a wide array of financial risk management products to be generated to help corporations deal with financial risk. Foreign exchange futures were first offered in May, 1972. Interest rate futures began trading in October, 1975. Options on U.S. Treasury bonds were introduced in October, 1982. Options on foreign exchange rates were introduced in

December, 1982. Additional futures, swaps and options, as well as combination products, quickly followed. These tools allowed financial institutions and other corporations to manage financial risk in the much the same fashion that they used for pure risks.

Unfortunately, these tools were not always used wisely or effectively. Since financial risk management was generally not handled by the traditional risk management department, many of the standards for managing risk were not followed in this area. In 1994 alone, due to an unexpected rise in interest rates, the following losses from derivatives occurred (Smithson, 1998):

> Codelco, Chile's national copper trading company, lost $207 million
> Gibson Greetings lost $20 million
> Procter and Gamble lost $157 million
> Mead lost $7 million
> Air Products lost $60 million
> Federal Paper lost $19 million
> Caterpillar lost $13 million

Even more serious losses from the misuse of derivatives include (Jorion, 2001, Holton, 1996):

> Barings Bank went bankrupt in 1995 as a result of $1.3 billion in losses in futures and options trading based on the Nikkei 225 and Japanese bonds
> Metallgelsellschaft lost $1.3 billion on oil futures contracts
> Orange County lost $1.8 billion in 1994 from leveraged interest rate contracts
> Daiwa lost $1.1 billion from unauthorized derivatives trading
> Sumitomo lost $1.8 billion from concealed trading in copper and derivatives on copper by the head trader

In many cases, these losses occurred due to the failure to follow common risk management practices, such as not having transactions verified by an independent authority, not setting limits to potential losses or failure to understand the risks to which the organization was exposed. Managers and boards of directors were, in some cases, reluctant to question individuals who were providing, or at least reporting, impressive profits in a new area of financial transactions, and were willing to provide authority to these individuals without adequate oversight. The fear was that the normal level of oversight, if exercised in these areas, would drive a person with extraordinary talent away from their firm. Thus, they were lured into risk areas they neither understood nor would have accepted.

Imagine the approach that would have been taken if a traditional risk manager, newly hired by a firm, claimed to be able to provide insurance coverage through a self-funding strategy at half the price that the current providers were charging. What if this risk manager wanted to take control of the funds for managing risks and wanted to be the person in charge of handling, and reporting, all monetary transactions involving this fund, but would not provide details about the fund to the company? Despite the apparent cost savings, I doubt that any firm would be foolish enough to disregard its oversight process in this situation, or to provide this person with performance bonuses based on the apparent cost savings. Traditional risk management has developed a series of checks and balances to prevent such obvious abuses. Financial risk management did not initially have this level of expertise. One reason for this failure is because traditional risk managers abdicated the area of speculative risk, exposing many organizations to disastrous losses.

The basic rule of risk taking, whether it is hazard risk, financial risk or any other form of risk, is that if you do not fully understand a risk, you do not engage in it, regardless of what profits are claimed or reported. This basic rule is, unfortunately, violated by individuals consistently. Promises of impressive returns entice many individual investors to participate in fraudulent investment schemes. Unfortunately, many corporations fell into this trap as well.

The losses of the mid-1990s led organizations to realize the importance of financial risk management. The financial instruments that were developed to deal with financial risk were complex, and often only understood by those in the financial areas of the firm. Thus, the use of these tools to manage financial risk was generally not coordinated with the approach used to manage other risks. This lack of coordination resulted in a number of problems, including the development of a different terminology from that used in traditional risk management, different measures of risk and different goals. For example, traditional risk managers frequently focus on the probable maximum loss, the largest loss that could reasonably be expected to occur. If that loss exceeds the ability of the firm to cope with, then steps are taken to manage that risk, by transferring some of the risk to other parties, by reducing loss severity through loss control steps or other standard practices. Instead of adopting this approach, financial risk managers developed a measure termed the Value-at-Risk (VaR). This value indicates the loss that the firm would expect to have occur over the selected time interval (for example, daily) the selected percentage of the time. Thus, the daily VaR at the 1% level is the loss that can be expected to occur once every 100 days. This is not the largest loss that is likely to occur, so it does not provide the same level of information as probable maximum loss. The daily VaR at the 5% level, which is expected to occur once every 20 days, is smaller than the 1% value. VaR indicates what losses to expect, not what losses could occur. Even the time frame is different, as the traditional risk manager is likely dealing with loss probabilities over an annual basis, or over the term of an insurance contract, while VaR is often based on daily or weekly price movements.

Another difference between hazard risk and financial risk is the degree of independence among separate elements. In hazard risk management, risks are frequently independent of each other. Thus, the calculation of the number of accidents that a pool of vehicles is likely to be involved in during a year is determined by assuming that each accident is independent of every other accident. Financial risks, on the other hand, are not considered to be independent. In many cases, the correlation between different financial transactions forms the basis of the risk management strategy. Financial risk management considers the relationships among different financial variables to construct hedges. For example, a firm exposed to long term interest rate risk might use futures on short term instruments, due to the high correlation between short and long term interest rates, to hedge their interest rate exposure. Financial risk management approaches can lead to difficulty when the historical relationships between financial variables shifts. For example, the hedge fund Long Term Capital Management lost 92 percent its value (approximately $4.5 billion) in 1998 when historical patterns between variables, including yields on U.S. and Russian bonds, changed significantly.

Thus, the Board of Directors and other managers that are determining the overall risk management strategy of the firm are likely to receive different types of information on financial risk and on hazard risk. The risks are different, the terminology is different and the measures of risk are different. This makes the task of coordinating the firm's overall exposure to risk more difficult. In addition to desiring a common approach to hazard and financial risks, these decision makers have also envisioned incorporating other forms of risk, including strategic and operational, into the same approach. It is this vision that has led to the creation of enterprise risk management.

Other Factors Leading to Enterprise Risk Management

A number of other factors have also contributed to the development of enterprise risk management. Recent advances in computing power provide the powerful modeling tools necessary to perform sophisticated risk analysis for hazard risks, such as catastrophes, for financial risks, such as interest rate movements, and for other risks. Also, the availability of extensive data bases of financial and other information allows users to examine historical information to determine trends, correlations and other relationships among variables that is essential to enterprise risk management.

Insurers are also developing an expertise in, and a focus on, financial risk management. Some insurers are beginning to provide policies that coordinate financial and pure risk. One insurer has offered a policy that provides protection against foreign currency losses within it insurance coverage (Banham, 1999). Another insurer provided protection for a utility in which the amount of coverage is a function of rainfall, which affect utility income (Taylor, 2001).

Insurers are beginning to utilize the financial markets themselves through the securitization on insurance risk. Several types of insurance securitization have been developed (ISO, 1999). The first was the use of exchange traded derivatives. Both futures and options on catastrophe risk have been traded on the Chicago Board of

Trade. Trading in futures began in 1992 based on an index of catastrophe losses paid by a number of insurers reporting to ISO. In 1995 the index was changed to catastrophe losses reported by Property Claim Services, and trading in options was instigated. Although neither of these instruments is traded currently, their existence provided an impetus for insurers to learn about financial risk management tools and encouraged subsequent development of other approaches. The second approach is through contingent capital. One form of this is termed a Cat-E-Put, or catastrophe-equity-put. Under this contract, an insurer purchases a contract under which the counterparty agrees to purchase equity in the firm, at a predetermined price, in the event of a catastrophe as defined in the contract. This is, essentially, a put option that is triggered by a catastrophe. A third type of securitization is termed risk capital, in which an insurer, through an intermediary, issues debt on which the repayment of interest and principal is dependent on catastrophe loss experience. The debt is not fully repaid if a certain level of catastrophic losses occur. As a result of these innovations, insurers have been able to tap the capital markets to help spread catastrophic losses. The successes in this area are encouraging additional growth into the financial risk management field.

Insurers and risk managers have a significant role to play in the field of financial risk management. From the point of view of the firm, the risk of a fire that costs the firm $1 million has the same impact on the firm's financial position as a loss in its bond portfolio of $1 million. Protection is available against both of these risks. A coordinated approach to an organization's risk would be preferable to a segmented approach.

After the shocks of mismanaged financial risks, the failed investments in interest rate derivatives, Nikkei 225 stock index futures, and the later success that financial risk management has had in reducing such exposure, corporations have begun to question whether other risks can be handled in a similar, integrated approach.

The Skills Required for Enterprise Risk Management

Although enterprise risk management represents a return to the roots of risk management, in order to be involved with enterprise risk management, traditional risk managers will need to obtain some additional skills. The starting point is to learn the terminology of finance and financial risk management. Due to their importance as potential investments and the growing use of this form of financing, often involving insurance guarantees, the role of asset backed securities should be given special attention. Although new instruments for financial risk management are constantly being generated, they can generally be broken down into their basic components of forwards, futures, swaps and options to be more easily understood. Traditional risk managers also need to learn about VaR in order to engage any comprehensive risk management process. Knowledge of portfolio theory as a method for dealing with correlated risks is also critical. Simulation and modeling are also important aspects of enterprise risk management. The ability to locate, and exploit natural hedges, those conditions that affect different aspects of an organization in offsetting ways, is vital as well. For example, telephone companies have a natural hedge against major disasters (Molnar, 2000). When a disaster strikes, the company will suffer a loss to its property, but the higher volume of telephone traffic that typically follows a major disaster will help offset this loss. However, the basic approach of identifying, measuring, evaluating, selecting and monitoring risk remains the same. The primary challenge to traditional risk managers is to examine all risks that an organization faces, and not just focus on those that are insurable.

Since enterprise risk management involves so many different aspects of an organization's operations, and integrates a wide variety of different types of risks, no one person is likely to have the expertise necessary to handle this entire role. In most cases, a team approach is used, with the team drawing on the skills and expertise of a number of different areas, including traditional risk management, financial risk management, management information systems, auditing, planning and line operations. The use of a team approach, though, does not allow traditional risk managers to remain focused only on hazard risk. In order for the team to be effective, each area will have to understand the risks, the language and the approach of the other areas. Also, the team leader will need to have a basic understanding of all the steps involved in the entire process and the methodology used by each area.

In assessing the potential losses an organization could experience, many items not covered under hazard risk or financial risk emerge. The company could suffer a significant loss if the chief executive officer were to step down and an adequate replacement could not be found. If the reputation of one of the company's key products is tarnished by a serious loss (Firestone tires, for example), the company could incur significant monetary losses. If the firm is found liable for underpaying taxes by losing a tax dispute, the required payment could be extremely large. A labor dispute could severely impact a firm's operations. A failed merger could have repercussions that puts the firm into a worse financial position than it was in before the negotiations commenced.

Although these risks are both present and significant, the ability to quantify such exposures is far less sophisticated than the approach that can be used for most hazard and financial risks. The lack of data and the difficulty in predicting the likelihood of a loss or the financial impact if a loss were to occur make it hard to quantify many risks a firm faces.

One feature of enterprise risk management is the consideration of offsetting risks within a firm. Catastrophe losses are one example. A major hurricane increases the losses of an insurer, but after most disasters people are more likely to purchase insurance against future catastrophes. Thus, future earnings increase, which can offset, on an enterprise risk management approach, the increase in losses the firm has to pay.

The steps of enterprise risk management are quite familiar to traditional risk managers. Shawna Ackerman, a consultant at MHL/Paratus Consulting, lists these steps as (Ackerman, 2001):

Identify the question(s)
Identify risks
Risk measurements
Formulate strategies to limit risk
Implement strategies
Monitor results
And repeat…

Another consulting firm lists the steps as (ARI 2001):

Identify risk on an enterprise basis
Measure it
Formulate strategies and tactics to limit or leverage it
Execute those strategies and tactics

Monitor process

The steps of enterprise risk management are the same, expect for minor changes in wording, as those first enumerated by Mehr and Hedges in 1963. Enterprise risk management is risk management applied to the entire organization. The basic approach, the goals and the focus of enterprise risk management are the same as those that have worked so effectively for traditional risk managers since the field was first developed.

Conclusion

The impetus for enterprise risk management arose when the traditional risk manager and the financial risk manager began reporting to the same individual in a corporation, commonly the treasurer or chief financial officer. Each risk management specialty had its own terminology, its own methodology and its own focus. However, each dealt with risk the firm was facing. It quickly became apparent that a common approach to risk management would be preferable to an individual approach and an integrated approach preferable to a separatist approach. The evident success of first hazard risk management and later financial risk management has encouraged managers to try to include these and other forms of risk in an overall risk management strategy. Whether this approach succeeds will depend on the ability of those involved in the separate risk categories to develop an integrated approach and extend it to other areas of risk. This is not truly a new form of risk management, it is simply a recognition that risk management means total risk management, not some subset of risks. The new focus on the concept of enterprise risk management provides an opportunity for risk managers to apply their well established and successful approaches to risk on a broader and more vital scale than previously. This is an excellent opportunity to advance the science of risk management.

References

Ackerman, Shawna. 2001. The Enterprise in Enterprise Risk Management. *Casualty Actuarial Society Enterprise Risk Management Seminar.*

ARI Risk Management Consultants. 2001. Enterprise Risk Management: The Intersection of Risk and Strategy. http://www.riskadviser.net/Cases/case.htm

Banham, Russ. 1999. Understanding the Skepticism about Enterprise Risk Management. *CFO Magazine*. April 1, 1999.

Bernstein, Peter L. 1996. *Against the Gods: The Remarkable Story of Risk*. John Wiley and Sons, Inc. New York.

Casualty Actuarial Society Websites:
http://www.casact.org/research/ermsurv.htm
http://www.casact.org/CONEDUC/specsem/erm/2001/handouts/handouts.htm

D'Arcy, Stephen P. 1999. Don't Focus on the Tail: Study the Whole Dog! *Risk Management and Insurance Review*. 2(2):iv-xiv.

Davenport, Edgar W. and L. Michelle Bradley. 2000. Enterprise Risk Management: A Consultative Perspective. *Casualty Actuarial Society Discussion Paper Program* p. 23-42.

Deloach, James and Nick Temple. 2000. *Enterprise-Wide Risk Management: Strategies for Linking Risk and Opportunity.* Financial Times Management.

Doherty, Neil A. 2000. *Integrated Risk Management.* McGraw-Hill New York.

Friedel, Wolfgang F. 2001. Enterprise Risk Management - Fad or Fact? *Casualty Actuarial Society Enterprise Risk Management Seminar.*

Guthrie, Vernon H., David A. Walker and Bert N. Macesker. 1999. Enterprise Risk Management. 17th International System Safety Conference. ABS Group Inc. Risk & Reliability Division and United States Coast Guard Research and Development Center (vguthrie@abs-group.com; dawalker@abs-group.com; bmacesker@rdc.uscg.mil).

Holton, Glyn A. 1996. Enterprise Risk Management. *Contingency Analysis*. (http://www.contingencyanalysis.com/_frame/frameerm.htm)

Hull, John C. 2000. *Options, Futures, and Other Derivatives* (Fourth Edition). Prentice Hall. Upper Saddle River, NJ.

Insurance Services Office. 1999. *Financing Catastrophe Risk: Capital Market Solutions*.

Institute of Internal Auditors. 2001. Risk Management Readings (http://www.theiia.org/ecm/guide-ia.cfm?doc_id=1604)

Jorion, Philippe. 2001. *Value at Risk* (Second Edition) McGraw-Hill New York.

Kawamoto, Brian. 2001. Issues in Enterprise Risk Management: From Theory to Application. Casualty Actuarial Society Spring Meeting.

Lam, James. 2000. Enterprise-Wide Risk Management and the Role of the Chief Risk Officer. *Erisk* March 25, 2000. Erisk.com

Markowitz, Harry M. 1952. Portfolio Selection. *Journal of Finance* 7:77-91.

Mehr, Robert I. and Bob A. Hedges. 1963. *Risk Management in the Business Enterprise*. Richard D. Irwin, Inc. Homewood, IL

Miccolis, Jerry and Samir Shah. 2000. Enterprise Risk Management: An Analytic Approach. Tillinghast - Towers Perrin Monograph.

Molnar, Michele. 2000. More Companies Embrace Enterprise Risk Management. *Office.com.*

Shimpi, Prakash A. 1999. *Integrating Corporate Risk Management.* Swiss Re New Markets

Smith, Michael L. and Robert C. Witt. 1985. An Economic Analysis of Retroactive Liability Insurance. *Journal of Risk and Insurance* 52:379-401.

Smithson, Charles W. 1998. *Managing Financial Risk: A Guide to Derivative Products, Financial Engineering, and Value Maximization* (Third Edition) McGraw-Hill New York.

Taylor, Gary. 2001. New Developments in Enterprise Risk Management in the Energy Industry With a Specific Focus on the Weather Risk Management Market. Casualty Actuarial Society Spring Meeting.

Tillinghast - Towers Perrin

BUILDING RELATIONSHIPS ▪ PRODUCING RESULTS™

Enterprise Risk Management

An Analytic Approach

A Tillinghast – Towers Perrin Monograph

Foreword

Business Risk Management…Holistic Risk Management…Strategic Risk Management… Enterprise Risk Management. Whatever you choose to call it, the management of risk is undergoing fundamental change within leading organizations. Worldwide, they are moving away from the "silo-by-silo" approach to manage risk more comprehensively and coherently.

This heightened interest in Enterprise Risk Management (ERM) has been fueled in part by external factors. In just the last few years, industry and government regulatory bodies, as well as institutional investors, have turned to scrutinizing companies' risk management policies and procedures. In more and more countries and industries, boards of directors are now required to review and report on the adequacy of the risk management processes in the organizations they govern.

And internally, company managers are touting the benefits of an enterprise-wide approach to risk management. These benefits include:

- reducing the cost of capital by managing volatility
- exploiting natural hedges and portfolio effects
- focusing management attention on risks that matter by expressing disparate risks in a common language
- identifying those risks to exploit for competitive advantage
- protecting and enhancing shareholder value.

ERM is actually a straightforward process. And, in most cases, the requisite intellectual capital and business practices needed to carry out ERM already exist within the company. But an accurate, useful ERM process is based on sound analytics. Without valid measurements, managing risk is effective and efficient only by chance.

In the following pages, we hope to add analytical rigor to the public discourse on ERM. Drawing from our client experiences, we offer a rational, scientific approach — one grounded in sound principles and practical realities.

"Risk," by definition and by nature, cannot be eliminated. Nor do leading organizations wish it gone. Rather, they want to manage the factors that influence risk so that they can pursue strategic advantage. How to identify and manage these factors is the subject of this monograph.

It is our intention to periodically update this document. We would be most interested in readers' comments and suggestions.

Contents

I Introduction

Purpose of this monograph

Pressure to adopt ERM has increased from both internal and external forces. Although optional in most cases, a formalized risk management culture and its benefits have gained recognition and have fueled interest in the process.

With this monograph, we intend to add analytical rigor to the public discourse on ERM by presenting a scientific approach grounded in sound business principles and practical realities.

In this document, we will:

- define the ERM process
- discuss what motivates organizations to adopt ERM
- describe our conceptual ERM framework and outline the process steps
- detail a comprehensive, analytic approach to ERM
- discuss methods by which organizations implement ERM.

Definition and objective of ERM

We define ERM as follows:

> ***ERM is a rigorous approach to assessing and addressing the risks from all sources that threaten the achievement of an organization's strategic objectives. In addition, ERM identifies those risks that represent corresponding opportunities to exploit for competitive advantage.***

ERM's objective — to enhance shareholder* value — is achieved through:

- improving capital efficiency
  - providing an objective basis for allocating resources
  - reducing expenditures on immaterial risks
  - exploiting natural hedges and portfolio effects
- supporting informed decision making
  - uncovering areas of high-potential adverse impact on drivers of share value
  - identifying and exploiting areas of "risk-based advantage"
- building investor confidence
  - establishing a process to stabilize results by protecting them from disturbances
  - demonstrating proactive risk stewardship.

Motivation for considering ERM

External pressures

Some organizations adopt ERM in response to direct and indirect pressure from corporate governance bodies and institutional investors:

- In Canada, the Dey report, commissioned by the Toronto Stock Exchange and released in December 1994, requires companies to report on the adequacy of internal control. Following that, the clarifying report produced by the Canadian Institute of Chartered Accountants, "Guidance on Control" (CoCo report, November 1995), specifies that internal control should include the processes of risk assessment and risk management. While these reports have not forced Canadian-listed companies to initiate an ERM process, they do create public pressure and a strong moral obligation to do so. In actuality, many companies have responded by creating ERM processes.
- In the United Kingdom, the London Stock Exchange has adopted a set of principles — the Combined Code — that consolidates previous reports on corporate governance by the Cadbury, Greenbury and Hampel committees.

* In this monograph, the emphasis is on shareholders rather than the broader category of stakeholders (which also includes customers, suppliers, employees, lenders, communities, etc.). Though some observers prefer to define the scope of ERM to include the interests of all stakeholders, we believe this is not pragmatic at the current evolutionary state of ERM and would result in too diffuse a focus. While shareholder value is not directly relevant to some organizations (e.g., privately held and nonprofit entities), the concepts and approaches developed in this monograph clearly apply to those organizations.

This code, effective for all accounting periods ending on or after December 23, 2000 (and with a lesser requirement for accounting periods ending on or after December 23, 1999), makes directors responsible for establishing a sound system of internal control, reviewing its effectiveness and reporting their findings to shareholders. This review should cover all controls, including operational and compliance controls and risk management. The Turnbull Committee issued guidelines in September 1999 regarding the reporting requirement for nonfinancial controls.

- Australia and New Zealand have a common set of risk management standards. Their 1995 standards call for a formalized system of risk management and for reporting to the organization's management on the performance of the risk management system. While not binding, these standards create a benchmark for sound management practices that includes an ERM system.
- In Germany, a mandatory bill — the Kon TraG — became law in 1998. Aimed at giving shareholders more information and control, and increasing the accountability of the directors, it includes a requirement that the management board establish supervisory systems for risk management and internal revision. In addition, it calls for reporting on these systems to the supervisory board. Further, auditors appointed by the supervisory board must examine implementation of risk management and internal revision.
- In the Netherlands, the Peters report in 1997 made 40 recommendations on corporate governance, including a recommendation that the management board submit an annual report to the supervisory board on a corporation's objectives, strategy, related risks and control systems. At present, these recommendations are not mandatory.
- In the U.S., the SEC requires a statement on opportunities and risks for mergers, divestitures and acquisitions. It also requires that companies describe distinctive characteristics that may have a material impact on future financial performance within 10-K and 10-Q statements. Several factors broaden the requirement to report on the risks to the organization, leading to setting in place an enterprise-wide approach to risk management:
  - The report, "Internal Control — An Integrated Framework," produced by the Committee of the Sponsoring Organizations of the Treadway Commission (COSO), favors a broad approach to internal control to provide reasonable assurance of the achievement of an entity's objectives. Issued in September 1992, it was amended in May 1994. While COSO does not require corporations to report on their process of internal control, it does set out a framework for ERM within an organization.
  - In September 1994, the AICPA produced its analysis, "Improving Business Reporting — A Customer Focus" (the Jenkins report), in which it recommends that reporting on opportunities and risks be improved to include discussion of all risks/opportunities that:

    — are current

    — are of serious concern

    — have an impact on earnings or cash flow

    — are specific or unique

    — have been identified and considered by management.

    The report also recommends moving toward consistent international reporting standards, which may include disclosures on risk as is required in other countries.

Institutional investors, such as Calpers, have begun to push for stronger corporate governance and to question companies about their corporate governance procedures — including their management of risk.

Internal reasons

Other organizations simply see ERM as good business. For example:

- The Board of Directors at a large utility mandated an integrated approach to risk management throughout the organization. They introduced the process in a business unit that was manageable in size, represented a microcosm of the risks faced by the parent and did not have entrenched risk management systems. This same unit was the focus of the parent's strategy for seeking international growth — a strategy that would take the organization into unfamiliar territory — and had no established process for managing the attendant risks in a comprehensive way.

- The CFO of a manufacturing company with an uninterrupted 40-year history of earnings growth embarked on ERM. This step followed the company's philosophy of "identifying and fixing things before they become problems." The movement was spurred by the company's rapid growth, increasing complexity, expansion into new areas and the heightened scrutiny that accompanied its recent initial public offering.
- A large retail company's new Treasurer, with the support of the CFO, wanted to "assess the feasibility of taking a broader approach to risk management in developing the organization's future strategy." As part of this effort, she hoped to "evaluate our hazard risk and financial risk programs and strategies, to identify alternative methods of organizing and managing these exposures on a collective basis."
- The Chairman of the Finance Committee of the Board at a manufacturing company complained about reports from Internal Audit that repeatedly focused on immaterial risks. His concern led to formation of a cross-functional Risk Mitigation Team to identify and report on processes to deal with risks within an ERM framework. The team now reports directly to the finance committee on a quarterly basis.

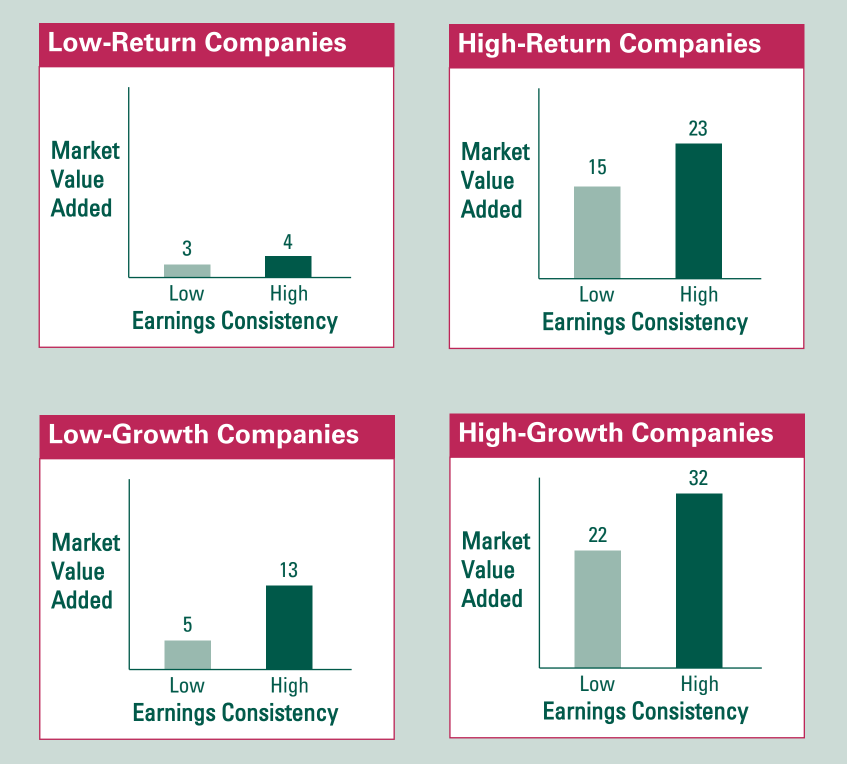
These organizations view systematic anticipation of material threats to their strategic plans as integral to executing those plans and operating their businesses. They seek to eliminate the inefficiencies built into managing risk within individual "silos." And they appreciate that their cost of capital can be reduced through managing volatility.

Some observers argue that investors do not put a premium on an organization's attempt to manage volatility. These observers maintain that investors can presumably achieve this result more efficiently by diversifying the holdings in their own portfolio. They argue further that investors do not appreciate, and do not reward, an organization that spends its resources on risk management to smooth results on investors' behalf.

Our research into the link between performance consistency and market valuation, however, indicates otherwise. We found that consistency of earnings explains a high degree of difference in share value (specifically, "market value added") among companies within an industry. This is true even after allowing for other influences such as growth and return (see *Figure 1* and Appendix A). Investors assign a higher value, all else equal, to organizations whose earnings are more consistent than those of their peers. This clearly reduces the cost of capital for these organizations.

In summary, organizations can use ERM to enhance the drivers of share value: growth, return on capital, consistency of earnings and quality of management. ERM can identify and manage serious threats to growth and return while identifying risks that represent opportunities to exploit for above-average growth and return. Achieving earnings consistency is, of course, a central goal of ERM. And institutional investors increasingly define management quality to include enterprise-wide risk stewardship.

FIGURE 1



Companies with higher earnings consistency tend to have much higher stock valuations than their similarly situated competitors. Details and definitions are presented in Appendix A.

II Framework for ERM

Company information and procedures already in place can make the ERM process efficient and effective. Our conceptual framework for ERM consists of four elements.

Assessing risk

Risk assessment focuses on risk as a threat as well as an opportunity. In the case of risk-as-threat, assessment includes identification, prioritization and classification of risk factors for subsequent "defensive" response. In the case of risk-as-opportunity, it includes profiling risk-based opportunities for subsequent "offensive" treatment.

Shaping risk

This "defensive track" includes risk quantification/modeling, mitigation and financing.

Exploiting risk

This "offensive track" includes analysis, development and execution of plans to exploit certain risks for competitive advantage.

Keeping ahead

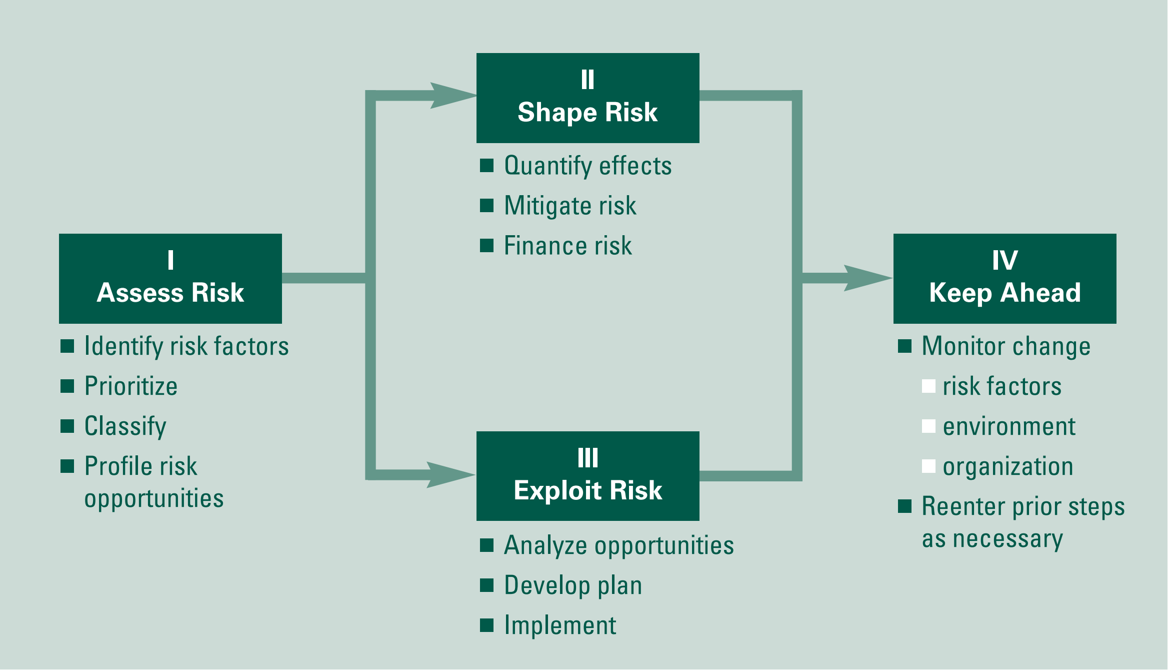
The nature of risk, the environment in which it operates, and the organization itself change with time. The situation requires continual monitoring and course corrections.

The chapters that follow provide a fuller description of the above elements (outlined in *Figure 2*).

The larger part of the discussion in this monograph is on the first two elements — risk assessment and risk shaping — as these create the foundation for the remaining elements. Accordingly, there will be more focus on the defensive track of ERM.

