

# Scientific Decision-making

# CHAPTER 1: WHY SCIENCE MATTERS



THE SCIENTIFIC METHOD

In our complicated world, our lives depend on many, many events and decisions outside of our immediate control as well as many within our control. Science as a way of making rational, evidence-based decisions about the natural world offers the best method we have of ensuring those decisions achieve what we want.

## SECTION 1

# Society's Responsibilities

Modern society is full of potential hazards. For many of those hazards we need public policies that protect us.

Virtually everything we do to better ourselves has a downside. Industries that make goods cause pollution, whether it be a copper smelter releasing toxic gasses or runoff from a corn field contaminating ground water with pesticides and fertilizer. Consider the innumerable issues we face in our society:

**food additives and processed food**

**medical practices and drugs**

**agricultural practices: pesticides, fertilizers, genetically modified organisms**

**pollution: chemical, radioactivity**

**environmental hazards of consumer products (freon, electromagnetic fields, lead in gasoline)**

**product safety (autos, electrical appliances, infant seats)**

**transportation safety**



Cholesterol	Less than	300mg	300mg
Sodium	Less than	2,400mg	2,400mg
Total Carbohydrate		300g	375g
Dietary Fiber		25g	30g

bakery products. We invite your comments and questions.

N85038.0B\_MP CHOCOLATE CHIP MINI MUFFINS

INGREDIENTS: SUGAR, ENRICHED BLEACHED WHEAT FLOUR [FLOUR, MALTED BARLEY FLOUR, REDUCED IRON, "B" VITAMINS (NIACIN, THIAMINE MONONITRATE (B1), RIBOFLAVIN (B2), FOLIC ACID)], WHOLE EGGS, CANOLA OIL, WATER, CHOCOLATE LIQUOR, CORN SYRUP, EGG WHITES, COCOA BUTTER, SOY LECITHIN. CONTAINS 2% OR LESS OF: MODIFIED CORN STARCH, MONO AND DIGLYCERIDES, WHEY, SODIUM STEAROYL LACTYLATE, LEAVENINGS (BAKING SODA, SODIUM ALUMINUM PHOSPHATE, ALUMINUM SULFATE), SOYBEAN OIL, SALT, WHEAT GLUTEN, POLYSORBATE 60, GLUCOSE, NATURAL AND ARTIFICIAL FLAVOR, NONFAT MILK SOLIDS, XANTHAN GUM, CALCIUM ACETATE, DEXTROSE, GUAR GUM, CITRIC ACID, SORBIC ACID (TO RETAIN FRESHNESS) 517501  
CONTAINS WHEAT, EGG, MILK AND SOY.  
MAY CONTAIN WALNUTS  
239124F\_MP



This list barely scratches the surface of the issues that affect us, and indeed, it only gives broad categories of possible hazards.

Many hazards can only be controlled at the level of society: as individuals, we cannot prevent farmers from using pesticides and fertilizer, prevent other consumers from releasing freon, from using leaded gasoline, etc. Nor is it a good idea to wait to address problems after the fact – when public health has suffered – if we can anticipate and thereby prevent them. We thus expect our government to protect us and to make decisions accordingly.

Our government does make those decisions. From an environmental perspective, we have banned or otherwise curtailed use of DDT, of (some types of) freon, of leaded gasoline, and of some second-hand exposures to tobacco smoke. Yet we continue to use many other things that are potentially harmful: carbon emissions are not taxed or regulated despite evidence of their contribution to global climate change, many other pesticides are still in use, and many aquifers and grasslands vital to the national interest are being exploited beyond their capacity to recharge. Likewise, most attempts by industry to market new drugs are prevented by the government because the drug is deemed harmful.



# People Believe Weird Things

Our perceptions cannot be trusted.

In our society, there is a disconnect between how we expect our government to behave regarding decisions (and how it does behave) relative to how individuals behave. There is a stunning, large fraction of U.S. citizens who say they believe in some aspect of the "paranormal" and other scientifically unfounded ideas.

CONCEPT	% CLAIMING TO BELIEVE IT
Astrology	52%
ESP	46%
Witches	19%
Aliens have landed	22%
Atlantis	33%
Dinosaurs with humans	41%
Communication with dead	42%
Had a psychic experience	67%
Ghosts	35%

(based on a 1991 poll of 1,236 Americans; Gallup, G.H. Jr, and F. Newport. 1991. *Skeptical inquirer* 15:137-147). Likewise, even many of us in this class at least suspect there is some validity to several of these ideas (our first-day survey).

Belief in magic, aliens, and recent dinosaurs is undoubtedly harmless in most cases and can even be entertaining – people rarely carry such beliefs to extremes that might harm themselves, and believing in astrology can take some of the dullness out of life, just as most of us read the fortune in our fortune cookie. There could be many social repercussions when a large fraction of a population does not know how to decide what is real -- wholesale criminal convictions of innocent people, failures to make medical and technical advances, failures to make other improvements in the standard of living, a decay in education systems, and much more. To a large extent, however, we do not expect the government to follow such beliefs.

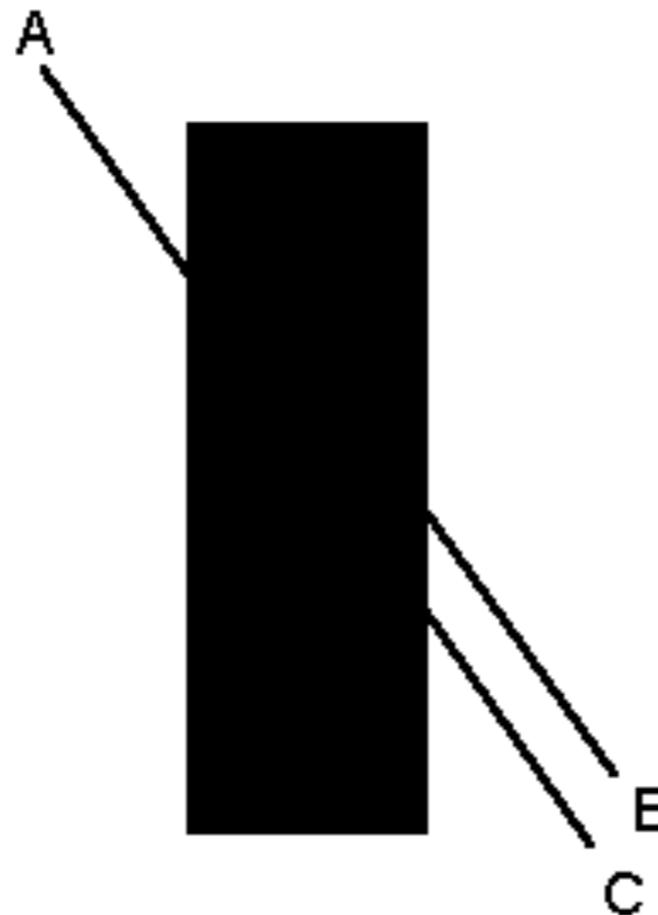
# Our Brains are Inherently Flawed

It is perhaps not our fault that we don't automatically know how to interpret Nature. There has been a moderate flood of books recently on the many fallacies of our brains:

TITLE	AUTHOR
Predictably Irrational	Dan Ariely
Thinking Fast and Slow	Daniel Kahneman
The Folly of Fools	Robert Trivers
The Myth of Repressed Memory	Elizabeth Loftus

These books offer many examples of fallacies in our thinking, most of which we are not aware of when we commit them. Some of the examples are stunning if not frightening: judges exhibiting a profoundly higher rate of granting paroles immediately after lunch, and advertisers exploiting our subconscious tendencies to be swayed in predictable directions by the choices offered – merely adding an unattractive choice can drive us to prefer the more expensive options. Loftus has researched for years the fallibilities of our memories, especially its impact on eyewitness testimony. Trivers argues that our brains have in fact been evolved to deceive us, rather than convey an unbiased perception of our world.

One of the easiest and most convincing demonstrations of the hard-wired imperfections of our brains is optical illusions. The figure below is easily displayed on a black and white page. The straight line appears to be AB but is really AC. Magic tricks often rely on illusions of sorts. Lecture will offer others.



# Understanding Nature

is not about being clever, it is about evidence.

To an outsider, this point is perhaps counter-intuitive, yet it defines the essence of the scientific method. Humans have a long history of failing to explain or predict Nature from first principles. The central pillar of science is that we need to observe Nature to know Nature – we need to look at the evidence. In mathematics, it is possible to prove result: a proof begins with a set of assumptions that defines the rules, and from those it becomes possible to define strict outcomes. In the real world, we never know the rules that bound a problem. So we make up approximate rules, work out the consequences and see if Nature fits. And it never fits perfectly, only approximately. When Nature actually does appear to fit, however, it does not mean that the rules are true or complete. At best it means that the rules are approximately true for that circumstance, but when we move to a new problem, the set of rules may be somewhat different. Science is like a patchwork quilt of evidence and the stories we have built to explain that evidence.

Thus with respect to science, it does not actually matter that our brains are flawed. Our perceptions could be accurate, and understanding Nature would still not be automatic. Our logic could be perfect and that would not be enough. We would still need a methodology for comparing our guesses and ideas to the evidence and then deciding whether an imperfect match between them is close enough.

# Decisions Involve More Than Just Evidence

All of us exhibit a range of beliefs – how strongly we accept something differs from one subject to another and from person to person. Our individual beliefs typically depend on a combination of factors, including but not limited to the following

1. the evidence – what you know about the issue
2. compatibility with your world view
3. reliability of the source
4. consequences of accepting/doubting

As these factors will usually vary from person to person, we can easily understand how two people will differ in their belief on any topic. Furthermore, it may often be appropriate that personal decisions be influenced by a variety of factors. Yet when the chosen option needs to be the one that lies closest to the natural truth, we want to go with the evidence – the one supported by science.

# What follows in this book

This book is about a method that empowers people and institutions achieve their specific goals. This method is widely known as the scientific method, though this term is a misnomer. Not only do scientists solve problems using this method, but it is also the mainstay of improvement in business and industry, and it provides a unique perspective on social institutions. Our goal in this class is to teach you how to use the scientific method and apply it to everyday health and social issues both for personal matters and to be informed about decisions made by our government. If your career is one in which you will be called on to solve problems, whether in business, law, or government, this style of thinking should be helpful in those areas as well.

However, because of this goal, the class emphasizes critical thinking rather than memorization of facts. In teaching you to tackle novel situations, we will teach you to analyze arguments and descriptions of new findings. For example, you will be given short news articles and asked to interpret the articles and to identify whether the research has certain features. So if your goal in taking a nonmajors biology class is to obtain an encyclopedic knowledge of biological facts, this class is not for you. But if you want to know how to identify weaknesses of a study or how to identify potential science frauds and cons, then this class should serve that purpose. Below, we list a few more examples of the ways that this class might help you as a nonscientist to think about everyday problems.

EXAMPLE	ISSUE
Being tested for illegal drugs	Do you know what testing practices best ensure your civil rights against erroneous test results?
A new study claiming that modest alcohol consumption improves longevity	Could you tell whether this study indicates that you should drink alcohol?
A 3-year study showing that 1 of every 200 university students carries HIV (the AIDS virus)	What does this number indicate about the chance that your partner is infected?
A juror being asked to decide the guilt of a rape suspect based on DNA evidence	How might the prosecution and defense each present a biased appraisal of the evidence?

# External Links

[Astrology Debunked - Richard Dawkins in Enemies of Reason](#)

[Reverse to Smaller Agriculture](#)

## CHAPTER 2: A TEMPLATE FOR SCIENTIFIC INQUIRY



Despite the complexities of studies conducted in different scientific fields, there is an underlying structure common to all. This structure involves 5 basic elements: goals, models, data, evaluation, and revision.

# Five Elements

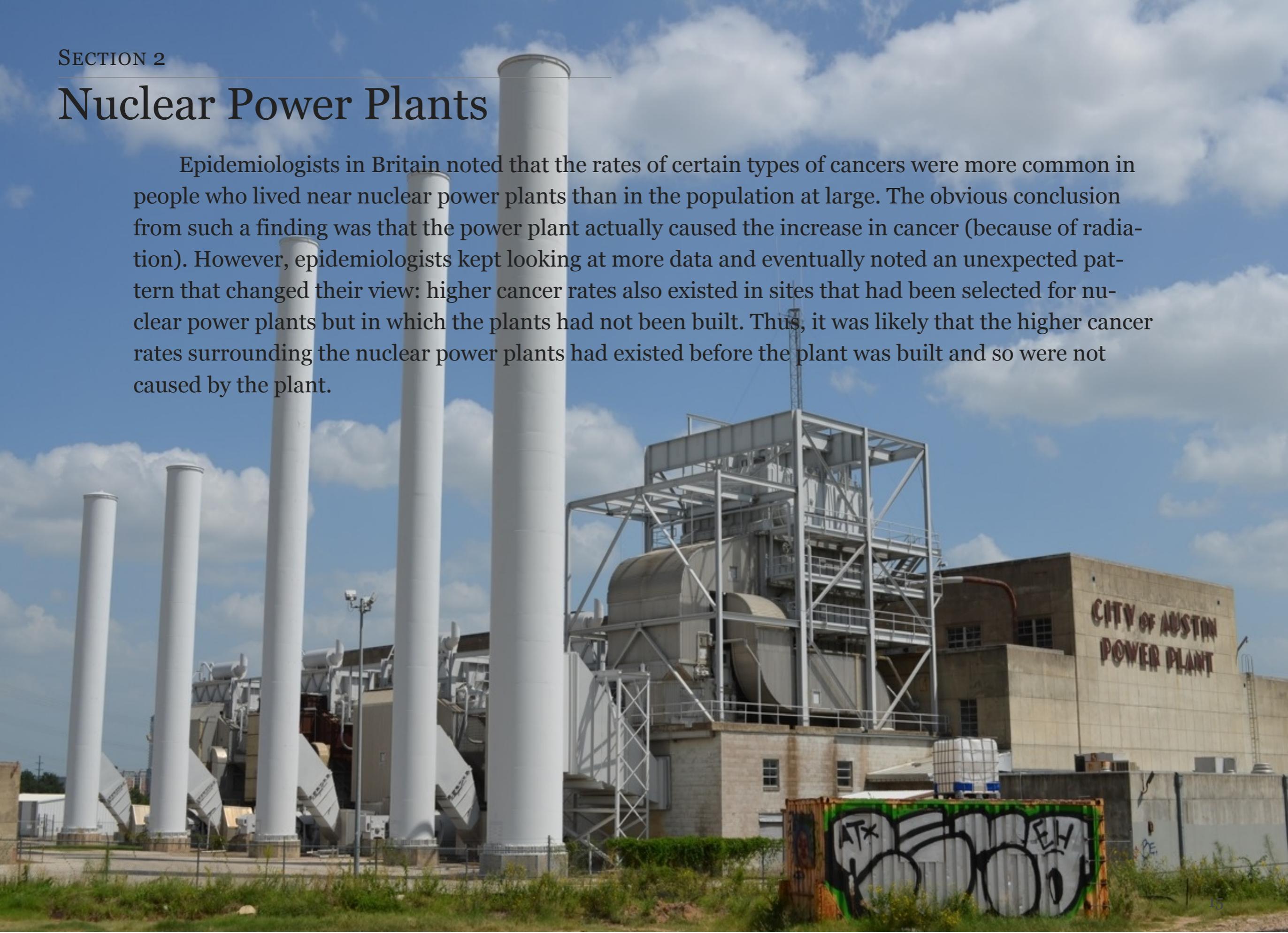
Five elements are found in most applications of the scientific method. Understanding these elements will enable you to understand both how to use the scientific method and its limitations. In template form, these 5 elements are:

## Scientific Method Template

GOAL	the objective of doing the study
MODEL	any and all abstractions of what is being studied or manipulated
DATA	observations made to represent "nature" for testing the model
EVALUATION	comparing the model to the data, to decide if the model is okay
REVISION	changing the model if it is not okay

# Nuclear Power Plants

Epidemiologists in Britain noted that the rates of certain types of cancers were more common in people who lived near nuclear power plants than in the population at large. The obvious conclusion from such a finding was that the power plant actually caused the increase in cancer (because of radiation). However, epidemiologists kept looking at more data and eventually noted an unexpected pattern that changed their view: higher cancer rates also existed in sites that had been selected for nuclear power plants but in which the plants had not been built. Thus, it was likely that the higher cancer rates surrounding the nuclear power plants had existed before the plant was built and so were not caused by the plant.



This simple example contains all five elements of our template:

**Goal:** The goal is to identify (and ultimately reduce) environmental causes of cancer.

**Models:** The most conspicuous model - and the one of greatest concern - is that a nuclear plant is the cause of increased cancer rates. It is a model because it is a description of what might be occurring in and around the nuclear power plants.

**Data:** The data are merely the cancer rates in people living in different locations. Data were analyzed in two sets, however. Set 1: cancer rates in people living near power plants and cancer rates in the population at large. Set 2: cancer rates in people living at sites selected for construction but where the plant was not yet built.

**Evaluation:** The model is fairly specific about which groups of people should show elevated cancer rates, so the evaluation can be performed without any sophisticated analysis: the model can only explain higher cancer rates around existing power plants. The first set of data is consistent with the model, whereas set 2 is not consistent with the model.

**Revision:** We reject the model because it is not consistent with both data sets. In the next round of applying the scientific method, we would consider an alternative model such as: increased cancer rates are caused by something associated with the sites chosen for nuclear power plants.

Using our template:

<b>SCIENTIFIC METHOD TEMPLATE</b>	
GOAL	identify environmental causes of cancer
MODEL	power plants cause cancer
DATA (1st set)	higher cancer rates near power plants
DATA (2nd set)	higher cancer rates at proposed sites
EVALUATION	the model is consistent with data set 1 but inconsistent with data set 2
REVISION	reject the model and choose an alternative

# Fetal Alcohol Syndrome

As little as 30 years ago, excessive alcohol was known to be a health risk to the drinker, but there was no public awareness of its possible impact on the fetus developing in a pregnant woman. In short, no one worried about it, and people were willing to assume that alcohol consumption by the mother was not a problem for the fetus. In the language of science, we would say that this public indifference was in fact an implicit model: alcohol consumption had no lasting effect on the fetus.

The first scientific studies on this topic were published in the early 1970s and demonstrated that women who drink a lot of alcohol during pregnancy have a much higher-than-average chance of producing an offspring suffering from mental retardation and various facial deformations. These data thus rejected that original model in favor of a model in which excessive alcohol consumption caused birth defects.

Later, in the 1980s, it was discovered that drinking even modest amounts of alcohol could cause the child to suffer learning disabilities and to be comparatively inept at certain physical tasks. At this point, the data supported a model in which consumption of moderate as well as excessive doses of alcohol could cause birth defects, with the dose corresponding to the degree of consumption.

Even with this progress, questions remain unanswered today, including whether drinking less than two drinks per day has any effect, and whether drinking in the first month of the pregnancy has a different effect than does drinking in the second and third months. That is, we aren't able to discriminate between models in which sporadic, light consumption of alcohol has slight, lasting effects on the fetus versus models in which such consumption has no effect.

Using our template:

<b>SCIENTIFIC METHOD TEMPLATE</b>	
GOAL	determine the impact of maternal alcohol drinking on the fetus
MODEL	alcohol has no effect on birth defects
DATA (1970s)	obvious birth defects are associated with excessive maternal consumption
EVALUATION	the model is inconsistent with the data
REVISION	the model is rejected, and a new one is adopted in which maternal drinking causes birth defects

# The Wright Brothers

An example of historical interest to Americans is the invention of powered, flying aircraft carrying a passenger. This invention is widely credited to Wilber and Orville Wright, on December 17, 1903. In today's culture of world travel by jet, it is virtually incomprehensible that, only 101 years ago, flight had not been achieved. It is even more stunning that such an important "first" in human invention was accomplished by a pair of bicycle shop owners with no formal training in science or engineering. Yet, a careful investigation of the steps leading to this invention reveals that the Wright brothers relied heavily on the scientific method (as described, for example, in the 1990 book, *Visions of a flying machine* by P.L. Jakab, Smithsonian Institution Press, Washington DC, USA). Engineers of the day had made little attempt to create flying machines, and there was little supporting scientific work.

The Wright brothers assault on this problem so closely fits the scientific method at many steps of the process that it is one of the clearest examples we can offer:

<p>GOAL</p>	<p>To make a powered aircraft that will carry a passenger. The 1990 Jakab book makes the point that the Wright brothers never deviated from this goal. They made many critical advancements to the science of flight, but they were never sidetracked into purely academic pursuits as they advanced the field. Each time they accomplished what they needed, they moved on toward the goal of a flying machine. It is sobering that it took them less than 5 years to achieve this goal from the time they started – about the length of time you will be an undergrad.</p>
<p>MODELS</p>	<p>They used lots of models. At first they worked with kites shaped like miniature airplanes. They then spent 3-4 years working with gliders – aircraft that could carry a person but had no propulsion system. These kites and gliders were tested at Kitty Hawk, on the North Carolina coast (lots of wind; long, flat stretches of sand, suited for a glider). At home, they worked with other types of models: small wind tunnels, devices to measure the lift and drag of wing shapes in those wind tunnels, mathematical models, and even a cardboard box – which gave Orville the idea of using wing-warping to control the direction of the airplane. Strange as it may seem, their understanding of bicycles was important to some of their steps, which means that the bicycle was used as a model of an airplane.</p>
<p>DATA, EVALUATION &amp; REVISION</p>	<p>The Wright brothers’ method involved continual testing, evaluation, and revision – we would call it trial and error. With each implementation of a new model into their glider, they would test it (= data), decide whether it worked well enough to move on (= evaluation), and modify it if needed (revision). Because the testing was done at Kitty Hawk, which they visited only once a year, they could not make improvements as fast as was desirable, although much of this trial and error was performed at Kitty Hawk as they were testing the glider (and powered craft in 1903). The modifications that they added over the years included changes in wing shape (to achieve a better lift/drag ratio), adding rudders to the rear of the glider, increasing wing length, and modifying the axels for the propellers.</p>

# A Continual Process of Improvement

Science is a process, and our ideas keep changing. These changes may be merely refinements of earlier ideas, or they may be complete overhauls in our understanding. Probably the most important single feature to remember about the scientific method is that it is a means by which we can achieve progress. The scientific method is used when we are trying to improve something, whether it be to cure cancer, design a new vaccine, increase profits or build a better airplane. Each success breeds new expectations, so that there is rarely any point at which we stop the process. Improvement and progress is measured by the turnover of models: better models allow us to better achieve our goals. The following figure helps to illustrate the dynamics that underlie this progress.