FIGURE 2

The Conceptual Approach to ERM



I
Assess Risk
- Identify risk factors
- Prioritize
- Classify
- Profile risk opportunities

II
Shape Risk
- Quantify effects
- Mitigate risk
- Finance risk

III
Exploit Risk
- Analyze opportunities
- Develop plan
- Implement

IV
Keep Ahead
- Monitor change
  - risk factors
  - environment
  - organization
- Reenter prior steps as necessary

The conceptual approach to ERM is straightforward.

III A Rational Approach to Assessing Risk

Overview

We approach risk assessment believing that managing risk effectively requires measuring risk accurately — and that accurate risk measurement requires well-formulated risk modeling. Such measuring and modeling:

- allow senior management to see a compelling demonstration of the "portfolio effect," i.e., the fact that independent and/or favorably correlated risks tend to offset each other without the organization having to invest in explicit hedges
- promote the proper allocation of capital resources to risks that really matter
- permit sizing of investments in risk remediation
- provide an objective framework for systematic risk monitoring.

Do all risks that face an organization need modeling? And isn't model-building on this scale daunting?

The answer to the first question is: "No." Methods to prioritize risk factors can screen for those that require modeling. These methods are qualitative; we focus on these later in this chapter.

The answer to the second question is: "Not typically." These models often have been built and exist in some form somewhere in the organization. This will be the focus of Chapter IV.

Before we discuss the steps in risk assessment, we should distinguish risks from the risk factors underlying them. Here we focus on the negative side of risk — as a threat, not as an opportunity. In this context, risk is the possibility that something will prevent — directly or indirectly — the achievement of business objectives. Risk factors are the events or conditions that give rise to risk. Loss of market share is a risk; lack of preparedness for the entry of new competitors is a risk factor. Risk is not something that can be directly managed or controlled. Risk factors, however — the causes of risk — can be. Therefore, managing risk, and particularly assessing risk, requires focusing on its causes rather than its manifestations.

STEP 1
Identify risk factors

In this initial step, a wide net is cast to capture all risk factors that potentially affect achieving business objectives. Risk factors arise from many sources — financial, operational, political/regulatory or hazards. The key characteristic of each is that it can prevent the organization from meeting its goals. In fact, if a risk factor does not have this potential, it is not truly a risk factor under an enterprise-wide interpretation of risk. Thus, the first "screen" through which a candidate risk factor must pass is materiality.

In identifying risk factors, we favor a qualitative approach — gathering material from interviews with experts and reviewing documents. The interviews typically span the organization's:

- Senior management
- Operations management
- Corporate staff, including:
  - Finance
  - Treasury
  - Legal
  - Audit
  - Strategic Planning
  - Human Resources
  - Risk Management
  - Safety
  - Environmental.

These interviews solicit informed opinion on:

- how the business works, and the way components of the business — the interviewees' realms of responsibility — mesh
- key performance indicators used to manage the business and its components
- tolerable variation in key performance indicators over relevant time horizons
- events or conditions that cause variations beyond the risk tolerances, and the probable frequency and possible maximum effect of these.

Often we find it helpful to supplement internal interviews with interviews among the organization's external partners, their counterparties (banks, insurers, brokers), analysts, customers, and — on occasion — competitors.

We also review the organization's strategic plans, business plans, financial reports, analyst reports and risk stewardship reports.

From all these data and information, a picture emerges of the organization's:

- corporate culture
- objectives
- forms of capital (human, financial, market and infrastructure)
- business processes (which convert the capital into cash flows)
- control environment
- roles and responsibilities
- key performance measures
- risk tolerance levels
- capacity and readiness for change
- preliminary list of risk factors.

Importantly, this approach starts with the business, not a checklist of risks — far different from an audit-type approach. In other words, this approach goes from the top down and not the bottom up. Such an organic method is strongly preferable because preconceived checklists of risk factors are usually incomplete. Further, the most crucial risk factors are usually unique to each organization and its culture. This alone makes generic checklists far less relevant than a business-first approach.

STEP 2
Prioritize risk factors

The resulting list of risk factors (typically several dozen long at this stage) is not yet useful or actionable, although each factor has passed the materiality screen. It now requires prioritizing.

In Step 1 (Identify risk factors), we compiled information on each risk factor's likelihood, frequency, predictability and potential effect on the organization's key performance indicators. We also examined the quality of the process, systems and cultural controls in place to mitigate these factors. At this stage, the information is subjective, but quite sufficient. Now, the objective is to cull the list of these factors into a manageable number for senior management. The attributes of each factor can be combined in an overall score that, when combined with subjective judgment on the timing and duration of the financial impact, can be expressed as a "net present value" score. In the example in *Figure 3,* this "NPV" score is on a scale of 1 (low) to 5 (high). Once scores are assigned, we can sort the risk factors from low to high and produce a prioritized list.

A team of risk management experts typically does this evaluation and scoring. They often collaborate with representatives of management. In addition, we find a follow-up questionnaire or focus group(s) extremely helpful for cross-validation purposes. In these, the interviewees view the collective results of the identification step — the full list of risk factors, the consensus view on key performance indicators and risk tolerances, etc. Then, with this richer context and some facilitation, they can prioritize risks. We compare the results of this exercise with those from the independent prioritization conducted by the expert team, and the differences are reconciled.

The number of risk factors that will ultimately pass through the prioritization screen is often known before the process begins. Given the demands on senior management, expecting them to concentrate on a dozen or more "top priority" risk factors is unrealistic. Generally, six or less is manageable, but this depends on the organization. Also, natural breakpoints in the prioritized list and strategic links among the risk factors can influence the ultimate number. The short list should, however, contain items deserving of consideration at the highest levels of the organization — factors that should influence the strategic plan and the affected business plans, alter the day-to-day priorities of business unit managers and affect the behavior of the rank and file.

STEP 3
Classify risk factors

Still, any list of risk factors, however short and prioritized, is a sterile device. Organizing this information to clearly indicate what type of risk-shaping action is necessary comes next.

We have used several classification schemes in our work, some more detailed than others, each tailored to the client organization. One general scheme that may have nearly universal relevance is described below (see *Figure 4*). Additional refinements can be added as appropriate.

In this scheme, high-priority risk factors are of two types. One is characterized by the fact that the environment in which they arise is familiar to the organization, and the skills to remedy those risk factors are already in-house. However, for some reason, these risk factors had not been given the attention they deserve. We label these "manageable risk factors." Other risk factors arise because the organization enters unfamiliar

FIGURE 3

When Prioritizing Risk Factors...

...subjective scoring is appropriate at this stage

Risk Factors	Likelihood	Severity	Quality of Controls	Aggregate "NPV" Score (1-5)
A. Strategy				
Informal planning, process and communications allow surprises	H	H	L	4.5
Market share and earning objectives are not aligned	H	L	L	3.0
⋮				
B. Growth				
Infrastructure is increasingly strained, will be difficult to retain culture and values with the changes that growth demands	H	H	L	4.5
Increased size creates more opportunity for mistakes	M	L	M	2.0
⋮				
C. Company Reputation				
Pressure to make numbers may prompt behavior that will impair company's credibility with financial markets	M	H	H	3.5
Adverse publicity (e.g., business practices, ethics) can affect image across multiple brands	L	H	H	2.5
⋮				
D. Human Resources				
⋮				
J. Systems				
⋮				

Risk factors can be prioritized using a subjective process.

FIGURE 4

When Classifying Risk Factors...

...use a scheme that implies action

"Manageable" Risk Factors	"Strategic" Risk Factors
■ Known environment	■ Unfamiliar territory
■ Capabilities and resources on hand to address	■ Capabilities or resources may not be in place
■ Fell between the cracks?	■ Major change in market or business
Just get on with it	Requires allocation of capital or shift in strategic direction

Proper classification clearly implies the appropriate risk-shaping action.

business territory (due, perhaps, to a major acquisition, a powerful new competitor or a significant change in customer buying patterns), or the organization lacks the skills necessary to respond. These are considered "strategic risk factors" and may require significant capital outlay and/or a major change in strategic direction.

Manageable risk factors in our experience include:

- "The R&D division is not keeping pace with the demand for new products."
- "Contingency planning is weak in the critical production facilities."
- "Mid-level employees are dissatisfied with their opportunities for advancement."

Strategic risk factors we have encountered include:

- "The share value is dependent on continuing uninterrupted earnings growth; this growth must come from top-line revenue growth; and opportunities for top-line growth are limited without branching out of the organization's product line and/or niche market."
- "Needed infrastructure changes clash with the current success formula and culture."

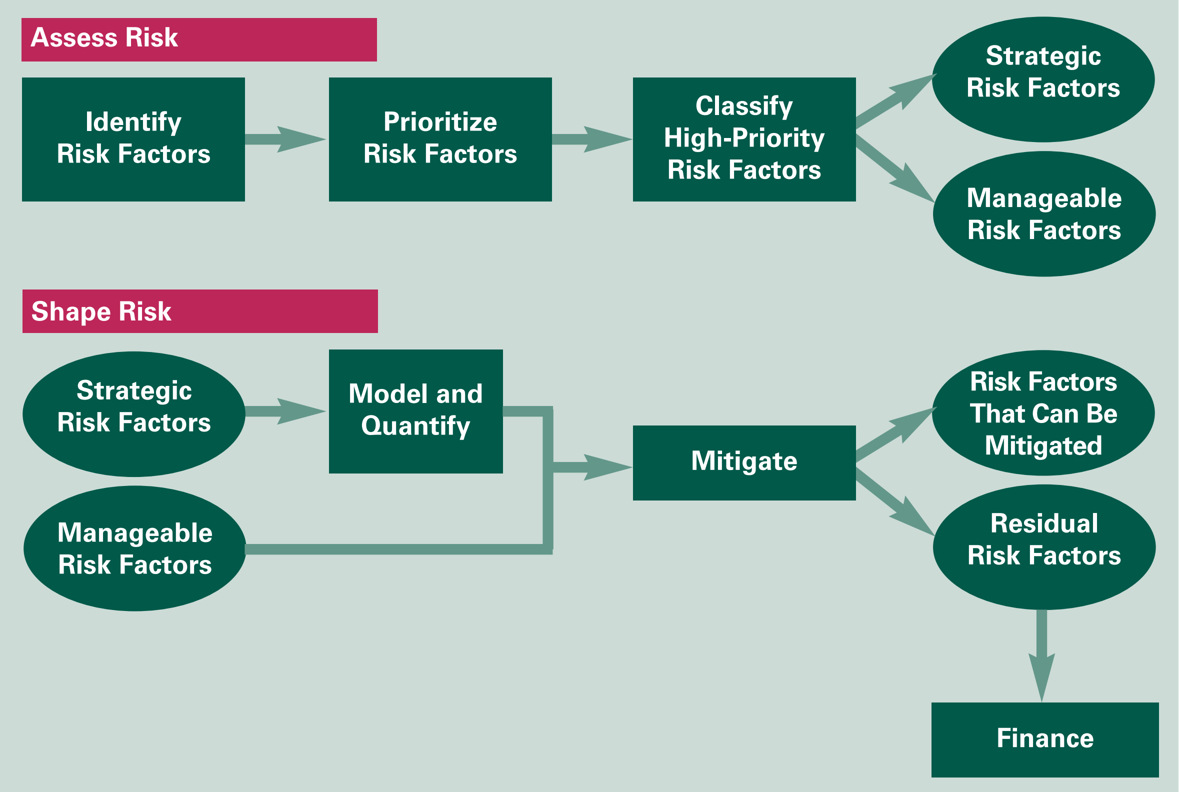
The proper response to manageable risk factors is to "just get on with it" — in other words, deal with them. The relevant skills already exist; they just need to be refocused on these high-priority items. Strategic risks, however, require greater analysis; this is covered in Chapter IV.

Recap… and segue

The steps described above are illustrated below (*Figure 5*). This graphic also illustrates the follow-on steps — the risk-shaping steps — that are the subject of the next chapter. The graphic demonstrates that not all risk factors need to be quantified and modeled, nor do all risk factors need to be financed. Risk factors needing quantification are those that pass through the "triple screen" — they are material, high-priority *and* strategic. Risk factors that need to be financed pass through the first two screens and cannot be fully mitigated through other means.

Underlying our approach to risk shaping — described in Chapter IV — is the premise that modeling, quantifying and formulating the strategy for mitigation and financing can be carried out simultaneously.

FIGURE 5



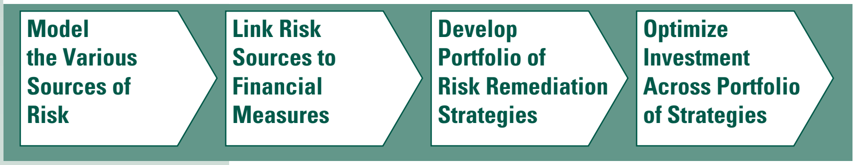
Triple screening in risk assessment creates efficiency in risk shaping.

IV A Scientific Approach to Shaping Risk

Overview

In this section, we will describe our approach to shaping risk and provide illustrations of its application. The approach to risk shaping relies heavily on Operations Research methods such as applied probability and statistics, stochastic simulation and portfolio optimization. To our knowledge, no organization has implemented this approach in its entirety as of the date of this publication, although we know of several that use portions of it in their incremental pursuit of ERM. (In Chapter VI, we describe how some of these organizations have gotten started.)

The Four Steps in Our Approach


Model the Various Sources of Risk
Link Risk Sources to Financial Measures
Develop Portfolio of Risk Remediation Strategies
Optimize Investment Across Portfolio of Strategies

In the first step, each source of risk is modeled as a probability distribution, and the correlation among the risk sources is determined. These probability distributions are typically expressed in terms of different operational and financial measures. The second step links these disparate distributions to a common financial measure (e.g., Free Cash Flow) through a stochastic financial model. These two steps represent the bulk of the analytical effort. At this stage, we have a holistic financial model of the business that can be used to:

- measure the volatility of the financial metric(s) under current operating conditions
- analyze the impact of risk management decisions through "what-if" scenarios.

The third step involves developing risk remediation strategies to be evaluated using the stochastic financial model. This basket of strategies represents a portfolio of risk management investment choices. In the final step, the ERM budget is allocated optimally across these strategies using portfolio optimization methods. Each step is described in greater detail below.

To illustrate this approach, we will introduce a hypothetical company (let's call it HypoCom) facing a broad array of strategic risks and show how the company would implement this approach in shaping these risks. Assume that HypoCom is a manufacturing company and has the following profile:

- Sells its product to retailers in the United States and Europe — with limited competition
- Has production plants in France, Mexico and Indonesia that deliver products to retailers through HypoCom's own distribution network
- Faces the following risks in the next fiscal year:
  - fire at a warehouse
  - volatility in the price of the raw materials used in the production process
  - possible employee union strike at the plant in France
  - possible new competitor entering the market.

While a real company, similar to HypoCom, would face many risks, we have limited their number here for the sake of simplicity. Please note, however, that the risks were selected to span those that are traditionally considered within the domain of risk management (hazard and commodity price risks) and those that are not (operational and competitor risks).

Again, to keep the example simple, we assume a one-year time horizon. At the end of this section, however, we discuss extending these steps to a more typical multi-period decision horizon.

STEP 1
Model various risk factors individually

Generate probability distributions

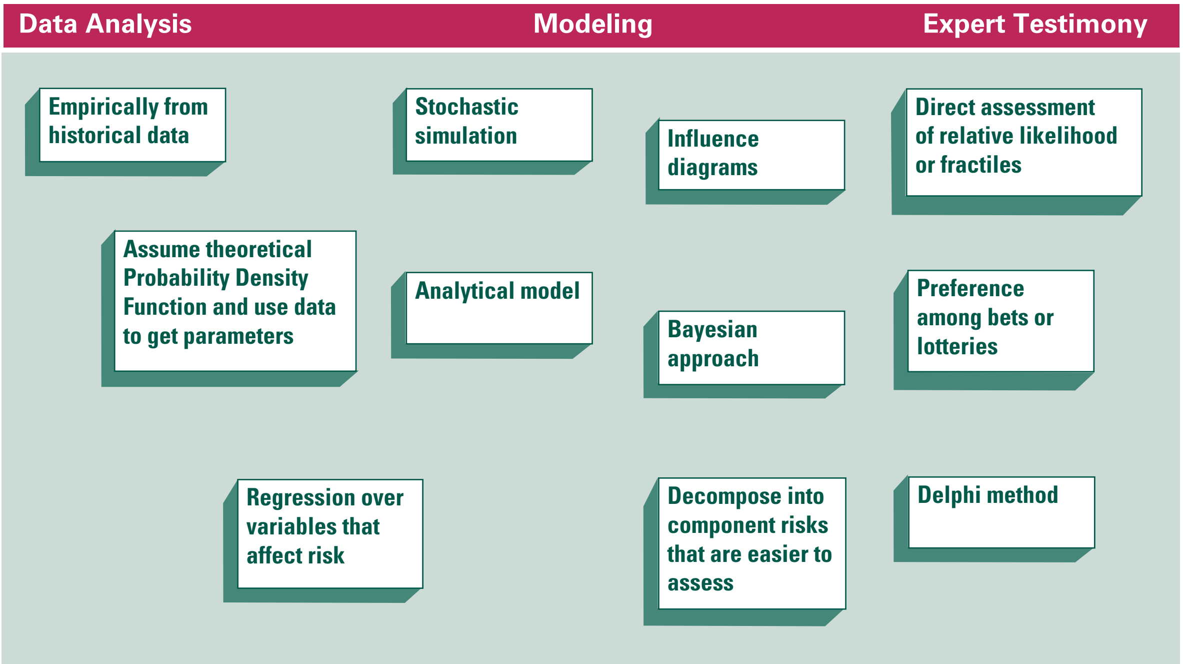
In Chapter III we outlined the approach for identifying which risk factors need to be modeled. Each risk factor contains uncertainty about how, when and to what degree it will manifest itself. This uncertainty is represented as a probability distribution. No one approach for developing probability distributions can be used for all the risks that an enterprise faces.

Risks that fall within the traditional domain of risk management — for instance, insurable risks or risks that can be hedged in the financial markets — are typically modeled using statistical methods that rely on the availability of historical data. However, when the domain is extended to enterprise-wide risks, it is unlikely that enough historical data exist to employ the same methods. Here, it is more likely that assessment of the uncertainty will be based entirely on expert testimony. Also, some risk sources will have to be modeled based on historical data combined with assumptions set by experts. Extending risk management to enterprise-wide risks suggests a continuum of methods for developing probability distributions. Such a continuum ranges from relying entirely on data to relying on expert testimony.

Figure 6 identifies methods for assessing probability distributions along this continuum. Readers of this monograph are likely to be familiar with methods based primarily on historical data (leftmost section of Figure 6). Therefore, instead of describing them, we have included references to source documents at the end of this monograph. At the opposite end of the continuum, there are formal methods developed and used by decision and risk analysts to elicit expert testimony for assessing uncertainty. We have provided brief descriptions of some of these in Appendix B. In the middle of the continuum, stochastic simulation modeling predominates for combining historical data and assumptions set through expert testimony. We will use this method to model the risk associated with an employee union strike at the HypoCom production plant in France.

(continued on page 16)

FIGURE 6


Data Analysis
Modeling
Expert Testimony
Empirically from historical data
Stochastic simulation
Influence diagrams
Direct assessment of relative likelihood or fractiles
Assume theoretical Probability Density Function and use data to get parameters
Analytical model
Bayesian approach
Preference among bets or lotteries
Regression over variables that affect risk
Decompose into component risks that are easier to assess
Delphi method

A continuum of methods for developing probability distributions ranges from those relying on data to those that rely on expert testimony. The positions of the methods identified above suggest which to use depending on the availability of data.

HypoCom – developing probability distributions for the four risks

Risk 1

Fire

A fire at a plant or warehouse can result in direct and indirect loss of sales volume. Direct losses result from destruction of inventory and work in progress. Indirect losses result from a prolonged interruption of production, through loss of short-term sales and perhaps through loss of market share. These risks have been insurable for a long time. Reliable methods exist for measuring the frequency and severity of losses based on review of historical data and business interruption worksheets. We will assume that for HypoCom, the frequency distribution is negative binomial and the severity distribution is lognormal (see references in Chapter VII for descriptions of these distributions).

Risk 2

Volatility in price of raw materials

Historical price data for commodities can be obtained from HypoCom's own purchase data or through financial markets if the commodity is traded on a futures exchange. Given the availability of data, several methods exist for developing the probability distribution. These are:

- Use empirical distribution
- Assume lognormal distribution using the sample mean and standard deviation
- Assume a stochastic process (e.g., jump diffusion) and use simulation to generate distribution of price movement.

An example of a stochastic process is the Schwartz-Smith two-factor model for the behavior of commodity prices (Schwartz & Smith 1999). The two-factor approach models both the uncertainty in the long-term trend and the short-term deviation from that trend.

For the sake of this example, we will assume that HypoCom faces a lognormally distributed price with a 2% standard deviation from the current price.

Risk 3

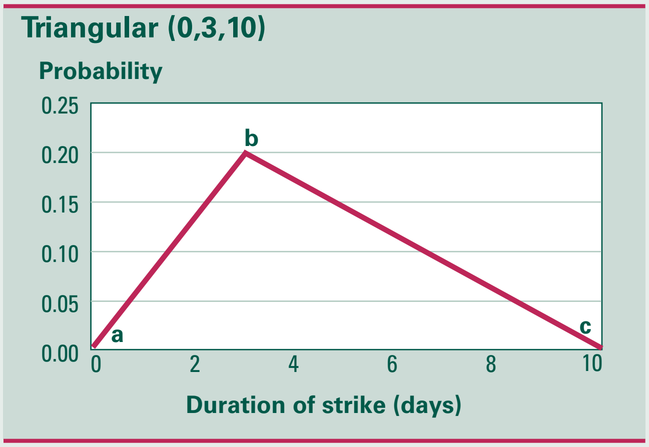
Employee union strike

An employee strike at the plant in France results in losses in sales volume. HypoCom services its European and U.S. markets from production at three plants (France, Mexico and Indonesia). This strike would result in a temporary shutdown of the plant in France. If the other two plants have capacity to increase production quickly enough to satisfy all demand, then there is little risk of loss in sales. But if all three plants are already running at high utilization (a more likely scenario), then the loss of one plant would result in longer lead times to market — the time from order placement to delivery. The strike would then affect HypoCom's ability to satisfy orders and lead-time commitments or expectations; this would result in a short-term loss of sales or possibly market share.

The probability distribution for the sales volume loss can be developed in three steps. First, determine the probability distribution for the length of the strike. It's quite likely that development of this distribution will have to be based almost entirely on expert testimony. As illustrated in Figure 6, there are several methods for assessing probabilities based on expert testimony: the Delphi method, eliciting preferences among bets or lotteries, and directly assessing relative likelihood or fractiles (see Appendix B for details on these methods). The labor relations manager(s) at HypoCom can be interviewed using one of these methods to determine the probability distribution for the length of the strike. For example, the result may be a triangular distribution as illustrated in *Figure 7*.

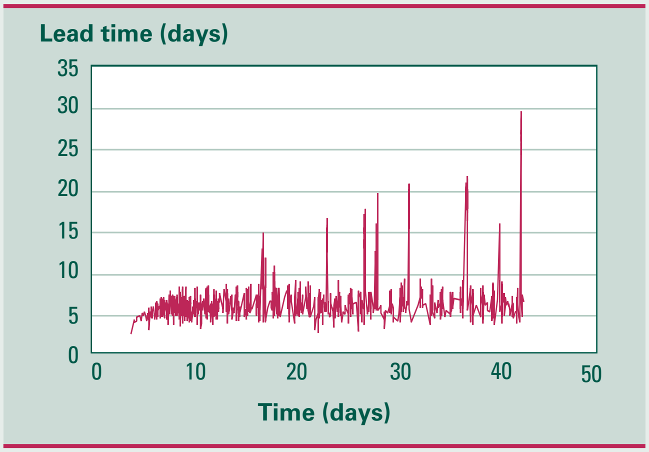
Second, develop a distribution on lead times conditioned on the length of the strike. We have developed a discrete-event stochastic simulation model of HypoCom's distribution network, using graphical, animated simulation software called ProModel®. The simulation modeled stochastic arrival of demand based on

FIGURE 7



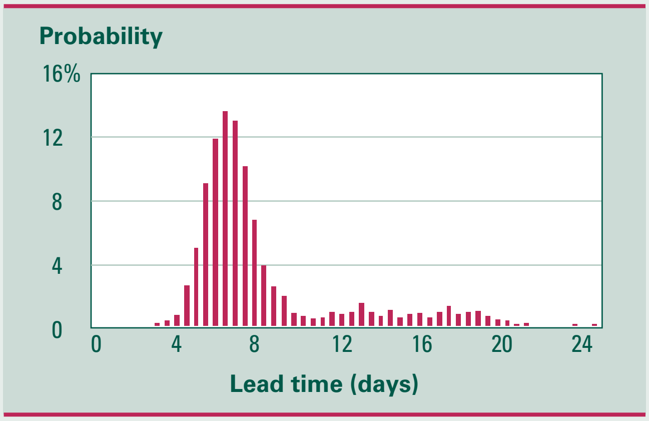
Triangular probability distribution with parameters minimum, mode and maximum (a, b and c, respectively). The expected value is (a+b+c)/3 and the standard deviation is $(a^2 + b^2 + c^2 – ab – bc – ac)/18$. This distribution is used often as a rough model when there is little historical data.

FIGURE 8



The chart shows the impact of a strike on lead times from one of the simulation runs. The strike starts on the 20th day and can last anywhere from 1 to 10 days, based on the probability distribution in Figure 7. You can see that the impact of the strike is felt long after the strike is over.

FIGURE 9



Discrete probability mass distribution generated from the lead-time data in Figure 8. The extended tail toward longer lead times is a consequence of an employee strike.

historical data, production rates at each of the plants and the logistics of distribution from the plant to regional distribution centers and then to retailers. It incorporated a distribution policy of supplying those distribution centers with the greatest backlog of orders. Inputs to this model are typically easy to get; in fact, many organizations already have a stochastic supply chain model used to optimize the logistics of their distribution network. The effect of the strike was simulated by shutting production at the plant in France and recording the increase in lead times. The chart of individual lead times in *Figure 8* is an output from a simulation run.

We usually run simulations a statistically valid number of times to attain a high level of confidence in the results. An empirical distribution of lead times based on these simulated data is shown in *Figure 9.*

Finally, determine the loss in sales conditioned on the increase in the lead times. With information in hand on the increase in the lead times, the sales and marketing managers at HypoCom would assess the effect on sales. One of the probability assessment methods for expert testimony described in Appendix B would be used here. The assessment would reflect contractual agreements with retailers as well as lead-time expectations and the competitive environment. So the final distribution on the decrease in the number of sales may be represented by a triangular distribution with parameters min. = 0, most likely = 4 million, max. = 10 million.

Risk 4
New competitor

Expert testimony provides the entire basis for the assessment of uncertainty associated with a new competitor. This process entails interviewing sales and marketing managers of HypoCom either individually or as a group. Any method described in Appendix B could be used here.

Here we develop a probability distribution on how new competition affects sales volume loss. It is helpful to dissect risk events into conditional causal events. For HypoCom, the causal events are illustrated in *Figure 10.*

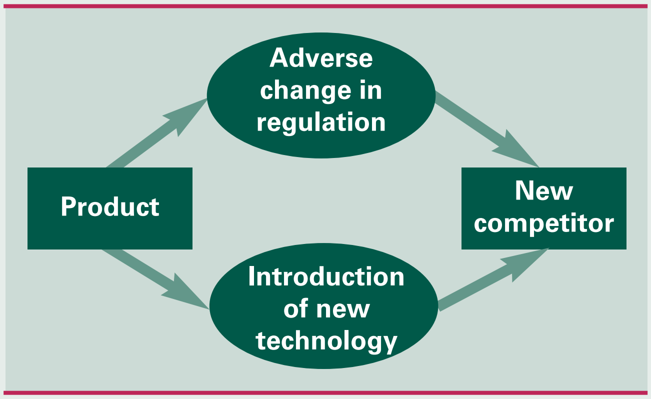
The probability of loss in sales volume due to competition, P(C), can be decomposed into:

$$P(C) = \Sigma_i \, P(C_i \mid R_i, T_i) \, P(R_i, T_i)$$

where i is the product index, $P(R_i, T_i)$ is the joint probability of an adverse change in regulation (R_i) and introduction of new technology (T_i) and $P(C_i \mid R_i, T_i)$ is the conditional probability of a loss in sales volume for product i due to new competition. If regulatory changes and introduction of new technology are not highly correlated, then $P(R_i, T_i)$ can be decomposed into the product of $P(R_i)$ and $P(T_i)$.

Instead of assessing P(C) directly, it is easier to ask different experts to assess the

FIGURE 10



Given the product, the possibility for change in regulation or introduction of new technology could influence the loss in sales due to competition.

conditional and joint probabilities. Company lobbyists are interviewed to assess the probability of adverse regulation for a specific product, $P(R_i)$, using one of two methods: preference among bets or judgment of relative likelihood (see Appendix B).

Managers of the Research and Development function are interviewed to assess the probability of introduction of new technology, $P(T_i)$. Finally, sales and marketing managers are interviewed to assess the probability of a new competitor, given the state of new regulation and technology, $P(C_i \mid R_i, T_i)$. Of course, experts may be interviewed as a group using the Delphi method (see Appendix B) instead of separately. This process is applied over all products of interest and the results summed according to the formula indicated above.

Determine correlation among risk sources

It is not enough to develop probability distributions on individual risk sources. One primary benefit of managing risks on an enterprise-wide basis is being able to take advantage of natural hedges and to explicitly reflect correlation among risks. Therefore, it is necessary to develop a matrix of correlation coefficients among pairs of risks that would be used in the next step to link the individual risk sources to a common financial measure.

It is unlikely that relevant data will exist to develop correlation among risks that span an enterprise. Thus, it is likely that this will have to be developed based on professional judgment and expert testimony. In some cases, it may be easier to develop correlations between risks implicitly by analyzing their correlation with a common linking variable. This process also ensures that a correlation matrix is internally consistent.

FIGURE 11

	Fire	Commodity Price	Union Strike	New Competitor
Fire	1.0	0.0	0.2	0.0
Commodity Price	0.0	1.0	0.0	-0.5
Union Strike	0.2	0.0	1.0	0.7
New Competitor	0.0	-0.5	0.7	1.0

Correlations among risks are modeled using correlation coefficients among risk pairs. For example, the risk due to commodity price fluctuations is negatively correlated with a new competitor entering the market.

For HypoCom, we would expect a negative correlation between the commodity price movements and a new competitor entering the market. If the commodity price increases, it creates a greater barrier to entry into the market for a new competitor and vice versa. However, a union strike is probably positively correlated with competition. Finally, there may be some slight correlation between a union strike and the incidence of fire.

It is unlikely that correlations would be determined with a high degree of precision. Rather, it is more likely that they could be judged in fuzzy terms such as high, medium or low. These terms suggest some natural ranges for correlation coefficients such as: high correlation = .70 to .80, medium correlation = .45 to .55, low correlation = .20 to .30. Within these ranges, there should be little sensitivity on the results. The inclusion of correlations should have a significant impact on the results, but the error within these ranges should have little impact. Using these as guides, a Correlation Coefficient Matrix can be developed for HypoCom as shown in *Figure 11.*

STEP 2

Link risk factors to common financial measures

Select financial metrics

The prior step provides a set of probability distributions representing enterprise-wide risks. Note that the probability distributions were expressed in terms of different units. We modeled the union strike as a probability distribution on lead time and then sales volume. Commodity price risk was modeled in terms of the price of raw materials. Other risks would be modeled in terms of the operational and financial measures that they directly affect. In this step, all these risks are combined and linked to one financial measure.

Managers of different organizations vary in their preference and propensity for the financial measures by which they manage the business. The financial measure will also vary depending on the objectives and goals of the organization. Above all, it is important that there is general agreement on the financial measure selected. For this document, we will use Free Cash Flow (FCF) to capture the impact of risk on both the income statement and balance sheet.

Develop a financial model to link risks to financial metric

Once a financial measure is selected, we can then model the aggregate impact of the sources of risk on the financial measure. We can construct a pro forma FCF model by decomposing each element in the calculation of FCF into its constituent metrics. See *Figure 12* for an illustration of this. The elements should be broken down to the level of the operational and financial measures used for modeling the individual risks in Step 1.

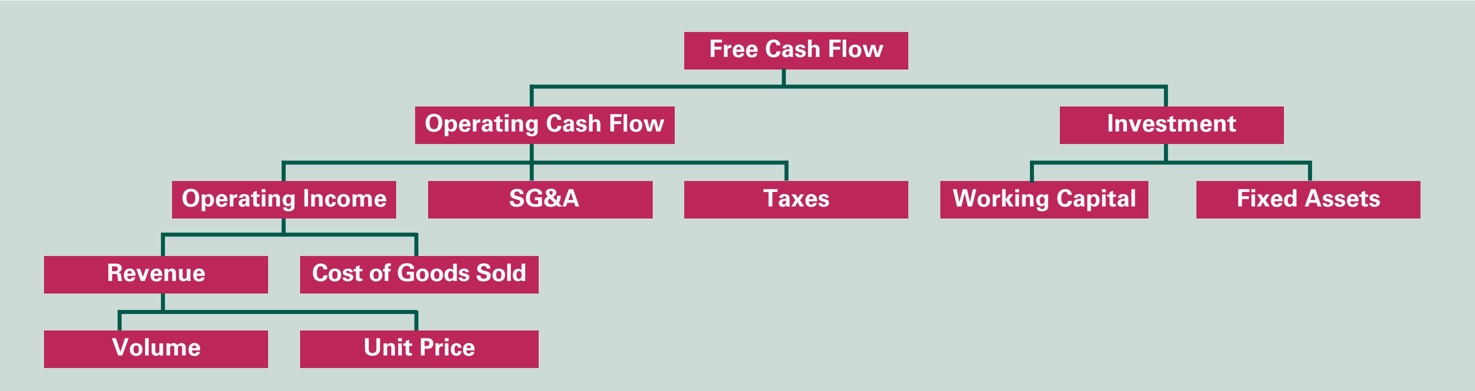
Some elements of the FCF model may be stochastic without consideration of the risks from Step 1. For example, there is some inherent uncertainty in product demand and price as well as cost of goods sold. These measures may fluctuate based on supply and demand economics. These inherent uncertainties are included in the base FCF model. The probability distributions from Step 1 are then added to the corresponding elements of the model. Finally, the Correlation Coefficient Matrix (from Step 1) is added to the model to reflect the interaction among the sources of risk. The resulting stochastic pro forma financial model links all the risks to FCF, the financial measure by which the risk remediation strategies will be evaluated in the next two steps.

Measure current level of enterprise risk before mitigation strategies

Before proceeding to risk remediation strategies, however, it is worth taking note of the value of the model thus far. At this point, we have a financial model that can be used to determine the current level of volatility in FCF. This information by itself would be extremely valuable in budgeting and financial planning. This analysis helps move managers' thinking away from the one-dimensional certainty of typical budgets and toward the range of possible outcomes and managing probable rather than definite outcomes.

(continued on page 21)

FIGURE 12



Free Cash Flow is decomposed into its elements: Operating Cash Flow and Change in Investment, which are further decomposed. Each element is broken down into its constituents until all operational and financial measures used for the distributions in Step 1 are isolated.

For HypoCom

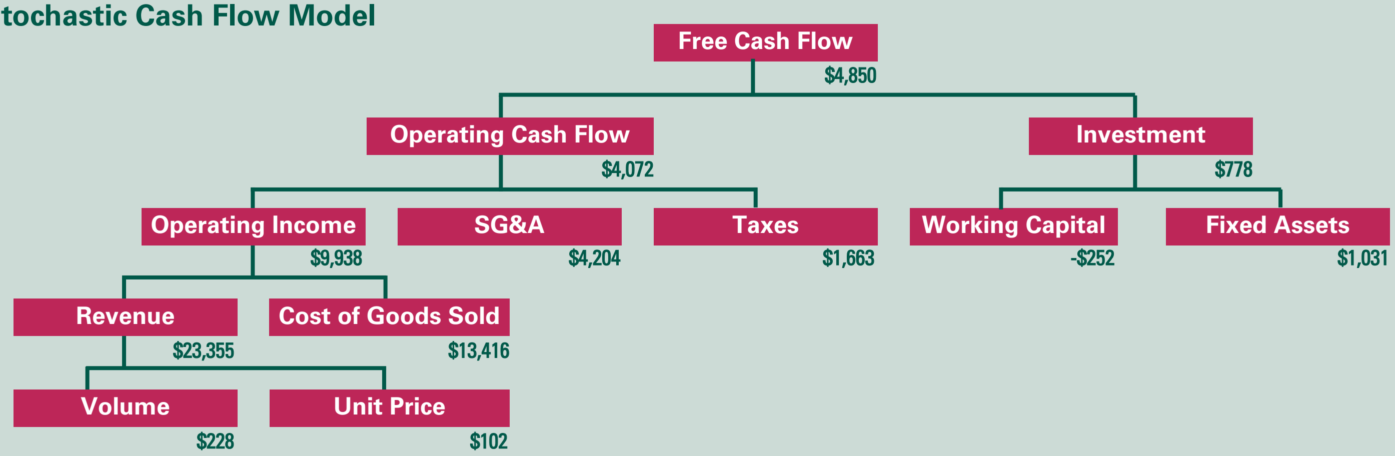
We developed an FCF model (see *Figure 13*). This model includes inherent uncertainty in volume, price and cost of goods sold. It also includes a correlation of -0.7 between volume and price, and a correlation of +0.5 between price and cost of goods sold before inclusion of the four risks from Step 1.

The fire risk effect on FCF was modeled by layering on the probability of loss in Volume developed in Step 1 (see *Figure 14A*). Also, an adjustment was made to Working Capital and Fixed Assets to reflect loss of inventory and the investment in rebuilding the plant destroyed by fire. The size of this adjustment was a function of the loss in Volume (i.e., the magnitude of the loss due to fire). The other risks were incorporated similarly — as shown in *Figures 14B, 14C* and *14D.*

(continued on page 20)

FIGURE 13

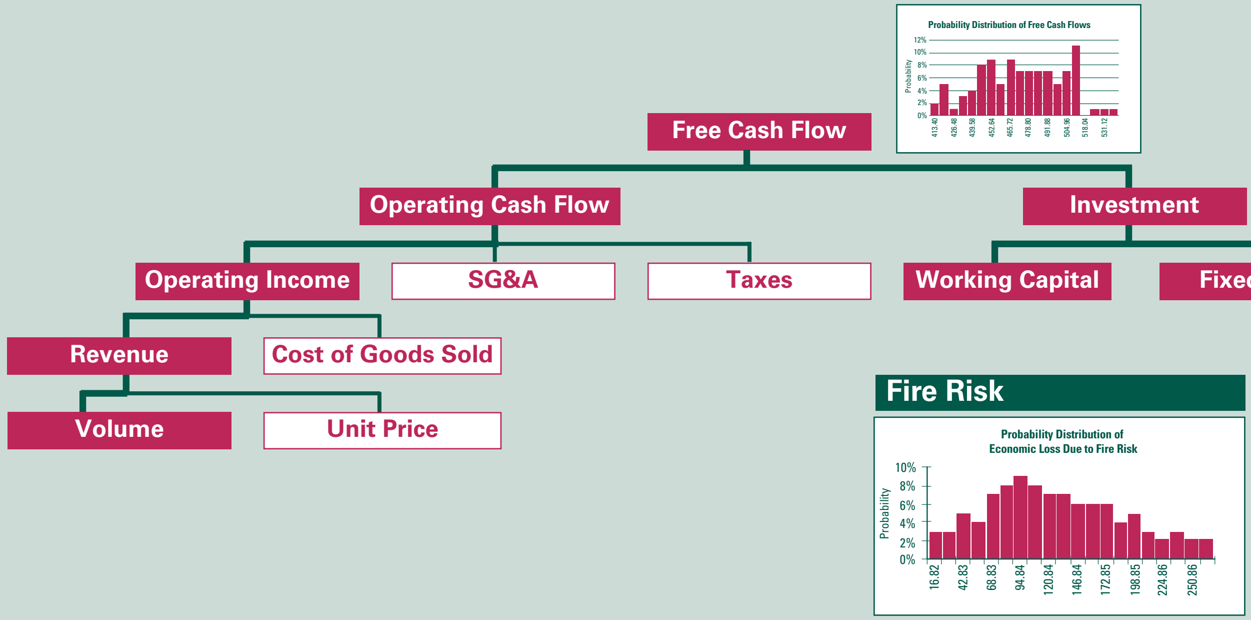
Stochastic Cash Flow Model



Stochastic Free Cash Flow for HypoCom. Volume, Unit Price and Cost of Goods Sold are represented as random variables with specified probability distributions and correlations.

Risk profiles are linked...

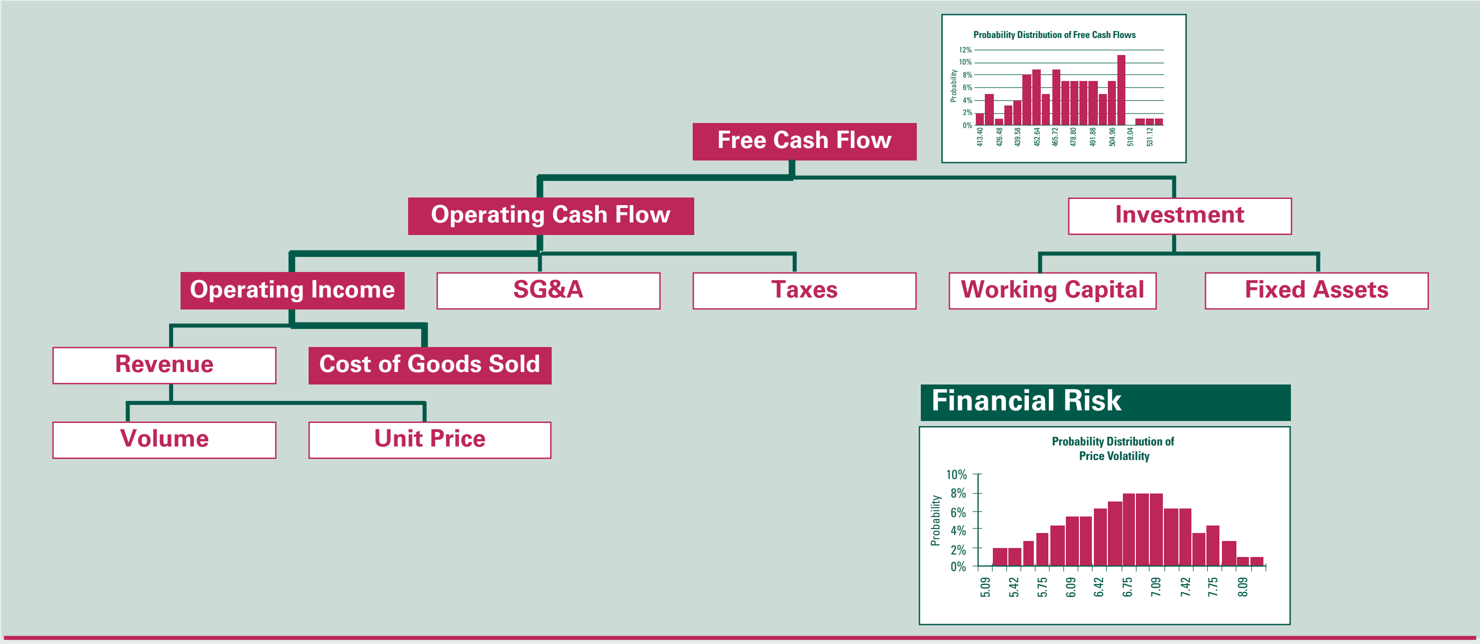
FIGURE 14A



The probability distribution for fire risk is linked to FCF through its effect on sales volume, working capital and fixed assets.

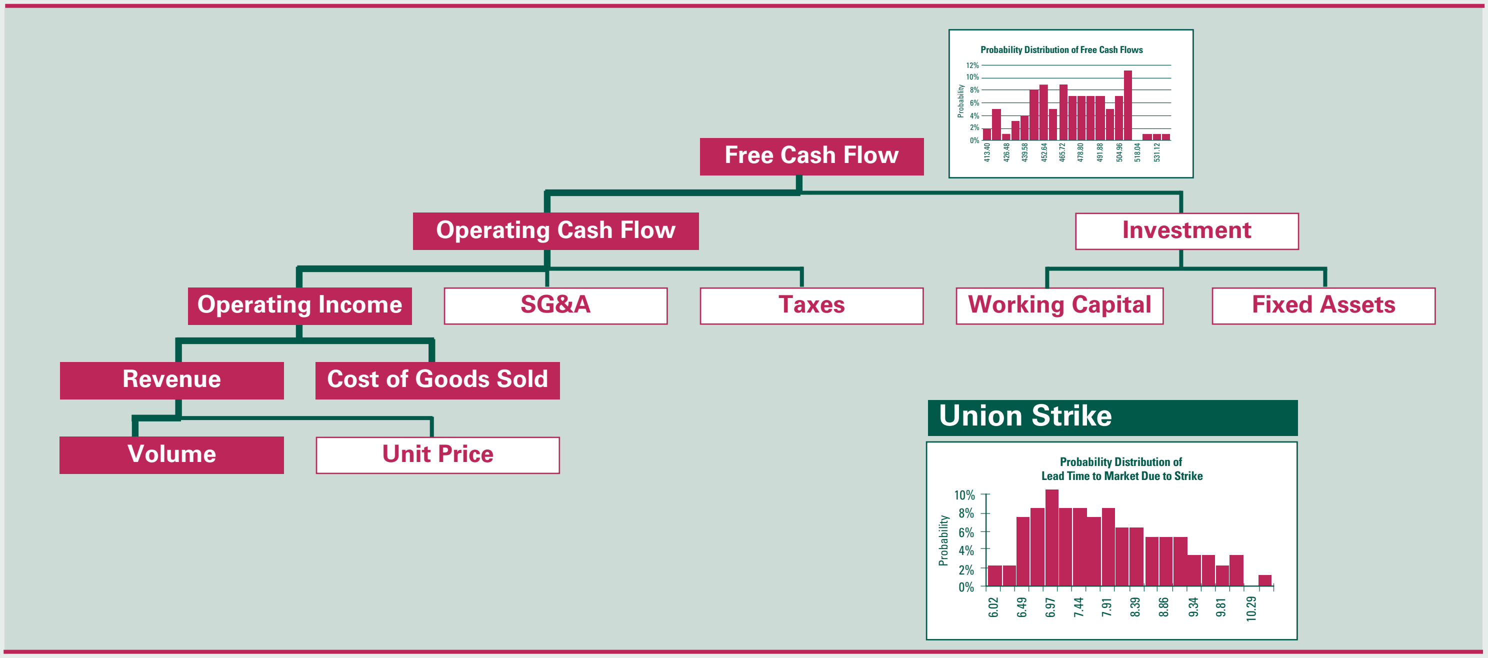
Risk profiles are linked... (cont'd)

FIGURE 14B



The probability distribution for commodity price risk is linked to FCF through its effect on cost of goods sold.

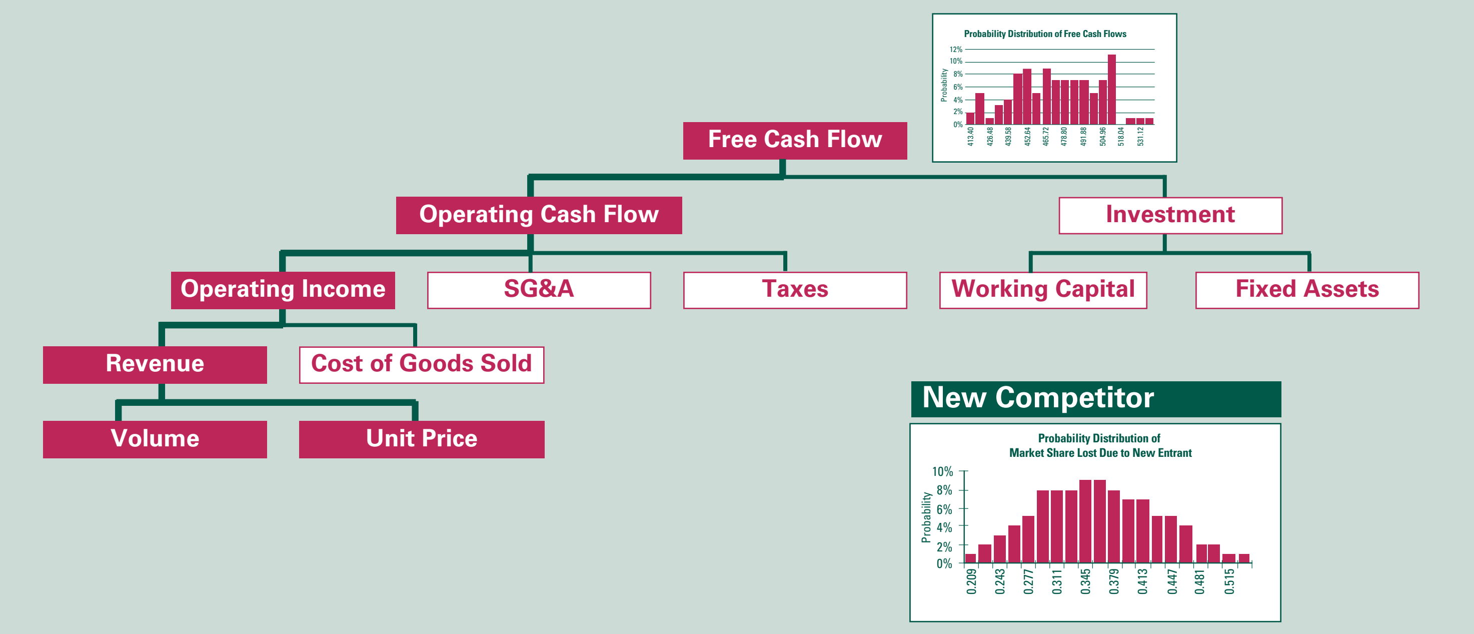
FIGURE 14C



The probability distribution for risk due to a union strike is linked to FCF through its effect on sales volume and cost of goods sold.

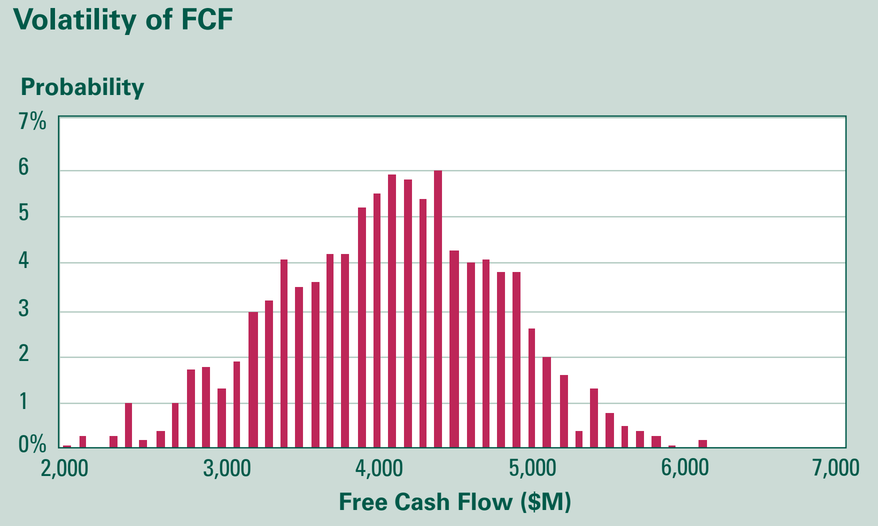
Risk profiles are linked... (cont'd)

FIGURE 14D



The probability distribution for new competitor risk is linked to FCF through its effect on sales volume and unit price.

FIGURE 15


Volatility of FCF

Volatility of Free Cash Flow for HypoCom. This reflects the aggregate impact of all four risks without inclusion of any remediation strategies.

The size of the FCF model and the number of risks modeled for HypoCom were kept small to simplify describing our approach. This way, we could construct this model in MS Excel™ and run simulations using @RISK™ software. However, in practice, models are built using specialized, industrial simulation and optimization software. The aggregate impact of all four risks on FCF is shown as a probability distribution in *Figure 15.*

STEP 3

Set up a portfolio of risk remediation strategies

The steps in the analysis thus far have produced information on the current level of risk for Free Cash Flow or any other financial measure selected for this analysis. Steps 3 and 4 outline a course of action to mitigate the current level of risk based on management's risk preferences. In Step 3, a portfolio of risk remediation strategies is developed as follows.

Identify risk remediation strategies

With a measure of riskiness of the FCF established, we can now determine how to reduce this risk. We can consult domain experts on strategies for mitigating each source of risk. This is a collaborative brainstorming effort among internal and external experts on the topic. Strategies are not restricted to financial remediation through insurance or financial derivatives; in fact, for many business risks, it may be impossible to find either insurance or a hedge in the financial markets. All the risk remediation strategies together constitute a portfolio of investment choices. To determine the optimal allocation of investment, the cost and benefit of each combination of strategies must be calculated.

Model effect of each strategy on financial metric

Each strategy aims to shape the risk on FCF to suit the risk preferences of management and shareholders. Shaping the risk means altering the shape of the probability distribution for FCF. At least three meaningful ways exist to shape the probability distribution:

- Shift the first moment of the distribution, i.e., increase the expected value of FCF.
- Shift the second moment of the distribution, i.e., decrease the deviations from the expected value of FCF.
- Reduce the tail of the distribution on the down side, i.e., reduce the worst-case scenario of Cash Flow-at-Risk (CFaR). This is a Value-at-Risk (VaR) type measure that is commonly used in financial risk management. For FCF, this means increasing the 5th percentile FCF so that there is less than 5% probability of FCF falling below some threshold value.

Each risk remediation strategy will affect the probability distribution of FCF in at least one of the three ways enumerated above. Thus, the measure by which the strategies should be evaluated will be a function of these three measures — described in greater detail in Step 4.

The FCF model from Step 3 measures the effect of each combination of strategies on the distribution of FCF. Simulations are run for each possible portfolio or combination of strategies and the resulting probability distribution of FCF is recorded for use in the next step.

Keep in mind that remediation strategies focused on mitigating the effect of one risk source may create a new source(s) of risk. For example, hedging in the financial markets may create counterparty risks. These unintended sources of risks should be incorporated into the financial model if they are deemed significant.

There is typically a cost associated with implementing each strategy, which can be measured directly. The cost may vary depending on the degree to which the strategy is undertaken. For example, various levels of insurance can be purchased, each with a different premium.

For HypoCom

Strategies for mitigating each risk appear in *Figure 16.* Note that for risks falling in the traditional domain of risk management — namely, fire risk and commodity price volatility — the strategies are also conventional, i.e., insurance and financial hedging, respectively. For mitigating the risk due to a union strike, however, there are several alternatives:

- build up inventory
- contract with third parties to provide a supply of products
- satisfy some or all union demands.

Like most manufacturing companies, HypoCom's distribution centers and plants optimize their inventory and production policies to minimize cost. However, the company did this without considering the impact of a union strike. As noted above, one alternative is to build up inventory beyond optimal levels; this would certainly mitigate the strike's impact. If there is no strike, however, the buildup of inventory beyond optimal levels creates a holding cost that can be calculated directly.

Similarly, each strategy alternative listed in Figure 16 has a cost that can be measured directly. The benefit of each strategy is determined through simulations using the FCF model. There are three alternative strategies each for mitigating fire risk, commodity price risk and union strike risk. Loss of sales due to new competition has only two possible strategies in our illustration. (Note that in each case, one of the alternatives is a default "do nothing" strategy.)

Altogether, there are 54 (3 x 3 x 3 x 2) possible combinations or portfolio strategies. Each of the 54 possible portfolios was evaluated by running simulations using the FCF model and recording the resulting probability distribution on FCF. The cost/benefit information for each portfolio produced in this step will be used in the next step to determine the optimal portfolio.

FIGURE 16

	Classification of Remediation Strategies		
	Insure	**Hedge in Financial Markets**	**Mitigate Through Business Activity**
Fire	■ Full range of loss ■ Catastrophic loss		
Commodity Price Volatility		■ Upside hedge ■ Full hedge	■ Acquire supplier of commodity
Union Strike			■ Build up inventory ■ Contract with third parties for product
New Competitor			■ Reduce price

Portfolio of risk remediation strategy alternatives for HypoCom. For each risk, there is also the default strategy of "do nothing."

STEP 4

Optimize investment across remediation strategies

This step takes the results from the prior steps to determine the optimal allocation of investment to the risk management portfolio. To do this, we must formulate the decision as a portfolio optimization problem and solve it using optimization technology. The following will describe how to formulate and solve this portfolio optimization problem.

Identify optimization objective(s)

To compare portfolios of different combinations of strategies for risk remediation, first determine the criteria for the comparison. In optimization terms, this is called the objective function.

As indicated in Step 3, the risk remediation strategies alter risk in at least three meaningful ways:

- increase the expected value of FCF
- decrease the deviation from the expected value of FCF
- increase the 5th percentile of FCF distribution (CFaR) so that there is less than 5% probability of FCF falling below some threshold value.

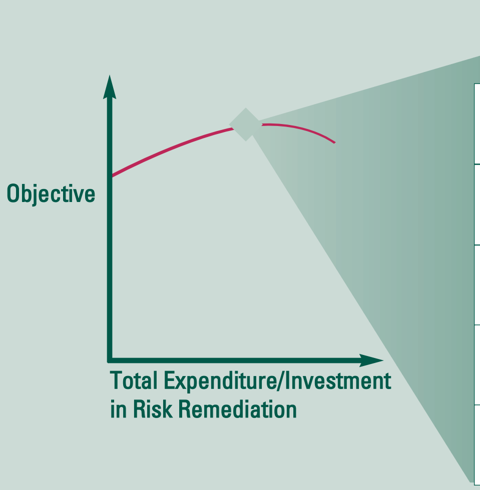
Therefore, one possibility is to use a weighted combination of these three measures as the objective function for comparing portfolios. The weightings would reflect the risk preferences of the decision-makers (who may be representing shareholder interest).

An alternative is to use expected utility of FCF as the objective function. First, a utility function must be developed that captures management's risk preferences for FCF. Development of a utility function is well documented in standard texts on decision analysis, two of which are included in the References (von Winterfeldt & Edwards 1986, Clemen 1996). The utility function is applied to the distribution of FCF to produce a distribution of utility or utiles. The expected value of this distribution is the expected utility. The relative preferences over the three measures of risk used in the prior method are captured in the shape of the utility function. One advantage of this method is that it easily extends to a multi-period objective using multi-attribute utility theory. This is explained further in a later section on multi-period risk management.

Either method can be used to develop the objective function of the portfolio optimization problem. The objective is to find the portfolio of strategies that maximizes this function.

Note that this method recognizes that management teams often differ in their risk preferences. We know that some companies are more aggressive than others in taking on strategic risks as a way of competing. Thus, the objective

FIGURE 17



	Insure	Hedge in Financial Markets	Mitigate Through Business Activity	Total
Fire	35%			35%
Increase in Commodity Price		10%		10%
Union Strike			25% Build up inventory 5% Contract with third parties	30%
New Competitor			25% Reduce price	25%
Total	35%	10%	55%	100%

The efficient frontier is a plot of all the portfolios that maximize the objective function given a fixed level of total risk remediation investment. Each point represents a unique allocation of the investment across the portfolio of strategies.

must be tailored to the unique risk preferences of the management team.

Identify constraints to optimization

Optimization may include some constraints on the optimum portfolio of strategies. A typical constraint may be a limit on the cost of implementing the portfolio of risk management strategies. There may also be constraints on the minimum/maximum level of insurance purchased, use of financial hedging, and/or the level of risk mitigated through business activity. Constraints on the downside risks to FCF may also be preferred. The constraints narrow the range of portfolios over which the objective function is maximized. Therefore, constraints have the effect of lowering the maximum value of the objective function.

Develop an efficient frontier of remediation strategies

The portfolio optimization problem as formulated above can be solved using optimization technology. Given a constraint on the size of the risk management budget, the optimization algorithms will determine the allocation of this budget to the alternative strategies that maximizes the objective function. This process can be repeated for varying levels of risk management budget. Plotting the results with the level of the risk management budget on the x-axis and the maximum value of the objective function on the y-axis produces a graph of the efficient frontier. The efficient frontier represents all the portfolios of strategies that constitute the optimal allocation of the risk management budget (see *Figure 17*).

For HypoCom

As mentioned at the end of Step 3, all 54 possible portfolios of strategies were simulated and the probability distribution of FCF was recorded. This information was then used to develop the objective function and the efficient frontier.

The objective function was based on a weighted combination of the three risk measures as follows:

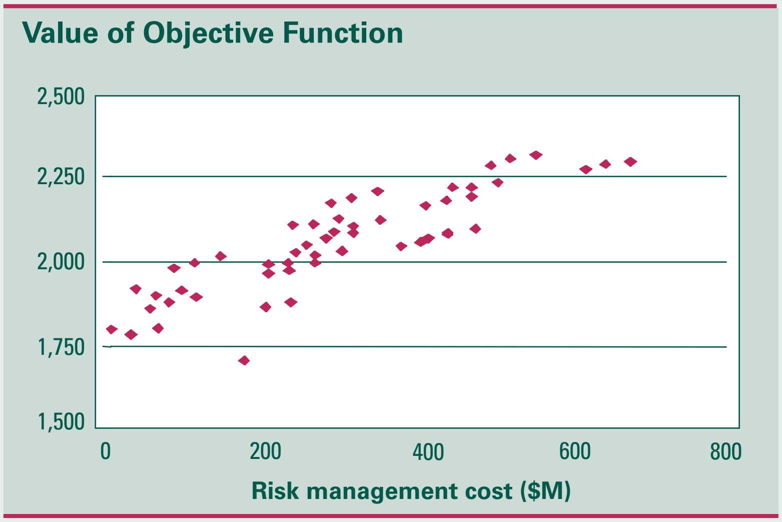
.40 * Expected FCF

+ .30 * Length of 90% confidence range of FCF

+ .30 * Value of FCF that has less than 5% probability of occurring.

Each of the 54 simulation runs produced a probability distribution of FCF. The objective function value was determined by applying the above formula to each of the runs. The results were plotted as an efficient frontier (see *Figure 18*).

FIGURE 18



Efficient frontier for HypoCom. Connecting all the points on top edge of the plot will produce an efficient frontier. Each point on the efficient frontier represents an optimum portfolio of strategies given the risk management cost. Portfolio points within the efficient frontier are suboptimal and should not be chosen.

Extension to multi-period risk shaping

Although the approach described above was based on a one-year decision horizon, in practice, most companies prefer a multi-year optimization analysis due to the strategic nature of this allocation. Fortunately, the method easily extends to a multi-year model.