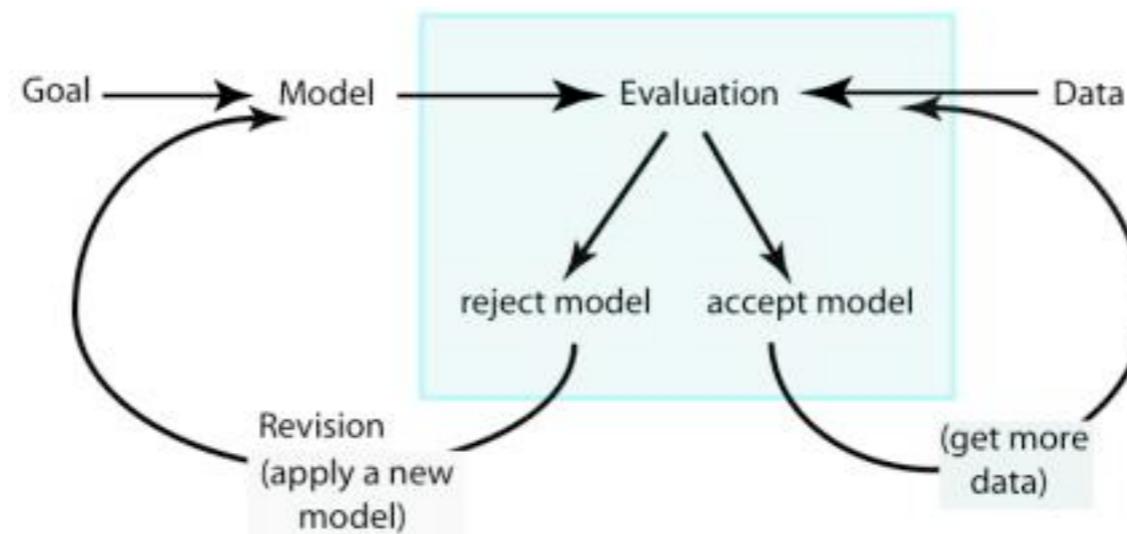


Figure. Pictorial view of the scientific method, showing the dynamics involving the different elements. Starting point in any inquiry is a goal. Then one develops a model of the process that will be studied or the phenomenon that will be manipulated. From there data are gathered (or one uses data that have already been gathered). The data and model are compared in a process of evaluation, which is simply a process of deciding if the model can make sense of the data. If the model does not perform well, it is revised -- either discarded completely or modified, and the process is repeated.

Except for the goal, each element in the scientific template is subject to change, so it is best to think of the scientific method as a cyclic process, repeated over and over. (For any given goal, the other 4 elements will be changing as progress is made toward that goal.) Thus, we start with one or more models of how we think nature works. These models are compared to data (the evaluation stage), and if the model is obviously at odds with the data, it is modified or replaced by a completely new one (revision). Whether the old model is retained or rejected, the process is continued with further refinements of data and evaluation.

Although we have formally dissected only two examples according to our template, there are many examples of progress achieved this century using the scientific method. Some of these are feats of engineering, as in larger buildings, bridges, and airplanes, or better electronic appliances such as stereos, televisions, and microwave ovens. Michael Shermer (1997, *Why people believe wierd things*, W.H. Freeman and Co., NY) offers an example of advances in the speed of man-made vehicles:

VEHICLE	MPH
1784 Stagecoach	10
1825 Steam Locomotive	13
1870 Bicycle	17
1880 Steam Train	100
1906 Steam Auto	127
1919 Early Aircraft	164
1938 Airplane	400
1945 Combat Plane	606
1947 X-1 Jet (Chuck Yeager)	750 (mach 1)
1960 Rocket	4,000
1985 Space Shuttle	18,000
2000 TAU Deep-space Probe	225,000

In biology, perhaps the greatest progress has been in genetics:

1900	Rediscovery of Mendel's Laws
1916	first proof of the chromosome theory of heredity
1944	demonstration that DNA was the material basis of heredity
1953	structure of DNA solved
1977	first entire sequence of a DNA genome (a bacteriophage)
1980s	genes identified for several inherited diseases
1990s	first gene therapy trials to correct genetic defects in humans
2001 - 3	completion of the human genome sequence

These events are only some of the more important advances; the science of genetics is filled with countless improvements of a lesser magnitude as well as many ideas overturned.

Progress (or at least change) is also evident in our understanding of the relationship between diet and health:

- 1916 The first USDA food guide was published. Other guides were published in subsequent years.
- 1956: USDA issued a diet recommendation consisting of four food groups that most of today's adults remember:
  1. meats, poultry, fish, dry beans and peas, eggs, and nuts;
  2. dairy products, such as milk, cheese, and yogurt;
  3. grains, and;
  4. fruits and vegetables

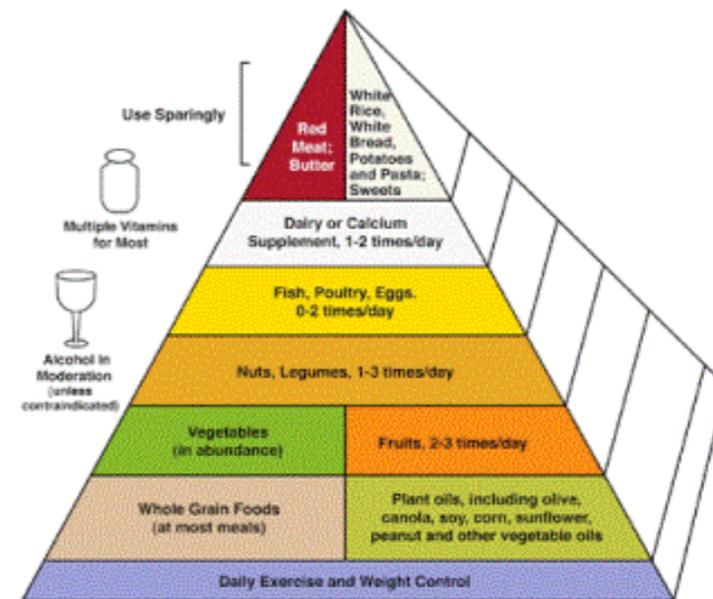
Recommendations for a balanced diet included foods from each of the four groups daily.

1992: In the decades following the release of the 1956 recommendations, it became increasingly clear that heart disease and some cancers were caused by certain types of fats found in dairy products and meats. In 1991, the USDA was about to release a new recommendation of four food groups that virtually omitted animal products and emphasized grains, vegetables, and fruits. After considerable politics in that year, the USDA issued a food “pyramid.” In contrast to the 1956 recommendation, this pyramid increased the emphasis on certain foods (grains, vegetables) and decreased the emphasis on others (fats, meats).



2003: In response to their disagreements with the USDA food pyramid, some members of the Harvard School of Public Health issued their own food pyramid, based on numerous studies (and perhaps less influenced by politics). Some of their main objections to the USDA food pyramid were (i) that many oils were known to benefit health, and (ii) simple carbohydrates (white flour, pasta) were not especially healthy and may have been contributing to an epidemic of obesity in the US. Their new food pyramid retained some elements of the 1992 food pyramid, but recommended reduced simple carbohydrates and increased vegetable oils.

## Healthy Eating Pyramid



This will not be the final advice on diets that you receive in your lives. Indeed, it is relatively common now to read commentary about the conflicting dietary advice we get. The latest craze is the Atkins diet, which in its extreme reduces carbohydrate intake to the point that the body goes into a metabolic state of ketosis, and obtains glucose from protein. No doubt you will see studies in the future that identify health complications of the Atkins diet. We also get lots of advice about supposed “magic bullet” foods – items that are minor components in any diet that may perform a special function, such as reduce heart disease or cancer and about commonly-eaten foods that may be exceptionally harmful (in the past, salt and eggs were given this distinction). One can merely hope that what we have learned to the present is an improvement over the past.

One of the difficulties in dealing with science and health is that diets are so complex and varied, that it is difficult to identify specific components of diets that are good or bad for you. Furthermore, a person’s genetic makeup and exercise habits also influence health, and those factors are not easily separated from diet. (This is an equally important problem in medicine – it is estimated that 100,000 Americans die each year because of complications with drugs they take. The problem is that not everyone responds to a medicine the same way.) So with diet, we have witnessed “progress” in the sense that new recommendations have replaced old ones, but we don’t yet know how much of an improvement is to be gained by adopting the new guidelines.

It might seem that improvement stops when the scientific method has achieved perfection. That is, we should be done once we have proved a model to be true, right? No. Science does not prove models to be true and does not achieve perfection. For example, we will never know all possible health risks to the fetus of maternal alcohol consumption or know all the environmental effects of a nuclear power plant. And computers continue to improve, as do airplanes.

The point of this book is to relate the scientific method to examples in everyday life - problems not traditionally regarded as science, and problems that will affect you regardless of your chosen career. The next chapter initiates that objec-

# Summary

You should conclude this chapter on the scientific method with 3 simple points:

- the scientific method has five elements (goals, models, data, evaluation, revision),
- the scientific method is cyclic,
- the scientific method is a means of achieving progress toward a chosen goal.

## CHAPTER 3: HOW NON-SCIENTISTS USE THE SCIENTIFIC METHOD

# 3

THE SCIENTIFIC METHOD

The scientific method is used unconsciously by many people on a daily basis, for tasks such as cooking and budgeting. The same elements present in traditional scientific inquiry are present in these everyday examples. Understanding how to apply the scientific method to these seemingly non-scientific problems can be valuable in furthering one's career and in making health-related decisions.

# Introduction

**This chapter captures the essence of this course:**

Its goal is to explain the workings of the scientific method in a familiar context. The last chapter introduced a formal framework using typical science examples. Yet the scientific method is not just for scientists, but is for lawyers, business executives, advertising and marketing analysts, and many others. We will discuss several examples and explain how each is composed of the 5 scientific method elements.



# Trial and error

In the simplest terms, common uses of the scientific method involve trial and error. Consider automobile repair. Every weekend handyman, and every high school student with a passing interest in autos knows about the method of trial and error. Your car is starting to run poorly, and you take matters into your own hands in an attempt to fix it. The first step is to guess the nature of the problem (your model). Acting on your hunch, you proceed to exchange a part, adjust a setting, or replace a fluid, and then see if the car runs better. If your initial guess is incorrect and the car is not improved, you revise your guess, make another adjustment, and once again test the car. With patience and enough guesses, this process will often result in a operable car. However, depending on one's expertise, quite a few trials and errors may be required before achieving anything remotely resembling success.

The methods scientists use to evaluate and improve models are very similar to the method of trial and error, and are the subject of this chapter. You may be reluctant to think that the bungling process of trial and error is tantamount to the scientific method, if only because science is so often shrouded in sophistication and jargon. Yet there is no fundamental difference. It might seem that scientists start with a more detailed understanding of their problem than the weekend car mechanic, but in fact most scientific inquiries have humble and ignorant beginnings. Progress can occur just as assuredly via trial and error as in traditional science, and the scientist isn't guaranteed of success any more than is the handyman: witness the failure to develop a vaccine for AIDS. One of the themes of this book/course is that the scientific method is fundamentally the same as these simple exercises that most people perform many times in their lives.

# Cooking From A Recipe

Another activity familiar to all of us is cooking. Although the microwave oven has reduced our dependency on preparing food for ourselves, many of us still face the need to perform rudimentary culinary skills. The preparation of most dishes begins with a recipe - a list of ingredients and instructions for mixing and cooking them. However, rare is the chef, whether budding or accomplished, that follows the recipe to the letter and does not taste and modify the dish during the cooking process. Modifications are attempted until the preparation meets the cook's approval, whence the food is served. Any significant alterations to the recipe may be adopted as permanent modifications, to become part of the recipe itself in the future.

Although it is likely that all of us can identify with this example, it may be less obvious how this example bears on our scientific method template. Returning to our template of 5 elements, we may dissect this example as follows:

SCIENTIFIC METHOD TEMPLATE	
GOAL	To prepare a food dish
MODEL	The Recipe
DATA	Tastings during preparation or when served
EVALUATION	Decisions on how it tastes
REVISION	Changes to the recipe

Let's consider each of these elements again. In the cooking example, the goal is to prepare a specific kind or quality of food dish. The model is simply the recipe you use. It is a model because it is an abstraction of the actual process used in preparing the food; it is essential, because you could not plan to prepare a specific kind of food dish without some guidance based on previous preparations. Here, the data are simply your tastings of the dish before or after it's finished. Evaluation is performed when you compare the actual taste (the data) to your idea of how the food should taste. If it tastes better (or worse) than you expect, you then try to figure out how to revise the recipe accordingly. These revisions may be short-term (how you modify the recipe on this particular occasion) or permanent changes to the written recipe.

The recipe example was chosen because it is commonplace. Yet it is extremely apt. The procedures that scientists use may be slightly more stereotyped and formal than those of the ubiquitous household chef, but the way you work with a recipe, garment pattern, and any of a number of other daily experiences are not fundamentally different than the way a career scientist operates. Lab chemistry and molecular biology is filled with just as many miserable failures as are our nations kitchens, and in both cases the mistakes are used to foster improvements for the future.

# Writing a News Story

A newspaper article about a murder starts as scribbled notes in the reporters notebook (first version of the model), then progresses to a rough draft (second version of the model), which is read by the editor and rewritten by the reporter to become the published article (third version of the model).

Using our template:

SCIENTIFIC METHOD TEMPLATE	
GOAL	Write an attention-getting article
MODEL	Current draft
DATA	Reactions of you and others to the draft
EVALUATION	Are the reactions achieved by your draft those you want to achieve?
REVISION	New drafts

Progress occurs as new drafts are written, in response to the reactions of the author and others (the data), and according to the author's intended responses (evaluation).

# Designing Advertisements

Advertising agencies use the scientific method explicitly to improve the effectiveness of the ads they compose. Ads are models that manipulate consumer behavior, and they are designed with a great deal of scientific input. Each ad has many dimensions that need be considered in detail, such as what headline to use, what size type to use, whether to use pictures, and how large the ad should be. All these questions can be answered using the principles of model evaluation and improvement.

The most useful evaluation of ads comes from mail order returns. To determine whether an ad with a picture sells more gizmos than one of the same size with only text, one simply has to gather some data: place one ad in half the copies of the February issue of a magazine, and the alternative ad in the remaining copies. Put different 800 phone numbers or P.O. Box numbers in the two ads, so you will know which ad generates more responses. The evaluation in this example comes when you compare the responses generated by the two ads, and the progress (model improvement) comes when future ads are changed to reflect the ad that generated the most responses. Again, in template form:

SCIENTIFIC METHOD TEMPLATE	
GOALS	Improve sales
MODEL	Current and modified ads
DATA	Responses to each ad in trials
EVALUATION	Deciding which ad most closely achieves your goal in numbers of responses
REVISION	Adopting an ad for general distribution

# Corporate Finances

Tangible examples of the scientific method also abound in business. Consider a corporation's financial planning. The most basic goal of the corporation is to survive economically. This goal requires a complicated, formal business plan, to control and monitor the company's finances. Data accumulate during the year in the form of actual revenues and expenditures, and these data are compared to the model (the model is evaluated) to determine whether further changes (revisions) are warranted:

<b>SCIENTIFIC METHOD TEMPLATE</b>	
GOAL	Increase profits
MODEL	A plan showing anticipated revenues and expenses
DATA	Actual revenues and expenses
EVALUATION	Comparison of plan to data
REVISION	Modifications of the plan in response to the evaluation

# Demonstrations

## **In-class examples: (1) Lamp switch; (2) Wheel of Fortune**

The scientific method template can be applied to any trial-and-error problem. The demonstrations used in class are but two of countless examples that can be offered. (You must attend this lecture to obtain the information.)

# Is THIS Science?

## **A shortcut to decide if something is science:**

*Is the use of evidence paramount?*

*Do the models keep improving?*

The template we have given has 5 components. You can get a good sense of whether a system obeys scientific methodology from two criteria. First and foremost is a strict and ruthless adherence to the evidence. If evidence is not used, is used selectively and sparsely or is downplayed, you can be sure that it's not strict science. Second is turnover of the accepted models – nearly everything in science undergoes change because new evidence and new models are continually introduced. You can think of this criterion as one of ongoing refinements. The changes may be few and slow of course. In many of our non-traditional examples above, the goal has a defined endpoint, so the turnover ends when the goal is met. (For example, writing a news story ends when the story is published.) So the 'continual turnover' criterion applies to problems large in scope but not necessarily small ones.

If an example fails on evidence or turnover, it's not good science. However, an example that passes this preliminary test may still fail on other criteria. Evidence and turnover merely provide a convenient first pass.

---

# Use of the 5 Elements by Various Institutions

It may be useful to understand how science works by considering institutions where it is used properly versus used improperly or not at all.

## **Criminal Trials:**

These come close to fulfilling all 5 elements. The jury has the goal of discovering whether the defendant is guilty or not guilty. This is the goal of deciding between the model advocated by the defense, and the model advocated by the prosecution. Data are presented by the defense and prosecution during the trial, and the jury evaluates the two models based on that evidence. The verdict (guilty or not guilty) is the jury's evaluation of which model best fits the data, with the proviso that in order to return a guilty verdict, the jury must find that the data presented support this model "beyond a reasonable doubt."

Criminal trials are weak on the strict use of evidence and on revision. Although trials routinely present evidence, critical evidence is sometimes excluded from the jury by the court; proper science would let the jury decide its relevance. And the jury is free to ignore evidence in reaching its decision, which apparently happens commonly (lawyers often appeal to a jury's emotions, an indication that evidence is not paramount). Revision is also weak, though not absent. The appeals process provides for limited revision. However, the types of model revision permitted on appeal are somewhat restricted. For example, after a defendant has been found guilty, it is very difficult to obtain a new trial and introduce into court factual evidence that exonerates him/her. Conversely, the prohibition against double jeopardy prevents the prosecution from reopening a case after a "not guilty" verdict has been returned, even in light of new data suggesting that the defendant was actually guilty. A strict scientific use of revision would mean that the verdict could be revisited (and potentially reversed) at any time based on new data or new analysis.

## **Astrology:**

Astrologers (psychics) claim to have ways of forecasting the future, if only in vague terms. There are books that specify how predictions are to be made (the models). A rigorous adherence to the scientific method would involve comparing predictions with outcomes, evaluating whether the predictions did better than random, and developing new predictors based on successes and failures of older predictors. Needless to say, those types of tests are not part of astrology, and the very suggestion of asking how often astrology predictions are held up is anathema to many astrologers (see the video in which Richard Dawkins attempts to put an astrologer to the test). So the example of astrology contains goals and models, but the other elements are absent.

## **Government Agencies:**

Nearly all government agencies are established with some specific (often lofty) goal. They are also provided with a set of rules (a model) of how that goal should be pursued. But there is rarely a formal procedure for evaluating whether the goal is achieved, and there is almost never a procedure for implementing a new model when the old one is deemed inadequate. Elected officials can and do sometimes bring about change, and the political climate now is more demanding of government accountability than in the past, but agencies generally are not established with the kind of built-in self-improvement system that underlies the scientific method. The federal and state constitutions DO specify how to implement a new model - via constitutional amendments.

Yet some agencies that are charged with making decisions do adhere to the scientific method rigorously. The FDA (Food and Drug Admin) is a good example. That agency is charged with approving new drugs and monitoring for problems with prior drugs. To obtain approval for a drug, a company must submit results of extensive (and often expensive) trials that are well documented and meet all the criteria of good science. The FDA is actually rigorous in its evaluation, although the submitting company does not have to provide all relevant data, and the FDA may thus receive biased information. The fact that the FDA continues to monitor for complications of approved drugs (some

## **Religion:**

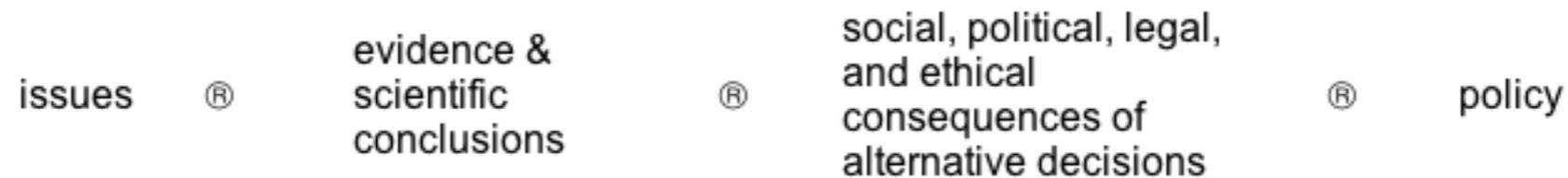
Religion is not science, nor does it pretend to be. Most religions are based on specific doctrines and codes of conduct that followers agree to accept. There is no attempt to "improve" religion by changing the mores every few years and assessing the impact.

An example in which the scientific method cannot be used:

Consider the difference between a gambler playing a card game versus a slot machine. Use of the slot machine is strictly random, a fully automated process of pulling a lever or pushing a button. It does not allow revision in how the game is played. That is, there is no alternative strategy possible in how the game is played (except in not playing the game). In contrast, the method one uses to play cards does admit the scientific method because there is a lot of strategy that can be adopted and altered by the player.

# Science is not the end-all, be-all in making decisions

This class will focus on how to apply and interpret the scientific method as a way of making rational decisions. It should not be construed that a strict adherence to scientific principles should be the sole criterion in reaching a decision. There are many factors that are relevant to our well being. An idealized view of the role of the scientific method in decisions of a societal level is:



Thus, science is (and should be) used to inform decisions, but there is no intent that it be the sole criterion. For example, ethical considerations may override the science, as has been the case with stem cell research in the U.S.

History abounds with examples in which science was ignored in reaching policy, and the policy was not made in the best interest of the people. A spectacular one was the Soviet suppression of genetics in the 1940s into the 1960s, leading to major agricultural failures. Genetics was at odds with the communist ideology that everyone was equal (recall the book *Animal Farm*), and T. D. Lysenko was given the authority to suppress Soviet research on genetics, which included imprisonment and eventual death of several prominent geneticists. (There is a UT connection here, in that Hermann Muller, who first showed that radiation caused heritable, genetic damage while he was at UT, moved from UT to the Soviet Union to show his support for communism. The reality of the Soviet regime led him to escape and ultimately return to the U.S., where he resided when he won a Nobel Prize for his earlier UT work.)

In general, scientific considerations may be overruled (or even ignored) due to a variety of factors:

- *political ideology*
- *financial interests*
- *religion*
- *legal precedents*
- *various alternatives: superstition, instinct, hunches*

Even when science is considered in making decisions, these factors can have a larger influence than they should.

The fact that science should not be used as the sole criterion in setting a policy or reaching a decision is fundamentally NOT the same as relying on poor science in reaching a decision. Poor science can give you the wrong answer (e.g., a dangerous drug appears to be safe). We need quality science to decide how the science should be used.

# External Links

[The Scientific Method](#)

[The Scientific Method Made Easy](#)

[Climbing as an metaphor of Scientific Method](#)

[Richard Feynman on Scientific Method](#)

[How to Conduct an Experiment Using the Scientific Method](#)

[Carol Sagan on the Scientific Method](#)

[Scientific Method Monty Python](#)

## CHAPTER 4: MODELS ARE THE BUILDING BLOCKS OF SCIENCE

# 4

MODELS

You know what a model airplane is. But models are ubiquitous. Advertisers manipulate you with models, and models determine your success in business or school. Because the scientific method is a way to think about models, if you are to understand the scientific method, you must be able to recognize models when you see them and appreciate their limitations.

# Models as Building Blocks and Substitutes

The model is the most basic element of the scientific method. Everything done in science is done with models. A model is any simplification, substitute or stand-in for what you are actually studying or trying to predict. Models are used because they are convenient substitutes, the way that a recipe is a convenient aid in cooking. This section of the book is dedicated to explaining what models are and how they are used.

Models are very common. The ingredients list on a bottle of ketchup is a model of its contents, and margarine is a model of butter. A box score from a baseball game is a model of the actual event. A trial over an automobile accident is a model of the actual accident. A history exam is a model designed to test your knowledge of history.

A model is a substitute, but it is also similar to what it represents. Thus the ingredients list is a fairly accurate guide to the contents of the ketchup bottle. Margarine looks and spreads like butter, and can substitute for it in many recipes. The box score contains most of the critical information about the baseball game---such as the winner, the final score, and the pitchers. Similarly, trials and history exams contain the essence of the events they model. In fact, models are more than just common, they are ubiquitous. Nearly everything we encounter is a model. To drive home this point, we list in Table 4.1 several objects or ideas that are models.