In essence, all model variables and parameters are indexed by time (e.g., years). Thus, in Step 1, the probability distributions are developed for each time period in the investment horizon. Similarly, linking individual risks to a common financial measure involves indexing the probability distribution of FCF by year. Thus, the riskiness of FCF may vary from year to year.

The evolution of risk over time is typically modeled using a scenario generation system. A scenario generator uses stochastic differential equations (SDEs) to generate thousands of possible paths that a variable may follow over time. An SDE typically expresses a change in the value of a variable (e.g., interest rate) over a small time period as the sum of a predictable change and an unpredictable change. The predictable change is typically a deterministic function of the current value of the variable, but can also be a function of other variables with which there is correlation. The unpredictable effect is represented as a random variable with a specified probability distribution. An SDE is used iteratively to produce a scenario of how a variable can change over time. Typically, the scenario generator will model several correlated variables together to develop scenarios that are internally consistent. These scenarios are then fed into a financial model to develop stochastic forecasts of financial metrics over time. (Please refer to Section VII, "References and Recommended Reading," for papers and texts that describe scenario generation and stochastic differential equations.)

The risk remediation strategies in Step 3 may involve phased implementation of the strategy or there may be a time lag between incurring the cost for a strategy and its impact on the volatility of cash flow. In particular, the time lag may extend to more than a year.

Finally, in Step 4, the objective function based on expected utility can be extended to a weighted sum of the expected utility for each year in the time horizon. The weights applied to each year's expected utility can be determined by applying methods based on multi-attribute utility theory. Furthermore, budget constraints may vary over time.

In the multi-year time horizon, the output of the analysis is a path of risk remediation investments over the time horizon rather than separate optimum portfolios and efficient frontiers — as in the single-year case. Dynamic programming determines the optimum path of investments in risk remediation strategies.

Recap

In summary, the four-step analytical process for managing risk across an enterprise includes:

- quantifying each risk source by applying the appropriate tool and method for developing a probability distribution
- linking all the risk sources to a common financial metric
- developing a portfolio of strategies to mitigate each risk
- selecting the optimal portfolio of strategies.

The first two steps represent the bulk of the analytical effort and provide crucial information on the underlying dynamics of the enterprise. Different tools and methods (see Figure 6) for probability assessment will quantify the risk source and develop correlation among risk sources, depending on the relative availability of relevant data and domain experts. Aggregating these risks by linking them to a common financial metric provides an assessment of the overall risk to the enterprise and provides a method for determining the relative contribution of each risk source to the overall risk. Examination of the results of these two steps provides valuable insight into the business dynamics of the enterprise.

The last two steps are necessary to determine the optimal total expenditure for risk management and the most efficient allocation of that capital. Optimization also reflects constraints imposed by exogenous factors — the timing of expenditures, level of insurance, level of financial hedging and value-at-risk. In combination, the four-step analytical process lays a firm foundation for management decision making with respect to ERM.

V A Brief Discussion of Exploiting Risk and Keeping Ahead

Risk has two faces. This monograph has focused on risk as a threat. But risk also represents an opportunity. In fact, organizations routinely pursue risk for the chance of increased reward. Companies achieve competitive advantage by correctly identifying which risks the organization can pursue better than its peers.

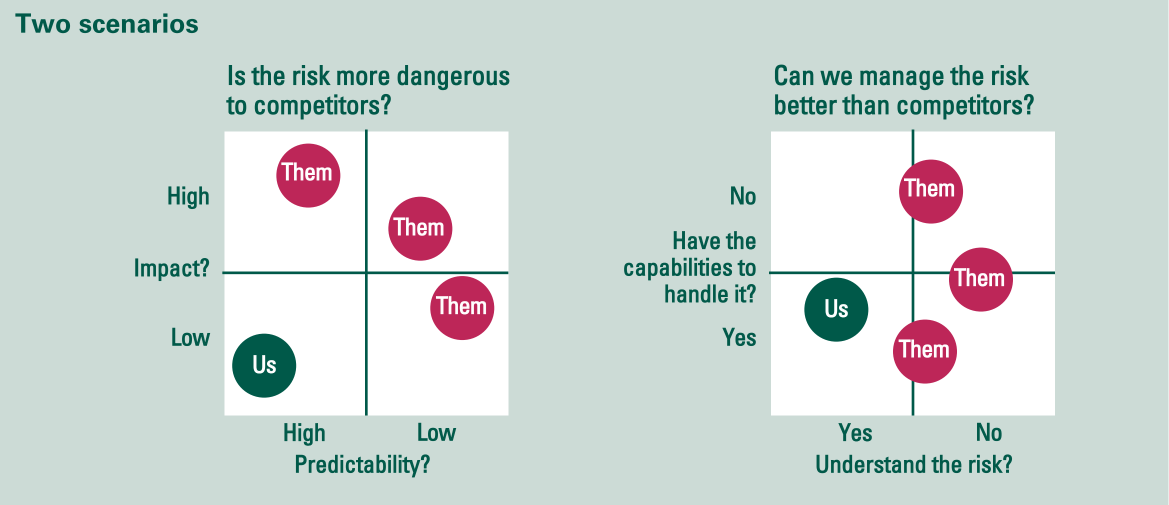
This advantage can arise in at least two ways (see *Figure 19*). The first relates to the nature of the risk itself. Certain risks, due to their predictability and/or effect on company financials, provide more of a risk to your competition than to your own organization. For example, currency translation risk is less of a concern to the organization whose distribution of cost of goods sold by country is similar to its distribution of revenue by country. The second way risk advantage arises relates to the organization's understanding of the risk and its capabilities to respond. For example, the oil company that, due to its hiring and training practices, has developed industry-leading capabilities in commodity risk analysis, can market these capabilities through a separate profit center.

A robust ERM assessment process will be alert to both faces of risk and will form the organization's strategic response accordingly.

In the dynamic risk environment, change is constant. It occurs in the organization's underlying risk factors, in the economic, political/regulatory and competitive landscapes within which the organization operates, and in the organization itself (e.g., its business objectives, the skill sets of its managers and key employees, and even its makeup after such events as downsizing, divestitures, mergers and acquisitions). Continual monitoring of this risk environment is therefore crucial if the organization's ERM program, however successful to date, is to remain relevant. Depending on the nature and degree of these inevitable changes, farseeing management reenters the ERM process at the appropriate step(s). Not surprisingly, several organizations make ERM an integral part of their business and strategic planning processes.

FIGURE 19

If You Understand Risk, It Can Be a Competitive Advantage


Two scenarios
Is the risk more dangerous to competitors?
High
Impact?
Low
Them
Them
Them
Us
High
Low
Predictability?
Can we manage the risk better than competitors?
No
Have the capabilities to handle it?
Yes
Them
Them
Us
Them
Yes
No
Understand the risk?

ERM includes identifying those risks that represent areas of competitive advantage.

VI Implementing ERM in Phases

Implementing ERM is clearly a challenge. Most organizations have therefore "started small," undertaking the implementation in discrete, manageable phases.

We can view ERM in three dimensions (see *Figure 20*). The first represents the range of company operations. Some organizations have started small by piloting ERM in one, or a small number, of their business units or locations, for real-time fine-tuning and eventual rollout to the entire enterprise. The second dimension represents the sources of risk (hazard, financial, operational, etc.). Some organizations confine the initial scope of their ERM to a selected subset of these risk sources, for example, property catastrophe risk and currency risk. Eventually, all sources of risk would be layered in, in sequential fashion.

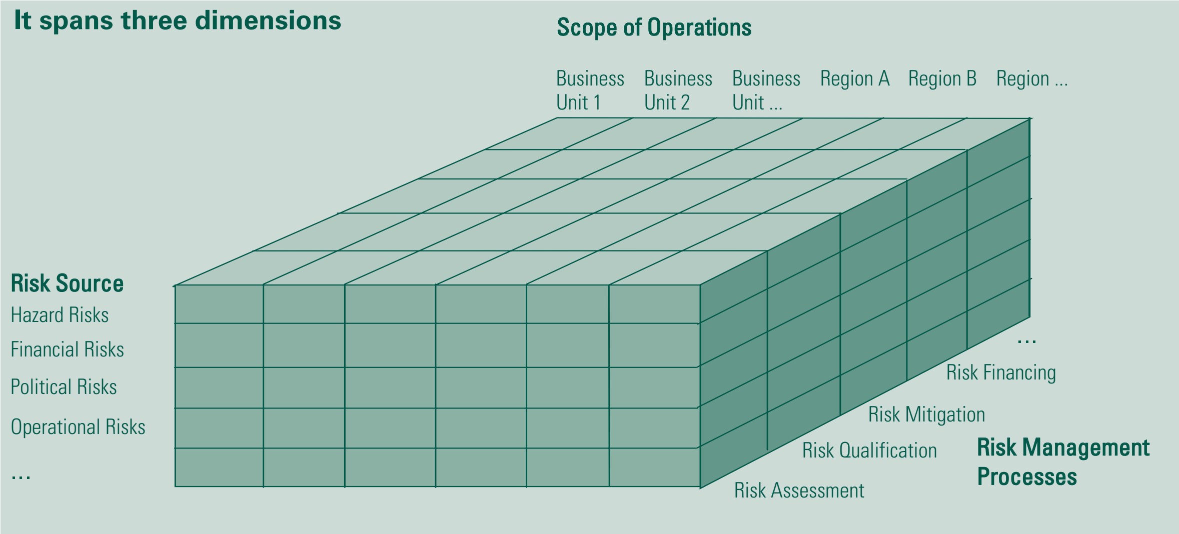
The third dimension represents the types of risk management activities or processes (risk identification, risk measurement, risk financing, etc.). Some organizations confine their initial vision to the identification and prioritization of enterprise-wide risks, with subsequent activities dependent on the results. Others begin by fashioning an integrated risk financing program around a subset of risk sources; these depend on the risk sources for which their financial service providers have integrated products. Still others begin by measuring and modeling virtually all sources of risk, regardless of their priority and the current availability of risk financing products.

While some of these approaches may appear more prudent than others, it is wise to reserve judgment. We believe no single best approach to ERM implementation exists that is appropriate for all organizations. Leading companies successfully employ a number of different phased approaches. The nature and sequence of these phases depend on the culture, strategic imperatives and management style of the organization. However, it is certain that for every organization a phased approach of some sort will be more successful than attempting to do too much, too soon.

Regardless of their starting point, many organizations include in their implementation plans the attempt to ingrain ERM into their cultures through communication, education, training and incentive programs. In some cases, these are coordinated in an extensive formal change management process to help impose the new order of things and achieve sustainable results. Clearly, to be successful, ERM needs to be more than a technique — and needs to be embraced by more than just management. These issues will be explored further in our subsequent publications.

FIGURE 20

The Universe of ERM Is Quite Large...



The scope of ERM is quite large. Organizations have variously "started small" by phasing in their implementation along one or more of ERM's three dimensions.

VII References and Recommended Reading

General risk management

Bernstein, Peter L. (1996). *Against the Gods: The Remarkable Story of Risk*. John Wiley & Sons.

Knight, Rory F. & Pretty, Deborah J. (1998). *The Impact of Catastrophes on Shareholder Value*. The Oxford Executive Research Briefings. A Research Report Sponsored by Sedgwick Group.

Probability assessment and utility theory

Clemen, Robert T. (1996). *Making Hard Decisions*. 2nd edition. Duxbury Press.

Linstone, H. A. & Turoff, M. (1975). *The Delphi Method: Techniques and Applications*. Addison-Wesley Publishing Company Inc.

Oliver, R. M. & Smith, J. Q. (1990). *Influence Diagrams, Belief Nets, and Decision Analysis (Wiley Series in Probability and Mathematical Statistics)*. John Wiley & Sons.

von Winterfeldt, D. & Edwards, W. (1986). *Decision Analysis and Behavioral Research*. Cambridge: Cambridge University Press.

Scenario generation and stochastic differential equations

Merton, Robert C. (1992). *Continuous-Time Finance*. 2nd edition. Blackwell Publishers, Inc.

Mulvey, John M. (1996). "Generating Scenarios for the Towers Perrin Investment System." *Interfaces, An International Journal of INFORMS*. March-April 1996, Vol. 26, Number 2.

Neftci, Salih N. (1996). *An Introduction To The Mathematics of Financial Derivatives*. Academic Press.

Schwartz, Eduardo & Smith, James E. (1999). *Short-Term Variations and Long-Term Dynamics in Commodity Prices*. Duke University, Fuqua School of Business.

Simulation and optimization

Bazaraa, Mokhtar S., Sherali, Hanif D. & Shetty, C. M. (1993). *Nonlinear Programming, Theory and Algorithms*. 2nd edition. John Wiley & Sons, Inc.

Fourer, R., Gay, D. M. & Kernighan, B. W. (1993). *AMPL, A Modeling Language for Mathematical Programming*. Duxbury Press (includes disk with AMPL and optimization solvers).

Hillier, Frederick S. & Lieberman, Gerald J. (1986). I*ntroduction to Operations Research*. 4th edition. Holden-Day, Inc., Oakland, California; Cambridge: Cambridge University Press.

Nelson, Barry L. (1995). *Stochastic Modeling: Analysis & Simulation*. McGraw-Hill, Inc.

Simulation and optimization software

@RISK and @RiskOptimizer, MS Excel add-in for simulation and optimization, developed by Palisade Corp., Newfield, New York.

Fourer, R., Gay, D. M. & Kernighan, B. W. (1993). *AMPL, A Modeling Language for Mathematical Programming*. Duxbury Press (includes disk with AMPL and optimization solvers).

Global CAP:Link, scenario generator for macroeconomic variables worldwide (such as interest rates, exchange rates and major asset classes), developed by Tillinghast – Towers Perrin, New York.

Service Model, discrete-event stochastic simulation software, developed by Promodel Corp, Orem, Utah.

VIII

Acknowledgments

The authors wish to acknowledge the invaluable assistance of their colleagues:

- **Robert Schneier** and **David Watkins,** who shared their expertise and ideas on strategy, organization and corporate culture — and their connection to ERM
- **Andrew Berry** and **Julian Phillips,** who contributed the corporate governance research
- **Peter Watson** and **Brian Merkley,** who analyzed the linkage between performance consistency and share value
- **Kathleen Waslov,** who contributed information on probability assessment methods based on expert testimony
- **Ravin Jesuthasan** and **Emory Todd,** who contributed to the development of the stochastic financial model
- **Anne McKneally,** who helped edit the text.

We also extend our heartfelt appreciation to the following individuals who constituted our Editorial Review Board:

- **Stephen D'Arcy,** Professor of Finance, University of Illinois
- **William Fealey,** Executive Director, Corporate Risk Management, Estée Lauder Companies, Inc.
- **Felix Kloman,** Editor & Publisher, Risk Management Reports
- **André-Richard Marcil,** Director, Control and Integrated Risk Management, Hydro Quebec
- **Robert Markman,** Vice President, Risk Management, United Parcel Service
- **John Mulvey,** Professor of Engineering & Management Systems, Princeton University.

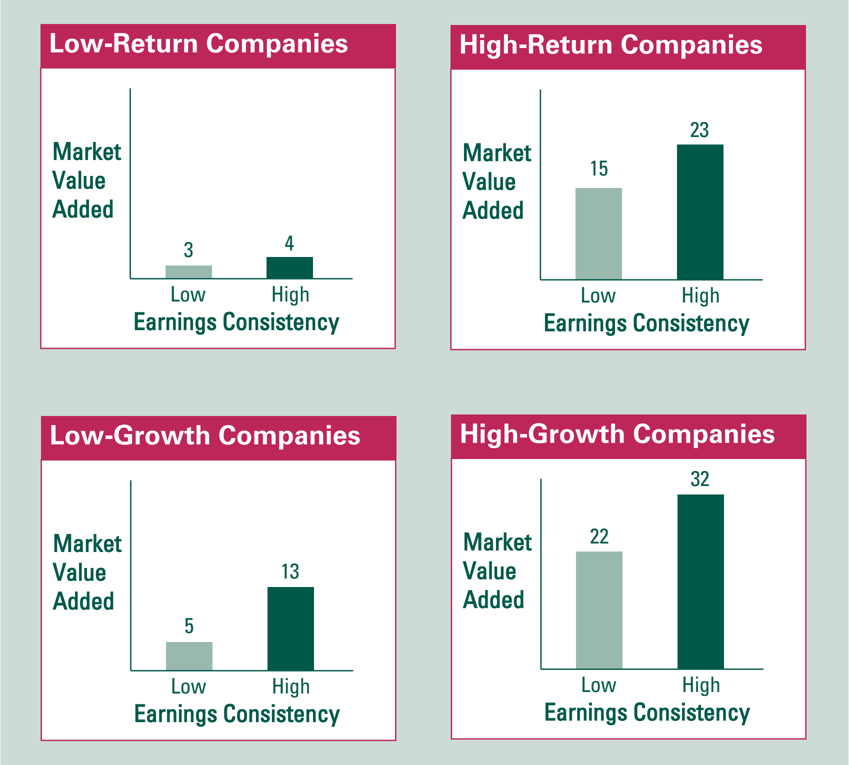
This monograph was immeasurably improved by the generous contributions of these individuals. Any errors that readers may find remain the sole responsibility of the authors.

Appendix

A

The Value of Consistency

- Earnings consistency typically explains 25% of annual change in share price
- Primarily affects premium over "warranted" multiple. Example (from the Integrated Petroleum Industry):



The market reacts to perceptions of how well risk is handled.

Source: Towers Perrin consistency analysis of selected industries (see following background information).

Background Information on Towers Perrin Consistency Analysis

Overview

Consistency analysis empirically estimates whether companies with more consistent earnings receive a premium market valuation relative to peers. Since many other factors — in addition to earnings consistency — shape market valuations, we use a series of basic analytic steps to attempt to control for the influence of other factors (e.g., earnings growth and return on capital) and isolate a consistency premium or discount. We use a relatively simple control process since (1) we find that more complicated methods introduce other sources of "noise" into the process and (2) consistency premiums are fairly robust across many industry groups and emerge readily with relatively simple control techniques.

A general description of the control process is provided below. For specific definitions and data sources used in the analysis, please see the Methodology section that follows.

Basic methodology

In performing consistency analysis, Towers Perrin's first step is to identify a relevant industry peer sample for a given company. Using an industry peer group helps filter out the effect of common industry factors (e.g., commodity price movements, regulatory risk) on market valuations. We typically use published industry groupings provided by Valueline or Standard & Poor's.

Next, we create a data set including a market premium measure, earnings growth rate, return on capital and earnings consistency for each peer. We employ historical growth rates and returns as surrogates for the future growth rates and returns that drive valuations. We calculate growth rates, using a least squares (regression) approach to avoid biases caused by point-to-point methodology, and average returns on capital over the measurement window (typically 10 years). To measure the market premium, we employ a standardized market value-added metric since it properly distinguishes between the capital that investors have placed in the business and the market value added to this capital.

Unlike market-to-book ratios, standardized market value added also captures the dollar growth in the value premium over time. Since the measure is standardized (indexed), it can be meaningfully compared across companies. Finally, Valueline's earnings predictability score (0%-100%) is used as the measure of earnings consistency.

We then calculate a median growth rate and return on capital for the peers and break the sample into "high growth" (growth ≥ median) and "low growth" (growth < median) and high-return (return ≥ median) and low-return (return < median) subsets.

The process is repeated one more time by calculating the median earnings predictability score for each of the four subsets and then further breaking each subset into a high earnings consistency (earnings predictability ≥ subset median) and low earnings consistency (earnings predictability < subset median). A total of eight subsets results from both steps.

Finally, an average market premium (standardized market value added) is calculated for each of the eight subsets, and the results are summarized in bar chart form.

Towers Perrin Consistency Analysis Methodology

Data Sources

- **Compustat PC Plus database**
- **Valueline Investment Survey (earnings consistency only)**

Performance Metric Definitions

"Return on Capital"

- **Definition**
  - 10-year (1989-98) average Return on Capital Employed (ROCE)
- **Formula**
  - (Income before Extraordinary Items + Special items) (Beginning Stockholders' Equity + Beginning Total Debt)
  - Perform same calculation for 10 years and take average
- **Comment**
  - Simplified return on invested capital definition (provides some adjustment for restructuring charges and other one-offs but makes simplifying assumption that special items receive no tax deduction)
  - Note: Compustat does not report after-tax special items

"Earnings Growth"

- **Definition**
  - 10-year (1989-98) least-squares EBIT growth rate
- **Formula**
  - Regress log adjusted operating income after depreciation against time to determine growth rate
- **Comment**
  - Growth rate based on regression more accurate than CAGR (which is biased by endpoints)

"Earnings Consistency"

- **Definition**
  - Valueline Earnings Predictability score as reported in Valueline Investment survey
- **Formula**
  - Valueline earnings predictability scoring based on stability of year-to-year comparisons, with recent years being weighted more heavily than earlier ones. The earnings stability is derived from the standard deviation of the percentage changes in quarterly earnings over an eight-year period. Special adjustments are made for comparisons around zero and from plus to minus.

"Market Premium"

- **Definition**
  - 1998 Standardized Market Value Added (MVA) based on 1988 ending invested capital base
- **Formula**
  - Std MVA = MVA % Capital x Indexed Capital = (M/C - 1) x Indexed Capital
  - M/C = (Stock price * Common shares outstanding + Preferred stock + Total debt)/(Shareholders' equity + Total debt)

    — All data reflect year-end 1998
  - Indexed Capital = (1998 Shareholders' equity + 1998 Total debt)/(1988 Shareholders' equity + 1988 Total debt)
- **Comment**
  - MVA captures value of growth (unlike M/B ratio) since it is measured in dollars. Standardizing MVA (by indexing every company's capital to same base year) corrects size bias of measure (so big companies with lots of capital but low M/C don't dominate smaller companies with higher M/C).

Appendix B

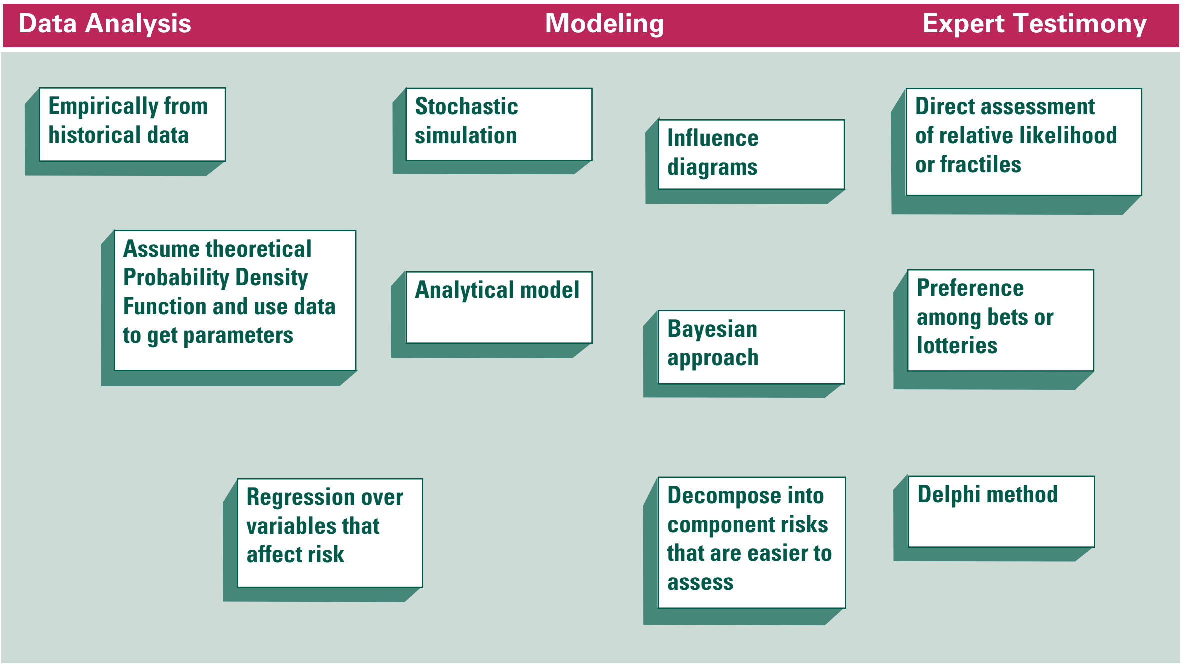
Probability Assessment Methods Based on Expert Testimony

Approaches to modeling risk

To model risk, it is necessary to understand the nature of risk itself. Risk arises from the fact that actual future results could differ from expected or projected results, often materially; one does not know with certainty what will happen in the future. In projecting into the future, one must consider a range of potential outcomes from a given event. Risk assessment aims to evaluate both the impact (financial, reputational, etc.) of each outcome and the likelihood or probability of each outcome occurring. The process develops a probability distribution that captures the impact and likelihood of given risk types or events.

There is a continuum of methods for developing probability distributions. These methods can be grouped into three principal categories: data analysis approaches, expert testimony and modeling (whose methods are often hybrids of methods from the other two categories). The choice of method depends significantly on the amount and type of historical data that are available. The methods also require varying analytical skills and experience. Each method has advantages and disadvantages over the other methods, so it is important to match the method to the facts and circumstances of the particular risk type.

Building a probability distribution of outcomes for each risk type is the first stage in developing an entire risk profile for the organization. In financial terms, each of these distributions needs to be combined with the others — taking into account correlations among risk types — and applied to the organization's financial value tree to develop a unique probability distribution of future financial results for that organization.


Data Analysis
Modeling
Expert Testimony
Empirically from historical data
Stochastic simulation
Influence diagrams
Direct assessment of relative likelihood or fractiles
Assume theoretical Probability Density Function and use data to get parameters
Analytical model
Bayesian approach
Preference among bets or lotteries
Regression over variables that affect risk
Decompose into component risks that are easier to assess
Delphi method

Estimating probabilities through expert testimony

Probability distributions for events for which there is sparse data can be estimated through expert testimony. A naive method for assessing probabilities is to ask the expert, e.g., “What is the probability that a new competitor will enter the market?” However, the expert may have difficulty answering direct questions and the answers may not be reliable.

Behavioral scientists have learned from extensive research that the naive method can produce unreliable results due to heuristics and biases. For example, individuals tend to estimate higher probabilities for events that can be easily recalled or imagined. Individuals also tend to anchor their assessments on some obvious or convenient number resulting in distributions that are too narrow. (See Clemen 1996 and von Winterfeldt & Edwards 1986 in the list of references for further examples.) Decision and risk analysts have developed several methods for accounting for these biases. Several of these methods are described below.

Preference among bets

Probabilities are determined by asking the expert to choose which side is preferred on a bet on the underlying events. To avoid issues of risk aversion, the amounts wagered should not be too large. For example, a choice is offered between the following bet and its opposite:

Bet	Opposite Side of Bet
Win $x if a competitor enters the market Lose $y if no new competition	Lose $x if a competitor enters the market Win $y if no new competition

The payoffs for the bet, amounts $x and $y, are adjusted until the expert is indifferent to taking a position on either side of the bet. At this point, the expected values for each side of the bet are equal in the expert’s opinion. Therefore,

$$\$x\ P(C) - \$y\ (1-P(C)) = -\ \$x\ P(C) + \$y\ (1-P(C))$$

where P(C) is the probability of a new competitor entering the market. Solving this equality for P(C):

$$P(C) = \$y/(\$x + \$y)$$

For example, if the expert is indifferent to taking a position on either side of the following bet:

Win $900 if a competitor enters the market

Lose $100 if no new competition

then the estimated subjective probability of a new competitor entering the market is $100/($100 + $900) = 0.10.

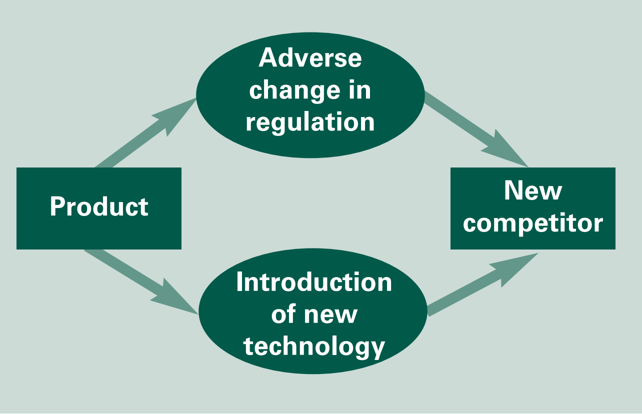
Judgments of relative likelihood

This method involves asking the expert to provide information on the likelihood of an event relative to a reference lottery. The expert is asked to indicate whether the probability of the event occurring is more likely, less likely or equally likely compared to a lottery with known probabilities. Typically, a spinning wheel (a software implementation of the betting wheels in casinos) is used on which a portion of the wheel is colored to represent the event occurring. The relative size of the colored portion is specified. The expert is asked to indicate whether the event is more, less or equally likely to occur than the pointer landing on the colored area if the wheel was spun fairly. The colored area is reduced or increased as necessary depending on the answers until the expert indicates that the two events are equally likely. This method is often used with subjects who are naive about probability assessments.

Decomposition to aid probability assessment

Often, decomposing an event into conditional causal events helps experts assess risk of complex systems. The structure of the conditional causal events can be represented by an influence diagram. Influence diagrams illustrate the interdependencies between known events (inputs), scenarios and uncertainties (intermediate variables) and an event of interest (output). An influence diagram model comprises risk nodes representing the uncertain conditions surrounding an event or outcome. Relationships among nodes are indicated by connecting arrows, referred to as arcs of influence. The graphical display of risks and their relationships to process components and outcomes facilitates visualization of the impacts of external uncertainties.

While this approach increases the number of probability assessments, it also allows input from multiple experts or specialists and helps combine empirical data with subjective data. For example, a new competitor entering the market may be decomposed using an influence diagram such as this one:



The probability of a new competitor, P(C) can be estimated, using a Bayesian approach. The approach uses Bayes' Rule, which is a formal, optimal equation for the revision of probabilities in light of new evidence contained in conditional or causal probabilities.

$$P(C) = \Sigma_i \, P(C_i \mid R_i, T_i \,) \, P(R_i, T_i)$$

where i is a product index, $P(R_i, T_i)$ is the joint probability of an adverse change in regulation and introduction of new technology, and $P(C_i \mid R_i, T_i)$ is the conditional probability of a new competitor entering a market for product i. This formula is useful when assessing the conditional probabilities $P(C_i \mid R_i, T_i)$ and is easier than a direct calculation of P(C).

Several different experts may be asked to assess the conditional and joint probabilities. For example, one expert (or group of experts) may assess the probability of adverse regulation for a specific product, another expert may assess probability of introduction of new technology, and yet a third may assess the probability of a new competitor given the state of new regulation and technology.

The Delphi technique

Scientists at the Rand Institute developed the "Delphi process" in the 1950s for forecasting future military scenarios. Since then it has been used as a generic strategy for developing consensus and making group decisions, and can be used to assess probabilities from a group of individuals. This process structures group communication and usually involves anonymity of responses, feedback to the group as collective views, and the opportunity for any respondent to modify an earlier judgment. The Delphi process leader poses a series of questions to a group; the answers are tabulated, and the results are used to form the basis for the next round. Through several iterations, the process synthesizes the responses, resulting in a consensus that reflects the participants' combined intuition, experience and expert knowledge.

The Delphi technique can be used to explore or expose underlying assumptions or information leading to differing judgments and to correlate informed judgments on a topic spanning a wide range of disciplines. It is useful for problems that can benefit from subjective judgments on a collective basis.

Pitfalls and biases

Estimating subjective probabilities is never as straightforward as implied in the description of the methods above. There are several pitfalls and biases to be aware of:

- None of the methods works extremely well by itself. Typically, multiple techniques must be used.
- To increase consistency, experts should be asked to assess both the probability of an event and separately the probability of the complement of the event. The two should always add up to 1.0; however, in practice they seldom do without repeated application of the assessment method.
- The events must be defined clearly to eliminate ambiguity. "What is the probability of a new competitor entering the market?" is not unambiguous. "What is the probability that a new competitor will take more than 5% market share of product A in the next two years?" more clearly defines the event.
- When assessing probabilities for rare events, it is generally better to assess odds. Odds of event E is [P(E)/P(complement of E)].

The Authors

Jerry Miccolis, a risk management consultant and consulting actuary with Tillinghast – Towers Perrin in its Parsippany, New Jersey office, has 20 years of consulting experience. He is a principal of Towers Perrin and is architect of several of Towers Perrin's multidisciplinary service offerings, including workers compensation cost management, strategic risk financing and enterprise risk management. He has served in a number of practice leadership positions, including practice leader for the worldwide risk management practice. He is a widely quoted speaker and author on risk management issues. A Fellow of the Casualty Actuarial Society (CAS) and a Member of the American Academy of Actuaries, Mr. Miccolis has served both groups on a number of professional committees, chairing several, and sitting on the Actuarial Standards Board. Mr. Miccolis also has authored and reviewed/refereed professional papers in actuarial literature and has served as an editor of CAS and Towers Perrin publications. He holds a B.S. degree in mathematics from Drexel University.

Samir Shah, a managing consultant with Towers Perrin's Strategy and Organization practice in the Washington, D.C. office, has over 15 years of consulting experience. He has provided a wide range of services to clients, including risk management, workforce planning, organizational design, process improvement and actuarial. He specializes in the application of Operations Research methods, such as computer-based simulation and optimization, to management decision making. Mr. Shah is a Fellow of the Society of Actuaries and holds an M.S. degree in Industrial Engineering and Management Sciences from Northwestern University. He is currently pursuing a Ph.D. in Operations Research with applications to Enterprise Risk Management at Northwestern. He is a member of the International Association of Financial Engineers, the Institute for Operations Research and Management Sciences, and the American Academy of Actuaries.

About Tillinghast – Towers Perrin

Tillinghast – Towers Perrin is a global firm that provides management and actuarial consulting to the insurance and financial services industries as well as risk management consulting to the public and private sectors. Tillinghast – Towers Perrin is part of Towers Perrin, one of the world's largest management consulting firms, with more than 8,000 employees and 80 offices in 23 countries.

If you would like to discuss specific aspects of this monograph in greater detail, or to explore the implications for your company, please contact:

Mr. Jerry Miccolis
Principal

Tillinghast – Towers Perrin
Morris Corporate Center II
Building F
One Upper Pond Road
Parsippany, NJ 07054-1050

Direct dial: 973-331-3524
Fax: 973-331-3576
E-mail: miccolj@towers.com

Mr. Samir Shah
Managing Consultant

Tillinghast – Towers Perrin
1001 19th Street North
Suite 1500
Rosslyn, VA 22209-1722

Direct dial: 703-351-4875
Fax: 703-351-4848
E-mail: shahsa@towers.com

Tillinghast - Towers Perrin

BUILDING RELATIONSHIPS ▪ PRODUCING RESULTS™

Argentina
Buenos Aires

Australia
Melbourne
Sydney

Bermuda

Brazil
Rio de Janeiro
São Paulo

Canada
Montreal
Toronto

China
Hong Kong

France
Paris

Germany
Cologne
Frankfurt

Italy
Milan

Japan
Tokyo

Malaysia
Kuala Lumpur

Mexico
Mexico City

The Netherlands
Amsterdam

Singapore

South Africa
Cape Town

South Korea
Seoul

Spain
Madrid

Sweden
Stockholm

Switzerland
Geneva
Zurich

United Kingdom
London

United States
Atlanta
Boston
Chicago
Dallas
Denver
Detroit
Hartford
Indianapolis
Irvine, Calif.
Jacksonville
Minneapolis
New York
Parsippany, N.J.
Philadelphia
St. Louis
San Francisco
Stamford
Washington, D.C.

Other Towers Perrin Locations
Auckland
Austin
Bern
Bristol
Brussels
Calgary
Canberra
Charlotte
Chesapeake, Va.
Cincinnati
Cleveland
Houston
Johannesburg
Los Angeles
Memphis
Miami
Milwaukee
Mississauga
Phoenix
Pittsburgh
Providence
Rotterdam
St. Albans
San Antonio
San Diego
São Paulo
Seattle
Tampa
Valhalla, N.Y.
Vancouver
Voorhees, N.J.
Wellington

Internet: www.tillinghast.com

Risk Management: An Overview

Alexander J. McNeil
Department of Mathematics
Federal Institute of Technology
ETH Zentrum
CH-8092 Zurich

mcneil@math.ethz.ch
www.math.ethz.ch/∼mcneil/   www.math.ethz.ch/finance/

Swiss Banking School, 17-18 September 2001

Module Overview

- Monday
  - ⋆ Overview and Introduction, McNeil
  - ⋆ Management of Market Risks, Bitz
  - ⋆ Management of Credit Risks, Haller
  - ⋆ Management of Operational Risks, Geiger & Piaz
- Tuesday
  - ⋆ Asset & Liability Management, Enderli & Spillmann
  - ⋆ Risk Management of a Private Bank, Hodler
  - ⋆ Risk Management of a Global Player, Guldimann
  - ⋆ Panel Discussion, Hodler, Guldimann, McNeil
  - ⋆ Closing Remarks, McNeil

How Should You Approach the Module ?

∗ Search for the common ideas. What concepts and concerns reappear when one discusses the various areas of financial risk - market, credit, operational ?

How Should You Approach the Module ?

- ∗ Search for the common ideas. What concepts and concerns reappear when one discusses the various areas of financial risk - market, credit, operational ?

- ∗ Appreciate the differences. What special problems do the various areas of financial risk management create ?

How Should You Approach the Module ?

- Search for the common ideas. What concepts and concerns reappear when one discusses the various areas of financial risk - market, credit, operational ?

- Appreciate the differences. What special problems do the various areas of financial risk management create ?

- Understand the role of regulation. Why is regulatory capital needed? What does the regulator require us to do? How will this change in the future ? What is likely impact of Basel II.

How Should You Approach the Module ?

- ∗ Search for the common ideas. What concepts and concerns reappear when one discusses the various areas of financial risk - market, credit, operational ?
- ∗ Appreciate the differences. What special problems do the various areas of financial risk management create ?
- ∗ Understand the role of regulation. Why is regulatory capital needed? What does the regulator require us to do? How will this change in the future ? What is likely impact of Basel II.
- ∗ Do all financial institutions face identical challenges? How does RM differ between a global player and a Swiss private bank?

Contents

A. Mathematical Finance and Risk Management at ETH

B. A Brief History of Risk Management

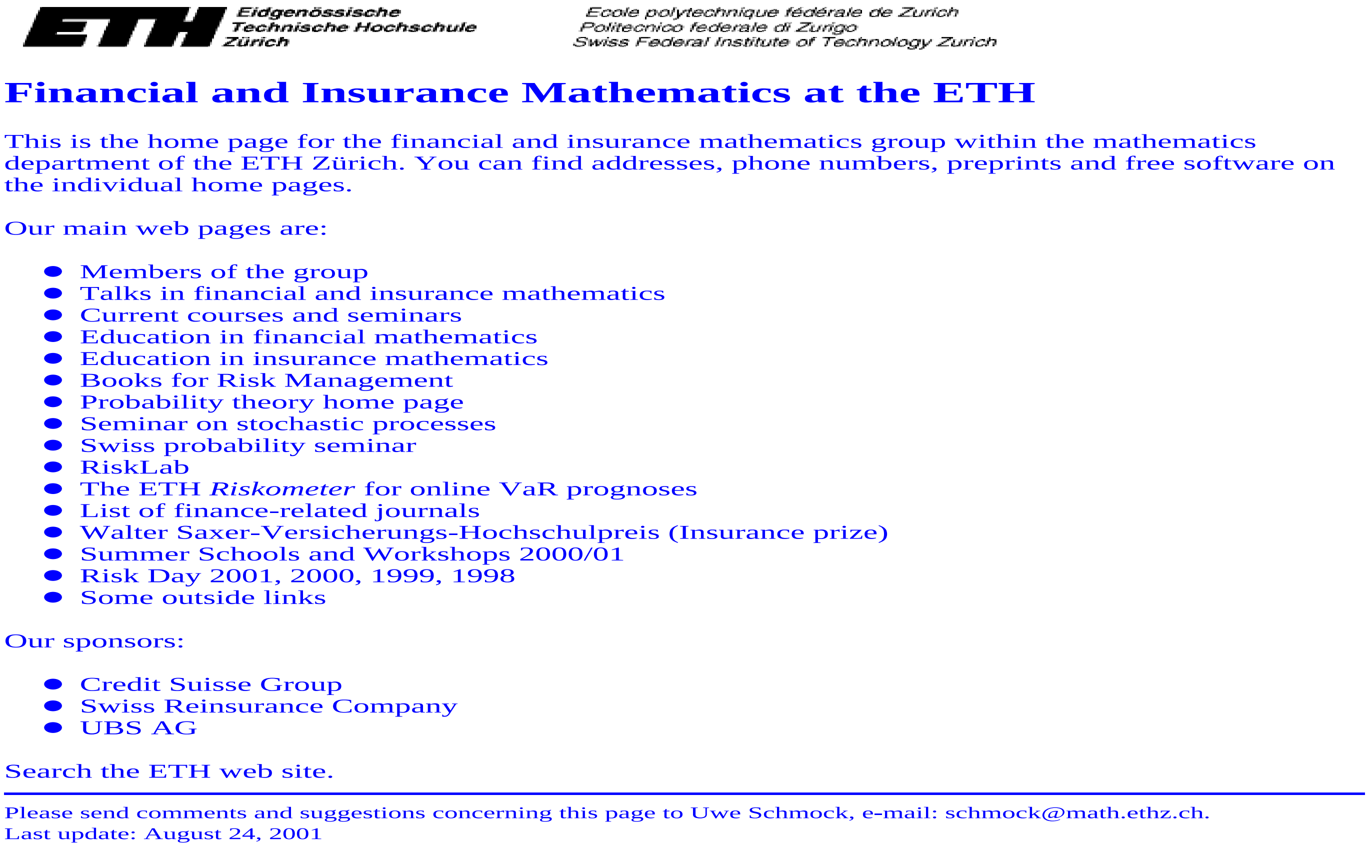
C. The VaR Concept

D. LTCM. Back to the Drawing Board?

E. The Need for Better Quantitative Methods

F. Where does Risk Management Stand?

A. Finance and Risk Management at ETH



Eidgenössische
Technische Hochschule
Zürich

Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo
Swiss Federal Institute of Technology Zurich

Financial and Insurance Mathematics at the ETH

This is the home page for the financial and insurance mathematics group within the mathematics department of the ETH Zürich. You can find addresses, phone numbers, preprints and free software on the individual home pages.

Our main web pages are:

- Members of the group
- Talks in financial and insurance mathematics
- Current courses and seminars
- Education in financial mathematics
- Education in insurance mathematics
- Books for Risk Management
- Probability theory home page
- Seminar on stochastic processes
- Swiss probability seminar
- RiskLab
- The ETH *Riskometer* for online VaR prognoses
- List of finance-related journals
- Walter Saxer-Versicherungs-Hochschulpreis (Insurance prize)
- Summer Schools and Workshops 2000/01
- Risk Day 2001, 2000, 1999, 1998
- Some outside links

Our sponsors:

- Credit Suisse Group
- Swiss Reinsurance Company
- UBS AG

Search the ETH web site.

Please send comments and suggestions concerning this page to Uwe Schmock, e-mail: schmock@math.ethz.ch.
Last update: August 24, 2001

Welcome to

RiskLab®
Switzerland

About RiskLab

Contact

Projects

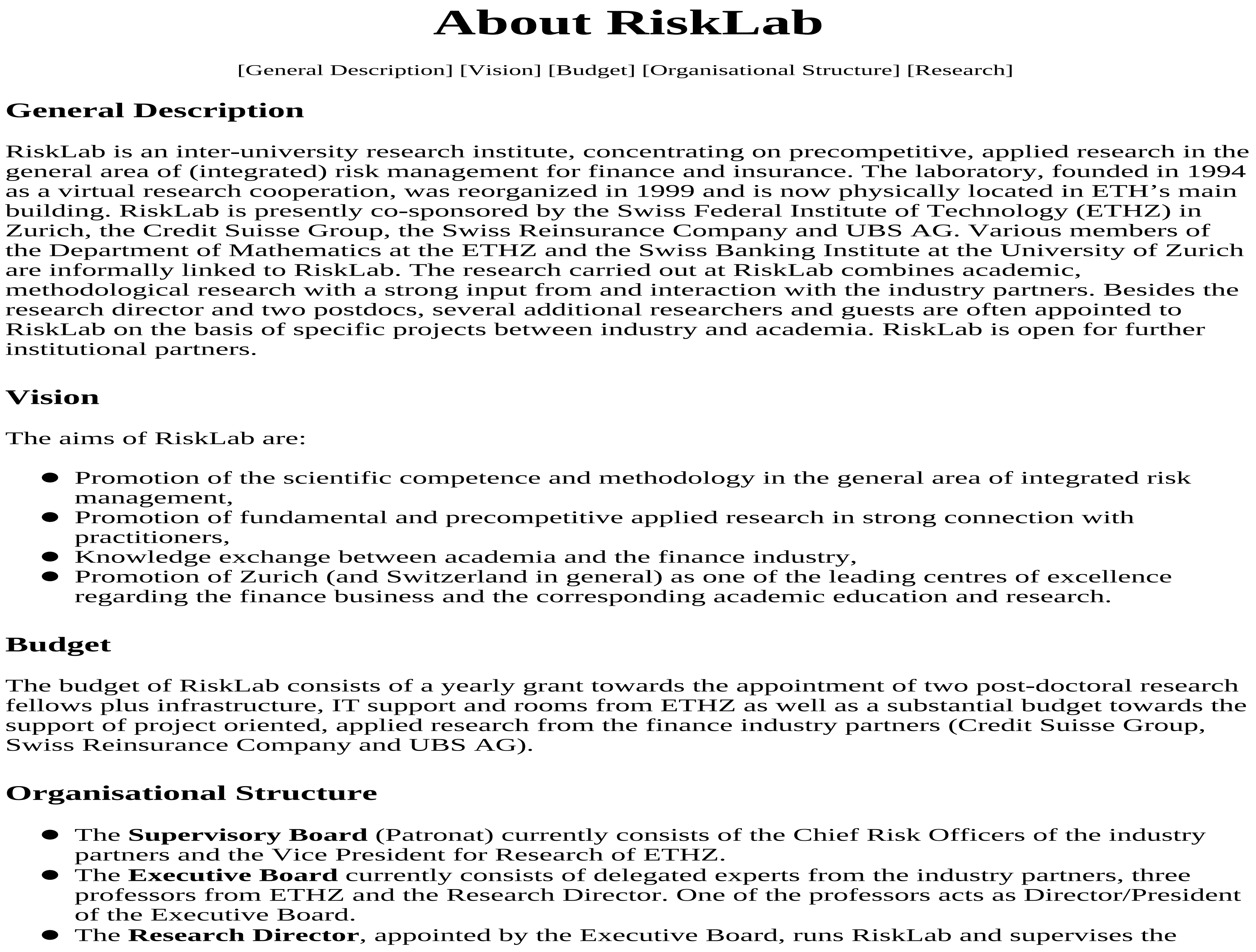
Papers

Talks

Links

The Press

`www.math.ethz.ch/risklab/`



About RiskLab

[General Description] [Vision] [Budget] [Organisational Structure] [Research]

General Description

RiskLab is an inter-university research institute, concentrating on precompetitive, applied research in the general area of (integrated) risk management for finance and insurance. The laboratory, founded in 1994 as a virtual research cooperation, was reorganized in 1999 and is now physically located in ETH's main building. RiskLab is presently co-sponsored by the Swiss Federal Institute of Technology (ETHZ) in Zurich, the Credit Suisse Group, the Swiss Reinsurance Company and UBS AG. Various members of the Department of Mathematics at the ETHZ and the Swiss Banking Institute at the University of Zurich are informally linked to RiskLab. The research carried out at RiskLab combines academic, methodological research with a strong input from and interaction with the industry partners. Besides the research director and two postdocs, several additional researchers and guests are often appointed to RiskLab on the basis of specific projects between industry and academia. RiskLab is open for further institutional partners.

Vision

The aims of RiskLab are:

- Promotion of the scientific competence and methodology in the general area of integrated risk management,
- Promotion of fundamental and precompetitive applied research in strong connection with practitioners,
- Knowledge exchange between academia and the finance industry,
- Promotion of Zurich (and Switzerland in general) as one of the leading centres of excellence regarding the finance business and the corresponding academic education and research.

Budget

The budget of RiskLab consists of a yearly grant towards the appointment of two post-doctoral research fellows plus infrastructure, IT support and rooms from ETHZ as well as a substantial budget towards the support of project oriented, applied research from the finance industry partners (Credit Suisse Group, Swiss Reinsurance Company and UBS AG).

Organisational Structure

- The **Supervisory Board** (Patronat) currently consists of the Chief Risk Officers of the industry partners and the Vice President for Research of ETHZ.
- The **Executive Board** currently consists of delegated experts from the industry partners, three professors from ETHZ and the Research Director. One of the professors acts as Director/President of the Executive Board.
- The **Research Director**, appointed by the Executive Board, runs RiskLab and supervises the

My Own Work

A book provisionally entitled Quantitative Methods in Risk Management is currently in preparation. Publication 2002-2003 ?
Authors: Paul Embrechts, Rüdiger Frey, Alexander McNeil

Aims:

- To provide practitioners of RM with a reference work on the quantitative (mathematical and statistical) tools their work often requires.

- To supply a course text for masters level courses on quantitative risk management, e.g. in a financial engineering programme. A joint University of Zurich and ETH programme starts Autumn 2002.

B. A Brief History of Risk Management

“Risk management: one of the most important innovations of the 20th century.” [Steinherr, 1998]

B. A Brief History of Risk Management

“Risk management: one of the most important innovations of the 20th century.” [Steinherr, 1998]

- The late 20th century saw a “revolution” on financial markets. Derivatives and other financial innovations.

B. A Brief History of Risk Management

"Risk management: one of the most important innovations of the 20th century." [Steinherr, 1998]

- The late 20th century saw a "revolution" on financial markets. Derivatives and other financial innovations.

- Large derivatives losses and other financial incidents followed.

B. A Brief History of Risk Management

"Risk management: one of the most important innovations of the 20th century." [Steinherr, 1998]

- The late 20th century saw a "revolution" on financial markets. Derivatives and other financial innovations.
- Large derivatives losses and other financial incidents followed.
- Banks became subject to regulatory capital requirements, internationally coordinated by the Basle Committee of the Bank of International Settlements.

Some Key Dates

- 1933. Glass-Steagall Act passed in aftermath of Depression prohibiting commercial banks from underwriting insurance and most kinds of securities. 20th century has seen many of these limitations gradually removed.

Some Key Dates

- 1933. Glass-Steagall Act passed in aftermath of Depression prohibiting commercial banks from underwriting insurance and most kinds of securities. 20th century has seen many of these limitations gradually removed.

- 1950s. Foundations of modern risk analysis are laid by work of Markowitz and others on portfolio theory.

Some Key Dates

- 1933. Glass-Steagall Act passed in aftermath of Depression prohibiting commercial banks from underwriting insurance and most kinds of securities. 20th century has seen many of these limitations gradually removed.

- 1950s. Foundations of modern risk analysis are laid by work of Markowitz and others on portfolio theory.

- 1970. The Bretton-Woods system of fixed exchange rates is abolished, leading to increased exchange rate volatility.

Some Key Dates

- 1933. Glass-Steagall Act passed in aftermath of Depression prohibiting commercial banks from underwriting insurance and most kinds of securities. 20th century has seen many of these limitations gradually removed.

- 1950s. Foundations of modern risk analysis are laid by work of Markowitz and others on portfolio theory.

- 1970. The Bretton-Woods system of fixed exchange rates is abolished, leading to increased exchange rate volatility.

- 1973. CBOE, Chicago Board Options Exchange starts operating.

Some Key Dates II

- 1973. Fisher Black and Myron Scholes, publish an article on the rational pricing of options. [Black and Scholes, 1973]
  Hitherto it had been pure guesswork.

Some Key Dates II

- 1973. Fisher Black and Myron Scholes, publish an article on the rational pricing of options. [Black and Scholes, 1973]
  Hitherto it had been pure guesswork.

- 1980s. Deregulation - the elimination of certain constraints on banks' activities; globalization - mergers on unprecedented scale; advances in IT.

Some Key Dates II

- 1973. Fisher Black and Myron Scholes, publish an article on the rational pricing of options. [Black and Scholes, 1973]
Hitherto it had been pure guesswork.

- 1980s. Deregulation - the elimination of certain constraints on banks' activities; globalization - mergers on unprecedented scale; advances in IT.

- 1999. Financial Services Act repealing many key provisions of Glass-Steagall. Bank holding companies will continue to expand the range of their financial services; further convergence of finance and insurance likely.

Consequences

Enormous growth in both volume and complexity of products traded on the financial markets.

Example 1

Average daily trading volume at New York stock exchange:
1970: 3.5 million shares 1990: 40 million shares

Example 2: Global market in OTC derivatives (nominal value).

	1995	1998
FOREX contracts	$13 trillion	$18 trillion
Interest rate contracts	$26 trillion	$50 trillion
All types	$47 trillion	$80 trillion

Source BIS; see [Crouhy et al., 2001]. $\$1$ trillion $= \$1 \times 10^{12}$.

First Problems Occur

The period 1993-1996 saw some spectacular derivatives-based losses:

- ⋆ Orange County (1.7 billion US$)
- ⋆ Metallgesellschaft (1.3 billion US$)
- ⋆ Barings (1 billion US$)

Although, to be fair, “classical banking” produced its own large losses.e.g. 50 billion CHF of bad loans written off by the Big Three in early nineties.

Classification of Risks

It is common to classify risks according to their source.

- **Market Risk** - risk associated with fluctuations in value of traded assets.

Classification of Risks

It is common to classify risks according to their source.

- **Market Risk** - risk associated with fluctuations in value of traded assets.
- **Credit Risk** - risk associated with uncertainty that debtors will honour their financial obligations.

Classification of Risks

It is common to classify risks according to their source.

- **Market Risk** - risk associated with fluctuations in value of traded assets.

- **Credit Risk** - risk associated with uncertainty that debtors will honour their financial obligations.

- **Operational Risk** - risk associated with possibility of human error, IT failure, dishonesty, natural disaster, terrorism etc.

Classification of Risks

It is common to classify risks according to their source.

- **Market Risk** - risk associated with fluctuations in value of traded assets.
- **Credit Risk** - risk associated with uncertainty that debtors will honour their financial obligations.
- **Operational Risk** - risk associated with possibility of human error, IT failure, dishonesty, natural disaster, terrorism etc.
- **Liquidity risk** - risk that positions cannot be unwound quickly enough at critical times due to lack of market liquidity.

Reactions in the Finance World

- Reaction of the Banks. Development of mathematical models for internal risk control, e.g. RiskMetrics by J.P.Morgan.
  First methodological progress on market risk front.

- Reaction of the Regulators

Reactions in the Finance World

- Reaction of the Banks. Development of mathematical models for internal risk control, e.g. RiskMetrics by J.P.Morgan.
  First methodological progress on market risk front.

- Reaction of the Regulators

  ⋆ 1988. Basle accord (BIS 88). First steps toward international minimum capital standard.

Reactions in the Finance World

- Reaction of the Banks. Development of mathematical models for internal risk control, e.g. RiskMetrics by J.P.Morgan.
  First methodological progress on market risk front.

- Reaction of the Regulators
  - ⋆ 1988. Basle accord (BIS 88). First steps toward international minimum capital standard.
  - ⋆ 1993. Seminal G-30 report making best-practice RM recommendations. VaR and stress testing emerge.

Reactions in the Finance World

- Reaction of the Banks. Development of mathematical models for internal risk control, e.g. RiskMetrics by J.P.Morgan.
  First methodological progress on market risk front.

- Reaction of the Regulators
  - ⋆ 1988. Basle accord (BIS 88). First steps toward international minimum capital standard.
  - ⋆ 1993. Seminal G-30 report making best-practice RM recommendations. VaR and stress testing emerge.
  - ⋆ 1996. BIS Amendment prescribing standardized model but allowing internal market VaR models for larger banks.

Reactions in the Finance World

- Reaction of the Banks. Development of mathematical models for internal risk control, e.g. RiskMetrics by J.P.Morgan.
  First methodological progress on market risk front.

- Reaction of the Regulators
  - ⋆ 1988. Basle accord (BIS 88). First steps toward international minimum capital standard.
  - ⋆ 1993. Seminal G-30 report making best-practice RM recommendations. VaR and stress testing emerge.
  - ⋆ 1996. BIS Amendment prescribing standardized model but allowing internal market VaR models for larger banks.
  - ⋆ 2001. Consultative process for new BIS Accord. Move toward internal credit models. Consideration of operational risk.

Why is the Regulator Concerned?

“Banks collect deposits and play a key role in the payment system. National governments have a very direct interest in ensuring that banks remain capable of meeting their obligations; in effect they act as a guarantor, sometimes also as lender of last resort. They therefore wish to limit the cost of the safety net in the case of bank failure. By acting as a buffer against unanticipated losses, regulatory capital helps to privatize a burden that would otherwise be borne by national governments.”
[Crouhy et al., 2001]

C. The VaR Concept

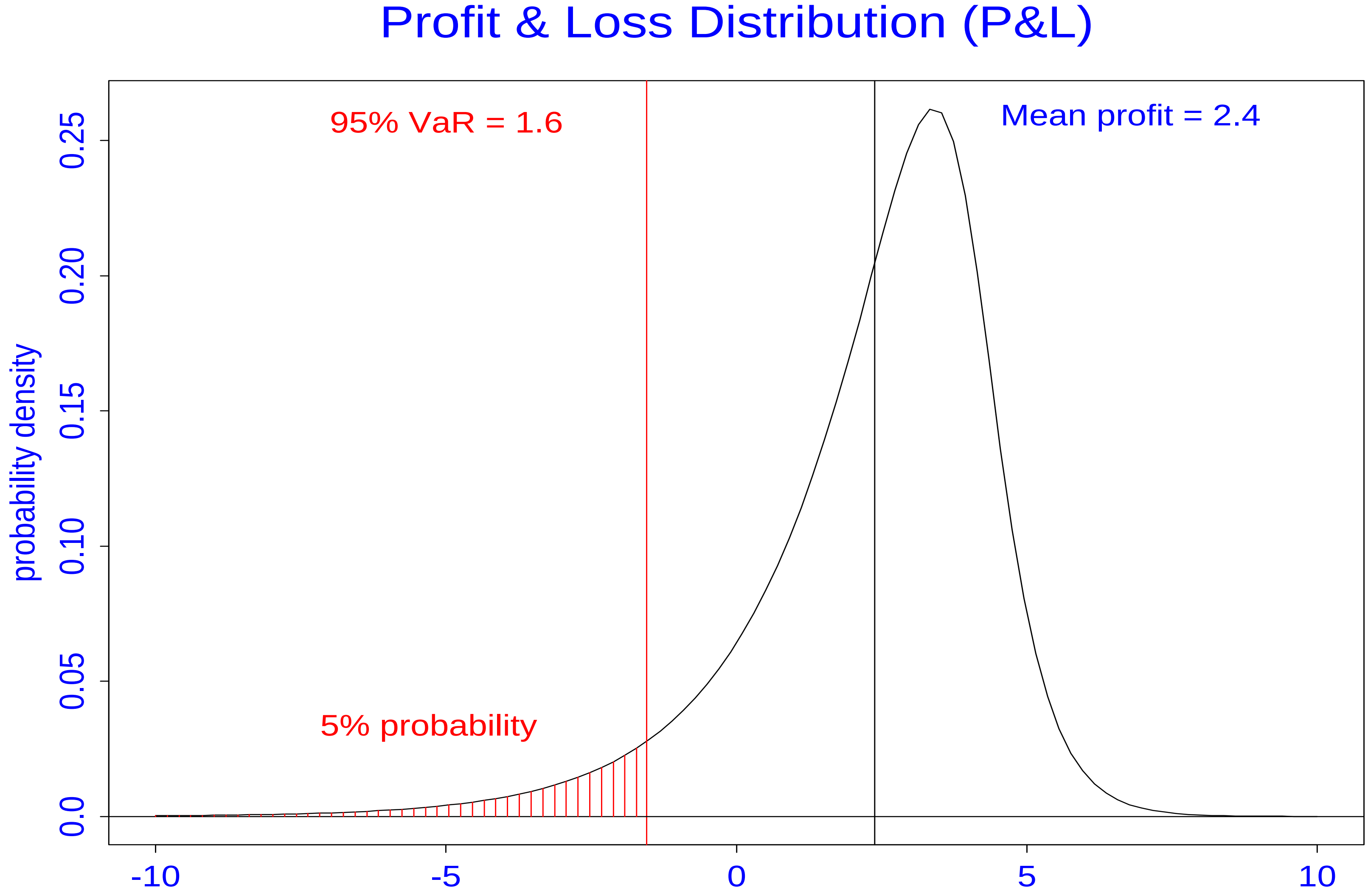
Consider a portfolio/position and potential profits and losses over a fixed time horizon - e.g. 1 day or 10 days.

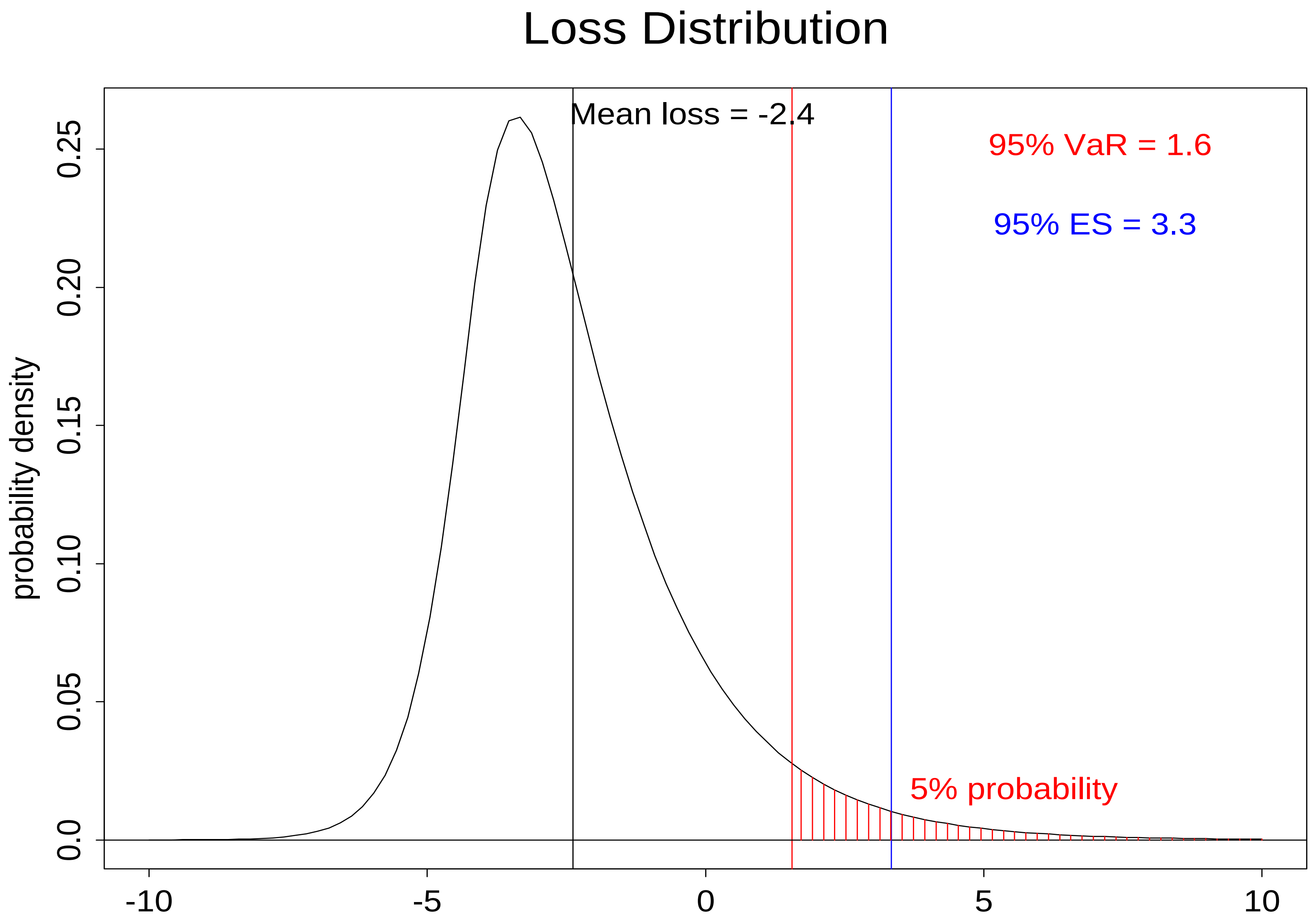
VaR is a percentile (or quantile) of the profit and loss (P&L) distribution with the property that, with a small given probability, we stand to incur that loss or more over the fixed time horizon.