MODEL	WHAT THE MODEL REPRESENTS
CAKE RECIPE	Process of making a cake
WEDDING PICTURES	The wedding
CHAPTER TITLE	Chapter contents
NEWS ARTICLE ABOUT CHICAGO CUBS' LATEST LOSS	The game itself
HOME VIDEO OF POLICE ARRESTING A MOTORIST	Police conduct in general
ROAD MAP OF MADISON, WISCONSIN	Paths of transit in Madison
HOUSEHOLD BUDGET	Household expenses and income
POLITICAL CANDIDATE'S CAMPAIGN PROMISES	Candidate's performance if elected
A STATISTICAL AVERAGE	Something close to what can be expected

# Models Inside Science

Scientific models are fundamentally the same as models outside of science, which will be introduced below. Many people think mistakenly that scientific models are always complicated, impenetrable mathematical equations. But in truth, many scientific models are just as understandable as are models found outside of science.

The USDA food pyramid, which recommends the proportions of different kinds of foods in a healthy diet, is a model of the thousands of scientific studies that have been undertaken on the relation among cancer, heart disease and diet. The figure summarizes these studies in a picture that recommends healthy diets. Thus, this figure is a substitute for the many scientific studies on diet, and it is also a substitute for an actual diet.

As a second example, when scientists use rats to determine whether a food additive causes cancer, the rats become a model of humans. Rats are convenient because they are relatively easy to raise in the lab (at least compared to humans), and one can perform experiments on them relatively quickly (in a matter of months rather than years). Moreover, most people find it more ethical to experiment on rats rather than humans.

# Hypotheses and Theories as Models/Big Models and Small Models

We've all heard about hypotheses and theories, especially in physics and chemistry. Theories usually comprise some idea that scientists have about how nature works, but that they aren't totally sure. Hypotheses and theories are merely particular kinds of models that we will refer to below as abstract models.

Even the most rudimentary science course contains some of the grand, all-encompassing, models that scientists have discovered. The periodic table of the elements is a model chemists use for predicting properties of the elements. Physicists use Newton's law to predict how objects will interact, such as planets and spaceships. In geology, the continental drift model predicts the past positions of continents. But these three models are atypical because they are immensely successful. Most models used are nowhere near so powerful or widely useful. But scientists use these less-successful ones anyway. Models are used at every turn in a scientific study. Samples are models. Ideas are models. Methods are models. Every attempt at a scientific study involves countless models, many of them small and of interest only to a small group of other scientists. The primary activity of the hundreds of thousands of U.S. scientists is to produce new models, resulting in tens of thousands of scientific papers published per year.

# Models Outside Science

Trying to enumerate all the models found in business, industry, and society is simply impossible. Models pervade all white collar jobs. Table 4.2 shows models from fields as diverse as advertising, architecture, finance and manufacturing. In this table we have chosen to give a single model from each of a number of fields. However, we could have just as easily picked one job, say retail sales, and listed 150 models associated with it.

<b>MODELS IN BUSINESS AND GOVERNMENT (TABLE 4.2)</b>	
<b>FIELD</b>	<b>COMMON TYPE OF MODEL</b>
Advertising	Response to an advertisement tested in a single city is a model of the national response to the ad.
Architecture	The plans for a new building are a model of the actual building.
Business	Past dealings with a client are a model of the trustworthiness and promptness you can expect from her/him in the next deal.
Education	A student's performance on a history exam is a model of everything learned about history since the last exam.
Finance	The rating <i>Morningstar</i> gives a bond fund is a model of the fund's future performance.
Federal Government	The federal budget is based on an economic model that predicts next year's revenues and expenditures.
Franchising	A company uses its existing stores to model the likely success of stores it is considering building.
Law	A criminal trial provides a model of the actual crime.
Manufacturing	Profit projections are based on a model of material and labor costs as well as sales price.
Medicine	Your doctor's diagnosis of the cause of your back pain is a model of its actual cause.
Prisons	A model, based on age, crime, and family status, is used to predict which prisoners are good candidates for parole.
Retail Sales	The December sales in 1995-2003 model the December sales expected in the coming year.

The ability to recognize, construct, and improve models gives you an advantage in many walks of life. A salesperson who recognizes that a sales pitch is a model can take steps to improve it. Other models are obvious but are so complicated that years of effort go into learning how to build them, as with the house, computer, and automobile models that are the trade of architects and engineers. Sometimes, the critical skill is not finding or building a model, but knowing how to improve an existing model, as with a budget or airline design.

Models are important outside of science because success in any professional endeavor involves accurately predicting or manipulating the future, and we need models to do this. Correctly predicting the stock market would net a person fame and fortune. The path to success in sales is only slightly less direct. If a salesperson can accurately predict how a particular client will respond to a particular pitch, the pitch can be modified to have maximum effectiveness, thereby increasing the probability of a sale or abandoning a non-buyer before wasting much time. Similarly, budgets predict the financial consequences of taking various actions, allowing the company to cut losses and increase profits.

The arts---whether an action movie like *Lethal Weapon III*, an abstract painting by Picasso, a historical novel by Michener, or Whitman's poetry---consist of models designed to evoke emotions and present unusual events or viewpoints. Because a scene from a Hollywood movie appears to be a plausible representation of the real world, it can make you frightened (a stunt man hanging out the window), or sad (a dead heroine), or anxious (an oncoming train). The protagonist of a historical novel substitutes for someone that actu-

# Models exhibit a one-to-many and many-to-one relationship

There is no such thing as just one model of something, nor is anything we use as a model necessarily useful as just one kind of model. A wedding will have many different models to remind us of the day: pictures, memories, wedding presents, and newspaper accounts -- all models of one event. At the same time, one of those wedding presents (e.g., a toaster) will be a model of the wedding but is also a model of other toasters, of the company that made the toaster (and its other products), and it may eventually become a model of electronic appliances when one of the kids (or parents) takes it apart to fix it or see how it works.

It is neither profound nor particularly useful to learn that everything is a model. If this was all we could say about models, there would be no call to focus heavily on them. The models we have discussed thus far were chosen to show that you are already familiar with models. In the remainder of this chapter, we describe models that are more subtle, and we explain how an understanding of models may be important to people outside of science.

# Classes of Models

Different kinds of models are used for different purposes. Table 4.3 lists three major types that will be used in this class: abstract, physical, and sampling models. Not all models fit neatly into these categories. Moreover, we won't bother to classify many of the models in this course. However, these three classes do accommodate many of models that we will focus on and discuss, so it is convenient to group them in this fashion.

<b>CLASSES OF MODELS (TABLE 4.3)</b>		
<b>CLASS</b>	<b>FAMILIAR TYPES</b>	<b>EXAMPLES AND COMMENTS</b>
Abstract	predictions, theories, hypotheses, many mathematical and computer models	Newton's laws in physics, plans, recipes, statements such as "taking anabolic steroids increases one's strength," or "smoking causes lung cancer."
Physical	organisms and their properties, replicas, structures, demonstrations	a globe is a physical model of the earth, each of us is a model for other humans, and the physical structures used in chemistry class are models of molecules
Sampling	random choice, personal preference	the sampling model refers to the way that subjects are chosen for a study and divided up among the different groups; sampling models are the subject of our section on Data.

# Summary

You should end this chapter with an understanding that models are a crucial element of the scientific method. A model is in some way a substitute for what is being studied. They are widely used, and there are many types of them. At this point in the book, you should be able to begin using the information being taught. For example, when reading news articles on topics relevant to scientific study, you should be able to identify models used in those studies and should be able to identify those belonging to the classes in Table 4.3.

# External Links

[Human Bohr Model](#)

[Bill Nye with a model of an Atom](#)

[Bohr Model Explained](#)

[Quantum Mechanics](#)

[Origami DNA Model](#)

[DNA Replication Model](#)

[DNA lecture \(28 minutes\)](#)

The Secret of Life -- Discovery of DNA Structure:

<http://www.youtube.com/watch?v=sfoYXnAFBs8>

Bacteria lecture:

<http://youtu.be/TDoGrbpJJ14>

Natural Selection model:

<http://youtu.be/GcJgWov7mTM>

The Universe Modeled:

<http://youtu.be/mwyTGcHP7kc>

# CHAPTER 5: ALL MODELS ARE FALSE. BUT SOME ARE USEFUL ANYWAYS.

MODELS

# 5

A newspaper account of a murder omits many of the grisly details, but it is still informative.

# Introduction

A useful fact about models is that they are all wrong in a strict sense, if only because they are incomplete. If a model is examined closely enough, it will invariably be found to differ from what it represents. Thus, a newspaper article about a murder tells you the victim's name, age and sex. It fortunately omits the exact locations of the knife wounds and the total volume of blood spilt on the floor, and omits details deliberately kept secret to aid in identifying the murderer. The list of ketchup ingredients tells you that the ketchup contains more tomatoes than sugar, but it conveniently overlooks certain details, such as the miscellaneous insect parts and feces in the bottle.

Never be surprised to learn that a model of interest to you is incomplete, hence is "false." Furthermore, it often takes very little effort to determine how a particular model is false. The reasoning we applied to a newspaper article and the ketchup ingredients could easily be repeated for the models given in the last chapter. Even with no previous exposure to the scientific method, it is usually easy to identify several important ways that any given model differs from what it represents.

The fact that a model is wrong does not mean it is useless. We continually use false models, such as the newspaper article and the ketchup ingredients discussed above. Because no model is one hundred percent correct, refusing to use a model merely because it is false is tantamount to refusing to use any models at all. The objective in using the scientific method is to distinguish useful models from useless ones.

# Models are useful because they simplify- they are false for the same reason

The main reason that all models are incomplete/false is that they are simplifications -- shortcuts. The ways in which they are simplifications may not be essential for certain purposes (the simplifications may in fact make the model useful). The budget for a corporation, for example, only approximates the actual expenditures, income and profits. Clearly, no matter how many accountants a corporation hires, and no matter how carefully these accountants work, it is impossible to prepare a budget that is completely correct. To predict income exactly, one would have to know exactly how many gizmos the company will make and sell in the next year, and what price each one would fetch. This prediction depends on details of the economy, on how each prospective customer will behave in the coming year, and so forth. Because these facts cannot be known when the budget is prepared, it is inaccurate, or false. Their faults notwithstanding, budgets are also universally used. A budget allows a corporation to make decisions and to plan, and thus to achieve higher profits than would be possible without the budget.

Although every corporation acknowledges that a budget plan will not predict exactly the future financial exchanges (hence the budget is a false model), there are limits to what kinds of false statements a corporation will tolerate in its budget. The omission of some events, such as a minor unexpected price increase of raw material may be of no consequence. But other false aspects create more serious problems, as with a serious underestimate of production costs, or a failure to pay employees enough to avoid a strike.

As a second example, consider a court trial over an automobile accident in which one car rear-ended another. The trial is a model of the accident itself. It is incomplete because eyewitnesses forget and make mistakes, because medical diagnoses of whiplash can be in error, and because a picture of an intersection will invariably differ from the intersection itself. These problems notwithstanding, the judge and jury, after listening to the evidence presented in a trial, will usually have a pretty good idea of what happened. Hence the trial is a useful model.

## "False" models as an integral part of science

The models that scientists use are no different from the models you use in everyday life. They are simultaneously false and useful. Learning even a small amount about scientific models can be quite useful in detecting major limitations of scientific approaches. This knowledge enables one to pose relevant questions to those who developed the model.

The Harvard food pyramid mentioned previously is useful as a model of a health-effective diet even though it condenses thousands of scientific studies about diet and health into a single picture (it is a summary model). This reduction must have resulted in the loss of substantial information, considering all the words, data, nuances and caveats in the original papers. Despite this, the food pyramid is an effective tool for communicating the results of a wide range of scientific studies to large numbers of people with varying backgrounds and levels of scientific sophistication. In fact, it is much more effective at this task than are the original scientific papers.

Biologists use animal models in developing new medical treatments (a type of physical model). Pharmaceutical manufacturers test the safety of new drugs using rats and mice before giving the drugs to humans, and heart surgeons develop new surgical techniques on dogs before trying them on humans. These models, however useful in preventing humans from taking unsafe drugs or having untried surgery performed on them, are nonetheless imperfect. Animals do not respond to drugs in exactly the same way humans do (think of how cats and humans respond to catnip). And while it might be useful for a heart surgeon to practice on a dog, performing surgery on a dog is clearly not the same as performing surgery on a human.

---

# Pieces and Parts as Models

It seems fairly straightforward to consider two Suburbans off the same assembly line as models of each other, or the renovated Ft. Davis Historic Site as a model of the cavalry fort of the 1800s. Many of us also accept without question that a picture of Abe Lincoln is a model of him. All of these models are obviously false: the picture of Abe is not the man (it is a cluster of silver oxide grains on paper); the modern Ft. Davis is not populated with cavalry soldiers, nor is it concerned with Indian raids. One Suburban is probably a good model of the other for many purposes, but there are countless differences between the two when it comes to how tightly bolts are fastened, which parts will fail first, inherent weaknesses in the materials, and in exact gas mileage.

Consider now, an airliner crash. TWA flight 800 exploded in mid-air just off Long Island only a few days before the Atlanta Olympics. Eyewitnesses reported seeing trails of orange behind the plane, suggestive of a missile. Early speculations focused on a bomb, both because airliners don't just blow up in mid-air by themselves, and because the Atlanta Olympics provided the kind of public focus that terrorists often target. Our airports switched to tightened security measures, and President Clinton and Congress responded by passing expensive legislation to increase airport security. In the long run, we don't know what caused the crash, but odds seem to favor an explosion caused by equipment malfunction rather than a bomb.

What, then, are models of the cause of this crash?

1. We would certainly want to include reconstructions of this crash -- the assembled wreckage and any computer simulations of how the same kind of plane explodes from a bomb.
2. We should also include as a model the deliberate detonation of another plane, which could be studied to understand how a plane breaks apart in mid-air (such a deliberate detonation has been contemplated).
3. A single piece of wreckage could give the valuable clue of bomb residue or metal twisted in a particular way, diagnostic of what went wrong.
4. Eyewitness accounts of the crash.
5. Data recorders from the plane.
6. Knowledge of the sources of baggage put on the plane (were bags from other flights transferred to TWA 800?).
7. Even the timing of the accident with the Olympic games are all pieces of information or pieces of physical evidence that could shed light on the cause of this crash.

They are thus all models of the cause, even though some seem to be more “obvious” models than others. It may seem strange to call a single piece of wreckage or information about baggage as a model, in that these particular models are clearly only portions of the entirety of the crash. Yet any model has countless differences from what it represents. A reconstruction of the whole plane from the recovered wreckage may seem more complete than all the pieces, and it may be more complete in many senses. Nonetheless, even the entire assembled wreck is a far cry from the actual accident -- it is not in the air, flying; there are no passengers, nor is there any way to retrieve the lost lives, and so on. So it is misleading to suppose that a single piece of wreckage or bit of information is too insignificant to be a model but that the sum of these insignificant parts is a legitimate model.

The important issue here is that some models are more USEFUL than others. One piece of wreckage may be more useful than a 1000 other pieces in understanding whether a bomb went off. What makes a model useful is explained

# ACU: Why we use particular models (Accuracy, Convenience, Uniformity)

The goal determines model usefulness. Models vary in their usefulness. Some are so different from what they represent that we just refuse to use them. Yet there is no such thing as an intrinsically good or bad model without considering its context. A model is judged against the goal, and a model may be good for some purposes and bad for others. Consequently, the standards we use to decide the utility of false models are extremely diverse. An algebra problem that your math instructor assigns is a model of the problems you will be asked to solve in some careers. Because your future boss will never ask you to work a problem exactly like those at the end of the chapters in your algebra book, each problem you work in class is a false model of this future need. The relevant question is not whether the model is false, but whether the false aspects of the model seriously degrade its usefulness. The answer to this question will vary depending on how you use the model. Thus an algebra course might be usefully false model for accountants, business managers and engineers, but a hopelessly false model for artists.

More generally, the usefulness of a model depends on the problem to which it is applied (the goal of the work). Thus, any model may be useful for some purposes, and it will invariably be useless for other purposes. When considering the value of a model, it is therefore essential to know its application. In many cases, this point is obvious - a nuclear physicist would not be the least interested in using GM's annual budget model to predict the behavior of elementary particles. However, the match between model and goal applies on a much finer scale as well. For example, the file of previous exams owned by a fraternity may be very useful for some classes but not others.

The criteria for model acceptability can be classified in many ways. Here we will recognize 3 criteria: Accuracy, Convenience, and Uniformity, or ACU. Acceptance of a model depends on a combination of all 3 criteria, though there is no universal rule for assessing the relative benefits of one criterion versus the others. And we are not looking for one model that simultaneously best satisfies all three criteria. In fact, we often use several different models of any one thing to overcome the limitations of any single model -- there is NOT a most useful model for any particular goal.

**Accuracy:** This is the most obvious criterion to use in accepting or rejecting a model. After all, if we are trying to represent something, we hope that our model does a good job of actually representing what is intended. Accuracy is the measure of how well the results from the model will enable us to predict the real situation. This criterion is thus easy to grasp, and we will move on to the next criterion.

**Convenience:** This third criterion covers time, cost, ease of application, and ethics. In an ideal world, we might imagine that cost is no problem. Yet, budgets dictate that we make the best use of the money. And time constraints dictate that we get answers soon as opposed to later. We use mice instead of monkeys for initial tests of foods and drugs because mice are more convenient than monkeys (in cost, time to results, and ethics). Virtually all models used to test products on a large scale are chosen with a heavy emphasis on convenience. Some such models seem ridiculous because so much accuracy has been sacrificed in favor of convenience (e.g., condom testing). But at some level, most models sacrifice accuracy to achieve convenience.

**Uniformity:** This criterion is the consistency of the model -- is the model uniform from one use to the next? Uniformity is important mostly with physical models instead of abstract models. (It is very important in sampling models, because sampling models usually involve physical models). Inbred strains of mice offer good models to study cancer-causing agents in humans because they possess uniformity -- thousands of mice of the same genotype can be tested, so that results can be compared for different chemicals. Likewise, methods for industrial testing of products are geared toward uniformity, because the test will be performed many times and the outcomes from different trials need to be comparable across the trials.

These are not the only criteria that are relevant to a model. Another factor is the repeatability of a model (can it be applied multiple times?) Attempts to understand unique events after-the-fact are often based on models lacking repeatability. Wreckage of a plane crash provides various models of the crash -- pieces of the plane, for example -- but these models lack or are weak in repeatability. Eyewitness accounts also lack this property, and as such, are limited in an important respect. However, we will limit ourselves here to the three criteria of Accuracy, Convenience, and Uniformity.

An example: Consider a detailed example of the conflict between accuracy and convenience. If our goal is to understand cancer in humans, we might use genetic studies of humans, monkeys, rodents, yeast, and/or bacteria. How do each of these models rank on the scales of accuracy and convenience?

<b>MODEL</b>	<b>ACCURACY RANK</b>	<b>CONVENIENCE RANK</b>
Humans	1	5
Monkeys	2	4
Rodents	3	3
Yeast	4	2
Bacteria	5	1

There is an inverse relationship between accuracy and convenience for these models. All of these model organisms are useful -- yeast and bacteria might be the most useful for some purposes because they are cheap, easy to manipulate, and do not raise ethical issues in experimentation (strong on convenience). The genetics of yeast is similar enough to that of humans (with respect to the control of cell division) that many breakthroughs in cancer research have come from them. Obviously, model accuracy is greatest with humans, monkeys next, and so on. So we can have very useful (convenient) models that are much less accurate than other models.

In research with plants, animals, and even yeast and bacteria, especially when treating them as models of humans, there is extraordinary emphasis on using strains. Why? The reason for doing an entire study with a single strain of an organism is to achieve uniformity. If we are investigating the effect of a chemical on the mutation rate of yeast, we want to minimize the variation due to causes other than from that chemical, so we use genetically uniform strains.

---

# Model Incompleteness is the Basis for Improvement

Most models deemed useful at some point in history have only a temporary life, i.e., they are replaced by better ones after awhile (returning to the point that science does not prove models to be true). For example, today's technology enables companies to assess financial status faster and more accurately than in the past, so that their budgets incorporate different components now that 20 years ago. In science, most successful models are improved through time as well. The better models simply address and overcome some of the falsity or incompleteness in their predecessors.

One can often anticipate how a model may eventually be improved merely by contemplating how it is incomplete. The goal is not to eliminate all incompleteness in a model, but rather to correct the more serious limitations. It is often possible to anticipate how a model may ultimately be rendered obsolete merely by thinking about its limitations. In teaching you to think about models, we will therefore emphasize their limitations. Table 5.1 illustrates how some common models can be trivially wrong versus seriously wrong.

<b>HOW SOME MODELS CAN BE FALSE (TABLE 5.1)</b>		
<b>MODEL</b>	<b>MINOR FLAW</b>	<b>MAJOR FLAW</b>
Medical diagnosis	No two patients are identical, although most individual details do not affect treatment	Incorrect diagnosis results in patient death or malpractice
Credit rating	Small details of personal finances are omitted	Omission of major debts or credits that have a big impact on personal finances
Test for heroin	The test assays various chemicals other than heroin itself, but these compounds are minor constituents in the body	Eating poppy seeds before the test gives the impression that you have illegally taken drugs
Income tax return	Small transactions are overlooked	Omission of large sources of income which can result in a financial penalty or incarceration
College exam score	A lucky guess results in a few points on a subject that the student was not prepared for	Exam score is totaled incorrectly
New car	Slight differences exist between different cars of the same model	The car you purchase is a lemon
Space shuttle flight	Each flight faces different problems, which are usually fixable	The shuttle explodes

# A Template for Models

To facilitate use of the information in this and the last chapter, we offer a template for models. This template is intended to help you think about the different aspects of models whenever you read (or think) about uses of the scientific method. Any news article describing medical research or a news article on studies into business practices can be analyzed from this perspective. In the next few chapters, we will describe some biological examples primarily from the perspective of models, and this template will be used in summarizing those presentations.

MODEL TEMPLATE
MODEL
KIND (Abstract, Physical, or Sampling)
APPLICATION (used as what?)
STATUS (Accepted, Rejected, or Uncertain)
LIMITATIONS

The kind of model is either abstract, sampling, or physical; if a model does not fall into any of these 3 classes, we will not worry about the class (there are too many ways to classify models for us to bother classifying all of them). The application is the problem for which the model is used. For example, a business (financial) plan is applied to managing company money, a monkey might be used as a model of humans in understanding AIDS, and so forth. The status of a model indicates whether it is currently regarded as useful (accepted), rejected, or is in dispute (undecided). For example, the model that X-rays directly cause bacterial mutation would be accepted for the set of experiments in which irradiation of bacteria leads to mutation but would be rejected for the experiments in which irradiation of Petri dishes alone leads to mutation.

The last item, limitations, is of interest only for models whose status is accepted or undecided. It is important to realize that all models have limitations and that these limitations may ultimately lead to the model's rejection when we have better data. By highlighting limitations of currently-accepted models, we should be constantly alert to possible revisions of the model that may be even more useful.

---

# All Models Must be Refutable

One of the most widely publicized features of the scientific method is that scientific models must be refutable or falsifiable. By this it is meant that observations can be imagined that would cause the model to be rejected. The criterion of falsifiability differs from our claim that all models are false. Typical examples of unfalsifiable models are of the form that various demons or spirits control all events in the world, or that some person has mystical properties. These models are not falsifiable, because it is not possible to even imagine data which would call for their rejection.

How does falsifiability fit into our framework? The goal of any application of the scientific method invariably involves predicting the unknown (explaining future results) or manipulating the future (increasing profits). But because an unfalsifiable model admits all possible outcomes (since nothing is inconsistent with it), an unfalsifiable model cannot be improved upon. Consequently, unfalsifiable models are useless.