Example. 10-day 99% VaR of 1M$

Interpretation.
If we hold our current portfolio position fixed for 10 days then
Probability (we lose 1M$ or more) $=$ 1%
Probability (we lose up to 1M$) $=$ 99%.

VaR in Visual Terms




Loss Distribution
probability density
0.0
0.05
0.10
0.15
0.20
0.25
-10
-5
0
5
10
Mean loss = -2.4
95% VaR = 1.6
95% ES = 3.3
5% probability

Var - badly defined!

The VaR bible is Philippe Jorion's book.[Jorion, 2001].

The following "definition" is very common:

"VaR is the $maximum$ expected loss of a portfolio over a given time horizon with a certain confidence level."

It is however mathematically meaningless and potentially misleading.
In no sense is VaR a maximum loss!
We can lose more, sometimes much more, depending on the heaviness of the tail of the loss distribution.

The VaR Discipline in Market Risk

Aside from problems of definition/interpretation, the VaR concept has been instrumental in introducing a culture of quantitative (statistical) risk analysis into banks.

1. Estimation of the distribution of future profits and losses for fixed holding period and portfolio
   - single position
   - trading book for a particular market
   - entire position of the bank
2. Estimation of risk measures (VaR) based on estimated P&L.
3. Use of these risk measures to manage enterprise.

A Simple Example: Portfolio of Equities

Today is day t. We are interested in a horizon h (say 10 days). We have an equity portfolio of 3 equities with value given by

$$V_t = \alpha_1 S_{1,t} + \alpha_2 S_{2,t} + \alpha_3 S_{3,t},$$

α_i is number of units of equity i, $S_{i,t}$ is price of equity i.

Our unknown profit/loss is given by $V_{t+h} - V_t$.

To estimate P&L distribution we use historical information concerning changes in the 3 underlying equity values. The underlying equities are known as the risk factors affecting the P&L.

The form of relationship between the risk factors and the value is known as the mapping.

VaR Estimation Methodology

A number of techniques are in widespread use:

- Analytic variance-covariance approach.

  Assumptions:

  - ⋆ Changes in risk factor values are assumed to have a (multivariate) normal distribution.
  - ⋆ Changes in value of portfolio are approximated by linear function of changes in risk factors.

  Problems: both normality and linearity.
  Why should risk factor changes be normal? Thin tails may underestimate risk.

VaR Estimation Methodology

- The historical simulation approach

  Observations from the P&L are simulated by examining what would happen if historical observed risk factor changes recurred.

  Problems: relies on availability and relevance of historical risk factor data.

- The Monte Carlo approach

  Assume more complex models for the risk factors and their dynamics. Simulate observations from resulting P&L using computer programs.

  Problems. computer intensive; what model to choose?

VaR: Deeper Problems

Aside from the statistical issue of how to estimate VaR, more fundamental issues have been raised. Many have asked

Is VaR the Right Risk Measure?

VaR: Deeper Problems

Aside from the statistical issue of how to estimate VaR, more fundamental issues have been raised. Many have asked

Is VaR the Right Risk Measure?

- ∗ VaR tells us nothing about the losses beyond VaR and may lead to false sense of security.

VaR: Deeper Problems

Aside from the statistical issue of how to estimate VaR, more fundamental issues have been raised. Many have asked

Is VaR the Right Risk Measure?

- ∗ VaR tells us nothing about the losses beyond VaR and may lead to false sense of security.
- ∗ VaR has poor aggregation properties (example to follow). It is said to be non-coherent [Artzner et al., 1999].

VaR: Deeper Problems

Aside from the statistical issue of how to estimate VaR, more fundamental issues have been raised. Many have asked

Is VaR the Right Risk Measure?

- ∗ VaR tells us nothing about the losses beyond VaR and may lead to false sense of security.
- ∗ VaR has poor aggregation properties (example to follow). It is said to be non-coherent [Artzner et al., 1999].
- ∗ The sophisticated trader may learn to game VaR - to assume positions that have a low VaR but are in fact extremely risky.

VaR: Deeper Problems

Aside from the statistical issue of how to estimate VaR, more fundamental issues have been raised. Many have asked

Is VaR the Right Risk Measure?

- ∗ VaR tells us nothing about the losses beyond VaR and may lead to false sense of security.
- ∗ VaR has poor aggregation properties (example to follow). It is said to be non-coherent [Artzner et al., 1999].
- ∗ The sophisticated trader may learn to game VaR - to assume positions that have a low VaR but are in fact extremely risky.

Alternative risk measures: expected shortfall (a.k.a. conditional VaR) - the expected size of a loss exceeding VaR.

Example of a VaR Paradox

Consider 100 corporate bonds, $X_1, \ldots, X_{100}$ with 1-year maturity. Each has face value 100, pays interest at 2% and has default rate 1%, per annum.

The P&L of a single bond is

$$X_i = \begin{cases} 2 & \text{with probability 99\%,} \\ -100 & \text{with probability 1\%.} \end{cases}$$

Now consider two portfolios:
A. 100 of bond X_1
B. One each of $X_1, \ldots, X_{100}$.

Which is riskier?

VaR Paradox II

The 95% 1-year VaR of portfolio A is -200.
(Informally: we are 95% certain of making a gain of 200 dollars.)

The 95 % 1-year VaR of portfolio B is > 0.

Paradoxically, the diversified portfolio B is riskier than A in VaR terms. This is clearly nonsensical.

This phenomenon relates to the non-subadditivity of VaR, which makes it poor for decentralized risk management.

Note, the trader who buys position A is ‘gaming the VaR’.

D. LTCM. Back to the Drawing Board?

The core strategy of LTCM was relative-value trades. The nature of the bet is to take long and short positions in closely related titles whose yields are expected to soon converge, e.g. German government bonds and Italian government bonds prior to EMU. Since the return is small leverage was used to create attractive returns. Before the crisis LTCM had leverage ratio of 25:1. Of the $125 billion on its balance sheet only $5 billion was equity; the rest was borrowed.

Unfortunately the Russian ruble crisis led to a flight to quality and the divergence of values that were expected to converge. The net result was huge losses - 4.4$ billion - of which 1.9$ billion was incurred by the partners and 2.5 by other investors (700M$ by UBS).

Was VaR to Blame?

LTCM actually used VaR methodology. According to LTCM the fund was structured so that the risk should have been no great than investing in the S&P 500.

"The non-fault bankruptcy". Myron Scholes in *Economist* 25.09.99.

"VaR, the product of portfolio theory, is used for short-run day-to-day profit and loss-risk exposures. Now is the time to encourage the BIS and other regulatory bodies to support studies on stress test and concentration methodologies. Planning for crises is much more important than VaR analysis. And such new methodologies are the correct response to recent crises in the financial industry." [Scholes, 2000].

VaR Wasn't to Blame

"The story of LTCM should not be taken as an indictment of VaR systems, which after all, performed reasonably well for the banking sector in 1998. Instead it provides a number of useful risk management lessons. First it illustrates the danger of optimization biases, or traders 'gaming the system'. LTCM's strategy can be interpreted as a constrained optimization, i.e. maximizing expected returns subject to a constraint on VaR. This strategy led to its demise, as it created huge leverage and extreme sensitivity to instability in the correlations." [Jorion, 2000]

Too Little Rocket Science?

“In a sense, maybe the problem wasn’t too much rocket science, but too little. Extreme, synchronized rises and falls in financial markets occur infrequently - but they do occur. The problem with the models is that they did not assign a high enough chance of occurrence to the scenario in which many things go wrong at the same time the “perfect storm” scenario.”

Business Week, September 21 1998.

First Lesson of all RM Disasters

First Lesson of all RM Disasters

$$\mathrm{RM} \approx \mathrm{MR}$$

E. Towards Better Quantitative Methods

Risk Management poses difficult quantitative problems. Much of conventional statistics is to do with “the average”, “the normal”, or “the expected”. Risk management has more to do with the extreme, the abnormal and the unexpected.

E. Towards Better Quantitative Methods

Risk Management poses difficult quantitative problems. Much of conventional statistics is to do with “the average”, “the normal”, or “the expected”. Risk management has more to do with the extreme, the abnormal and the unexpected.

Three technical issues are:

∗ How to model volatility?

E. Towards Better Quantitative Methods

Risk Management poses difficult quantitative problems. Much of conventional statistics is to do with "the average", "the normal", or "the expected". Risk management has more to do with the extreme, the abnormal and the unexpected.

Three technical issues are:

∗ How to model volatility?

∗ How to model extremes and stress events?

E. Towards Better Quantitative Methods

Risk Management poses difficult quantitative problems. Much of conventional statistics is to do with "the average", "the normal", or "the expected". Risk management has more to do with the extreme, the abnormal and the unexpected.

Three technical issues are:

∗ How to model volatility?

∗ How to model extremes and stress events?

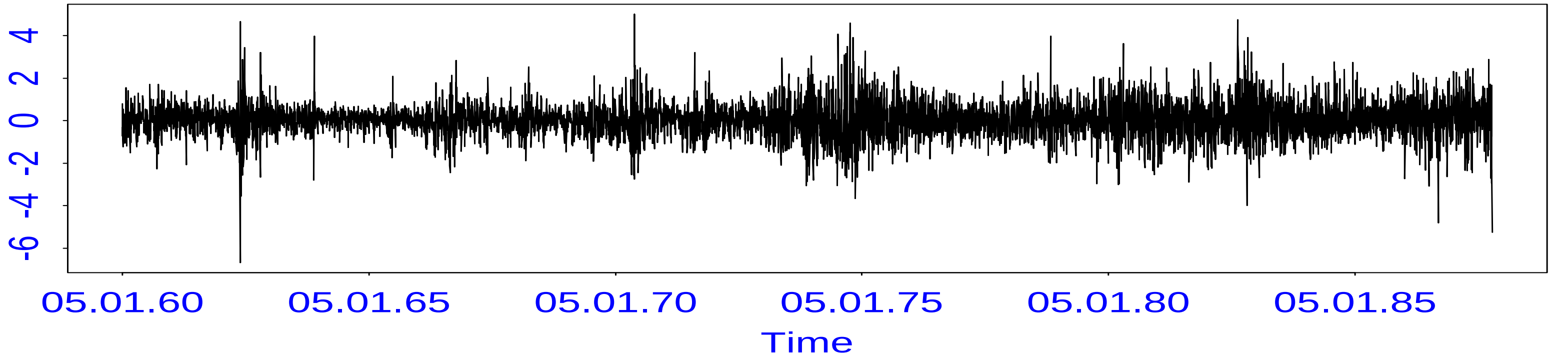
∗ How to model correlation and concentration risk?

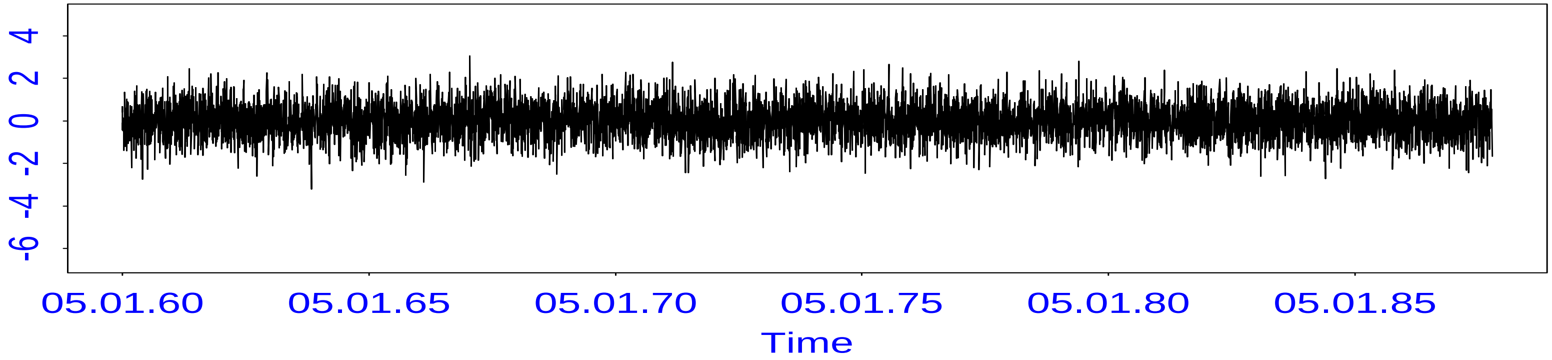
Volatility

Any financial asset with an element of market risk shows volatility. The scale of this volatility generally contradicts the standard model of finance - geometric Brownian motion - which is the basis of pricing theory.

The implication is that the models with which we measure risk, should probably be different to the models with which we price risky assets.

Stock-market data versus simulated normal





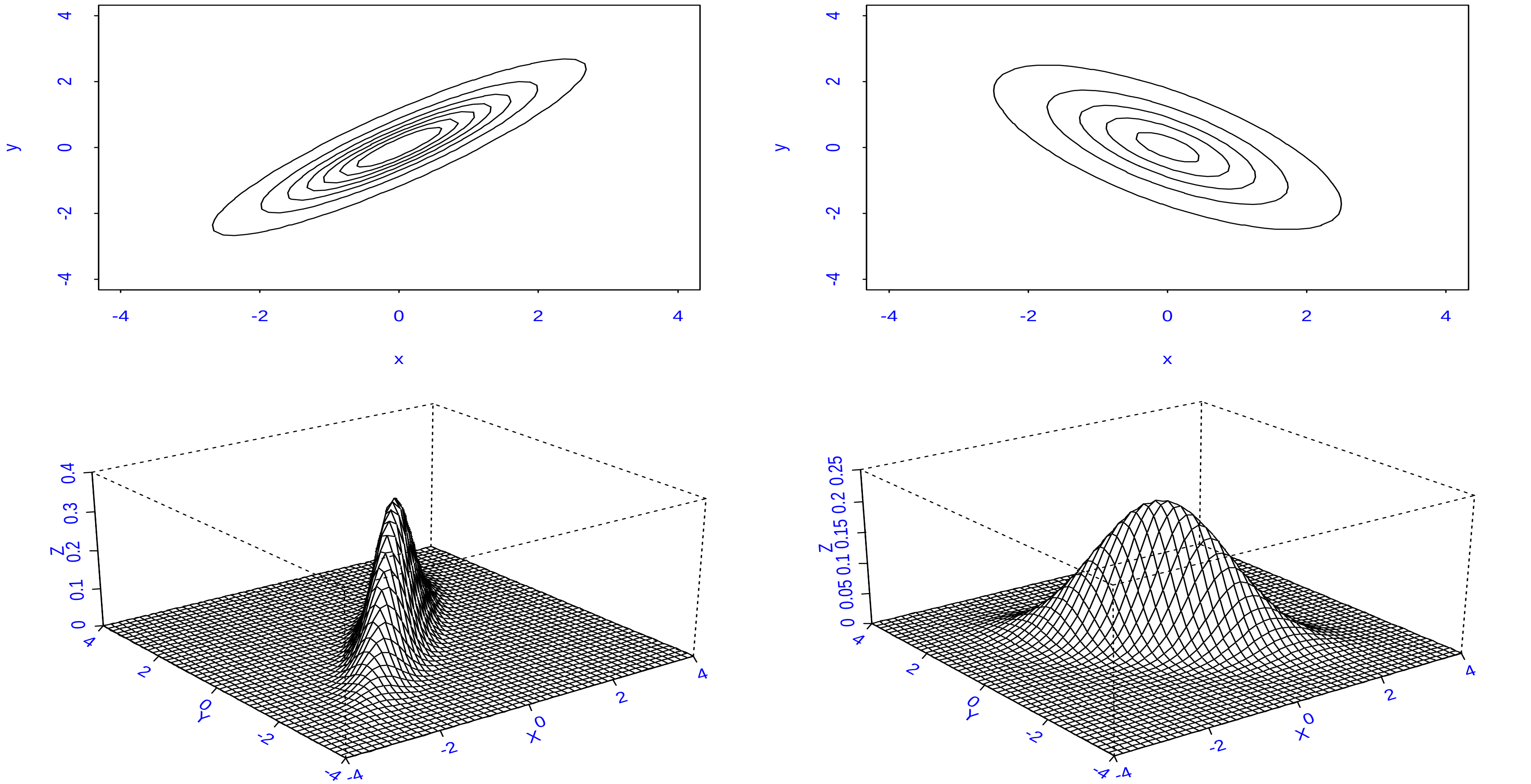
The Stylized Facts of Empirical Finance

Consider daily returns on a stock price, exchange rate, commodity price or other financial instrument, or portfolio of instruments.

We consistently observe the following stylized facts:

- Returns not iid but correlation low
- Absolute returns highly correlated
- Volatility changes randomly with time
- Returns are heavier–tailed than normal distribution
- Extremes appear in clusters

How Normal is the Normal Distribution ?



Standard (bivariate) normal distributions ($\rho = 0.9, -0.7$).

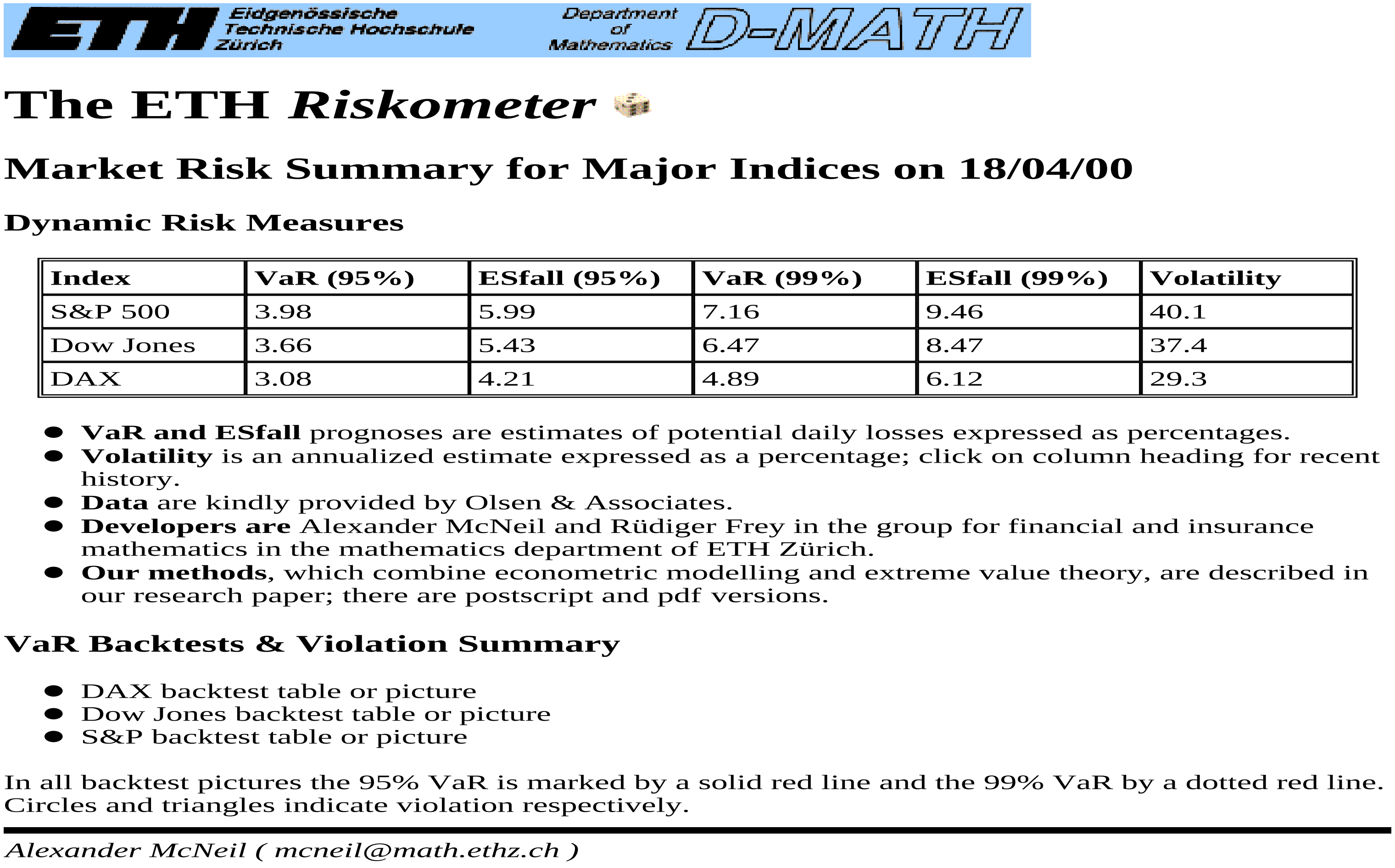
Extreme Values

Above and beyond this persistent background of volatility there is the phenomenon of extreme returns.

Econometric forecasting technology (such as GARCH models and stochastic-volatility models) can go some way to predicting at least short-term volatility development.

But the standard versions of these models (which assume normality of return shocks) fail to explain the frequency and severity of the most extreme movements.

By working with more realistic statistical distributions (heavy-tailed distributions) we can often get a truer risk appraisal. This is the essential idea of extreme value theory. The consideration of stress scenarios is also vital.



Eidgenössische Technische Hochschule Zürich Department of Mathematics D-MATH

The ETH *Riskometer*

Market Risk Summary for Major Indices on 18/04/00

Dynamic Risk Measures

Index	VaR (95%)	ESfall (95%)	VaR (99%)	ESfall (99%)	Volatility
S&P 500	3.98	5.99	7.16	9.46	40.1
Dow Jones	3.66	5.43	6.47	8.47	37.4
DAX	3.08	4.21	4.89	6.12	29.3

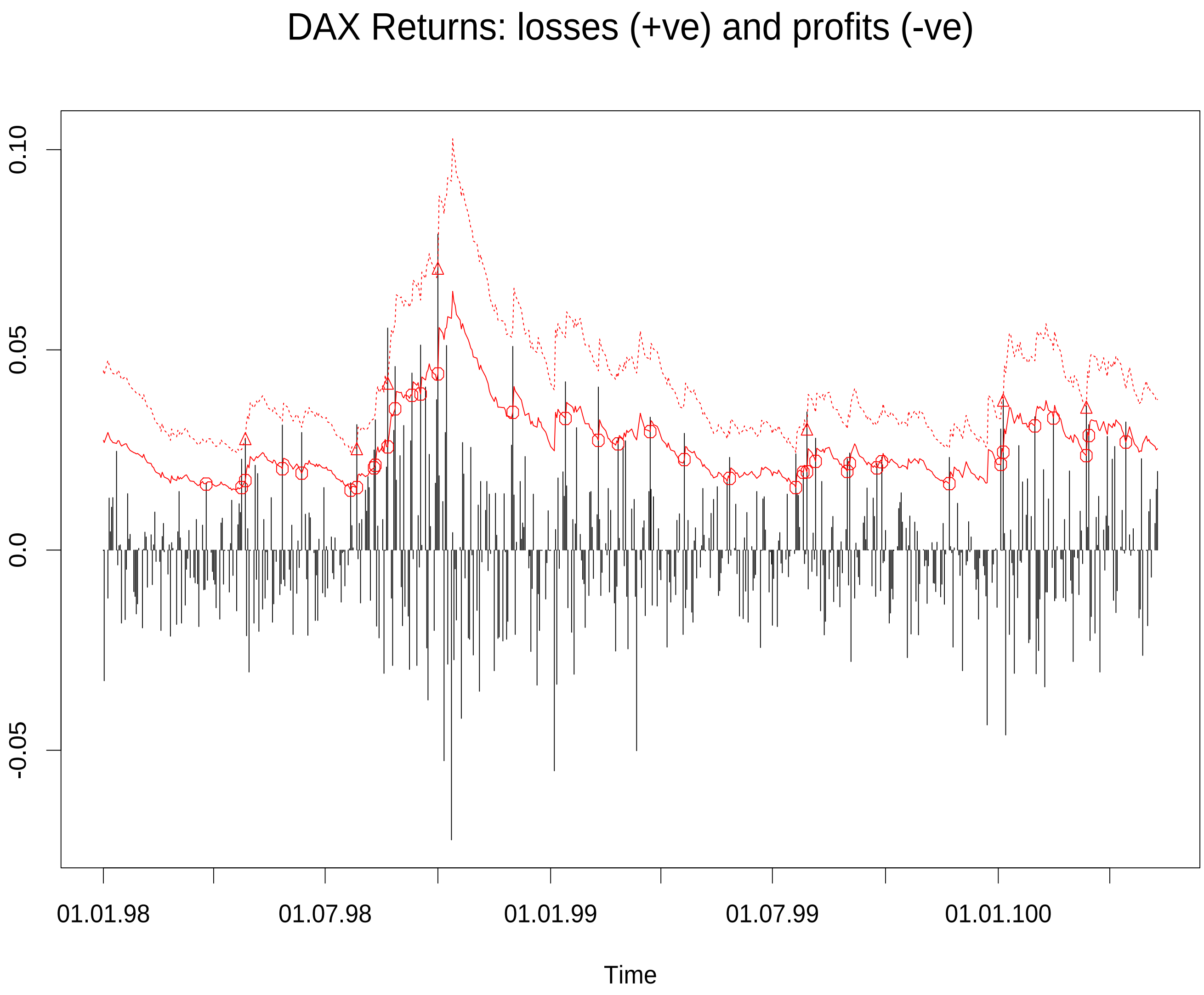
- **VaR and ESfall** prognoses are estimates of potential daily losses expressed as percentages.
- **Volatility** is an annualized estimate expressed as a percentage; click on column heading for recent history.
- **Data** are kindly provided by Olsen & Associates.
- **Developers are** Alexander McNeil and Rüdiger Frey in the group for financial and insurance mathematics in the mathematics department of ETH Zürich.
- **Our methods**, which combine econometric modelling and extreme value theory, are described in our research paper; there are postscript and pdf versions.

VaR Backtests & Violation Summary

- DAX backtest table or picture
- Dow Jones backtest table or picture
- S&P backtest table or picture

In all backtest pictures the 95% VaR is marked by a solid red line and the 99% VaR by a dotted red line. Circles and triangles indicate violation respectively.

Alexander McNeil (mcneil@math.ethz.ch)


DAX Returns: losses (+ve) and profits (-ve)
0.10
0.05
0.0
-0.05
01.01.98
01.07.98
01.01.99
01.07.99
01.01.100
Time

Correlation Confusion

“Among nine big economies, stock market correlations have averaged around 0.5 since the 1960s. In other words, for every 1 per cent rise (or fall) in, say, American share prices, share prices in the other markets will typically rise (fall) by 0.5 per cent.”

The Economist, 8th November 1997

Correlation Confusion

“Among nine big economies, stock market correlations have averaged around 0.5 since the 1960s. In other words, for every 1 per cent rise (or fall) in, say, American share prices, share prices in the other markets will typically rise (fall) by 0.5 per cent.”

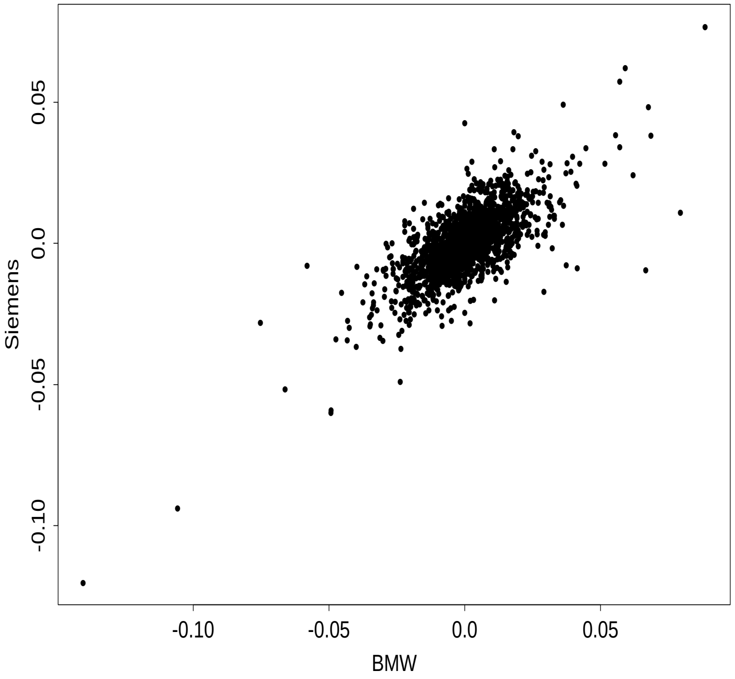
The Economist, 8th November 1997

“A correlation of 0.5 does not indicate that a return from stock-market A will be 50% of stockmarket B’s return, or vice-versa...A correlation of 0.5 shows that 50% of the time the return of stockmarket A will be positively correlated with the return of stock-market B, and 50% of the time it will not.”