Unfalsifiable models are common outside of science. Some market analysts have a flair for "explaining" the ups and downs of the market after-the-fact. Regardless of the market changes, these nightly reports profess to account for all the ups and downs, and there is no pattern that can't be explained. Of course, as long as these reports don't attempt to forecast the directions in the future, they can never be challenged (of course, some analysts do predict future trends). Prophecies also tend to lack falsifiability. These statements are often couched in extremely vague terms, and only in retrospect are people able to "interpret" them to make sense. To be falsifiable, a prophecy needs to be specific enough that we know in advance what to expect (e.g., the world will collide with a large comet on the morning of your 3rd exam).

---

# History as a Model of the Past

A senior (and now deceased) colleague of ours once commented that the subject of history became ever more interesting the closer one got to becoming a part of it. There seems to be some truth in that, at least as judged by the general lack of enthusiasm for the subject in high school and college. Nonetheless, there are several human endeavors that involve attempting to reconstruct the past. Aside from history per se, they include:

1. many sciences (evolution, astronomy, geology, anthropology)
2. criminal trials (reconstructing a crime to convince a jury that the defendant committed it)
3. religion (many teachings refer to events from the past)
4. medicine (who you contacted that gave you this disease, what you ate that led to your heart condition)

Reconstructing the past poses special problems to a scientific approach, because the past is gone and it is unique. This means we cannot study it directly, the same way we would study a blood sample or your computer software problem. However, we nonetheless reconstruct history, and a variety of models is used in doing so.

Consider “Billy the Kid.” Virtually everyone knows this name – a notorious, young outlaw who killed several people in the late 1800s, mostly during the Lincoln County wars in New Mexico. Our model of Billy the Kid is a list of episodes and deeds, derived from legal documents, testimonies, newspaper accounts, diaries and the like from the 1800s: his name (William Bonney), birth (1859, New York City), death (1881, shot by Pat Garrett, buried in Ft. Sumner), along with twenty one murders, escapes from jails, associations with other outlaws, and raids he conducted.

Each of these descriptors is a potentially false model of Billy the Kid. Doubts have been raised that William Bonney was actually killed by Pat Garrett; rather it has been suggested that Pat Garrett shot someone else. And, it is of course possible that some of the murders attributed to Billy the Kid were committed by someone else. It is not uncommon that notoriety begets further notoriety, whether deserved or not. (The infamous “serial killer” Henry Lee Lucas confessed to over 600 murders and had law enforcement agents accepting over 100 of the confessions, but nearly all of those were later shown to be false – Lucas may have killed as few as 3 people – and his death sentence in Texas was commuted to life in prison because he could not possibly have committed the murder for which he was convicted.)

Interest in Billy the Kid is inspired by a current controversy. If you drive the 100+ miles north of Austin to the Hill Country town of Hico, you will find a quaint, aging downtown area of antique shops in sight of a stately but defunct stone cotton gin, harkening back to the days when cotton was a major crop in the area. (There is also a great chocolate shop nearby.) One of the old stores harbors a museum of sorts, dedicated to keeping alive the possibility that the real Billy the Kid died there in 1950 at the age of 90. The person’s name was Brushy Bill Roberts, born in Buffalo Gap, TX in 1859. In his final year of life, he attempted to establish his identity as Billy the Kid, dictating the details of his acts and deeds, even requesting a pardon from the Governor of New Mexico (at which time he had suffered a stroke and could not communicate).

Brushy Bill Roberts and William H. Bonney were clearly different people. Who, then, was the real Billy the Kid? We will likely never have a satisfactory answer, at least not an answer that everyone can agree upon. There are many ways in which our Y2K (year-2000) model of Billy the Kid may be false – we may be associating the wrong name with the deeds, the deeds themselves may be different from legend, and the 21 murders may have been committed by more than one person. The legend of Billy the Kid is a model that may have little bearing on reality, or it may be accurate in some respects but not others.

The New Mexico town of Ft. Sumner, located at a crossroads along the Pecos River in eastern New Mexico, claims to be the final resting spot of Billy the Kid. Billy the Kid's grave is probably the main tourist attraction in this small town. At least a few people in this town are annoyed at Hico's insistence that the real Billy the Kid is buried nearby in Hamilton, TX. The dispute has escalated to the point that Ft. Sumner wants testing to establish that the bones in their grave have DNA related to the DNA in the bones of William Bonney's mother (whose grave is known). This test may indeed help establish that the body in the Ft. Sumner grave is that of William Bonney, but it will not resolve the more important issue of who did the killings.

# Summary: Main Points About Models

To wrap up these two chapters on models, we offer a list of points that you should understand now. Refer back to earlier text for elaborations.

1. Models are shortcuts; they simplify
2. Models exist in at least 3 general classes: physical, abstract, and sampling
3. Any one thing can be represented many models; conversely, one model may represent many things
4. All models are false
5. Model usefulness is judged by accuracy, convenience, and uniformity
6. Pieces and parts are models of the whole
7. Models must be refutable

## CHAPTER 6: MODELS OF SEX IN CONDOM TESTING

6  
MODELS

Condoms are made to withstand the rigors of sex. But the models used by governments to test condom durability have nothing to do with sex.

# Goals of Condom Manufacture: prevent disease transmission, prevent pregnancy, maintain sensation

Condoms (the male variety) have long been the mainstay of last-second, desperate appeals for contraception, although with the advent of AIDS and hepatitis B virus they have taken on additional prominence in the prevention of sexually-transmitted disease (the only purpose for which they were legally sold in many states just a couple decades ago). In either case, a condom merely serves as a barrier against microbes (sperm or pathogens). And to make sure that condoms indeed function in this capacity, governments of Western countries test condoms to meet minimal standards, to ensure that the condoms don't break or leak when in use.

Any male who has used a condom realizes that disease prevention and contraception are not the only goals in condom design - if they were, the simplest solution would be to avoid sex. An additional goal is to minimize the loss of sensation realized during sex. The goal of sensation is in direct conflict with the goal of preventing disease and sperm transfer, because the way to manufacture a condom to be a faultless barrier is to make it thick, whereas the way to maximize sensation is to make it thin. Manufacturers thus make condoms as thin as possible while maintaining minimum standards for condom strength. And it is precisely this compromise that leads to occasional condom failure.

The material used in condoms is latex (rubber). Advantages of latex are that it can be produced in thin sheets, and it can stretch greatly without breaking. But latex is a biological material, and it is profoundly sensitive to various environmental conditions. Even temperatures as high as body temperature degrade it, albeit slowly, and oils (as in vegetable oil and baby oil) degrade it rapidly. So even if condoms meet reasonable standards in the factory, they may fail under a wide array of conditions experienced in "the field."

There are many models of condom testing. You might ask why. The reason is merely that none of them are useful for each of the many goals we have in using condoms – they all have major limitations. Different models are used to overcome the perceived shortcomings of the other tests.

# Human Models

Condoms need to be tested to ensure that they meet certain standards. It is obvious that the truest model of condom performance is sex itself – that humans are the most accurate models. It is otherwise difficult to know how much sensation is being lost, or what the real transmission rate of disease is. Humans as models here have several drawbacks, however. First, they are not convenient. Governments are not about to hire people as condom testers, to have frequent sex with different partners, each with different sexually-transmitted diseases. We let people make their own choices about who to sleep with and whether they are willing to risk HIV, hepatitis, gonorrhea, etc., but we cannot ethically assign them to such risks. Any pregnancies arising from such experiments would raise problems as well.

An alternative to trained human “sex technicians” is human volunteers. This avoids the ethical/convenience complication, but it introduces a new one: lack of uniformity. Untrained people are notoriously inconsistent in how they use condoms, the condoms may be mistreated (e.g., they degrade rapidly when exposed to any kind of oils or Vaseline). So volunteers have their own problems. They are used, however, as we will come back to below.

A drawback of any use of humans to test condoms is time. If a batch of condoms was being tested for disease transmission or blocking pregnancy, it would take weeks to months after sex to determine whether the condom had done its job.

Add up all these problems with humans, and you can begin to understand why condom testing is done with non-human (non-animal) physical models.

# Physical Models of Durability

**“Empty-condom” tests:** There are lots of ways to test condoms that avoid ethical issues while adhering to high standards of uniformity. These kinds of tests are what all governments use. **Some tests measure the durability of the entire condom, some measure only part of the condom, and some test for holes.** Some of the more common tests involve testing empty condoms (or putting water in them). We might think of these as “dickless” tests, but this is of course not a term that you’ll hear used by professionals:

- **Electrical conductance test:** This is a non-destructive test applied to all condoms. Each condom is tested to see if it blocks electricity. An intact condom should not allow electricity to pass through it.

The next tests are all destructive – a tested condom cannot be sold.

- **Water leak test:** Used by the Food and Drug Administration, this test involves filling a condom with 10 ounces of water and looking for leaks.
- **Tensile test (stretch test):** This method involves slicing a band from the shaft of a condom and testing its stretchability.
- **Airburst test:** A method used by many European countries, Canada, and now the U.S., inflates the condom with air until it bursts; the maximum volume of air tolerated is used as the measure of strength.

Other tests are of the packaging (package integrity test, lubricant test) and a simulated aging test by warming the wrapped condom in an oven at 70° C.

Regardless of which specific test is used, condom testing involves taking a sample of several condoms from a batch and calculating the fraction that pass the test. The condoms tested are thus a sample (sampling model) of the others in the batch. In the US, a batch of condoms cannot be sold if 5 or more condoms per 1000 fail the test.

So you can be relatively confident that any condom sold in the U.S. (and maintained under proper conditions) will survive the water test and airburst test. Should we be comforted with that knowledge? Only to the extent that condom survival in these tests reflects condom survival during sex. That is, only to the extent that a water test or airburst test is a good model of the rigors of sex. Furthermore, the fact that different batches of condoms pass the FDA test does not mean that all of them are equivalent. Consumer Reports has evaluated several brands of condoms using the airburst test and has ranked them accordingly (to be presented in class). Perhaps surprisingly, a few brands had failure rates of 10% or more.

There are several levels at which models apply, beginning with the lowly one in which a few condoms in a batch are treated as a model of the entire batch (otherwise, we would have to test every condom in the batch). At a higher level, we may regard one condom brand as a model of other brands (hence the advice from healthcare workers to "use a condom," which makes no statement about a brand). Then we have the government models of sex that are used to evaluate condom survival, such as the airburst test, the water test, and the stretch tests.

Validation of the "empty-condom" tests. It takes little imagination to understand how these models are limited and may be seriously in error. However, although the airburst test is not anyone's idea of sex, when properly calibrated, it might give us a good idea of whether a condom will hold up during sex. Not surprisingly, people have been interested in this question. The organization Family Health International (see website above) has been involved in several studies and has also evaluated others to determine just how accurately the empty-condom tests predict condom failures in humans. These tests involve (i) evaluating some of the condoms with the airburst test (most commonly), and (ii) using other condoms from the same batch to obtain breakage rates from volunteers. These tests have been the most illuminating when done with aged condoms, because breakage rates are higher with aged condoms. In general, the physical, **empty-condom tests are only mildly good at predicting condom failure rate** experienced by human volunteers.

## More Accurate Physical Models: Fake Penises

The foregoing physical models can be objected to on the grounds that they don't mimic realistic forces during sex. Over the last decade or so, there have been a few studies that tested condoms by simulating sex with various types of dildos. None of those studies have inspired widespread acceptance of test. The elaborate system developed a decade ago by the "Mariposa" Foundation of Topanga California (probably now defunct ) consisted of a rubber "vagina" through which water (at body temperature) is circulated and into which a dildo (model of a penis) is thrust with a piston device. The condoms are inserted over the dildo and subjected to several "cycles" of piston thrusting. An ejaculation was also simulated. The various parameters have been established by a number of methods (quotes to be read in class). It is clear from the difficulties encountered by that organization (if only in calibrating their model) that creating a more realistic model that could be used as an industry standard would not be easy.

# Models of STD Transmission

Even if the models used to test condoms are reasonable indicators of whether a condom will break during sex, and thus whether they will function adequately in preventing sperm from reaching the female's reproductive system, but they may be rather poor indicators of whether a microscopic pathogen can pass from one partner to the other. For example, the water test can detect holes only as small as 5 mm, but this sized hole is many times the size of sexually-transmitted viruses and even of the bacterium Chylamidia. Similarly, the airburst test is insensitive to small holes. So here we find new limitations of existing methods of testing condoms: these models don't give us a good understanding of the barrier to pathogens afforded by a condom. That is, these models have serious limitations when considering condoms as barriers to infectious disease.

Other models have been tried. Several involve filling a condom with a pathogen (in water, for example) and determining whether the pathogen escapes to the outside - a passive transmission test. Some tests use the sexually-transmitted pathogen itself, which is the best model of a pathogen. But those tests are expensive because they require special facilities for working with pathogens. So other, simpler tests use the harmless bacterial virus fX174, which is somewhat smaller than the smallest sexually-transmitted pathogen (Hepatitis B Virus) and is easily assayed on bacterial plates. These tests can also be made more realistic by subjecting the condom to various forces, such as might be encountered during sex.

Volunteers. When we want definitive data on how condom use influences disease transmission rates, there is no substitute for accuracy – using sexually active humans known to be exposed to STDs, necessarily volunteers. There have been several studies of this sort in the last decade, mostly with HIV. These studies use “discordant” couples, in which one partner is not infected and the other is infected. Couples were rated (after the fact) as to whether they used condoms consistently or inconsistently (the latter category including those who didn't use them at all). Overall the studies suggest that condom use greatly decreases the risk of HIV infection.

NUMBER OF COUPLES IN WHICH THE HIV- PARTNER BECAME HIV+ DURING THE STUDY	
CONSISTENT USERS	INCONSISTENT USERS
0/123	12/122 (10%)
3/171 (2%)	8/55(12%)
<2%	12% (300 total couples)
0%	10% (250 total couples)

1996 study from Haiti: estimated a conversion rate of 1/100 person years with consistent condom use; 6.8/100 with inconsistent use.

(a higher female to male than male to female infection rate was observed in this study as well)

1994 study from Europe lasting 20 months: 0/124 consistent users converted; conversion rate of 4.8/100 person years among inconsistent users.

In an interesting twist on these and similar data, the CDC (Centers for Disease Control) public position on condom use shifted from emphasizing the benefits of STD prevention afforded by condoms to one that now emphasizes that condoms do not always prevent STD transmission (see the comparison of 'now and then' at <http://www.advocatesforyouth.org/PUBLICATIONS/iag/condom.htm>). It is suspected that the change in condom advocacy was from political pressure.

# Summary of Material in Templates for Models

MODEL	KIND	APPLICATION	LIMITATIONS
Passive transmission test	physical	pathogen passage through condom during sex	neglects wear and tear during sex
Volunteers	physical, sampling	condom integrity and breakage during sex	poor uniformity and compliance by untrained people; time to complete studies; other risk factors
discordant couples	physical, sampling	STD prevention by condoms	poor uniformity and compliance by untrained people; time to complete studies; other risk factors
PhiX174	physical	a sexually-transmitted pathogen	PhiX174 may pass through condoms differently than pathogens
Airburst test	physical	condom integrity and breakage during sex	lacks the complexities of sex; does not test porosity
electrical conductivity	physical	condom integrity and breakage during sex	lacks the complexities of sex; does not test durability of the condom
Stretch test	physical	condom integrity and breakage during sex	tests only part of a condom and lacks the complexities of sex; does not test porosity
A few condoms	physical or sampling	model of entire batch	variation exists between condoms of one batch
One brand	physical	model of other brands	different brands have different properties

<b>GOAL</b>	to provide a barrier against STDs, pregnancy	to provide a barrier against STDs, pregnancy	to provide a barrier against STDs, pregnancy
<b>MODEL</b>	airburst test	volunteers	airburst test
<b>IS A MODEL OF</b>	sex between humans	sex between humans	sex between humans
<b>ACCURACY</b>	+	-	+
<b>CONVENIENCE</b>	+	+/-	+
<b>UNIFORMITY</b>	+	-	+
<b>HOW USEFUL</b>	overall measure of condom integrity, convenient, uniform	accuracy – we can find out whether condoms work for their intended goal	not useful – the poor accuracy of the airburst test is too severe for this goal
<b>LIMITATIONS</b>	not address STD passage, not fully reliable indicator of breakage during sex (both due to poor accuracy)	source of failures unknown, limited sample sizes, slow turnaround (poor on convenience, uniformity)	does not enable assessment of sensitivity (due to poor accuracy)

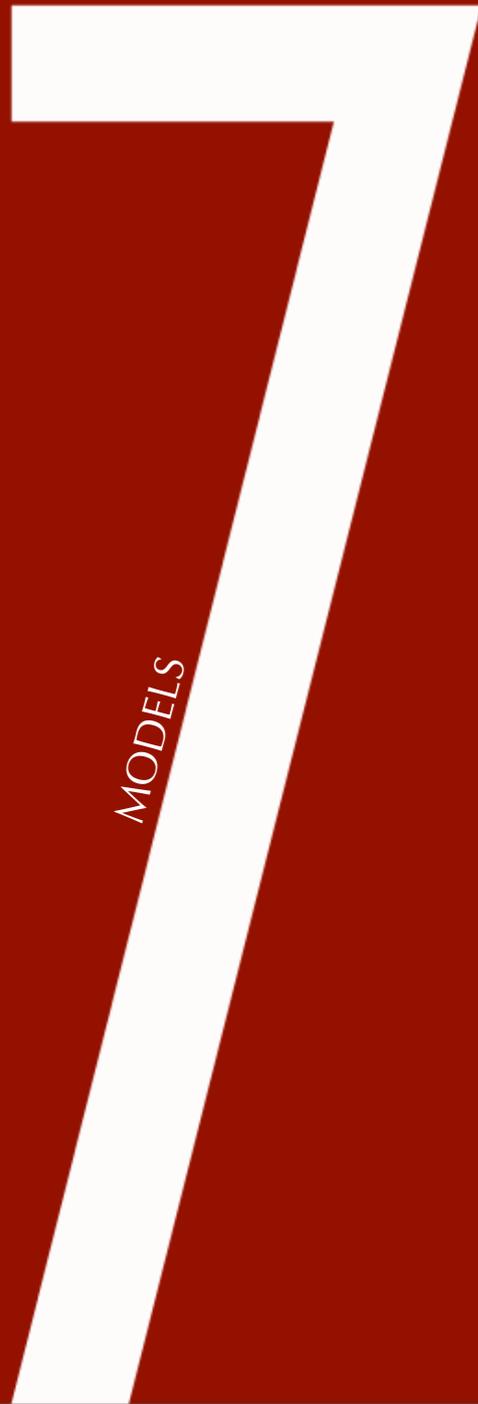
# External Links

[See how LifeStyles Condoms Are Made and Tested](#)

[National Geographic on Condom Testing](#)

[Condom Testing from Consumer Reports](#)

## CHAPTER 7: ARE YOU TOO INTOXICATED TO DRIVE SAFELY?



Drivers under the influence of alcohol are thought to be responsible for half the fatal traffic accidents in the U.S. To curtail this loss of life, it is imperative to have a means of detecting whether drivers are impaired. Several models of driving under the influence are used in Texas.

# The Problem: Lots of deaths from a combination of alcohol and driving

For people your age (18-24) in the U.S., 51% of the 8242 traffic deaths in 2001 involved alcohol. About 32% of that age group was in college, so an estimated 1349 traffic deaths of college students involved alcohol. The rate was 15 per 100,000, or about 7 per UT population. (This value is about twice that of the national average for suicides.) From another perspective, there is about a 1% chance that at least one of your classmates in your Bio301D section will die in an alcohol-related traffic accident this semester.

To put these numbers into perspective, the traffic deaths in your age group is about the same level as for U.S. personnel who died each year during the main 8 years of the Vietnam War. The turmoil caused by U.S. protests against the Vietnam War caused probably the most extreme social disruption of the post WWII generation. There has never been much of a protest against the same magnitude of traffic fatalities.

## **THE SOLUTION: DETERRENCE**

Alcohol is such an integral part of our society, especially in social gatherings, that people have not voluntarily avoided driving after drinking. In 2001, nearly 1/3 of college students in the U.S. reported driving while under the influence of alcohol. To increasingly discourage “driving under the influence,” we have increased the chance of being caught and increased the penalties. Nor are these tactics limited to the U.S. – Canada and many European countries are very aggressive about catching impaired drivers.

# The Ideal Model of Impairment

Once it is decided that driving under the influence (DUI) is unacceptable (i.e., criminal), we face the problem of establishing criteria for being impaired while driving. From our perspective in this class, we need models of DUI. The main issue is a person's ability to drive safely, so if we were to consider the most accurate model of DUI, it would include the driver's performance in:

- coordination
- judgement
- reaction time

It would be great to have a model of DUI that included each of these criteria, but we don't, although as we will see below, one model adopts some of these criteria. (Note that there is a legal distinction in Texas between DUI and DWI – the latter means driving while intoxicated – but for our purposes here, we are not concerned with the distinction. DWI is the more serious offense; DUI is reserved for drivers under 21 and does not require the same level of proof as DWI.)

The reason that we don't have the perfect model of DUI is the usual problem with all of our models – all models have limitations. In particular, it is not practical to administer a test that covers all of these criteria, and it would probably be difficult to measure these behaviors objectively. But you might live to see a test of this sort in the future, administered as a video game in a police car to test your ability to drive in a simulation. While such a scenario might seem far-fetched, the concept of a breathalyzer was equally unimaginable forty years ago.

# Texas law

The law in our state is both vague and specific about what constitutes impaired driving. The law (penal code 49.04) considers a driver to be legally impaired when:

1. not having normal use of physical faculties or mental faculties, or
2. having a blood alcohol concentration (BAC) of 0.08% or greater.

Older laws in some states used a BAC threshold of 0.15, later down to 0.10%, and most now use 0.08%. In Texas, a BAC may be measured in blood (gm EtOH in 100 mL of blood), breath (gm EtOH in 210L breath), or urine (gm EtOH in 67mL urine).

The vague model in this law is (i), lacking “normal use of physical or mental faculties.” It is vague, because there is no criterion for “normal use.” The test that is used to assess these behaviors is the Standardized Field Sobriety Test (SFST), although it is not part of the penal code. It typically consists of 3 parts administered where the driver is apprehended:

- (A) The Walk and Turn test (WAT)
- (B) The One Leg Stand (OLS)
- (C) Horizontal Gaze Nystagmus (HGN)

We will return to these tests in the section on Data, but for now the WAT test consists of walking along a straight line for 9 steps, turning around in a specific way, and returning along the line for 9 steps. The OLS test consists of standing on one leg, arms at sides, for about 30 seconds, while counting. These two tests are tests of coordination and ability to follow directions. The HGN test is a measure of the involuntary behavior of your eyes as they track an object to the side of your field of vision. Each of these tests is scored according to a strict set of criteria that includes following directions.

# Appraising Models of DUI

Many of you can relate to the limitations of the models of DUI, but it is also important to acknowledge the benefits:

MODEL	STRENGTHS	LIMITATIONS
BAC of 0.08% using blood	easy to obtain accurate reading; is an objective criterion	one threshold does not produce the same level of impairment in all people
BAC of 0.08% using breath	easy to obtain accurate reading of breath alcohol; is an objective criterion	one threshold does not produce the same level of impairment in all people; breath concentration may differ from blood concentration
SFST	performance is relevant to driving impairment; easy to administer- no equipment required	scoring is subjective; performance is affected by many factors unrelated to driving (road surface, physical properties of the person, age, shoes); no baseline data exist for each person

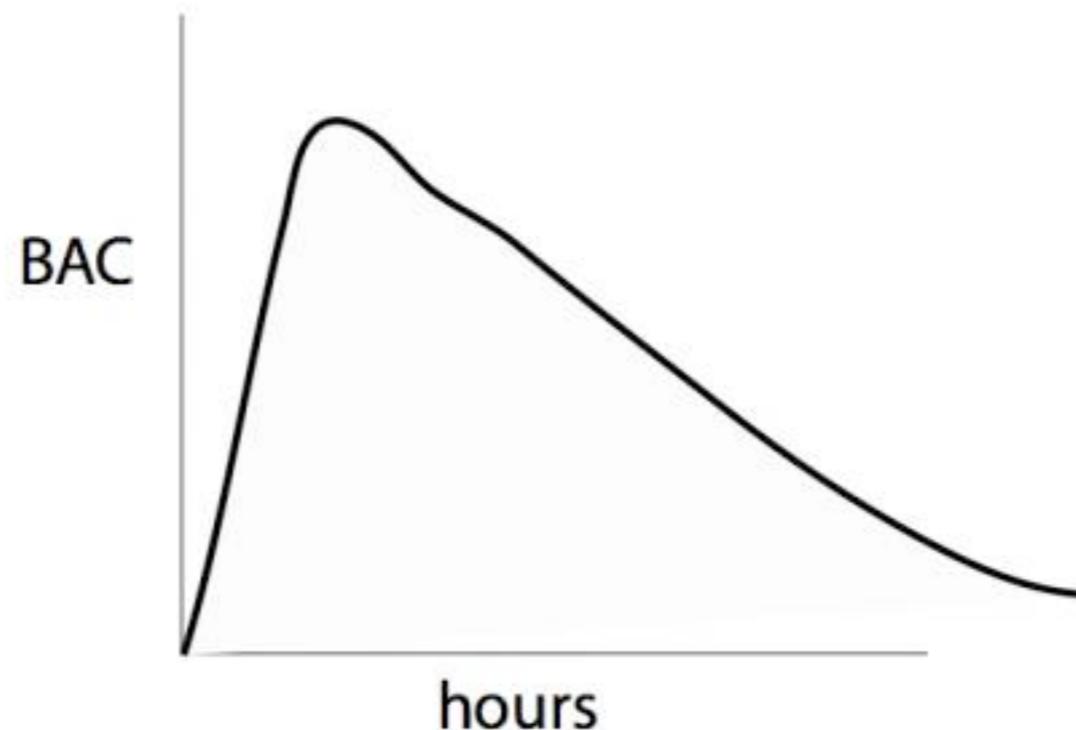
Perhaps the major limitation of any legal definition of DUI is that there are no gradations, because the legal system either finds you guilty or not. Impairment is instead a graded property of a person's behavior. A person at 0.070% may be legal to drive, but they obviously will not be as safe as at 0.04%, and even that will not be as safe as 0.01%. The law needs to set a threshold, but that threshold is a compromise which allows some impairment, at least in some people. So the legal definition of DUI is not an accurate model of impairment.

## New and Scarier Models

Suppose you are stopped while driving soon after leaving a restaurant, where you had a meal and 2 beers (or 1-2 glasses of wine). The SFST ordeal takes two hours. You pass (maybe they don't tell you this.) But then, in a moment of overconfidence, you blow for the breathalyzer. It comes out at 0.04, and you breathe a sigh of relief. Off the hook, you think. Not necessarily ....

A new tactic in some parts of Texas and perhaps throughout the U.S. is to back-calculate your BAC at the time you were stopped. If you were 0.04 some 2-3 hours after you last consumed alcohol, you may have been over 0.08 when you were stopped even if you were under 0.08 when you blew.

When people quickly consume alcohol on an empty stomach, a common pattern is that the BAC spikes soon after the alcohol is consumed and then the BAC beings a linear decay toward zero (this pattern is called a Widmark plot):



You can easily see how simple it would be to do the back calculation if you had one measure out near the right end (and knew the slope of the line). And this is what the courts are starting to do.

In this case, model accuracy is critical. Use of the Widmark plot is certainly 'convenient,' but if it is not accurate an accurate backwards measure of BAC, then most of us would think it should not be used to decide someone's guilt or innocence. It may come as no surprise to you that the model is not an accurate representation of BAC, largely because people and the circumstances under which it is applied, are not uniform. Measurements of breath alcohol content from people given known amounts of alcohol (of various types) have revealed that

1. the Widmark curve does not always apply, and even when it does,
2. there is considerable variation in the time of the peak and slope of the decay.

From this work, it seems almost impossible to make reasonably accurate back calculations. This limitation of the model has not stopped its use in court; in Texas, higher courts have even overturned lower courts' rejections of the method.

What this means is that a person could maintain their BAC well below 0.08 and still be convicted of DWI. Of course, this same outcome could happen from poor performance on the SFST.

# External Links

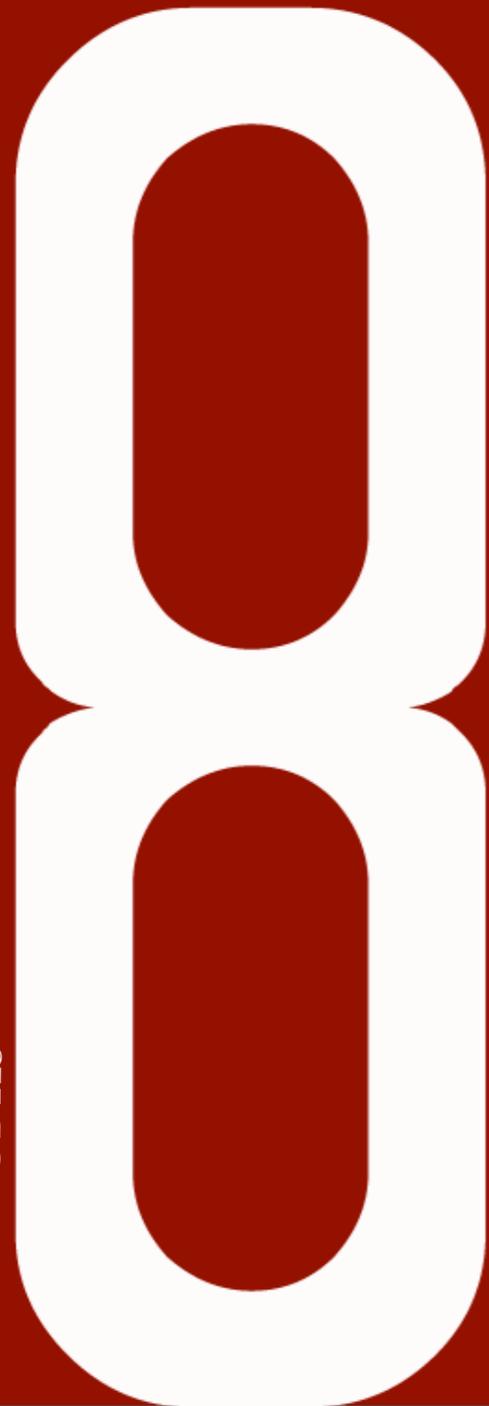
[Test Your Intoxication](#)

[The Man With Two Brains](#)

[Reno 911 DUI stop](#)

## CHAPTER 8: ERADICATING INFECTIOUS DISEASES

MODELS



The toll from infectious diseases in the U.S. was huge not long ago. Vaccines and drugs have been profoundly successful in reducing this scourge. Yet we have been successful in eradicating only one disease, smallpox. Simple mathematical models explain why it is difficult to eradicate a disease and why some diseases are more difficult than others to eradicate.

# Introduction

Our society has had many scares recently from new infectious diseases:

- HIV
- Ebola
- SARS
- West Nile
- Bird Flu

There are two different bases to our fears. One is how likely we are to get it. The other is how bad it will be if we do get it. In the above list, only HIV and West Nile have become permanent (endemic) infections in the human population, but we are or have been worried that the others might invade and start spreading. All but West Nile have high mortality rates.

No matter how bad a disease is, if we can prevent it from spreading in the population, it will die out. And even if we can't make it die out, reducing its spread means that fewer people will be hurt by it. Some very simple math models underlie the principles used to control infectious diseases.

# Basic Reproductive Number,

# $R_0$

In epidemiology, the most fundamental property of a disease is its basic reproductive number  $R_0$ , or the average number of new infections started by the first infection in a population. When an infection first enters the population, its spread is like a chain reaction: 1 infection becomes  $R_0$  new infections, those each start new infections to make  $R_0^2$ , then those become  $R_0^4$ , and so on. If  $R_0 > 1$ , the disease spreads in what is called an epidemic, and the larger the value of  $R_0$ , the faster the spread. The table below gives some estimates of  $R_0$  for some common infectious diseases.

DISEASE	$R_0$
Measles	5 - 18
Chicken Pox	7 - 12
Polio	5 - 7
Smallpox	1.5 - 20
HIV	2 - 12
SARS (crowded)	2.2 - 3.6
SARS (community)	1.2

We can see that the estimates have a lot of variation. In part, this is because  $R_0$  is a limited model – it is not the same across all environments. This point was important in the SARS outbreak. SARS showed moderately good transmissions in high-density apartment buildings and hospitals, but (fortunately) it was poorly transmitted in normal community settings.

From this epidemiological perspective, the goal is to reduce  $R_0$  to be  $<1$ . When this happens, the infection will die out, perhaps gradually. Intuition suggests two ways we might reduce  $R_0$ . One is to block transmission, as by cleaning up the environment, wearing masks or gloves, etc. The other is to use a vaccine to reduce the number of people who can get infected.

# Epidemiological Models

Some simple but elegant math underlies disease eradication. The main result uses an equation for the rate at which the number of infected individuals (I) changes over time:

change in number of infected individuals = new infections – loss of old infections from death & recovery

$$\mathbf{\Delta I = BI - I(r + d)}$$

SYMBOL	VALUE
I	number of infected individuals
delta-I	change in number of infected individuals
S	number of susceptible individuals
B	the infection rate parameter (disease-dependent)
r	the rate at which infected individuals recover
d	the rate at which infected individuals die

The goal is to make  $\Delta I < 0$ , and with 2 steps of algebra, this condition becomes:

$$R_0 = BS / (r + d) < 1$$

We haven't shown that  $BS/(r+d)$  is  $R_0$ , but it is when all individuals are susceptible ( $I = 0$ ).

Let's now consider how to use this result. The parameters  $b$ ,  $r$ , and  $d$  are all properties of the infection.  $S$  is a property of the population (the number of susceptibles). If we wore masks and cleaned up the environment, we would reduce  $B$ . A drug that hastened recovery would increase  $r$ . Vaccination would reduce  $S$ , because as more people are vaccinated, there are fewer susceptibles about.

Suppose now that  $R_0$  was 3 in an unvaccinated population. This formula would say we needed to reduce  $S$  to  $1/3$  (or vaccinated  $2/3$ ) to cause extinction. But if  $R_0$  was 30, we would need to vaccinate  $29/30 = 97\%$  of the population. If a vaccine was only 90% effective (which is pretty good), we could never reduce  $S$  to the 3% and achieve the extinction threshold with just the vaccine.

---

# Model Strengths and Weaknesses

Foremost, the mathematical model allows us to identify the properties of the infection and population that are important to control. No matter how much hands-on work we did with an epidemic and a vaccine, we might never understand why eradication did or did not work without some kind of model of the epidemic process. Furthermore, we can work through the model quickly, without having to watch thousands get sick and die year after year.

The model also gives us an idea of what to change to reduce infection rates: shorten the term of infection (increase  $r$ ), reduce  $b$ , reduce  $S$ . The latter two are obvious, but the first is not necessarily. Nor would we have a quantitative appreciation for what to do without the model.

The model has many limitations. It does not fit sexually-transmitted diseases even approximately. It assumes a constant  $b$ , which clearly varies from place to place. It assumes “mass action” (by counting the number of new infections as  $bSI$ ), which means that a susceptible person in New York has the same chance of getting the infection as a person in Langtry, TX, regardless of where most of the infections are found. We know this isn’t right. Indeed, when smallpox was finally eliminated, it was achieved by tracking down the last few infected individuals and ensuring that their contacts were vaccinated. Poliovirus still exists in part of Africa and Asia despite high levels of vaccination in the rest of the world.

# Flu

The latest infectious disease to cause significant worry is bird flu. Bird flu is caused by a virus designated influenza A. Bird flu virus is the same general type of influenza virus that circulates in the human population every year, for which we have vaccines. Flu is generally a respiratory infection, causing fever and aches, but one rare type of flu infects the membranes of the eye. In typical years, influenza kills an average of 36,000 Americans, mostly old ones.

Influenza poses several types of problems:

- First, we have influenza vaccines, but we need to be re-vaccinated every year. Why? The reason is that influenza viruses keep evolving to escape our immune system, so that old immunity becomes progressively less effective. It's not that our immunity wanes, rather the viral targets of that immunity keep changing. But this change in the virus is slow and gradual. The situation is similar to a car manufacturer making slight changes to a model every year, to keep pace with changing consumer preferences. The newer vaccines keep pace with newly-evolved viral structures.
- There are several types of influenza strains. These have very different properties from each other with respect to our immune system (much greater than the differences that evolve within a type). They are designated with letters and numbers, such as H3N1. This system refers to 2 of the types of proteins in the virus (hemagglutinin and neuraminidase) that are important to immune recognition. Currently, the human population has H3N1 and H1N1 circulating. Birds have several types (up through H9) that are not found in humans. There are also influenza strains found in other animals (pigs, horses, ...) Occasionally, a strain from a species with another H-type or N-type successfully humans and starts an epidemic referred to as a "pandemic." Pandemics are problems because they kill more than the usual numbers of people.

Three pandemics from the 1900s are recognized:

- i. 1918 – the first H1N1 (Spanish flu), killing 20,000,000-40,000,000 people worldwide (500,000 in the U.S.)
- ii. 1957 – the first H2N2 (Asian flu)
- iii. 1968 – the first H3N2 (Hong Kong flu)

H1N1 died out but was re-introduced by a botched vaccine in Russia. H2N2 is gone now, and we have H1N1 and H3N1 in our population.

- The bird flu we hear lots about is H5N1. So far it spreads rapidly in birds, but dies out whenever it jumps into humans (thus  $R_0 < 1$  for humans so far). Our population has no prior exposure to it, which may explain why the mortality rate is about 50%, but lack of immunity is probably not the sole cause of this high death rate. The big worry is that, if we don't monitor it carefully, the virus will jump into humans, spread among a few people with close contact, and evolve to the point that  $R_0 > 1$ . When this happens, we are in trouble, unless there is a vaccine. Although we have known about H5 infecting humans for nearly a decade, making a vaccine has been difficult because other influenza viruses for vaccines are grown in chicken embryos. H5N1 kills the embryos, so we can't get enough virus for a vaccine. This problem has been overcome, but initial vaccine results have been discouraging because the vaccine did not elicit high antibody levels.

# External Links

[Eradicating Polio](#)

[Bill Gates on Eradicating Polio](#)

[D.A. Henderson, Leader in Global Smallpox Eradication](#)

[D.A. Henderson, MD, Chronicles the 10-year fight to eliminate smallpox](#)

[Eradicating Rinderpest](#)

# CHAPTER 9: EXTRAPOLATING HEALTH RISKS

39

MODELS

# Introduction

## AT&T TO CUT WORKFORCE 120 PERCENT

NEW YORK, N.Y. ([SatireWire.com](#)) AT&T will reduce its workforce by an unprecedented 120 percent by the end of 2003, believed to be the first time a major corporation has laid off more employees than it actually has.

AT&T stock soared more than 12 points on the news.

The reduction decision, announced Wednesday, came after a year-long internal review of cost-cutting procedures, said AT&T Chairman C. Michael Armstrong. The initial report concluded the company would save \$1.2 billion by eliminating 20 percent of its 108,000 employees.

From there, said Armstrong, "it didn't take a genius to figure out that if we cut 40 percent of our workforce, we'd save \$2.4 billion, and if we cut 100 percent of our workforce, we'd save \$6 billion. But then we thought, why stop there? Let's cut another 20 percent and save \$7 billion...

Of course, it is immediately obvious that this "news" piece from [SatireWire.com](#) is not serious. One cannot cut more than 100% of a workforce. Furthermore, the supposed motivation for cutting the entire workforce is to save five times the money that would be saved by cutting only 1/5 of the workforce.

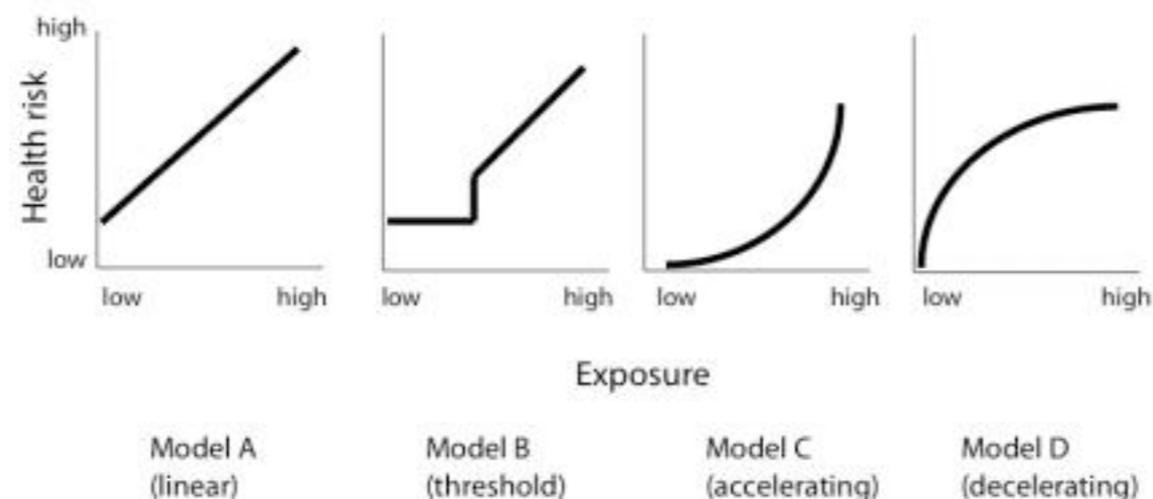
This example introduces the concept of dose extrapolation. We deal with it every day, one way or another. The above article is humorous because we realize that some things cannot be extrapolated proportionately: adding 10 tablespoons of salt to your kettle of soup does not make it taste 10 times better than 1 tablespoon of salt. (One of our friends, the well-known population biologist Bruce Levin, once tried to speed the baking of a sponge cake by raising the oven temperature to its maximum; the cake did not turn out as expected.)

Dose extrapolation enters most calculations of risk. Our government pays a lot of attention to environmental hazards that kill us or give us disease. The people who do these calculations worry a lot about small risks, because a small risk multiplied by 250 million (the approximate U.S. population) can add up to a lot of death or disease. (It is of some passing interest that second-hand tobacco smoke is estimated to kill about 3000 Americans each year – the approximate number killed in the World Trade Center attacks, yet the country has been slow to take on this health risk.) Time and again, however, we estimate small risks by extrapolating from high risks. Many of these extrapolations may be just as flawed as in the AT&T spoof, but we don't know it.

# Extrapolation Across Doses: The Abstract Models

Too much of almost anything can kill you – even the things we eat and drink daily. But by now, most of us know to avoid the things that are likely to kill us outright – we call them poisons, accidents, etc. There are many other exposures in life that won't kill us immediately, but they might have a long-term, cumulative effect that eventually does us in. Cancer is one such concern. We increase our cancer risks by smoking, eating poorly, being exposed to the sun, driving, getting X-rays and other types of radiation exposure (the risk from radiation exposure is almost too small to measure), and by inhaling or ingesting any of many man-made or natural chemicals in our environment or the workplace. In most of these cases, we not only want to know how to avoid large risks but also how to avoid small risks – and what the risks actually are.

Small risks are usually calculated from large risks, and to determine the small risks, we need to extrapolate. Fortunately or unfortunately, the natural laws of risk extrapolation do not necessarily follow straight lines. There are several basic forms of extrapolation that need to be considered:



Four abstract models of dose extrapolation. They differ in the health risks associated with increasing exposure to the health hazard. Without knowing which model applies, we do not know how to estimate the risk at low doses from observed risks at high doses.

It doesn't matter whether the vertical axis is death rate, survival rate, cancer rate, or other measure, the problem is basically the same. For concreteness, we will assume that we are trying to determine the relationship between cancer rate and exposure to some agent (e.g., radiation). All four graphs are models of how radiation might lead to cancer, and we want to know which ones are correct.

- **Model A** is the easiest to understand. Under this model, your cancer rate is simply proportional to your dose. This model could not apply across all conceivable doses, because the cancer rate can't exceed 100% - just as AT&T can't lay off more than 100% of its workforce – but it could apply to the range of exposures that interests us.
- **Model B** is possibly counter-intuitive: increases in exposure do not increase cancer rates until the exposure reaches some threshold. We have drawn a linear response to exposures above the threshold, but our interest in this model lies chiefly in the existence of a threshold regardless of the form of the curve beyond the threshold. In a variant of this type of model, the flat part of the curve (at low doses) actually decreases with increasing exposure – a phenomenon known as hormesis.
- **Model C** is one in which cancer rate increases more steeply with increasing dose. That is, doubling your exposure more than doubles your increased risk, so we call it accelerating.
- **Model D** also shows increasing cancer levels with increasing exposure, but the rate of increase declines with exposure (decelerating).

The statistical difficulties. In principle, there needn't be anything unusually difficult about telling apart different curves – you have no difficulty figuring out that a little bit of salt goes a long way, whereas you have a far wider tolerance range for sugar (unless you are diabetic). Why, then, is it so difficult to determine health risks? Four reasons. First, it is usually difficult to get large numbers – we are often concerned with rates on the order of 1 in ten thousand or less, so we may not have more than a handful of positive cases. For example, there is a grand total of 5000 factory workers exposed to reasonably high levels of dioxin in the US. That's it -- no chance of obtaining a bigger sample. Second, we are usually addressing rates of something, whereby everyone is either affected or not. By analogy to a coin flip, every individual is either a “head” or a “tail”, and the rate we are concerned with is a probability that lies in between heads and tails. The statistical probabilities of rates are harder to estimate with precision than are quantities such as IQ scores, and speeds. We have much greater statistical power when the values we measure can taken on any value across a continuum than when they can only take on either of two values. Third, the different dose extrapolation models differ chiefly in how they deviate from linearity. Our statistical methods are most useful at fitting straight lines to data. It requires

relative few points to demonstrate the slope of a relationship, but it requires many points to show a deviation from a linear relation. (Fundamentally, this is also a sample size problem) Fourth, we often have incomplete information about how potential poisons and carcinogens act inside the body. What are the metabolites? Are the metabolites more or less toxic. What are the half-lives? What cell-receptors does the chemical bind to? What cascades does this trigger? We don't have this information because the science just has not yet advanced far enough. Indeed, if we understood this much, we would not require large samples to figure out the form of dose extrapolation.

# Animal Extrapolations

We are all familiar with the LD50 concept – the dose of a substance that will kill half those exposed to that dose. Virtually everything we eat has an LD50 – table salt, sugar, water, and many more everyday items. Animals, mostly mice and rats, are used to establish lethal doses and other biological effects, but other common testing animals are rabbits, guinea pigs, birds, and one person at UT even uses turtle embryos.

At present, there is no alternative for determining the lethal effects of most substances, although we may be approaching an era in which lethal and other harmful effects can be ascertained in cell cultures. The determination of a lethal dose in mice, however, does not necessarily extrapolate directly to humans. Thus, if we determine that the LD50 of some compound is 1 milli-gram of stuff per kilogram of rat, it does not mean that the human LD50 will also be 1 milli-gram per kilogram of human. A correction is often needed to go from the LD50 in the test animal to the LD50 in humans.