The Economist (letter), 22nd November 1997

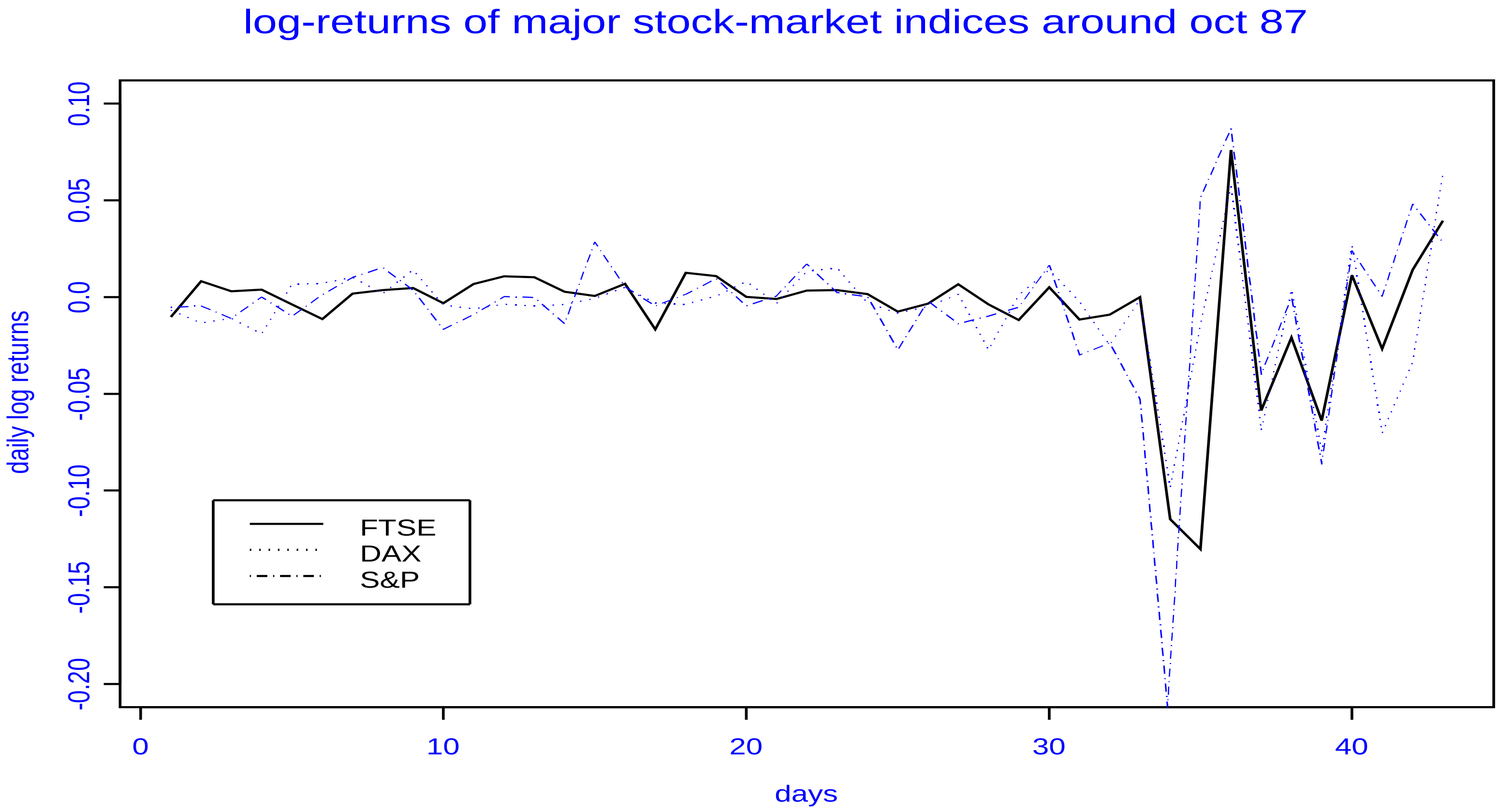
Concentration Risk: Extremes Occur Together

“Correlations are higher in stress periods than in normal periods.”

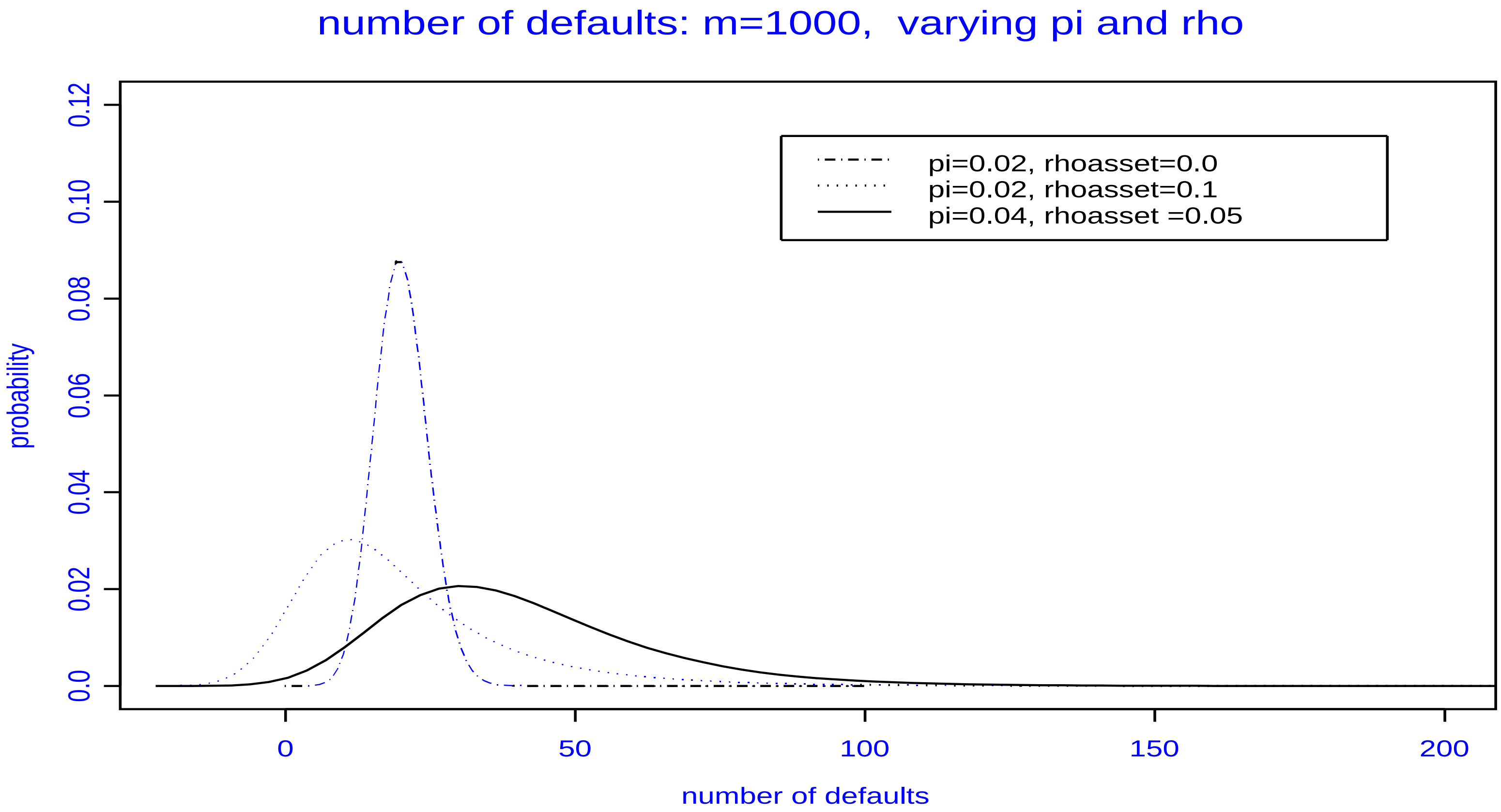


This multivariate stylized fact may express the observation that extreme moves of many financial assets are synchronous.

Extremes Occur Together II



Dependent defaults and credit losses



Distribution of number of defaults in portfolio of 1000 firms.
Dependence between defaults has a large influence on distribution.

Further Technical Reading

- On Extreme Values [Embrechts et al., 1997]
- On Volatility and Extremes [McNeil and Frey, 2000]
- On Dependence and Correlation [Embrechts et al., 2001]
- On Correlation and Credit [Frey and McNeil, 2001]

F. Where does Risk Management Stand?

- Market Risk. Subjected to much of the early effort; a feeling that this is well-understood. Still room for improvement.

F. Where does Risk Management Stand?

- Market Risk. Subjected to much of the early effort; a feeling that this is well-understood. Still room for improvement.

- Credit Risk. Methodology now available, but often poorly understood and implemented. Even more room for improvement.

F. Where does Risk Management Stand?

- Market Risk. Subjected to much of the early effort; a feeling that this is well-understood. Still room for improvement.
- Credit Risk. Methodology now available, but often poorly understood and implemented. Even more room for improvement.
- Operational Risk. On the agenda, but less amenable to quantitative approaches.

F. Where does Risk Management Stand?

- Market Risk. Subjected to much of the early effort; a feeling that this is well-understood. Still room for improvement.
- Credit Risk. Methodology now available, but often poorly understood and implemented. Even more room for improvement.
- Operational Risk. On the agenda, but less amenable to quantitative approaches.
- Liquidity Risk. Very topical since LTCM, but extremely challenging.

F. Where does Risk Management Stand?

- Market Risk. Subjected to much of the early effort; a feeling that this is well-understood. Still room for improvement.
- Credit Risk. Methodology now available, but often poorly understood and implemented. Even more room for improvement.
- Operational Risk. On the agenda, but less amenable to quantitative approaches.
- Liquidity Risk. Very topical since LTCM, but extremely challenging.
- Risk Integration. Market-credit integration has been addressed, but hardly mastered.

References

[Artzner et al., 1999] Artzner, P., Delbaen, F., Eber, J., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9:203–228.

[Black and Scholes, 1973] Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

[Crouhy et al., 2001] Crouhy, M., Galai, D., and Mark, R. (2001). *Risk Management*. McGraw-Hill, New York.

[Embrechts et al., 1997] Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin.

[Embrechts et al., 2001] Embrechts, P., McNeil, A., and Straumann, D. (2001). Correlation and dependency in risk management: properties and pitfalls. In Dempster, M. and Moffatt, H., editors, *Risk Management: Value at Risk and Beyond*. Cambridge University Press.

[Frey and McNeil, 2001] Frey, R. and McNeil, A. (2001). Modelling dependent defaults. Preprint, ETH Zürich. available from `http://www.math.ethz.ch/~mcneil`.

[Jorion, 2000] Jorion, P. (2000). Risk management lessons from long-term capital management. *European Financial Management*, 6:277–300.

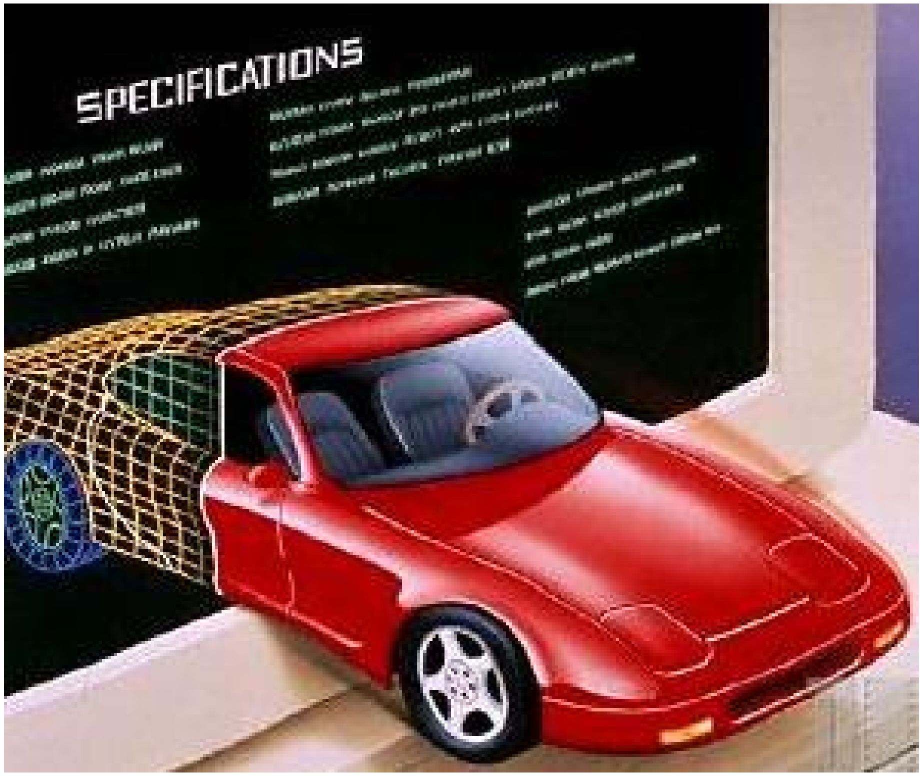
[Jorion, 2001] Jorion, P. (2001). *Value at Risk: the New Benchmark for Measuring Financial Risk*. McGraw-Hill, New York.

[McNeil and Frey, 2000] McNeil, A. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7:271–300.

[Scholes, 2000] Scholes, M. (2000). Crisis and risk management. *American Economic Review*, pages 17–22.

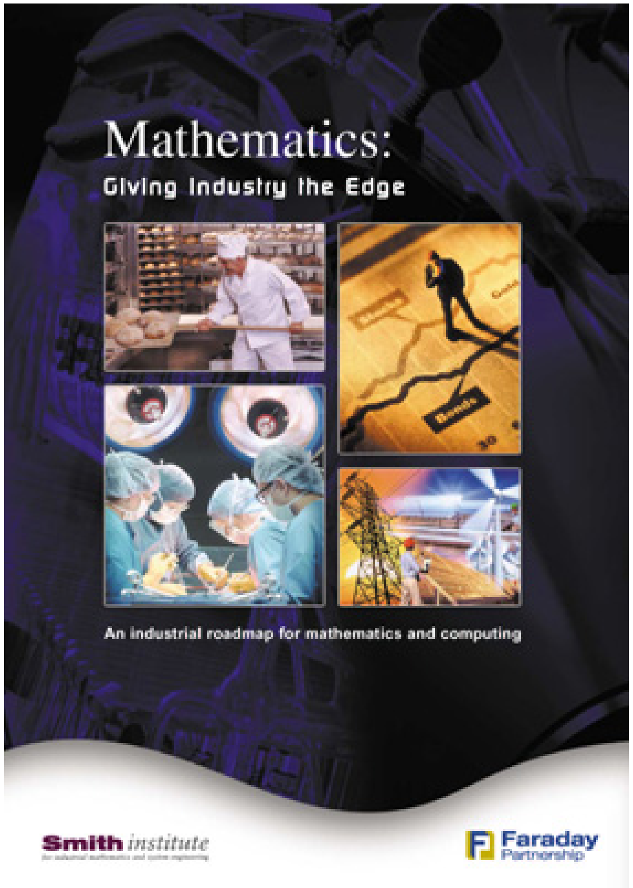
[Steinherr, 1998] Steinherr, A. (1998). *Derivatives. The Wild Beast of Finance*. Wiley, New York.

Mathematics in Industry and Government



Douglas N. Arnold

Institute for Mathematics and its Applications



Mathematics is the most versatile of all the sciences. It is uniquely well placed to respond to the demands of a rapidly changing economic landscape. Just as in the past, the systematic application of mathematics and computing to the most challenging industrial problems will be a vital contributor to business performance. The difference now is that the academic community must broaden its view of mathematics in industry and its expertise must be managed in more imaginative ways.

Mathematics now has the opportunity more than ever before to underpin quantitative understanding of industrial strategy and processes across all sectors of business. Companies that take best advantage of this opportunity will gain a significant competitive advantage: mathematics truly gives industry the edge.

The Odom Report

Academic mathematics is insufficiently connected to mathematics outside the university. One of the greatest—and most difficult—opportunities for academic mathematics is to build closer connections to industry.

Academic mathematical science must strike a better balance between theory and application. At one extreme, a narrowly inward-looking community will miss both the opportunities that arise outside the mathematical sciences and the opportunities that are part of scientific and technological developments. At the other extreme, an exclusive concern with applications and collaborative research would severely limit the mathematical sciences and deprive the scientific community of the full benefits of mathematical inquiry. At present, the balance is tilted too far towards inwardness.

A narrow vision of mathematics in academic departments translates into a narrow education for graduate students, most of whom are orinted toward careers only in academic mathematics.

Observations and opinions

- The potential impact of contemporary mathematics on science, on technology, and on industry is vast.
- Unfortunately, the actual impact—though great—is no where near as large as it should be.
- In significant part, this results from the decision of many mathematicians to address themselves to internally generated challenges rather than to the challenges that arise from the complexities of the modern world.
- Industrial mathematicians almost always face problems coming from outside mathematics.
- Industrial managers are convinced of the power of mathematics. . . they hire 25% of mathematics doctorates.

A problem from outside mathematics

Planning for and responding to the deliberate release of infectious agents is a clear example of a problem that mathematics cannot solve, but to which it can contribute immensely.

For a smallpox attack for example, many critical decisions have to be made. Examples:

- who to vaccinate (direct contacts of infected, neighborhoods of infected, essential personnel, the city, the country,. . . , healthy, at-risk, young, old, . . .)
- prophylactic vaccination?
- quarantine policy
- value of early detection
- value of diagnostic testing
- dealing with uncertainty

Math can help!

SIR model of mathematical epidemiology

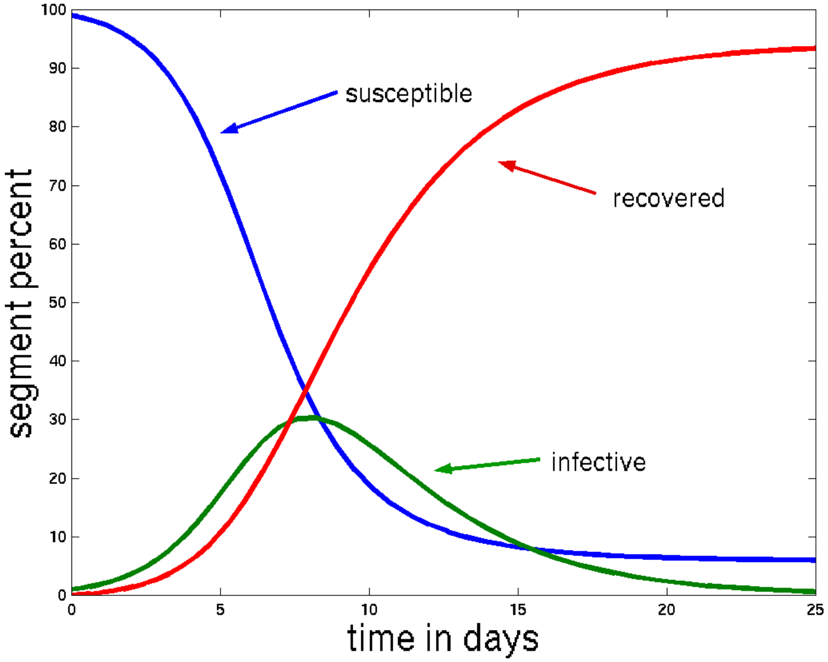
Daniel Bernoulli published a mathematical study of smallpox spread in 1760. In the 1920's Kermack and McKendrick formulated the SIR model:

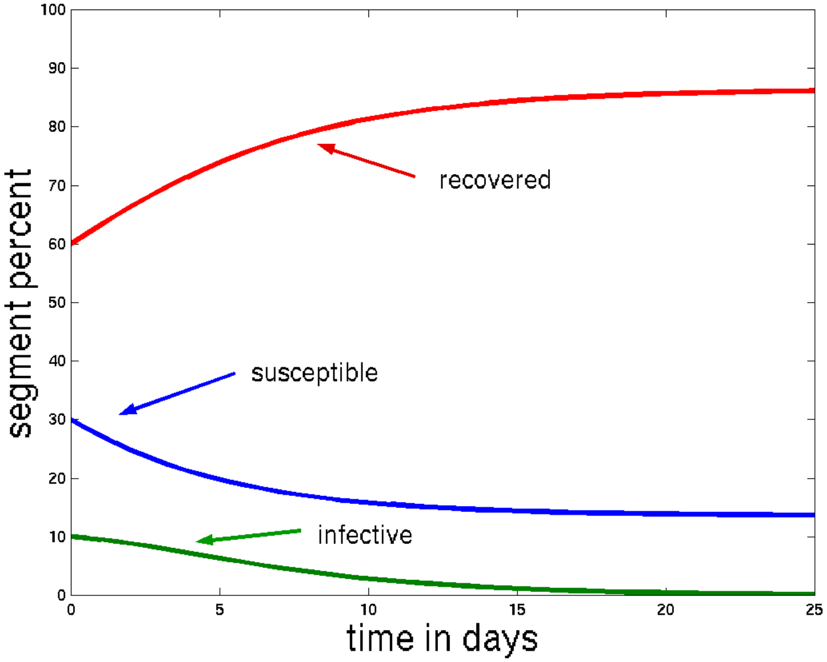
$$\frac{dS}{dt} = -\beta S I, \quad \frac{dI}{dt} = \beta S I - \gamma I, \quad \frac{dR}{dt} = \gamma I,$$

where $S + I + R = 1$ give the division of the population into susceptible, infective, and recovered segments, $\beta > 0$ the *infection rate*, $\gamma > 0$ the *removal rate*.

Threshhold theorem

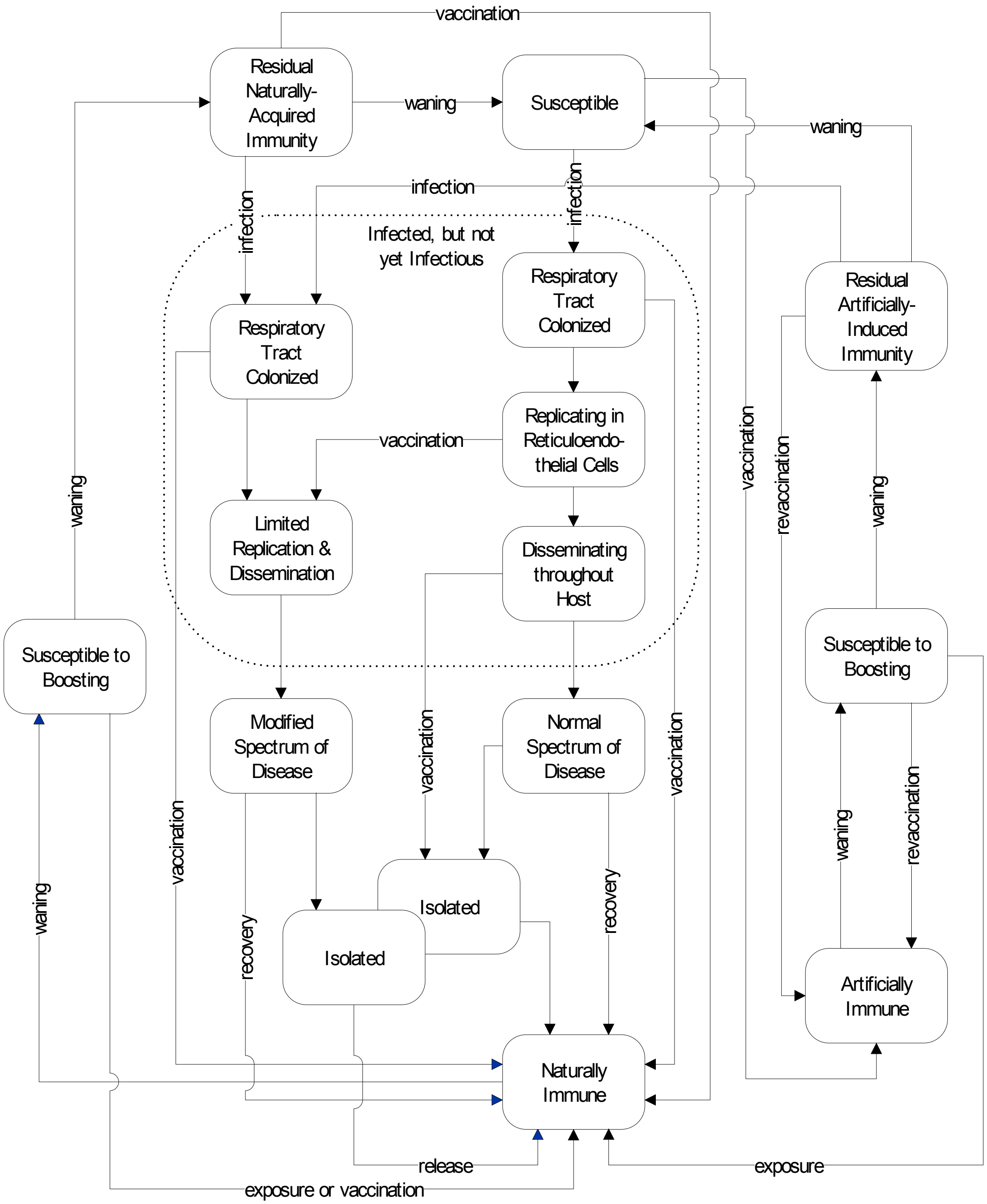
Theorem. *Let $S(0), I(0) > 0$, $R(0) = 1 - S(0) - I(0) \geq 0$ be given. For the solution of the SIR model with $S(0) > \gamma/\beta$, $I(t)$ increases initially until it reaches its maximum value and then decreases to zero at $t \to \infty$. Otherwise $I(t)$ decreases monotonically to zero as $t \to 0$.*





herd immunity

Smallpox modeling at the Center for Disease Control



vaccination
Residual Naturally-Acquired Immunity
waning
Susceptible
waning
infection
infection
infection
Infected, but not yet Infectious
Respiratory Tract Colonized
Respiratory Tract Colonized
Residual Artificially-Induced Immunity
Replicating in Reticuloendothelial Cells
vaccination
Limited Replication & Dissemination
Disseminating throughout Host
waning
vaccination
revaccination
waning
Susceptible to Boosting
Susceptible to Boosting
Modified Spectrum of Disease
vaccination
Normal Spectrum of Disease
vaccination
vaccination
waning
revaccination
Isolated
Isolated
recovery
recovery
waning
Artificially Immune
Naturally Immune
release
exposure
exposure or vaccination

Plague modeling at Dynamic Technology, Inc.

Multi-patch generalization of Keeling and Gilligan, 2000

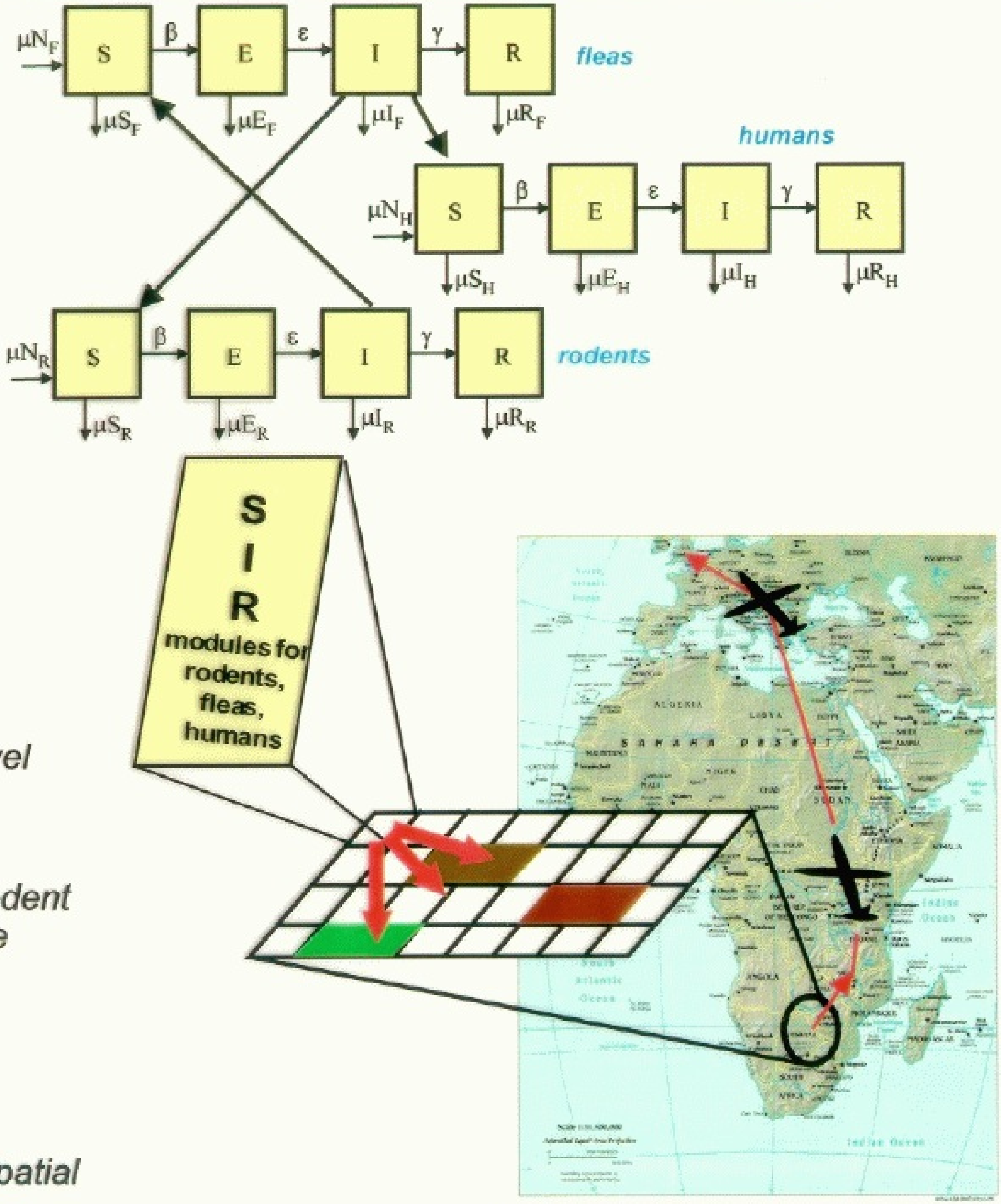
- *Treating patch-patch heterogeneity by Lloyd and May's (1996) approach*
- *Incorporating spatial spread (city-city to transnational) by modeling transportation networks, rates via Rvachev et al. (1977) approach*
- *Including human pneumonic transmission term*

Includes essential dimensions of plague epidemiology

- *Human, rodent and flea interactions*
- *Patch-patch ecological variation*
- *Regional, national and international travel and migrations*
- *Climatology and meteorology*
- *Effects of vaccination, rodent control, rodent genetic resistance to Y. pestis, pesticide application, and others*

Evaluation to include

- *Single-patch incidence, prevalence, R_0*
- *Patch-patch disease propagation and spatial spread*



www.nytimes.com
The New York Times
ON THE WEB

July 7, 2002

U.S. to Vaccinate 500,000 Workers Against Smallpox

By WILLIAM J. BROAD

The federal government will soon vaccinate roughly a half-million health care and emergency workers against smallpox as a precaution against a bioterrorist attack, federal officials said. The government is also laying the groundwork to carry out mass vaccinations of the public - a policy abandoned 30 years ago - if there is a large outbreak.

Until last month, officials had said they would soon vaccinate a few thousand health workers and would respond to any smallpox attack with limited vaccinations of the public. Since 1983, only 11,000 Americans who work with the virus and its related diseases have received a vaccination, according to the Centers for Disease Control and Prevention.

The plan to increase the number of "first responders" who receive the vaccination to roughly 500,000 from 15,000 and to prepare for a mass undertaking of vaccinations in effect acknowledges that the government's existing program is insufficient to fight a large outbreak.

Mathematical techniques relevant to bioterrorism

- mathematical epidemiology
- ODE, dynamical systems
- PDE
- numerical analysis, scientific computation
- probability, statistics
- graph theory, network analysis
- game theory
- control theory
- optimization
- . . .

Industries using mathematics

Aerospace

Automation and control

Automotive

Computing

Defense

Energy

Transportation

Financial services

Geosciences

Healthcare

Information Technology

Manufacturing

Telecommunication

Shipping

scores of others and increasing

Areas and Applications (MII '98)

Mathematical Area	Application
Algebra and number theory	Cryptography
Computational fluid dynamics	Aircraft and automobile design
Differential equations	Aerodynamics, porous media, finance
Discrete mathematics	Communication and information security
Formal systems and logic	Computer security, verification
Geometry	Computer-aided engineering and design
Nonlinear control	Operation of mechanical and electrical systems
Numerical analysis	Essentially all applications
Optimization	Asset allocation, shape and system design
Parallel algorithms	Weath modeling and prediction, crash simulation
Statistic	Design of experiments, analysis of large data sets
Stochastic processes	Signal analysis

Are all mathematical fields of interest to industry?

Just about, but some more so than others.