---

# Extrapolations Across Related Hazards: TEF & rad/rem

In the section below on radiation, we will find that harmful effects of vastly different types of radiation are measured on a single scale (rad or rem). The same concept has been developed for some classes of related chemicals: the toxic equivalent factor or TEF. There are over 200 chemically distinct dioxins, furans and PCBs (depending on where the Cl atoms go on the benzene rings. And dioxins, furans and PCBs differ only in how the two benzene rings are attached). All dioxins, furans and PCBs are thought to act biologically only via binding to the Ah receptor. Evidently, the different congeners differ in their binding affinity, and hence in their biological effects. The TEF for each of these chemical congeners is an estimate of how toxic it is relative to 2,3,7,8 tetrachlordibenzo-dioxin (the most toxic of them all, with a TEF of 1.0 by definition). The TEFs for the others are often orders of magnitude smaller (see table at end of <http://www.epa.gov/ncea/pdfs/dioxin/part2/drich9.pdf>). Thus, TEFs were created to solve the problem of extrapolating from one chemical to another. The example shows how molecular biology will eventually shed light on extrapolation problems.

# Rodent Models of Cancer

Drugs and food additives for human consumption need to be tested for their possible abilities to cause cancer. One of the important tests involves feeding the substance to rodents (rats or mice) and assessing cancer rates.

Consider how to use an animal model efficiently and ethically when testing whether a food additive causes cancer. There are several considerations in setting up a study:

1. What organism to use -- humans are the most accurate, but experimentation is costly, lengthy, and often considered unethical (depending on the experiment and country. Some pharmaceutical manufacturers have gone overseas to gather data using human subjects using experimental protocols that would be unethical and hence impossible to implement in the US.). Bacteria and yeast are inexpensive and free of ethical considerations, but they are single-celled and cannot become cancerous. Rodents offer a good compromise between humans and simpler organisms and thus are used for much of testing.

Given the choice of a rodent model, other issues arise:

2. How many rats - fewer is cheaper. (The UT Animal Resources Center lists costs per cage for different types of animals – which does not include the cost of the animal. A medium rat cage costs about \$0.50/day and can house 2-3 animals. Thus a study using 1000 rats would cost about \$250.00 per day.) Fewer is also more ethical from the perspective of animal welfare, but an inadequate sample size may fail to detect an effect (larger sample sizes will be able to detect smaller effects), which has negative ramifications for humans.
3. How long to monitor them -- cost increases with the duration of the study but so does the ability to detect harmful effects
4. The dose – if small doses cause cancer, then higher doses will cause even more cancer and possibly cause cancer faster.

The compromise achieved amid these conflicting issues has often been to feed rodents large amounts of the substance -- enough that a calculated number of the animals actually die from it.

Bruce Ames has pointed out a possibly serious limitation of this model -- that the cancer-causing potential of the substance in humans may be overestimated by the rates in mice. High doses typically kill lots of cells in the mouse, even if the mouse survives. The mouse responds by replacing those cells -- by dividing at higher than normal rates. These increased cell division rates, all by themselves, can lead to increases in cancer. (Cancerous cells have an abnormally high cell division rate, and cells become predisposed to become cancerous when they divide.) Feeding mice so much of an otherwise harmless substance that their bodies must produce lots of new cells may give the mice an elevated cancer rates; people would not eat high enough doses to have their cancer rates affected.

The difficulty with this model arises from dose extrapolation – do low doses cause cancer at a rate proportional to the cancer rates from high doses? In essence, Ames is arguing that many chemicals with cancer-causing activity at high doses are not cancerous at low doses (an extreme form of the “accelerating” model). The matter is unresolved.

---

# Fetal Alcohol Syndrome

We can go back nearly 100 years to find the first observations that children born to alcoholic mothers exhibit certain facial abnormalities, have reduced motor skills, and have reduced mental capacities. The characteristics across different afflicted children are similar enough (though varying in degree) that the term “fetal alcohol syndrome” is used to describe the condition. Widespread public awareness of the possible ill effects of maternal drinking during pregnancy has come only in the last 3 decades. Alcoholic beverages now display warnings for pregnant women, something that was not done as little as 30 years ago.

The first warnings to the public recommended no more than 2 “drinks” a day for a pregnant woman. Now the advice is to abstain entirely during pregnancy. The difference between the “2 drinks a day” threshold and the “zero tolerance” rule is the difference between acceptance of a threshold model (or an extreme form of the accelerating model) on the one hand, versus acceptance of the linear (or decelerating) model on the other hand. Recent studies demonstrate effects on child behavior down to as low as an average of 1 drink per week during pregnancy. The enhanced awareness of the risks of fetal alcohol merely represents better and more data from low doses – direct observations at low doses rather than extrapolations.

---

# Secondhand Smoke

The admonitions against smoking given out in high schools 4 decades ago did not extend to smoke from someone else's cigarettes: the risk of smoking was only to those who puffed, we were told. Jump forward to the present: second-hand smoke is estimated to kill 3,000 people in the U.S. annually. For lung cancer, living with a smoker elevates your risk by 30% (the smoker's risk is elevated 1000%).

This risk has been estimated two ways: directly, and also as a linear extrapolation. Both estimates agree reasonably well. But how is this done? First of all, how do you even determine the level to which a non-smoker is exposed to tobacco smoke? This much is a prerequisite for determining the risk from second-hand smoke. A breakdown product of nicotine is cotinine, whose levels can be assessed in urine (just like other drug tests). Its persistence time is long enough to be useful for measuring exposure to nicotine, which is used as a surrogate for tobacco smoke (although there are now several products available that can give you nicotine without the smoke). Cotinine levels in your body are thus a model of exposure to tobacco smoke. From there, one measures lung cancer rates in large groups of people and estimates the risk. Initially, the risk of second-hand smoke was estimated by a linear extrapolation from smokers' risks: if you know the amount of smoke that a non-smoker receives compared to a smoker, and you know the excess lung cancer rate of a smoker, then you can easily estimate the elevated lung cancer rate of the smoker. Studies that look at excess cancer rates in non-smokers who receive significant amounts of second-hand smoke have supported the risk estimates.

---

# Pesticides in Your Food

A topic that occasionally hits the news is pesticide (and other chemical) residues in food. Not too long ago, the controversy was about Alar in apples, which isn't even a pesticide, but is used for cosmetic purposes (to keep the produce looking nice). A few celebrities took issue with the fact that this chemical served no "purpose" but that it was estimated to cause a handful of cancers in the U.S. population each year. The same issues underlie pesticide residues in food – that produce must have such incredibly low residual levels of pesticides by the time the food reaches market. Public concern over pesticide residues in food is driven by the fact that high doses can cause cancer in lab animals, and this concern has spawned a significant demand for "organic" produce. (One of the largest retail suppliers in the U.S. is Whole Foods, a company that started in Austin.)

Acceptable pesticide residues in food are based on extrapolations of risk from higher doses. Very high doses of pesticides can kill you, but sub-lethal doses can cause cancer. Here, there is great uncertainty regarding the actual cancer-causing effect of low pesticide residues, and the best that the EPA (Environmental Protection Agency) can do is a linear extrapolation. However, the data are simply so sparse here (and there are so many confounding factors with diet – plants are full of gobs of natural pesticides) to know whether these risks are anything close to real.

---

# Dioxins and Animal Extrapolations

Just a few decades ago, dioxin, a chemical widely used in the electronics industry, was found to be incredibly toxic. There was already enough of it in our environment to cause problems, and indeed, many people then and now have detectable levels of it in their blood. But determining the risks to humans has not been easy and it now appears that the early concern about extreme toxicity of dioxin was due to an inappropriate animal extrapolation. For some lab animals, dioxin is very toxic in acute doses—by weight, the LD50 for guinea pigs is at the 1-billionth level (1 billionth of a gram of dioxin per gram of body weight). But the guinea pig is the most sensitive species found. Hamsters can tolerate 3000 times this dose. Different rat strains vary in sensitivity by about a thousand-fold. The LD50 for acute doses in humans is unknown, but we are certainly at least 10 times and maybe hundreds of times less sensitive than guinea pigs; indeed, no cases of acute mortality (sudden death from exposure) are known in humans – there have been some occupational exposures. There are, of course, long-term effects of dioxin beyond acute mortality (e.g. cancer). Some biochemical indicators in cell

# A Detailed Example: Radiation

It's in this century that we've come to understand radiation, and it offers many technical advances:

- nuclear bombs & nuclear power
- medicine (X-rays, radiation treatment of cancer)
- tracers in applied and basic research
- agriculture (screw worm eradication)

Yet long before radiation found wide application, scientists were aware of its harmful effects. In 1927, H.J. Muller reported that X-rays caused inherited lethal mutations in *Drosophila* (a small fly raised in the laboratory). Initial concern about radiation as a human health hazard thus focused on its potential in causing inherited defects, but most attention nowadays concerns cancer. Despite society's perhaps excessive concern about cancer from radiation, the scientific foundation here has numerous uncertainties.

A preceding chapter (Dose Extrapolation) addressed the abstract models evaluated when assessing the risk of low exposures. This chapter looks at the many physical models that assist us in understanding these risks and offers some details about the true complexities of making risk calculations.

### **(1) No Animal Extrapolation: Humans Only**

We know from studies of insects that radiation causes genetic damage. But insects don't develop cancer. Even with mice (which can develop tumors), their biology is sufficiently different from human biology to doubt their utility as a model of human cancer. Humans thus make the best models and perhaps the only satisfactory models when studying cancer. And virtually all of the data used come from humans. However, because experiments in which people are deliberately exposed to radiation are not permitted, the humans we have this data for are limited to a few groups. Examples include: Japanese survivors of atomic bomb blasts, U.S. soldiers exposed to radiation during atomic tests, and victims of diseases whose treatment requires large radiation exposures. So the radiation data are based on the most suitable organisms (humans) but are limited in the extent to which a wide range of sources of human exposure have been included.

## **(2) Extrapolation Across Types of Outcomes:**

### **Few Cancers Can Be Quantified**

**(A) Leukemia:** There are dozens of kinds of cancer, and radiation may well contribute to all of them. Yet most of the radiation-cancer data are for leukemia (an over-proliferation of white blood cells). One reason is that leukemia is a relatively common form of cancer, especially so for children. Second, the time lag between radiation exposure and the resulting cancer is shorter for leukemia than for many other cancers, which also contributes to the ease of studying it. The limitation of relying so heavily on leukemia as a "model" cancer is that it may not represent all cancers.

**(B) Chromosome Breaks:** In view of the enormous sample sizes required for detecting increases in cancer from modest increases in exposure, it is useful to have alternative physical models that give us insight to the risks from radiation. A simple assay that can be applied to large numbers of people is the incidence of chromosome breaks in white blood cells. A sample of blood is drawn, the white cells are cultured, and the chromosomes of the dividing white cells are spread out on a microscope slide. Chromosomes whose "arms" are broken can be identified quite easily. So this assay can be performed on thousands of people, and it has a potential to be quite sensitive, because tens of thousands of chromosome arms can be screened per person, enabling detection of slight increases in the rate of chromosome damage. Even so, studies have concentrated on people who have received large doses of radiation: Japanese bomb survivors, nuclear shipyard workers in Scotland, uranium miners, and victims of ankylosing spondylitis (treated with 15 Gy). The limitation of this assay is that the form of the relationship between radiation and chromosome breaks need not be similar to that between radiation and cancer.

### (3) Extrapolation Across Hazards: Many Types of Radiation

We tend to talk about all types of ionizing radiation as if they were the same. Well, not quite. We worry about exposing our skin to ultra-violet light, and we realize that a hat, shirt, or sun-screen will protect us. We do not assume that we can be so easily protected from other types of ionizing radiation. But the very use of the word radiation to include these various classes of physical phenomena reflects our lumping them together. In fact, there are many types of ionizing radiation, and we are familiar with most of them. They fall into two main classes:

1. Electromagnetic (photons): ultra-violet light (UV), X-rays, g rays, cosmic rays
2. Atomic particles: b particles (electrons), a particles (helium nuclei), neutrons

Each of these types carries enough energy to enter cells, and by colliding with atoms and molecules in the cell, they can cause harm to the cell's chemistry. In general, a high-energy photons (X rays, g rays, cosmic rays) can sometimes go through us and some types do so with only a small probability of causing damage per photon -- the very fact that we can use X-rays to expose film when we stand between the X-ray source and film illustrates that many of the X rays pass through us. The particles, however, tend to be stopped more easily. Furthermore, the kind of damage each type of radiation causes is somewhat different: g rays and cosmic rays are much higher in energy than X-rays and UV, hence can do different types of damage. So the cancer-causing effect of a dose of radiation will vary at least slightly from one type of radiation to another.

Measures of Radiation The science of radiation and cancer would be perhaps unbearably complex were it not for the fact that physicists and biologists have figured out ways to compare the biological effects of radiation on a common scale. That is, the biological effect of an exposure to X-rays can be equated to the biological effects of a particular exposure to gamma rays or b radiation, and so on. One reason that such a common measure is desirable is that people as a group are exposed to multiple kinds of radiation, and it is easier to keep track of overall doses than to monitor each type separately. (The last chapter introduced the concept of TEF, or Toxic Equivalent Factor, which attempted to achieve the same extrapolation across different toxic molecules.) The following list explains some measures of radiation.

- A roentgen is a measure of radiation according to the number of ions which are produced in a standard mass of air. This measure is from physics.

The following measures are attempts to convert physics measures of radiation into biological measures.

- A rad (roentgen absorbed dose) is a measure of absorbed energy/roentgen.  $100 \text{ rads} = 1 \text{ Gray (Gy)}$
- A rem (roentgen equivalent man) is the biological response in man which equals one roentgen of x-rays. At low exposures the rad and rem are roughly equivalent.  $100 \text{ rems} = 1 \text{ Sievert (Sv)}$

There are obviously many possible ways we could have decided to measure the effects of radiation, but these measures have been adopted worldwide and will be with us until a better alternative is found. In the context of this chapter, we can think of each way of measuring radiation itself as a model. Two people who both receive exactly 10 rads of radiation can have different biological responses. Other factors partially determine the biological consequences of 10 rads of radiation, such as a person's age, whether the radiation was received in one large dose or many small doses, and whether the radiation was in the form of  $\beta$  particles or X rays. Hence, the statement that someone "received 10 rads of radiation" is a summary, or model, of the radiation that that person actually received. That model is an attempt to allow us to combine the effects across the different types of radiation.

The difficulty of the problem. Our understanding of the cancer risk from above-background radiation is based on haphazard models when it comes to the types of radiation involved in the exposures. As noted above, we do not experiment with people to determine the cancer risk of radiation, so we must rely on medical, military and occupational exposures. For the most part, these kinds of exposures were not measured at the time, and there is a fair bit of guesswork in calculating the types of radiation involved. Two conceivable problems arise from this dependence on uncontrolled exposures. First, and as already discussed, the different kinds of radiation may have different effects in causing cancer. If the rem and rad indeed accurately collapse the differences in cancer risk from diverse types of radiation, then this problem is not serious; but the rad and rem are not calculated from cancer risk directly, so this problem may be real. Second, the calculated exposures may themselves be in error in these studies.

For example, with the Japanese bomb survivors - the largest and most extensive database for cancer risk from radiation - there is now controversy over whether most of the radiation from the blast was neutrons (quite harmful to cells) or photons (less harmful to cells), and earlier calculations had assumed mostly photons. (After all, a neutron bomb is simply an atomic bomb whose plutonium core was compressed more than normal, to change the types of

#### (4) Extrapolation across doses

We do not live in a radiation-free environment. Each of us is exposed to radiation throughout life. To appreciate what excess radiation exposure means, it is useful to understand what the baseline exposure is. An average American can expect to receive is about 300 milli-rem/year (mrem/yr). You might be surprised that this estimate is fully three times what we thought in the 1980s – we didn’t know how common our exposure to radon was until recently. The breakdown is as follows:

Radon	~ 200 mrem
Rocks/Soil	~ 28 mrem (90 in Colorado; 23 on the East or Gulf Coast)
Cosmic	~ 28 mrem
Internal (K-40)	~ 26 mrem
Medical	~ 10 mrem
Dental	~ 5 mrem
Fallout from Weapons Testing	< 0.3 mrem
<b>TOTAL</b>	<b>~ 300 MREM</b>

[SOURCE LINK](#)

(for perspective, a single dose of 350 rads is serious enough to require medical care and have effects weeks into the future, but it is not quite enough to kill most people; that is about 1000X the annual average).

These values are averages. Your own exposure to medical and dental sources depends on whether you get X-rays (a chest X-ray adds 30 mrem to your annual dose, but a mammogram and gut fluoroscopy each add 200 mrem.). Your exposure to terrestrial sources depends on where you live. Your exposure to cosmic rays depends both on your elevation and your occupation, because the atmosphere screens out most of the cosmic rays. For example, pilots and flight crew members are exposed to excess radiation, but the excess is less than 25mrem/year, and there is no evidence for elevated cancer rates in these occupations. And there is an 80-fold increase associated with smoking; however it is due mostly to alpha particles and is limited to the lining of the lungs. (Is 80-fold the same as 80%?).

High doses. Most of the effort in studying the cancer risk from radiation has gone into groups of people who have received large doses. This is not to suggest that we should be complacent about smaller doses, but rather the cancer risk even from large doses is small enough that it takes years of work and tens of thousands of people to detect statistically-significant increases in cancer rate. So we focus on people who have received large exposures and extrapolate to low doses. Of course, by focusing on people who have received large doses, our data do not tell us about which of the dose extrapolation models apply. The problem is a classic "catch-22": we want to know about the cancer risk from low doses of radiation, but we need to study people who have received large doses in order to measure the effect. Yet these data don't necessarily tell us what we want to know.

The Japanese Database. One of the first, and still perhaps the most extensive database is from Japanese residents of Hiroshima and Nagasaki who survived the atomic bomb blasts. After the Japanese surrendered, the surviving victims in Nagasaki and Hiroshima were interviewed. Approximately 6,000 survivors exposed to 100 rads of radiation and 40,000 survivors exposed to 1 rad of radiation were identified, based on their statements of how close they had been to ground zero at the time of detonation. These individuals and their families were then monitored. To appreciate the difficulty of studying increased cancer rates, consider the annual number of excess deaths per million people per .01 Gy of exposure from the bomb:

leukemia	4 (in 1952)	1 (in the 1970's)
other cancers	2 (in 1952)	4 (in 1972)

The medical treatment for ankylosing spondylitis involves an accumulated exposure of 15 Gy (given over many years), and this group has been used as well.

### **Not much definitive resolution on dose extrapolation**

Putting it all together, there isn't much we can say about the shape of the association between radiation and cancer. Low doses don't elevate cancer rates very much. The only cancer with adequate data is leukemia. For this one type of cancer, the accelerating model applies – big doses are proportionately worse than small doses. For the rate of chromosome breaks, the linear model applies.

<b>“CANCER”</b>	<b>MODEL SUPPORTED</b>
leukemia	Accelerating
chromosome breaks	Linear

# Power Plant Accidents and Public Attitudes

**3-Mile Island:** When most of you were too young to remember (which includes not being born), the U.S. had an accident at one of its nuclear power plants: in March, 1979, one of the reactors overheated at the 3-Mile Island power plant near Harrisburg, PA, and some radioactive gas (approximately 10 Curies) was released into the atmosphere. You might ask why we haven't studied cancer rates in those citizens exposed during that accident, thus augmenting the Japanese data. The reason is that the exposures were trivial:

- the largest exposure was to 260 nearby residents = 20-70 mrem
- the average increased exposure within a 10 mi radius = 6.5 mrem

Assuming equal exposure to radon, the average exposure per year for a resident of Harrisburg is 116 mrem, and of Denver is 193 mrem. Thus, the worst increased exposure from this accident for these Pennsylvania residents was on the order of a summer-long visit to Denver. The accident shut down the unit for over a year and was very costly, so it was not economically trivial. But we will never be able to see an elevated cancer rate in residents near 3-Mile Island resulting from this accident. However, public reaction to the accident was outrage and near hysteria.

**Chernobyl:** In April of 1986, a true meltdown and rupture of a nuclear power plant occurred at Chernobyl in the Ukraine. Nuclear power plants cannot explode in the sense of an atomic bomb, but because the core is housed in water, overheating can create such pressure that the cement containment vessel bursts. The amount of radioactivity released was 50-250 million Curies, which is 5-25 million times the amount of radiation released at 3-Mile Island. 30 people died in the accident, mostly workers at the plant who knowingly went outside to inspect the damage (the effect was to kill their skin cells within days). The accident shed radioactive ash and gases (radioactive iodine was the main radioactive gas released). Much of the ash fell nearby, but the gases were distributed widely and exposed people across parts of Europe.

The long-term impact of this accident on human health is difficult to monitor. All residents within a 10 km radius of the plant were permanently evacuated. Residents from the nearby town of Pripyat, only a couple miles away and the only population center near the reactor, was bused away to locations unknown, but not before 24 hours after the accident, so there is no easy way to monitor cancer rates in them. Their exposures are unknown, but the iodine from the accident was said to be so thick in the air that it could be tasted.

A recent UN report has suggested that there is no scientific evidence of any significant radiation-related health effects to most people exposed to the Chernobyl disaster. There is a significant rise in thyroid cancer (a consequence of the radioactive iodine exposure), and the report points to some 1,800 cases of thyroid cancer. But "apart from this increase, there is no evidence of a major public health impact attributable to radiation exposure 14 years after the accident. There is no scientific evidence of increases in overall cancer incidence or mortality or in non-malignant disorders that could be related to radiation exposure." There is yet little evidence of any increase in leukemia, even among clean-up workers.

Some U.S. biologists have been studying the wildlife in what has become known as the "10km zone" around the reactor. Despite the high levels of radioactivity (which at the time of the accident were enough to kill trees in some areas), the wildlife populations today flourish. Indeed, they report that they have observed more wildlife in the 10km zone than in all other parts of the former U.S.S.R. that they have visited. While these observations should not be considered as evidence that radiation is harmless, they do point to the tremendous impact humans have on wildlife populations – no one is allowed to live inside the 10km zone these days.

# External Links

[Berkeley Lecture on Radioactivity](#)

[Use of Radioactive material in Makeup Test](#)

[Declassified U.S. Nuclear Test Film #55](#)

[Fukushima Radiation Not Safe!](#)

## CHAPTER 10: WHY ERROR IS UNAVOIDABLE

# 10

EVIDENCE

Data are models, and as such, are never perfect. But there are a few standard types of errors to watch out for.

# Why Data Are Important

Whether it be the advance warning of hurricanes, an impending military conflict, or customer response to a new product, we have no difficulty appreciating the importance of good data. Good data do not solve all problems, but they help in making decisions. Had the designers of the Titanic understood the ship's true limitations, they would likely have provisioned the ship with an adequate number of lifeboats. The crucial data were obtained on the ship's first - and last - voyage. And knowing that a hurricane is 24 hours from destroying your dwelling may not allow you to save the dwelling, but you can at least escape personal injury.

Data constitute our only link to the world we study - they are the best models of nature we have. As a consequence, they hold a supreme position in all applications of the scientific method. When predictive models are found to be at odds with good data, we keep the data and discard the predictive models. Any set of data is necessarily a model, with all the inherent limitations of models, but data comprise very special models. Ultimately, progress in science, business, and industry rests on the data.

Much of progress in science is simply the gathering of better and better data. Compared to even twenty years in the past, we live in an information age, and we now have access to data on countless phenomena. Satellites give us information about weather and climate, opinion polls indicate how the public thinks on all sorts of topics, journals and newspapers give us information about the latest products, various statistics on the economy and social events impact our optimism and pessimism about the future. The information we have today is more extensive, more detailed, and arrives faster than at any time in the past. We refer to the changes leading to this increase in information as technology, and the scientific method has been the basis for much of the improved technology we enjoy today. But improved technology does not guarantee better data. Improved technology does indeed enable us to gather better data, if we know how, but we can just as easily use improved technology to gather poor data.

The quality of data is important for a simple reason. As we noted in an earlier chapter, science does not prove models to be true. The ultimate reason behind this conclusion is that any set of data is consistent with many models. What determines the quality of a data set is the number of models that can be rejected by it. Poor data sets are consistent with many different models, hence are of little use.

There are several issues that concern data. The one addressed here is simply how to gather accurate data -- to get the measurements as exact as needed and to minimize the error associated with the data. When we address the topic of Evaluation in subsequent chapters, we address the further problem of gathering data that enables a researcher to test particular models. This latter dimension of data concerns the interpretation of data rather than their accuracy and is relevant to experimental design.

# Data As Models -- With the Usual Problems

Data represent a special type of model, one that is central to the scientific method. We use data to tell us about the phenomenon we are studying. Abstract models, such as theories and hypotheses are models that help us simplify nature. But data are our surrogates of reality.

Like all models, data are "false." No matter how hard we may try, the data will never exactly match what we think they represent. Instead of referring to data as being "false," however, we say that data are "measured with error." In this context, "error" does not imply blunder (as in baseball), rather it means "variation."

Another way to look at data is this. There is one fundamental issue that underlies data collection in all applications of the scientific method: if the data were to be gathered more than once (and by someone else), would they turn out the same each time? We say that data are measured with error to describe the extent that attempts to record the same phenomena differ. That is, any variation that causes our measurement of something to be inexact is error.

A universal goal when using the scientific method is to reduce the error/variation so that you know what the data represent (as closely as you need). This claim may seem to contradict the statement above that error is unavoidable. But, in fact, there are ways to reduce the error. Understanding how this error can arise is the first step in reducing it.

# Four Types of Error

When someone makes a claim that you find hard to believe or at least need to know whether you can trust it, there are three critical questions to ask:

1. What's the evidence? (what are the data?)
2. How was it obtained? (are the data any good?)
3. Who obtained it? (can the source be trusted?)

To some extent, all 3 questions pertain to this section of the course (“data”), although (3) is also addressed by the last part of the book. This chapter focuses specifically on (2), doing our best to ensure data quality.

The basic problem is that several things can “go wrong” with data. Another way of saying this is that there are different sources of error in data. Although error cannot be completely eliminated -- a single coin flip can only be 100% heads or tails, even though the probability of heads may be 50% -- there are safeguards and precautions that can reduce many types of errors. However, different types of error require different safeguards. Understanding the different types of error is thus the first step in understanding those precautions.

## **Rounding, Precision, and Accuracy Error**

Some kinds of measurements can never be made exactly, so we have to "round off" the value at some quantity that is less than exact. When a machine fails to provide a value beyond a fixed number of decimal places, we call it precision error. Consider the weight (mass) of a penny. To the nearest tenth, a penny is 2.5gm. To the nearest 0.01, it is 2.54. Using the finest balances, we could measure the mass to many more decimal places. But at some point, we would reach the limit of precision for our scale and thus be left with a rounded-off value. Or we would reach the point that the scale was no longer accurate enough to consistently give us the correct weight (accuracy error). We can never measure the mass exactly - not even to millions of decimal places, much less to an infinity of decimal places. In many branches of science, this type of error is specifically included in measurements by providing a measurement  $\pm$  (plus or minus) some smaller value, such as  $101 \pm 0.23$  meters. The number behind the  $\pm$  indicates the level of error that can be expected in the number preceding the  $\pm$ .

Precision and rounding error apply to many kinds of measurements - those in which we are not simply counting numbers of things: time, speed, weight, energy, volume, distance, and many others. For most non-technical applications of the scientific method, however, this kind of error is unimportant because we don't care about the value beyond a few decimal places. In economics, for example, a company is not likely concerned about the cost to produce an item to the fraction of a cent. And our monetary system forces each of us to accept rounding error because we cannot pay in fractions of a cent. Rounding error even applies to the estimation of percents and probabili-

## Sampling Error

*random deviation from an average*

Another source of error comes from sampling only some of the data in which we are interested. Consider again a coin toss. If the probability of heads was exactly  $1/2$ , and we tossed the coin 4 times, there is only a  $3/8$  chance that we would get 2 heads and 2 tails ( $1/8$  of the time we would obtain all heads or all tails). The reason is sampling error. As a second example, we might be interested in the percent student attendance in lecture. The average attendance might be 60%, but attendance on some days would certainly be higher than on other days. Again, we would attribute this variation to sampling error. In both cases, the data we gather in one trial would not generally match exactly the data we gathered in other trials. The issue here is not in our ability to count accurately -- we know how many heads and tails we got or how many people attended class. Rather, the error lies in the fact that what actually happens one time is not the same as what happens the next, even though the underlying rules or probabilities are the same.

Sampling error is a widespread phenomenon that is often ascribed to random "noise" and unmeasured variables. In the case of a coin toss, the outcome of the toss is usually attributed to random noise. In the case of student attendance, there would undoubtedly be reasons why each non-attending student missed class, but the reasons would be too diverse to measure and thus be attributed to unmeasured variation.

Sampling error is universal, although its importance may vary greatly from case to case. The way to reduce sampling error (discussed in the next chapter) is to make many observations and to obtain an average that swamps out most of the sampling error made in each observation. Sampling error is a big problem in studies of environmental hazards (e.g., cancer-causing agents), because only a low percentage of people develop any specific kind of health problem, so we need large samples to overcome the sampling error. For example, if we observe 1 excess case of cancer in one million people who eat bacon and 0 excess cases of cancer in people who avoid bacon, we can't infer that the cancer rates differ between the two groups because sampling error would give us this result 50% of the time if there was no difference between the groups. We would need a sample size about 10 times larger than this to overcome sampling error.

## **Technical and Human Error**

Our machines and our abilities to record data are not foolproof. Technicians handling hundreds of tubes, loading samples, and labeling samples can and do make mistakes. A common example occurs in televised football games, in which an official misreads a play and inappropriately assigns penalties. And a machine which has been calibrated wrong or whose calibration has drifted will also give erroneous data - the Hubble space telescope gave fuzzy pictures during the first few years of its operation due to faulty assembly.

Some machines and people are obviously less error-prone than others, and indeed, some technicians may never actually make any mistakes in their career. But there is always the possibility of error, and no amount of observations on any machine or person can show that a mistake is impossible (recalling our points about sampling error above).

While some instances of RPA, sampling error, and unintentional bias (next) may be errors caused by humans, our category of human and technical error is used here to describe errors that do not fall into those other categories.

## Unintentional Bias

Biases are consistent differences between the data gathered and what the data are thought to represent. In particular, bias is a tendency of the data to fall more on one side of the average than the other, which distinguishes it from sampling error. Whereas sampling error tends to balance itself out in the average as more observations are gathered, bias persists -- when data are biased, gathering bigger samples means that the average of the data is certain to differ from the expected average (or the true average). For example, opinion polls are often conducted over the telephone. Data gathered in these surveys do not represent people who lack telephones, and those data would be biased if people lacking phones had consistently different opinions than people with phones. Or consider the frequency of people carrying the AIDS virus. At this time, the frequency in the U.S. population is thought to be something like 1 in 200. But the frequency of people with this virus would be much higher in some groups than in others (prostitutes versus nuns, for example). The data for one subgroup would then be a biased model of the population at large; this bias would be important when calculating the chance of acquiring the virus from a sexual encounter with someone who had lots of other sexual partners, for example. Another, similar example comes from a kit once marketed to allow couples to "choose" the sex of their child. The kit involved instructions on the timing of sex as well as some buffers to supposedly influence the relative success of sperm with an X versus sperm with a Y chromosome (the chromosomes of the one lucky sperm indeed determines the sex of the child). The evidence in support of this method was a collection of letters from parents who wrote to the physician who developed the method, and a majority of letters reported success. It does not take much imagination to figure out that this sample of letters was likely biased -- parents whose baby was not the chosen sex were undoubtedly less inclined (and perhaps even reluctant to) write about their "failure." The FDA was not fooled, however, and the kit was withdrawn soon after it was marketed.

Unintentional bias is easy to confuse with sampling error. Remember that bias represents a deviation consistently to one side. As an analogy, think of sighting-in a rifle. If the rifle sights are mis-aligned, the average of the bullets will consistently lie to one side of the bull's-eye, no matter how many shots are fired. This is analogous to bias. Where-ever the sights are set, however, bullets will lie in a cluster around the average point of impact; this scatter around the average is akin to sampling error.

Biases may occur by mistake or deliberately. In this chapter, we restrict attention to accidental or unintentional bias. In a subsequent module, we deal with the problem of deliberate bias, as when people intentionally attempt to deceive.

# External Links

Responsible Conduct in Data Management

# CHAPTER 11: REDUCING THE ERROR, A TEMPLE FOR IDEAL DATA



EVIDENCE

You presumably wouldn't mistake a junker car for a Porsche. But can you tell junker data from Porsche data? The quality of the data has a huge impact on the conclusions you can draw.

# Introduction

Unless your job requires you to design experiments, you may rarely have the opportunity to gather data using the ideas we discuss in this chapter. Hence you may view this exercise as being pointless. But even if you don't use these methods, others that manipulate your behavior and opinion do. Advertising agencies design experiments using these rules to determine which of several ads is most effective in manipulating your behavior, and politicians use polls to adjust their positions. Furthermore, the ideas we discuss should be used (but often aren't used) to gather scientific evidence introduced in rape and murder trials, and to evaluate the safety and effectiveness of new drugs. In short, even if you never conduct an experiment using the principles we discuss, your life has been and will continue to be influenced by these principles. Because data are so important, your knowledge of the quality and limitations of data is also important.

There are five features that affect the accuracy of data:

THE IDEAL DATA TEMPLATE
REPLICATES
STANDARDS
RANDOM
BLIND
EXPLICIT PROCEDURES

Our goal in this chapter is to explain each of these features briefly in the context of an example. The goal in this example is the hypothetical one of assessing the fraction of the Texas A&M student body whose blood-alcohol levels exceed the legal intoxication threshold the night of the A&M-Texas football game (at midnight). In particular, we want to know how to avoid errors in making this measurement. Each of the four features will be shown to be relevant.

# Replicates

What are replicates?	Replicates are multiple observations taken under similar conditions.
Why use replicates?	They reduce sampling error, and reduce or allow detection of some human and technical error.
When to use replicates?	Any time variation is expected to arise from these 2 kinds of errors.

The most basic requirement of any data set is that the data be replicated -- doing things more than once. Replication includes any of the following-- taking the same measurement more than once, using more than one subject, using multiple groups, undertaking multiple studies.

Our hypothetical sampling of A&M students for intoxication levels should be based on a large number of students. If we surveyed only 10 students, then we could expect to be no closer than within 5% of the true value (precision error), and because of sampling error, we could be even further off. With 100 students, we could expect to get much closer to the true value, and 1000 students would get us still closer.

As hinted at above, replication applies to many aspects of a study. Consider a study to test the effectiveness of a medication. Replication can be achieved by enlisting several patients, but replication can also be achieved by testing different batches of the same medication, by performing the study at different times of the year and in different years, and so on. For reasons that are not always clear, results of two replicates of the entire study do not always agree, so multiple levels of replication are usually required before we fully accept any result. In our A&M example, we might get different results from year to year, depending on the outcome of the game and attitudes in College Station about student drinking. No matter how much a study is replicated, however, there are always countless ways in which it is not replicated.

# Standards

What are standards?	They are observations offering a known point of comparison for a measurement.
Why use standards?	To verify a measurement and thereby detect or avoid technical and human errors (e.g., to establish that a machine or person is working correctly).
When to use standards?	To verify a measurement; whenever there is any reasonable possibility of human or technical error.

The most basic requirement of any data set is that the data be replicated -- doing things more than once. Replication includes any of the following-- taking the same measurement more than once, using more than one subject, using multiple groups, undertaking multiple studies.

Our hypothetical sampling of A&M students for intoxication levels should be based on a large number of students. If we surveyed only 10 students, then we could expect to be no closer than within 5% of the true value (precision error), and because of sampling error, we could be even further off. With 100 students, we could expect to get much closer to the true value, and 1000 students would get us still closer.

As hinted at above, replication applies to many aspects of a study. Consider a study to test the effectiveness of a medication. Replication can be achieved by enlisting several patients, but replication can also be achieved by testing different batches of the same medication, by performing the study at different times of the year and in different years, and so on. For reasons that are not always clear, results of two replicates of the entire study do not always agree, so multiple levels of replication are usually required before we fully accept any result. In our A&M example, we might get different results from year to year, depending on the outcome of the game and attitudes in College Station about student drinking. No matter how much a study is replicated, however, there are always countless ways in which it is not replicated.

The standards needed to verify measurements taken in a variety of settings.

STANDARD	UNKNOWN
Weight of a known mass on a scale	Weight of any other object on the same scale
Thermometer reading of boiling water at sea level	Thermometer reading of other substances
The reading of a sober person on a breathalyzer test	The reading of a suspected drunk on the test

A **proficiency test** is a test involving standards. A proficiency test is simply the submission of known samples (standards) to an individual or agency that takes measurements. A proficiency test enables one to measure the error rate in data.

**Reference databases** also represent standards for a population. Strictly speaking, a reference database is a collection of known values from different individuals in a population. Reference databases are especially important when measuring characteristics that differ from person to person (e.g., fingerprints, hair samples, odors, blood types). The reference database enables you to know, for example, how common each characteristic is in the human population. For example, a reference database would tell you whether a particular DNA type, fingerprint, or hair type was rare or common. Reference databases are not typically used to detect human and technical error, however.

Standards are similar to controls (in the Evaluation section). The only difference is that a standard is a type of control used to verify that measurements are being taken accurately. When we introduce controls in the Evaluation section, we will indicate that we are using them to evaluate a model, such as whether a treatment is having an effect (does Y change if we change X). You may think of a standard as a control for data measurement.

# Randomization

What is randomization?	It is the process of making choices according to some random process.
Why use randomization?	It destroys unwanted associations in the data and thereby eliminates many kinds of bias.
When to use randomization?	Any time a choice is made between two or more equivalent options.

It would not be possible for a limited task force to sample all A&M students on the night in question. Choices would thus have to be made about which students would be tested (we'll assume that we have access to all of them, even if they go home or hide out in their dorm room). The only acceptable method, if we want accurate data, is to choose randomly - to literally use random numbers or flip a coin and base the choice on these random numbers. (Random is not the same as haphazard.) Other methods of choosing the sample risk the possibility of biases. For example, were we to sample just fraternity members, we would likely get a different result than if we sampled dorms or the library. There are lots of methods that may seem to be random (closing your eyes and choosing a name from a phone book), which are not truly random, so in this class, we consider something to be chosen randomly only when it is stated as being chosen by a coin flip, roll of a die, using a random number table, or drawn (blindly) from a hat.

# Blind Data

What is meant by blind?	It is the gathering of data when the subjects and/or observer do not know the treatment each individual received.
Why use blind methods?	It prevents certain kinds of biases.
When to use blind methods?	Blind observations should be taken when there is any possibility of subjectivity in gathering or interpreting the results; blind subjects should be used when they can influence the results by knowing the treatment they have received.

Blind designs may have several dimensions to them:

- **Blind observers** The person gathering the data is unaware of the treatment of each subject
- **Blind subjects** (applies only when the subjects are humans)
  - Subjects are unaware that an experiment is being conducted
  - Subjects are aware of the experiment but unaware of the group to which they have been assigned. In this case, a “placebo” is used to fool the subjects, so that no one is sure which group they belong to.

**Blind Observers:** Protocols employing blind observers are needed when there is a large element of subjectivity in gathering the data. They prevent the observer's preconceptions from influencing the data gathered. For example, if you wanted to determine whether children fed candy were more hyperactive than children fed apples, you would not want the observer to know which children were fed candy and which apples; it would be too easy to unintentionally over-interpret the activities of candy-fed children. Doing the experiment blind prevents the observer's preconceptions from influencing the data. In our hypothetical study of A&M students, we would want blind subjects (students should not know who was going to be tested, and even better, should not know that the study was being conducted). We would also want the choice of students for testing to be done blind (random is even better),

<b>BLIND DATA</b>
Testing new drugs. Half the patients in the experiment receive the new drug and the other half receive a placebo. The doctor evaluating the patients does not know which patients received the new drug and which the placebo. (blind observer)
Student evaluations of professors. The professor does not know which student wrote any particular evaluation. (blind observer)
<b>DATA NOT COLLECTED IN A BLIND MANNER</b>
Expert witnesses in criminal cases. An expert witness knows what those who have hired him (either the defense or prosecution) want to show.
Endangered species surveys. The Endangered Species Act sometimes requires a landowner to hire a private biological consultant to determine if there are endangered species on his land. Before doing the survey, the consultant in all likelihood will know what his client wants to hear.
Grading exams by hand. The person grading a paper knows who wrote it. For numerical answers this is not a major concern, but with the more subjective grading of essay questions, there is the possibility that the graders preconceived notions about the students abilities could influence the grade assigned.

These tables offer several examples of observations made blind and others lacking this feature. Blind studies prevent the researcher's preconceived notions from influencing the data collected.

## **Blind Subjects:**

One further variation to the concept of a blind experimental design using humans is that the patient is not informed of the treatment being received. To render a study blind in which medicines are tested, the subjects in the control group receive a placebo (inert pill). Although a placebo may seem like an unnecessary precaution, experience teaches us that it is not superfluous. Patient attitude has a major effect on the recovery period for some illnesses and surgeries, and a convincing body of data shows that patient who know they are in a control group often do not recover as quickly as patients unaware that they are in the control group.

In some cases, it is impossible to conceal the treatment from the subjects but it may be possible to conceal from them that they are part of an experiment. This kind of design would arise in studying people's responses to some kind of experience, such as something they read or felt. For example, we might want to test the effect of having students listen to 30 minutes of soothing music versus hard rock prior to an exam. There would be no way to conceal from the person the type of music they were exposed to, but it would be possible to conduct the study without telling them that an experiment was involved. In this type of blind design, the subjects are prevented from modifying their response in anticipation of the outcome of the study, since they are not aware that a study is being conducted.

## **Double Blind:**

Designs with blind observers and patients are known as double-blind. As noted, there are circumstances in which some of these features are not relevant to a design. How does one tell when a feature is relevant or irrelevant? The relevance depends on the goal of the study and the model being tested. Once those dimensions have been specified, it is possible to indicate whether each feature of ideal designs is relevant.

# Explicit Protocols (Explicit Procedures)

What is an explicit protocol? A protocol is a procedure -- ANY procedure or set of methods. An explicit protocol is simply a formalized procedure for gathering data -- a set of specific rules.

Why use one? An explicit protocol enables different observations to be taken uniformly and specifies which of the other features of ideal data will be applied.

When to use explicit protocols? In all data gathered for some important purpose.

What are the consequences of failing to use an explicit protocol? The data may be gathered inconsistently and be unreliable or unrepeatable.

In any serious attempt to gather data, a formal protocol should be used to specify how the data are to be gathered. The simplest step to take towards creating an explicit protocol is decide to record the data in a systematic fashion (often merely by writing down the observations). For example, consider how a reporter would record the events at a city league softball game versus the way a casual spectator would record the events. The reporter would record specific items such as final score, winning and losing pitchers, inning-by-inning history of hits, errors critical to the final outcome, and so forth. By contrast, the casual spectator would probably remember the final score and possibly the pitchers, but many other details would go unrecorded. The protocol will minimally indicate which of the preceding four features are present (blind, random, replication, standards).

Use of a written protocol has two effects. First, it enables subsequent data to be gathered in similar fashion. It is essential that all observations be repeatable (for they are otherwise useless), and an explicit protocol allows data to be gathered consistently from one time to the next. Second, an explicit protocol is itself a model that can be subjected to evaluation and improvement using the scientific method. That is, formalizing a protocol is the first step in improving it.

The field sobriety test that police officers give suspected drunk drivers is based on a protocol. It consists of having the suspect,

- walk heel-to-toe,
- extend their arms and then touch their nose,
- balance on one foot,

and so on. Certain types of data gathered using this protocol are consistent with the hypothesis that the driver is drunk (e.g., failing these simple coordination tests), while other data are not. The protocol makes it easy to compare data gathered by different police officers, and it ensures that all relevant data are gathered from each suspect. Additionally, the protocol makes a police officer's case against a suspected drunk driver stronger in court than it would otherwise be - the officer can cite specific tasks that the suspect was unable to accomplish, instead of testifying that the suspect "looked and acted drunk."

<b>PROTOCOL</b>	<b>MODEL FOR WHICH DATA ARE GATHERED</b>
Job application and interview	Applicant's skills and suitability
Takeoff checklist	Safety of the flight
Car repair manual	Car function
Balancing a checkbook	Current account balance

## More Complicated Protocols

Corporations use protocols to prevent fraud and similar activities, as well as to gather data about their financial condition. These protocols cover everything from how clerks operate the cash registers, to how the money is transported from the store to the bank, to more abstract problems such as how depreciation and good will are treated on the company's books. The care with which the accountants develop and implement their protocol determines the quality of the data the corporation has for making financial decisions.

The rules that determine what kind of evidence can be admitted in a criminal trial also constitute a protocol. Although this protocol may appear to be completely different than the corporation's protocol for gathering financial data, it is similar in that both protocols systematize the gathering of data. The rules governing admissibility of evidence allow certain types of data to be presented to the court but prohibit others:

- no hearsay evidence,
- no evidence gathered contrary to the U.S. constitution and laws,
- all witnesses are subject to cross-examination.

One may argue about the desirability of particular features of this protocol (indeed, a fair amount of political debate does). But at least everyone involved knows exactly what types of data are permitted, and how the data can and can't be gathered.

To appreciate the importance of explicit protocols, consider the day-care workers who were accused and convicted of sexually-abusing the children under their care (and are now being released). No records were kept of the psychological interviews with the children, so it was not possible to know whether the allegations came unsolicited from the children or were instead suggested by the psychologist conducting the interviews. More recently, it has been suggested that psychologists are capable of inspiring false memories in people about earlier events in their lives. Explicit protocols are thus obviously vital in evaluating these two different models of the source of the children's accusations.

### **Limitations of Explicit Protocols:**

A protocol is a model, and like all models, it has limitations. First, a protocol never contains all information about data gathering. That which is left out may be trivial or important, and the fact that a protocol contains lots of detail does not mean that all important features have been included. Second, it is usually impossible to follow any explicit protocol exactly, unless the protocol is worded so vaguely as to admit many different ways of gathering the data. Whenever reading a protocol to assess how data were gathered, there are two questions which should be asked in understanding the protocol's limitations:

- Was the protocol followed?
- What is omitted from the protocol?

---

# Protocols for Interpreting and Analyzing Data

The raw data are rarely presented. At the least, averages and standard errors are given. In DNA evidence, the lab may declare a match; they may even give the actual numbers obtained from a DNA sample, but the raw data in the database used for comparison are not presented.

More importantly, samples may be analyzed multiple times. Labs and investigators sometimes throw out data (sometimes with good reason), or they may downplay some data in preference to others. Every study has its unique points, and how the data are handled can have profound effects on the conclusion reached. In one of the early trials involving DNA evidence in the U.S. (New York vs. Castro), the lab declared a match between the victim and a blood spot found on the defendant's watch. Yet inspection of the raw data from the lab indicated that the two samples did NOT match (according to the lab's published protocols); there were many other cases in which their analysis of the data ignored discrepancies that should have caused them to reject a match.

The interpretation and analysis of data is thus an important step in the presentation of data. The explicit protocol should describe how the data were analyzed, as well as which data were omitted (and why). In some cases, we can make a clear distinction between errors in recording data versus errors in interpreting or analyzing data. For example, incorrectly reporting the result of a coin flip is clearly a mistake in gathering the data, whereas the discrepancy between the true proportion of heads and the observed proportion heads (in data properly recorded) is an error that affects interpretation.