> *What kind of mathematics is useful? Every kind, but at Kodak partial differential equations are useful more often than topology. – Peter Castro*

Industry hired 50% of the 2001 PhDs in statistics, 43% in numerical analysis, and 10% of those in geometry/topology.

Field considerations

Field of specialization is a secondary condition in industry. An academic mathematician very well may spend his career working around the area of their thesis, but an industrial mathematician almost never does.

> *We never know what kind of mathematics is the right kinds, so an "algebraist for life" is not the right kind of mathematician.*

An industrial mathematician must be a generalist, learning whatever kind of mathematics the problem calls for. She should be interested in all kinds of mathematics, and also in things other than mathematics.

Depth in one area is certainly a plus, especially if the area seems relevant to the industry, but breadth is more important.

What do mathematicians bring to industry?

- logical thinking
- the ability to abstract and recognize underlying structure
- knowing the right questions, recognizing the wrong ones
- familiarity with a wide variety of problem-solving tools

Problems never come in formulated as mathematical problems. A mathematician's biggest contribution to a team is often an ability to state the right question.

What can't mathematics do for industry?

Solve its problems.

There are countless problems in industry that require deep mathematics, but almost none that can be solved by mathematics alone.

> *The strength of the mathematical sciences is that they are pervasive in many applications. The challenge is that they are only a part of each application. – Shmuel Winograd*

∴ a mathematician in industry must be part of a team.

∴ communication skills and social skills matter (while, according to popular opinion, these are positively harmful for an academic mathematician).

Traits of successful industrial mathematicians

- skills in modeling and problem formulation
- flexibility to go where the problems leads
- breadth of interest, interdisciplinarity
- balance between breadth and depth
- knowing when to stop
- computational skills
- written and oral communication skills
- social skills, teamwork

IMA Industrial Programs

- Industrial Problems Seminar
- Industrial math modeling workshop
- IMA Industrial Postdocs
- Hot topics workshops
- IMA Participating Corporation program
- symbiotic relation with MCIM

Recent IMA Industrial Problems Seminars

- Infectious Disease Modeling (Dynamics Technology Inc.)
- Micromagnetic Modeling of Writing and Reading Processes in Magnetic Recording (Seagate Technology)
- Mathematics and materials (3M)
- Mathematical modeling in support of service level agreements (Telcordia)
- Global Positioning Systems (Honeywell)
- F. John's Ultrahyperbolic Equation and 3D Computed Tomography (General Electric)
- Mathematical Modeling of Mechanical and Fluid Pressures in Chemical-Mechanical Polishing (Motorola)

Industrial math modeling workshop 2002

10 days of intensive work in 6 teams of 6 w/ industrial mentor.

- Designing Airplane Engine Struts using Minimal Surfaces (Boeing) differential geometry
- Mobility Management in Cellular Telephony (Telcordia) discrete math and optimization
- Optimal Pricing Strategy in Differentiated Durable-GoodsMarkets (Ford) game theory
- Modeling of Planarization in Chemical-Mechanical Polishing (Motorola) differential equations
- Modeling Networked Control Systems (Honeywell) graph theory, control theory
- Optimal Design for a Varying Environment (3M) differential equations, optimization

Institute for Mathematics and its Applications

IMA Industrial Postdocs

Time and funding is split 50–50% between the IMA and an industrial sponsor. Mentors at both organizations.

- Network design and optimization (Christine Cheng, Telcordia, McGill)
- Modeling of epicardial ablation (Jay Gopalakrishnan, Medtronic, U. Florida)
- Multiresolution approach to computer graphics (Radu Balan, IBM, Siemens)
- Diffractive and nonlinear optics (David Dobson, Telcordia, U. Utah, Siliconoptics)

Hot topics workshops

- E-auctions and markets (Ford and IBM)
- Modeling and analysis of noise in integrated circuits (Motorola)
- Mathematical challenges in global positioning systems (Lockheed Martin)
- Text Mining (West Group)
- Scaling phenomena in communications networks (AT&T and Telcordia)

Closing remarks

- Industry provides a rich source of problems involving a wide range of advanced mathematics.
- A math job in industry can provide intellectual challenge, a good salary, and a chance for real impact.
- The distinction between industrial mathematics and academic mathematics is more one of attitude than content.
- Future potential is tremendous potential. Mathematics can, and should, have much greater impact in the future.
- Traditional graduate math training helps develop several skills useful in industry, but downplays others.
- Many grad programs are adapting. Many programs for students are available (workshops, internships, conferences).

Action item

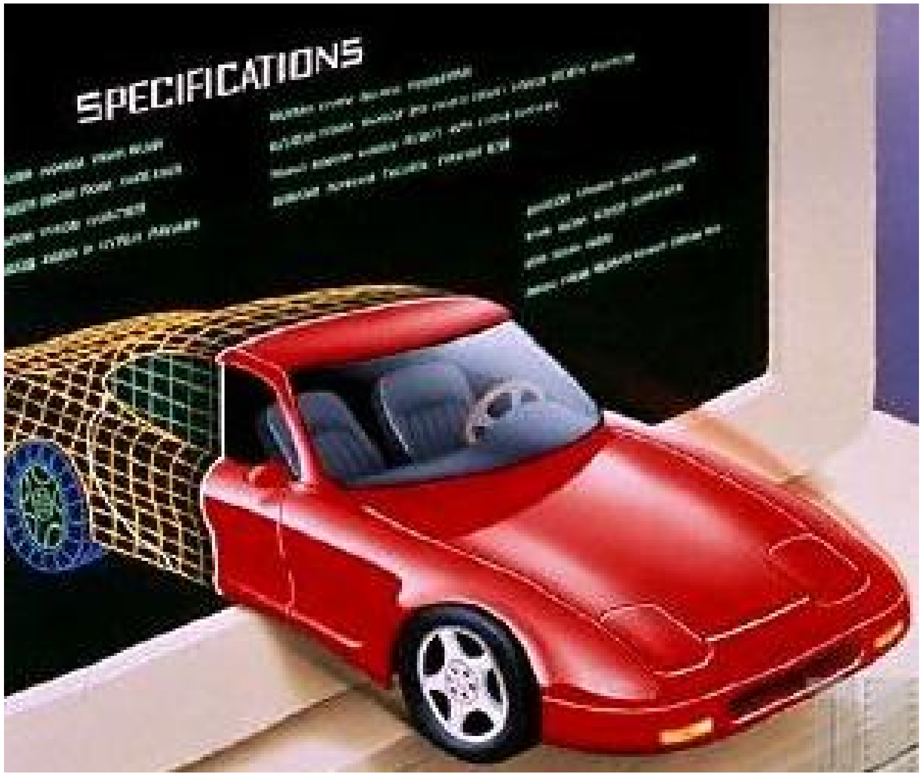
Encourage your students (and faculty) to think deeply about how they want to spend their lives, to collect information about the alternatives, to look outward as well as inward, to avail themselves of non-traditional and interdisciplinary programs, and to keep an open mind.

Two useful references

The SIAM Report on Mathematics in Industry (MII), 1998,
http://www.siam.org/mii/miihome.htm

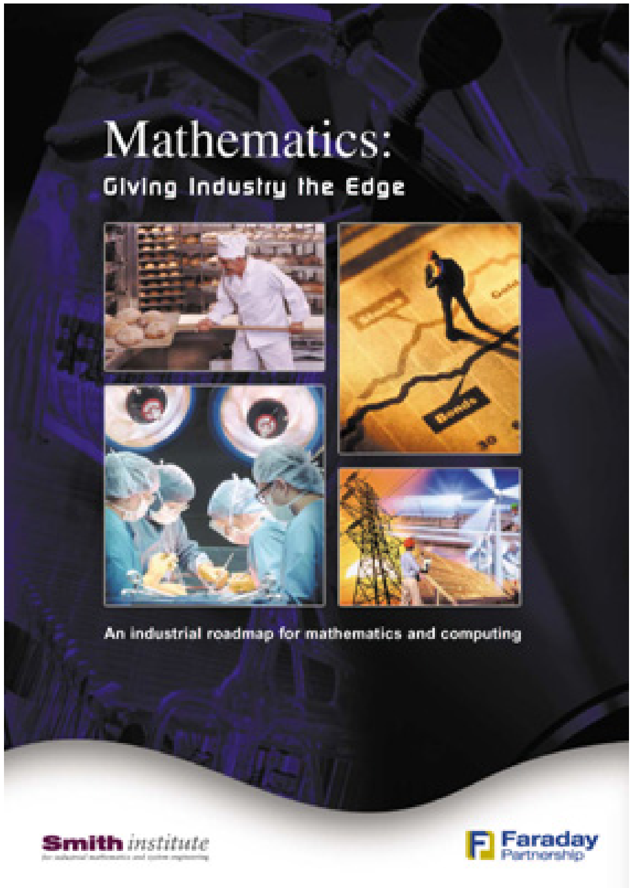
Mathematics: Giving Industry the Edge, 2002,
Smith Institute,
http://www.smithinst.ac.uk/news/RoadmapLaunch

Mathematics in Industry and Government



Douglas N. Arnold

Institute for Mathematics and its Applications



Mathematics:

Giving Industry the Edge

An industrial roadmap for mathematics and computing

Smith institute

Faraday Partnership

Mathematics is the most versatile of all the sciences. It is uniquely well placed to respond to the demands of a rapidly changing economic landscape. Just as in the past, the systematic application of mathematics and computing to the most challenging industrial problems will be a vital contributor to business performance. The difference now is that the academic community must broaden its view of mathematics in industry and its expertise must be managed in more imaginative ways.

Mathematics now has the opportunity more than ever before to underpin quantitative understanding of industrial strategy and processes across all sectors of business. Companies that take best advantage of this opportunity will gain a significant competitive advantage: mathematics truly gives industry the edge.

The Odom Report

Academic mathematics is insufficiently connected to mathematics outside the university. One of the greatest—and most difficult—opportunities for academic mathematics is to build closer connections to industry.

Academic mathematical science must strike a better balance between theory and application. At one extreme, a narrowly inward-looking community will miss both the opportunities that arise outside the mathematical sciences and the opportunities that are part of scientific and technological developments. At the other extreme, an exclusive concern with applications and collaborative research would severely limit the mathematical sciences and deprive the scientific community of the full benefits of mathematical inquiry. At present, the balance is tilted too far towards inwardness.

A narrow vision of mathematics in academic departments translates into a narrow education for graduate students, most of whom are orinted toward careers only in academic mathematics.

Observations and opinions

- The potential impact of contemporary mathematics on science, on technology, and on industry is vast.
- Unfortunately, the actual impact—though great—is no where near as large as it should be.
- In significant part, this results from the decision of many mathematicians to address themselves to internally generated challenges rather than to the challenges that arise from the complexities of the modern world.
- Industrial mathematicians almost always face problems coming from outside mathematics.
- Industrial managers are convinced of the power of mathematics. . . they hire 25% of mathematics doctorates.

A problem from outside mathematics

Planning for and responding to the deliberate release of infectious agents is a clear example of a problem that mathematics cannot solve, but to which it can contribute immensely.

For a smallpox attack for example, many critical decisions have to be made. Examples:

- who to vaccinate (direct contacts of infected, neighborhoods of infected, essential personnel, the city, the country,. . . , healthy, at-risk, young, old, . . .)
- prophylactic vaccination?
- quarantine policy
- value of early detection
- value of diagnostic testing
- dealing with uncertainty

Math can help!

SIR model of mathematical epidemiology

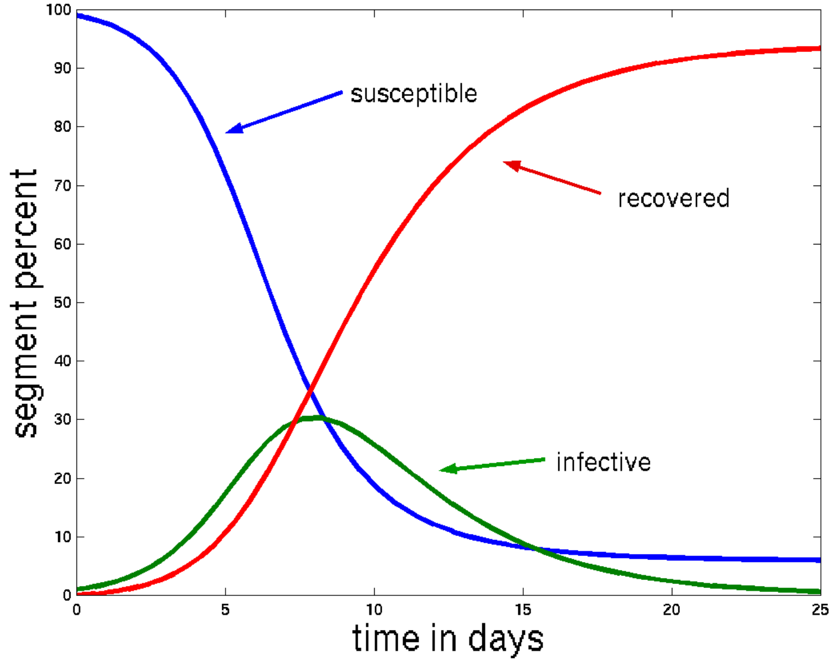
Daniel Bernoulli published a mathematical study of smallpox spread in 1760. In the 1920's Kermack and McKendrick formulated the SIR model:

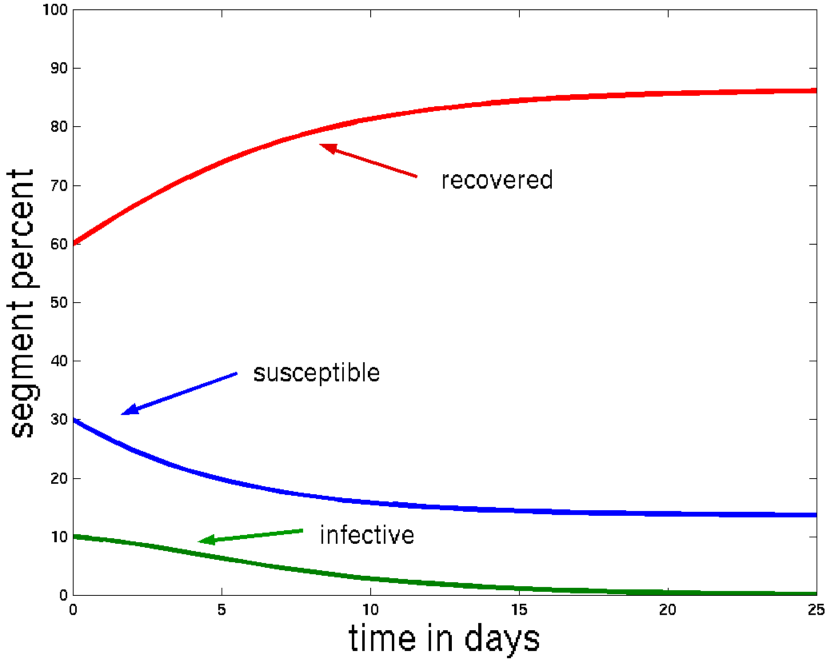
$$\frac{dS}{dt} = -\beta S I, \quad \frac{dI}{dt} = \beta S I - \gamma I, \quad \frac{dR}{dt} = \gamma I,$$

where $S + I + R = 1$ give the division of the population into susceptible, infective, and recovered segments, $\beta > 0$ the *infection rate*, $\gamma > 0$ the *removal rate*.

Threshhold theorem

Theorem. *Let $S(0), I(0) > 0$, $R(0) = 1 - S(0) - I(0) \geq 0$ be given. For the solution of the SIR model with $S(0) > \gamma/\beta$, $I(t)$ increases initially until it reaches its maximum value and then decreases to zero at $t \to \infty$. Otherwise $I(t)$ decreases monotonically to zero as $t \to 0$.*

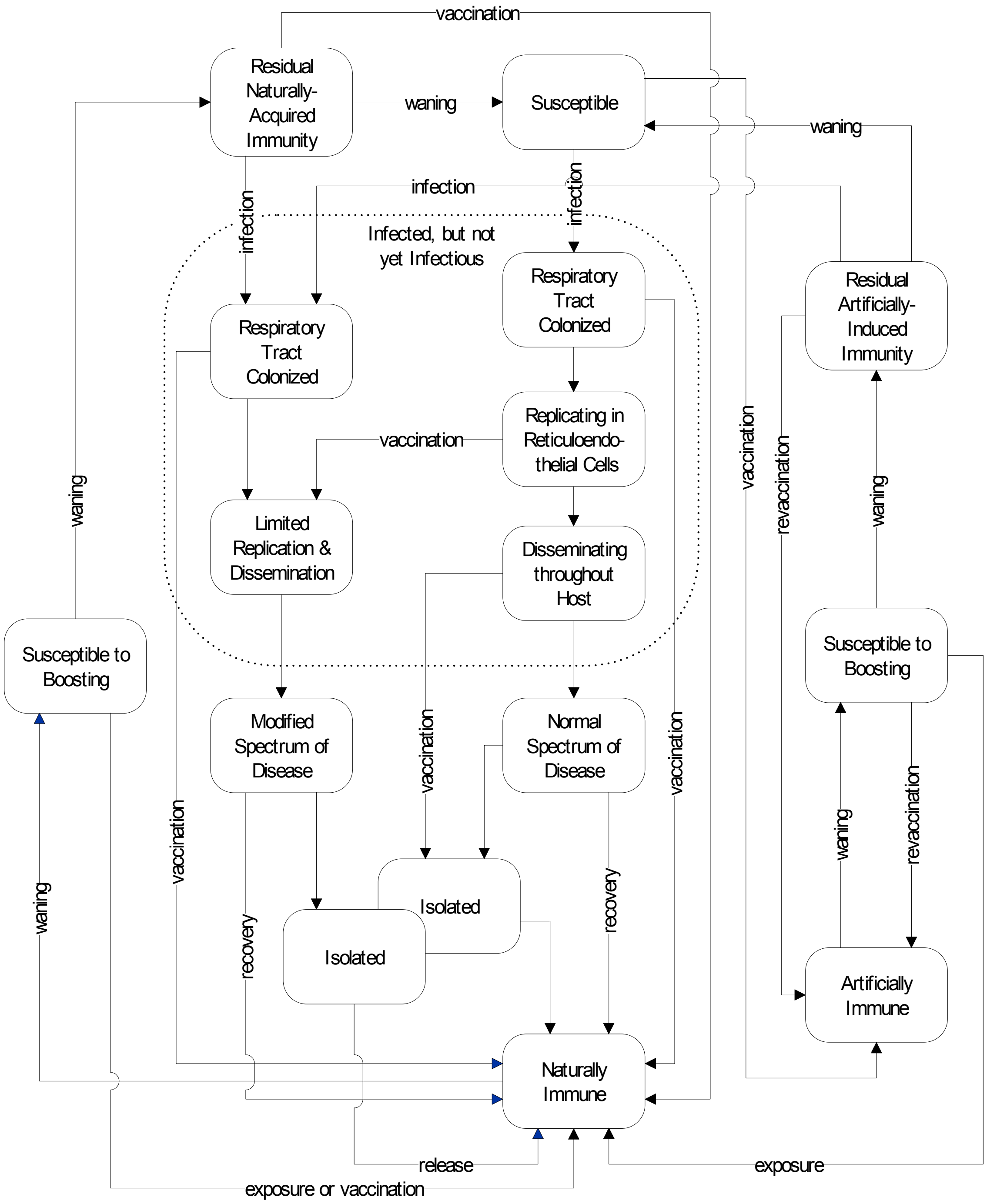




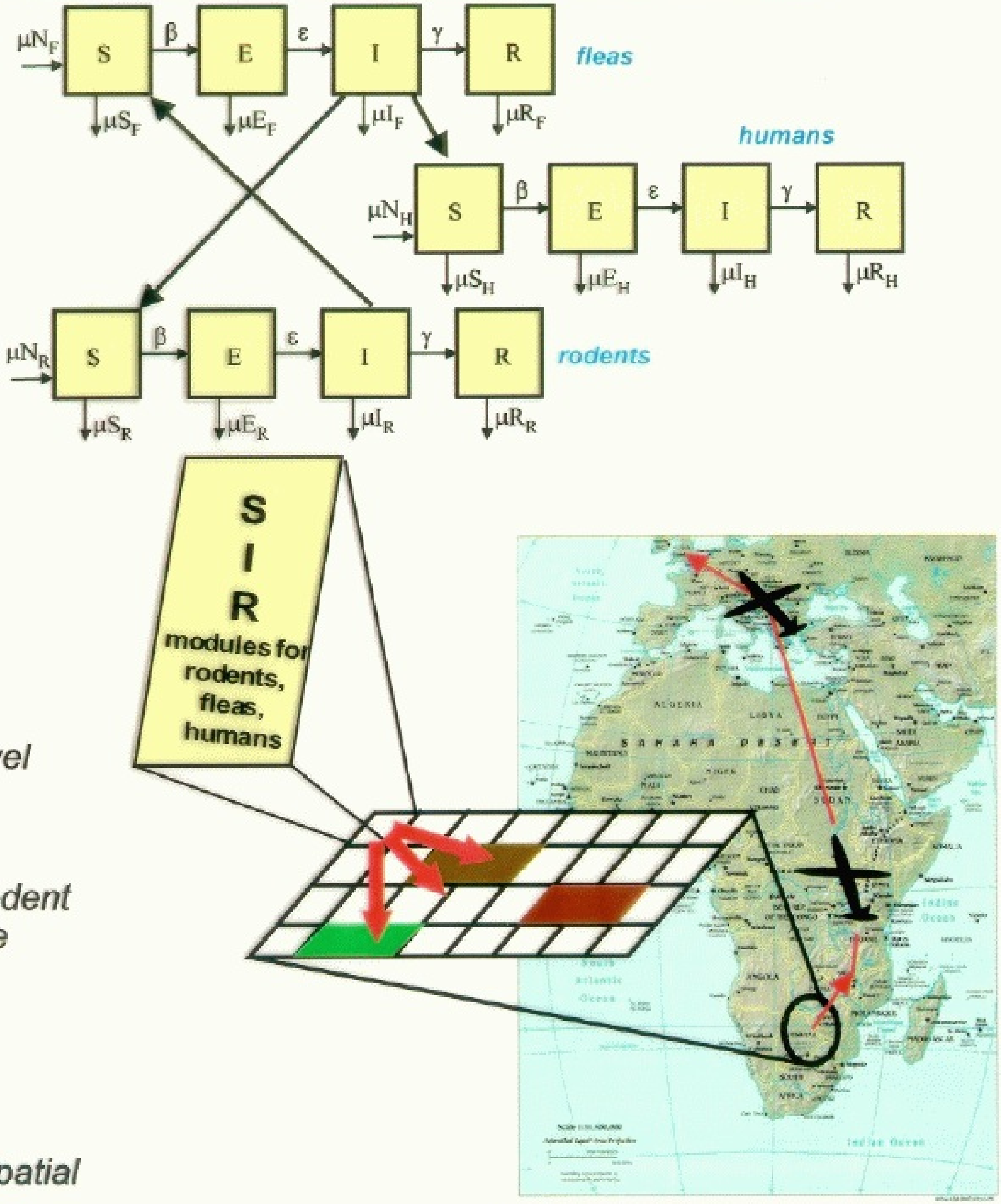
herd immunity

Institute for Mathematics and its Applications

Smallpox modeling at the Center for Disease Control


vaccination
Residual Naturally-Acquired Immunity
waning
Susceptible
waning
infection
infection
infection
Infected, but not yet Infectious
Respiratory Tract Colonized
Respiratory Tract Colonized
Residual Artificially-Induced Immunity
Replicating in Reticuloendo-thelial Cells
vaccination
Limited Replication & Dissemination
Disseminating throughout Host
waning
vaccination
revaccination
waning
Susceptible to Boosting
Susceptible to Boosting
Modified Spectrum of Disease
vaccination
Normal Spectrum of Disease
vaccination
vaccination
waning
revaccination
Isolated
Isolated
recovery
recovery
waning
Artificially Immune
Naturally Immune
release
exposure
exposure or vaccination

Plague modeling at Dynamic Technology, Inc.



- Multi-patch generalization of Keeling and Gilligan, 2000
  - *Treating patch-patch heterogeneity by Lloyd and May's (1996) approach*
  - *Incorporating spatial spread (city-city to transnational) by modeling transportation networks, rates via Rvachev et al. (1977) approach*
  - *Including human pneumonic transmission term*
- Includes essential dimensions of plague epidemiology
  - *Human, rodent and flea interactions*
  - *Patch-patch ecological variation*
  - *Regional, national and international travel and migrations*
  - *Climatology and meteorology*
  - *Effects of vaccination, rodent control, rodent genetic resistance to Y. pestis, pesticide application, and others*
- Evaluation to include
  - *Single-patch incidence, prevalence, R_0*
  - *Patch-patch disease propagation and spatial spread*

www.nytimes.com
The New York Times
ON THE WEB

July 7, 2002

U.S. to Vaccinate 500,000 Workers Against Smallpox

By WILLIAM J. BROAD

The federal government will soon vaccinate roughly a half-million health care and emergency workers against smallpox as a precaution against a bioterrorist attack, federal officials said. The government is also laying the groundwork to carry out mass vaccinations of the public - a policy abandoned 30 years ago - if there is a large outbreak.

Until last month, officials had said they would soon vaccinate a few thousand health workers and would respond to any smallpox attack with limited vaccinations of the public. Since 1983, only 11,000 Americans who work with the virus and its related diseases have received a vaccination, according to the Centers for Disease Control and Prevention.

The plan to increase the number of "first responders" who receive the vaccination to roughly 500,000 from 15,000 and to prepare for a mass undertaking of vaccinations in effect acknowledges that the government’s existing program is insufficient to fight a large outbreak.

Mathematical techniques relevant to bioterrorism

- mathematical epidemiology
- ODE, dynamical systems
- PDE
- numerical analysis, scientific computation
- probability, statistics
- graph theory, network analysis
- game theory
- control theory
- optimization
- . . .

Industries using mathematics

Aerospace

Automation and control

Automotive

Computing

Defense

Energy

Transportation

Financial services

Geosciences

Healthcare

Information Technology

Manufacturing

Telecommunication

Shipping

scores of others and increasing

Areas and Applications (MII '98)

Mathematical Area	**Application**
Algebra and number theory	Cryptography
Computational fluid dynamics	Aircraft and automobile design
Differential equations	Aerodynamics, porous media, finance
Discrete mathematics	Communication and information security
Formal systems and logic	Computer security, verification
Geometry	Computer-aided engineering and design
Nonlinear control	Operation of mechanical and electrical systems
Numerical analysis	Essentially all applications
Optimization	Asset allocation, shape and system design
Parallel algorithms	Weath modeling and prediction, crash simulation
Statistic	Design of experiments, analysis of large data sets
Stochastic processes	Signal analysis

Are all mathematical fields of interest to industry?

Just about, but some more so than others.

> *What kind of mathematics is useful? Every kind, but at Kodak partial differential equations are useful more often than topology. – Peter Castro*

Industry hired 50% of the 2001 PhDs in statistics, 43% in numerical analysis, and 10% of those in geometry/topology.

Field considerations

Field of specialization is a secondary condition in industry. An academic mathematician very well may spend his career working around the area of their thesis, but an industrial mathematician almost never does.

> *We never know what kind of mathematics is the right kinds, so an "algebraist for life" is not the right kind of mathematician.*

An industrial mathematician must be a generalist, learning whatever kind of mathematics the problem calls for. She should be interested in all kinds of mathematics, and also in things other than mathematics.

Depth in one area is certainly a plus, especially if the area seems relevant to the industry, but breadth is more important.

What do mathematicians bring to industry?

- logical thinking
- the ability to abstract and recognize underlying structure
- knowing the right questions, recognizing the wrong ones
- familiarity with a wide variety of problem-solving tools

Problems never come in formulated as mathematical problems. A mathematician's biggest contribution to a team is often an ability to state the right question.

What can't mathematics do for industry?

Solve its problems.

There are countless problems in industry that require deep mathematics, but almost none that can be solved by mathematics alone.

> *The strength of the mathematical sciences is that they are pervasive in many applications. The challenge is that they are only a part of each application. – Shmuel Winograd*

∴ a mathematician in industry must be part of a team.

∴ communication skills and social skills matter (while, according to popular opinion, these are positively harmful for an academic mathematician).

Traits of successful industrial mathematicians

- skills in modeling and problem formulation
- flexibility to go where the problems leads
- breadth of interest, interdisciplinarity
- balance between breadth and depth
- knowing when to stop
- computational skills
- written and oral communication skills
- social skills, teamwork

IMA Industrial Programs

- Industrial Problems Seminar
- Industrial math modeling workshop
- IMA Industrial Postdocs
- Hot topics workshops
- IMA Participating Corporation program
- symbiotic relation with MCIM

Recent IMA Industrial Problems Seminars

- Infectious Disease Modeling (Dynamics Technology Inc.)
- Micromagnetic Modeling of Writing and Reading Processes in Magnetic Recording (Seagate Technology)
- Mathematics and materials (3M)
- Mathematical modeling in support of service level agreements (Telcordia)
- Global Positioning Systems (Honeywell)
- F. John's Ultrahyperbolic Equation and 3D Computed Tomography (General Electric)
- Mathematical Modeling of Mechanical and Fluid Pressures in Chemical-Mechanical Polishing (Motorola)

Industrial math modeling workshop 2002

10 days of intensive work in 6 teams of 6 w/ industrial mentor.

- Designing Airplane Engine Struts using Minimal Surfaces (Boeing) differential geometry
- Mobility Management in Cellular Telephony (Telcordia) discrete math and optimization
- Optimal Pricing Strategy in Differentiated Durable-GoodsMarkets (Ford) game theory
- Modeling of Planarization in Chemical-Mechanical Polishing (Motorola) differential equations
- Modeling Networked Control Systems (Honeywell) graph theory, control theory
- Optimal Design for a Varying Environment (3M) differential equations, optimization

Institute for Mathematics and its Applications

IMA Industrial Postdocs

Time and funding is split 50–50% between the IMA and an industrial sponsor. Mentors at both organizations.

- Network design and optimization (Christine Cheng, Telcordia, McGill)
- Modeling of epicardial ablation (Jay Gopalakrishnan, Medtronic, U. Florida)
- Multiresolution approach to computer graphics (Radu Balan, IBM, Siemens)
- Diffractive and nonlinear optics (David Dobson, Telcordia, U. Utah, Siliconoptics)

Hot topics workshops

- E-auctions and markets (Ford and IBM)
- Modeling and analysis of noise in integrated circuits (Motorola)
- Mathematical challenges in global positioning systems (Lockheed Martin)
- Text Mining (West Group)
- Scaling phenomena in communications networks (AT&T and Telcordia)

Closing remarks

- Industry provides a rich source of problems involving a wide range of advanced mathematics.
- A math job in industry can provide intellectual challenge, a good salary, and a chance for real impact.
- The distinction between industrial mathematics and academic mathematics is more one of attitude than content.
- Future potential is tremendous potential. Mathematics can, and should, have much greater impact in the future.
- Traditional graduate math training helps develop several skills useful in industry, but downplays others.
- Many grad programs are adapting. Many programs for students are available (workshops, internships, conferences).

Institute for Mathematics and its Applications

Action item

Encourage your students (and faculty) to think deeply about how they want to spend their lives, to collect information about the alternatives, to look outward as well as inward, to avail themselves of non-traditional and interdisciplinary programs, and to keep an open mind.

Two useful references

The SIAM Report on Mathematics in Industry (MII), 1998,
http://www.siam.org/mii/miihome.htm

Mathematics: Giving Industry the Edge, 2002,
Smith Institute,
http://www.smithinst.ac.uk/news/RoadmapLaunch