---

# Adherence to Protocol has Become a Substitute for Data

The goals in specifying a protocol are (i) to minimize error, (ii) understand what types of errors may still be present, and (iii) allow others to gather data in a similar way. In the long run, an explicit protocol enables us to develop even better ways of gathering data (yet another realm of progress). In various bureaucratic and legal settings, however, the protocol assumes an even more important role: it becomes a surrogate for data quality. That is, a company, agency, or person is evaluated strictly on whether they are following the protocol, independently of the quality of data they produce. A recent audit of the FBI DNA crime lab, for example, was limited entirely to whether the proper documentation was being maintained (as specified by protocol). Thus, it did not matter if the quality of DNA typing was good or bad, only whether the lab was following procedure and filling out the requisite paperwork. As an analogy, you could imagine evaluating a company assembling computers. The company documents the fact that all parts get assembled correctly, but does not actually check on the quality of the parts it uses or whether the end product works the way it should. You can imagine just how “useful” such an evaluation procedure might be.

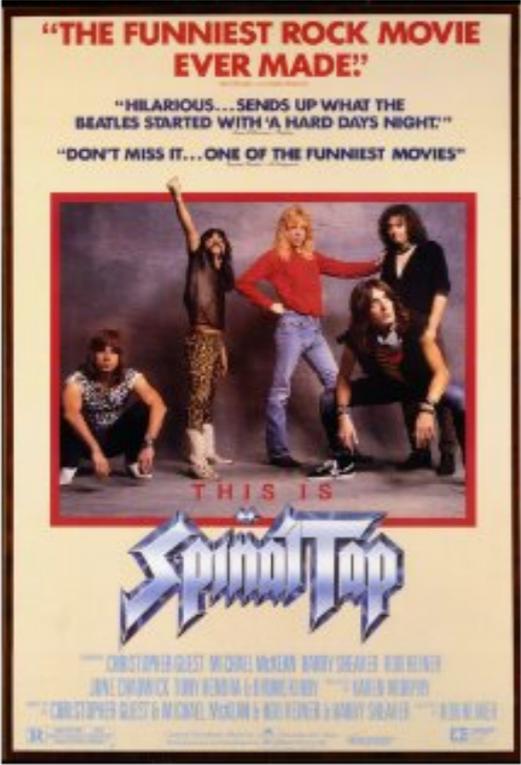
# Exercise

These 5 descriptions are merely introductions to complicated aspects of data gathering. Even in cases in which you do not understand the full complexities, you should be able to inspect the description of data gathering for the 5 elements in our template. Popular articles on topics that have been approached from a scientific perspective provide material that can be inspected for the presence or absence of these 5 features; in many cases, the descriptions of studies are not clear about certain features of data, in which case the information is ambiguous.

# External Links

Spinal Tap video

Spinal Tap IMDB rating goes up to 11



**"THE FUNNIEST ROCK MOVIE EVER MADE!"**  
"HILARIOUS... SENDS UP WHAT THE BEATLES STARTED WITH 'A HARD DAYS NIGHT'"  
"DON'T MISS IT... ONE OF THE FUNNIEST MOVIES"

**THIS IS Spinal Tap**

CHRISTOPHER GUEST MICHAEL MCKEAN BOBBY SHARER ROB REINER  
JANE CHARNOCK TONY RENDON GERRIE KIRBY CAROL BOOPPI  
CHRISTOPHER GUEST & MICHAEL MCKEAN & BOBBY SHARER & TONY RENDON & GERRIE KIRBY & CAROL BOOPPI

**This Is Spinal Tap (1984)**  [Top 5000](#)

**R** 82 min - [Comedy](#) | [Music](#) - [2 March 1984 \(USA\)](#)

**Your rating:** ★★★★★★★★★★ -/11

**8.0** Ratings: **8.0/11** from 68,689 users Metascore: 85/100  
Reviews: 300 user | 113 critic | 13 from Metacritic.com

Spinal Tap, the world's loudest band, is chronicled by hack documentarian Marty DeBergi on what proves to be a fateful tour.

**Director:** [Rob Reiner](#)

**Writers:** [Christopher Guest](#), [Michael McKean](#), and [2 more credits](#) »

**Stars:** [Rob Reiner](#), [Michael McKean](#) and [Christopher Guest](#) | [See full cast and crew](#)

# CHAPTER 12: DRUG AND DWI TESTING PROTOCOLS ENSURE DATA ACCURACY

# 12

EVIDENCE

The way in which data are gathered can affect whether you are wrongfully accused of crimes and can affect your success in challenging the evidence. This chapter considers the way in which drug tests should be and are required to

---

# Your Scientific Rights

The technology to identify criminal activity is finding increasing use in both the courtroom and the workplace. Whether the method be a breathalyzer test, a test for illicit drug use, DNA fingerprinting, or other, the average citizen (as juror, witness, or defendant) can easily be overwhelmed by the scientific, technological aura of these methods. Yet, technology does not guarantee accuracy, and the public should not place blind trust in the application of such technologies, especially when someone's liberty or life is at stake. As a society, we want the benefits of these new technologies, but at the same time, we want to make certain that they are applied in a fair manner.

Short of learning the technologies yourself, there are a few simple steps you can undertake to determine whether such technologies are being applied fairly. The law does not explicitly guarantee an individual's right to sound scientific practices when gathering evidence to convict them, but one's constitutional right to a fair trial can certainly be argued to include these rights. Any test used against you should have the five relevant features of ideal data. Unfortunately, although a good defense attorney might be able to insist on the criteria of ideal data in court, corporate use of scientific data to fire employees is not subject to such scrutiny: a drug test may be used as the basis of job termination without the employee's knowledge of the basis for dismissal. On the positive side, however, the Department of Transportation has instituted a lengthy and detailed protocol for lab-based drug-testing and alcohol-testing, which incorporates several features of ideal data. In what follows, we list the elements of what we regard as a fair drug test and note whether the features are included in the Dept. of Transportation rules (DOT rules). By and large, the DOT rules go a long way toward ensuring that ideal data criteria are met. The last part of the chapter deals with DWI data that are not part of lab testing under the DOT.

# A Fair Drug Test

The following sections consider each of our ideal data features, first describing our model of a fair drug test and then followed by the relevant features prescribed by the DOT. A summary of whether the DOT rules match the features of ideal data is offered below.

<b>GOAL: ASSESS WHETHER THE CONCENTRATION OF A DRUG EXCEEDS A THRESHOLD</b>		
<b>MODEL BEING TESTED: THE DRUG LEVEL IS BELOW THE THRESHOLD</b>		
<b>DATA FEATURE</b>	<b>RELEVANT</b>	<b>STATUS</b>
Explicit Protocol	Yes	Present
Replication	Yes	Present
Standards	Yes	Present
Randomization	Yes*	Present
Blind	Yes	Present

\* Randomization is relevant for certain civil rights issues but is not a major factor in data accuracy in this case

# Explicit Protocol

## *Ideally:*

The protocol for gathering, analyzing, recording and reporting the findings should be published before any drug tests are conducted. It should explicitly cover: 1) the procedures used to collect the specimens, 2) the criteria or methods used to decide which employees will be tested, 3) how the specimens will be handled between collection and being received by the testing laboratory, and 4) the experimental protocol employed by the testing laboratory, and 5) what information and specimens will be stored as a permanent record.

## *DOT rules:*

The Department of Transportation has developed a comprehensive protocol for its drug testing program. The details are presented in the Federal Register 49 CFR PART 40 and described in the 100+ page pamphlet "Guidelines for Implementing the FHWA anti-drug program" published by the DOT in 1992. These rules specify which records are to be maintained, how long samples must be kept, procedures for handling and transferring specimens, what kinds of tests are to be performed on which classes of employees, and many other features needed for a program of this nature. More important for our purposes, these rules specify a number of features that pertain to our five additional design features, as explained next.

# Standards

## *Ideally:*

Each batch of samples sent to the testing laboratory should contain several standards, some known to be non-positive and some known to be positive. The testing laboratory should not be told which samples are standards. The company should publicize the results of the standards.

## *DOT rules:*

Standards are included at two levels. The testing laboratory itself is required to include samples of known concentrations in each run. These standards enable them to calibrate the test procedure and thus to determine drug concentrations in the unknowns. As an additional standard, each manager whose employees are tested is required to include 3 known blank samples (lacking drugs) for every 100 unknowns; managers with more than 2000 drivers must also include some standards known to contain drugs. The rules do not specify publication of the results but they require that the DOT be notified of all "false positives," so the information should be accessible. Whether the data are, in fact, treated this way would be difficult to determine.

# Replicates

## *Ideally:*

Each sample should be divided into at least two tubes. One should be retained as a voucher for retesting, if necessary. The other should be sent to the testing laboratory. An even better design, however, would be to split the sample three ways, sending two for independent testing at different laboratories. After the tests are completed, the company should publicize what percent of the replicated samples came back with inconsistent results.

## *DOT rules:*

As of 1994, replication is required in the form of a "split sample" collection (at least for urine tests). Split sample collection simply involves the partitioning of the original sample into two vials. One vial is sent for testing, whereas the other is retained for retesting in the event of a positive result on the first vial. The rules specify that the retest is to be conducted by a different laboratory than did the first test. Retesting is not automatic, however: the person being tested must request the retest. In addition, the lab uses the original sample for multiple-stage testing - the specified procedure is that all samples are tested with an initial and rapid screen, and only the positive ones are analyzed more comprehensively. The rules specify that written records must be kept, but we are not aware that the results of replications must be published or made generally available.

# Random

## *Ideally:*

The decision about which employees are tested should be random. There may be special circumstances which invariably warrant testing so that no element of choice is involved (e.g., a driver involved in some kind of accident), but if a choice is to be made from a group of otherwise equivalent individuals, that choice should be random. Random choice here is not so much to reduce error but to avoid abuses.

## *DOT rules:*

One form of required testing is random. The rules specify that the choice in "random" testing must be made in a strictly random manner, with suggested methods being the use of random number tables or names on slips of paper drawn from a bin. The rules further specify that sampling with replacement must be used: an individual tested in the previous round is eligible to be chosen in the next round.

# Blind

## *Ideally:*

All samples should be labeled with a number, not a person's name, when sent to the testing laboratory. Standards should be indistinguishable from the unknowns so that the lab is not explicitly careful just when processing the standards. Additionally, the testing laboratory should receive no correspondence concerning which specimens the company suspects might be positive.

## *DOT rules:*

Blind testing is used at two levels. Most importantly, names are not included with the samples; the name of the person being tested is kept only on a single form (one of 3 copies) with the manager requesting the test. In addition, the standards are to be labeled in ways that make them appear as ordinary samples.

---

# Avoiding Deliberately Falsified Test Results

In science, the penalty for being caught falsifying evidence is complete and total ostracization from the scientific community. However, perhaps the major deterrent against fraud in academic is the high probability of being detected: any important discovery will be checked and verified by other scientists. Drug tests are different because there are no repeated cycles of evaluation and revision. That is, if your blood or urine tests positive by a fraudulent procedure, the authorities will not come to you later and take another sample to confirm the result. It is thus imperative that procedures be implemented to guard against falsification of test results. Of course, the use of replicates, standards, and blind procedures help in this respect. But the rules also further safeguard against falsifying results by requiring tamper-proof labels and requiring that all test results be sent to an independent 3rd party (the medical review officer) before being returned to the office originating the test. In general, the DOT rules are far more specific about matters of personal liberties and rights than is the typical procedure in science, largely because most matters of pure science do not impinge on people's rights and freedoms, and they can be verified at a later date.

---

# DWI Field Testing

Providing a blood sample for DWI testing should subject the analysis to the usual DOT rules because the sample is processed in a lab. However, providing breath samples or taking the SFST are done on site and don't necessarily have all those safeguards. At least in Austin, a breathalyzer test involves two types of standards, with known concentrations of alcohol and alcohol-free samples. Furthermore, the subject provides two breath samples (replication). So sampling error, human/technical error, and bias should be minimized; there is little opportunity for bias, so blind and randomization are not necessary.

A more problematic measurement is the score on the SFST. To ensure data accuracy, the SFST involves:

1. a formal protocol for scoring, including a point system
2. a formal protocol for giving instructions
3. replication in the form of 3 tests (one leg stand, walk & turn, horizontal gaze nystagmus), although failure on any one test (which requires two mistakes) is grounds for arrest
4. in many cases, a video tape of the driver's performance is taken, which provides a type of standard by enabling another observer (e.g., the court) to judge the performance.

When there is no video record of the test, the score will be almost impossible to challenge because there is no basis to evaluate the officer's scoring. This is a serious drawback of the test, although video cameras are now standard features of most patrol cars in this state. Perhaps the main drawbacks of the SFST when a video record is obtained are (i) the HGN test cannot be reliably interpreted on the video, and (ii) there are no baseline data for the SFST, either for the population at large or for the individual being tested. In the few rigorous studies of SFST performance, there are no data for our population that establish failure rates for sober people, which would be considerable, especially among older people. Such a database might work against using the SFST in some cases, because it is undoubtedly the case that thresholds for arrest are stringent. (I would speculate that a majority of people over 65 would fail it when sober because coordination declines with age.) Although baseline data for an individual could be provided at a later time, they never are. Nor would the performance of the arrested individual when sober be a blind measure of their ability, as there would be a strong incentive to perform poorly when establishing a baseline. However, as the HGN is involuntary, baseline data could be provided after the fact.

# External Links

[Nystagmus Field Sobriety Test](#)

[Spiders on Drugs](#)

[Ali-G Drugs](#)

## CHAPTER 13: DNA TYPING NOW AND BEFORE

# 13

EVIDENCE

Each of us is genetically unique, and there are many cases in which it is convenient to make use of our genetic individuality: for parentage analysis, identification of victims, and identification of criminals. DNA provides one of the most specific methods of "typing" a person, but many features of ideal data are being violated when evidence has been gathered for criminal prosecution.

# Motivation

Someone has committed a violent crime, and some blood was left at the crime scene. The blood type (presumed to be that of the assailant) is AB. The suspect also has blood type AB. Is this fact sufficient evidence of guilt? No, for several reasons. One reason is that AB is too common in the general population to warrant any conclusions as to the guilt of the suspect. If we have more information such as: the genotype of the blood at the scene is both AB and X1, and the suspect is both of these, it is now perhaps more likely that the suspect is the source of the crime-scene blood. However, it is necessary to know how common is the combination of AB and X1 in the population of possible assailants. If this combination is very common, then we still do not have much more information than we had with the AB blood type alone. We want enough information to know that the chance of finding a random person with that genotype is small. DNA gives us this specificity.

# What is DNA Typing?

DNA typing is a method in which our genetic material (DNA) is converted into a “barcode” that, ultimately distinguishes each of us from nearly everyone else on earth. DNA is easily recovered from many sources, so that criminals often unwittingly leave their DNA at crime scenes, and the DNA of victims is even sometimes carried away on the clothes of their assailants. By using DNA, we are thus often able to place individuals at crime scenes, and in the case of rape, are able to identify the man who “provided” the sperm.

Recent numbers. By 1990, DNA technology had been used in over 2000 court cases in the U.S., encompassing 49 states and Washington D.C. The October 12, 1991 Austin American Statesman reported that Williamson County's first use of DNA typing had just resulted in the conviction of a rape suspect, who was sentenced to 99 years in prison. Not all DNA typing has led to convictions, however, and the news nowadays more often reports the release of someone in prison (often having served more than 10 years) because DNA analysis of the old samples shows that he cannot have committed the crime. From any attempt to match a DNA fingerprint between suspect and forensic sample, three outcomes are possible. For the U.S. up to 1990, these outcomes (and their frequencies) were: (i) exclusion of the suspect (37%), (ii) inability to resolve the DNA fingerprint (20-25%), and declaration of a match (40%).

DNA typing is or can be used for many different crimes and circumstances: rape, assault and murder, body identification, and establishing parentage. It is also useful in conservation (establishing that meat came from an endangered species, for example, or that a set of antlers came from a deer poached on someone’s ranch). And a specific kind of DNA typing is used in molecular epidemiology, to identify the source of infectious agents.

# Sources of DNA

DNA occurs in all living cells of our bodies, with the exception of most red blood cells. However, because our blood also contains white blood cells, DNA can be obtained from blood samples, even tiny ones. DNA can also be obtained from saliva (saliva has cells from your mouth in it), so spitting on something or licking something allows your DNA to be typed. Hair has DNA. The root of the hair has cells adhering to it, which can be used to type a person, but the shaft of the hair has degenerate DNA which can be used for some DNA typing processes but not others. Skin has DNA, so touching an object can leave enough cells for DNA typing: someone who wielded a baseball bat in an assault was identified by the DNA left from holding the bat. Bone has DNA. Chances are that a criminal will not leave behind pieces of his bones, but bones are typically used to identify remains of bodies that may be even decades old. And last, even feces have your DNA – biologists often use “scat” samples to identify individuals (in studies of bears, for example).

DNA is extremely useful, but it does not last forever. Environmental samples can degrade, especially if wet. It is thus important for forensics labs to keep the samples frozen.

# The Typing Process

In the barely-20 years that DNA typing has existed, the technology of DNA typing has changed a great deal. Earlier versions of this chapter devoted many pages to the technology. We will abandon that emphasis on technology in this installment and restrict attention to the basics, as follows:

- i. Obtain a tissue sample and extract the DNA
- ii. “Xerox” the DNA with a technique known as PCR (polymerase chain reaction)
- iii. Determine the DNA type with either of two methods
  - a. **STR (short tandem repeat):** This method generates the typical DNA bar code and is based on variable regions of DNA from your chromosomes. Chromosome regions have been found that are highly variable among us (even the two chromosome sets you have differ in these regions). Typing involves the assessment of about 5 of these regions. The method measures the length of the DNA on your chromosomes at those regions but does not determine the actual sequence.
  - b. **Mitochondrial Sequence:** Each of your cells (except red blood cells again) contains hundreds to thousands of organelles known as mitochondria. Mitochondria ultimately evolved from bacteria, and they have their own miniature chromosome. Unlike the case with your (nuclear) chromosomes, all of the mitochondria in your body are inherited from your mother – so you have a single type. The sequence of your mitochondrial DNA matches that of your mother’s mitochondrial DNA and can be used as one form of a DNA type.

iv. Observe whether one sample has the same DNA type as the other sample (e.g., suspect). If so, calculate the odds of obtaining a match at random (as if the suspect had no association with the crime). The “random match probability (RMP) is typically less than 1 in a million for STR types, but it is much greater for mitochondrial DNA types (e.g., 1 in 100). For example, two full brothers (who aren’t identical twins) will certainly have different STR types, but they will also certainly have the same mitochondrial DNA sequence as each other, as their mother, and her mother, and as any other siblings, as any siblings of their mother, etc. Even so, mitochondrial types are often useful, and they can be determined from samples whose DNA is too degraded for STR determination (because mitochondrial DNA is present in so many more copies than is nuclear DNA).

# Errors

When DNA typing was a new technology, its introduction to the courts in the U.S. was hotly contested by some scientists. One objection was that the DNA typing process itself was not meeting ideal data criteria. Initially, there were NO rules for DNA labs, and there were no certification procedures. Databases for evaluating RMPs were inadequate. Many of the former problems have been resolved with database expansion and with technologies that removes the subjectivity in assigning DNA type to a sample, but problems still remain, at least for some labs. In summer of 2003, the Houston Crime Lab made the news by having such sloppy DNA procedures that even the local authorities recommended withdrawal of its certification. Dr. Larry Mueller's web page at U.C. Irvine (<http://darwin.bio.uci.edu/~mueller/> go to "Forensic DNA Resources" at the bottom of the left menu) lists some of the lab errors that he has encountered in his experiences as an expert witness for the defense. Another, more recent and comprehensive site is <http://www.scientific.org/>. Since most or all of these errors favored the prosecutions' cases until they were discovered, there is no incentive for the government to maintain a public record of them.

*The types of errors and problems most commonly encountered fall into a few types (sample mixups and bad data analyses are apparently the most prevalent):*

### **Sample Mixup:**

This is probably the most common source of false matches – the people in the lab mixed up the samples. Sample mixup is understandable simply because the technologies involve use of standardized tubes and other plasticware, and unless one is absolutely rigorous, it is very easy to accidentally grab the wrong tube, or load the wrong well with a sample. Ultimately, every sample is handled by a person before it gets processed, and this step of human handling is the vulnerable one.

### **Sample Contamination:**

Some cases of sample contamination are similar to sample mixup. In other cases, sample contamination occurs because an officer touches the material with his/her hands, or the contamination may occur when the sample is deposited (e.g., if a blood stain gets bacteria in it).

### **DNA Degradation:**

DNA degrades if it is not kept cold or dry. Thus, by the time the police arrive at a crime scene, the DNA in some of the samples may already be bad. Improper storage of samples also contributes to degradation. Degradation may lead to inaccurate DNA typing, though more so for the STR method than for the mitochondrial method.

### **Bad Data Analysis:**

The calculation of RMP may be straightforward in many cases, and some software automatically calculates it for each STR. However, unusual cases require a deep understanding of probabilities (and statistics), which is often lacking.

## Ideal Data: What's Missing?

Lab error rates are typically regarded as being around 2%, although the labs do what they can to conceal errors (as well as avoid them). If the RMP is as low as 1 in a million, a lab error rate of 2% dominates the considerations of the significance of a match, so labs need to be striving for vastly lower error rates than they have had in the past. As outsiders, it is difficult to know what all the causes of these errors are, but we can get an idea from past exposures of these errors. A big unknown is the extent to which a lab actually follows its own protocols. The written protocol is only a model of what is done, and if the technicians deviate from the written protocol, it is difficult to uncover that after the fact.

1. Absence of external, blind proficiency tests (inadequate standards). The only way a lab can begin to correct its mistakes is to know how often and why they occur. Blind proficiency tests are the surest way to know the lab's error rate. Few labs submit to external, blind proficiency tests, though all labs now submit to some form of proficiency testing. (A blind test means that the lab does not realize they are being tested on the sample; a blind test is good because it means that the technicians are being no more careful in testing that sample than in testing any other sample.)
2. Sample identification is known when processing occurs (bad protocol: absence of blind). By knowing which samples belong to which people (or crimes), it is far easier to unintentionally produce a false match (perhaps by sample mixup or contamination).
3. Samples from the same crime are often processed together, in the same lab (bad protocol). This greatly increases the chance of sample mixup going undetected.
4. Inadequate replication (bad protocol). With the use of PCR, a single sample can be processed many times (which was not true of past methods). Ideally, samples should be split and sent to different labs for testing, which would greatly reduce sample mixups going undetected. Cost is probably the biggest impediment to this kind of replication.
5. Bad protocols for data analysis. People analyzing DNA data have not usually been trained adequately for assessing the true RMP. It is thus common for the RMP to be miscalculated (and the error may go in favor of or against the defendant).

# External Links

DNA Fingerprinting:

[Example 1](#)

[Example 2](#)

[Example 3](#)

## CHAPTER 14: SCIENCE AND THE CRIMINAL JUSTICE SYSTEM

# 14

EVIDENCE

There has been a landslide of problems exposed in the criminal justice system that stem from faulty science being used to convict the innocent. There are too many examples to dismiss them as exceptions, or as technical problems that apply only to specific methods. Instead, there are basic, fundamental problems in the way that scientific data are being gathered, used, and presented in our courts. It is a crisis. It means that you can live a righteous life, abiding by the law, and yet become a victim of a faulty prosecution. It also means that the real criminals are going free, to commit more crimes and spread the word that they can get away with it.

# Introduction

On paper, the foundation of the US criminal justice system appears to be a triumph for the innocent. Concepts such as “innocent until proven guilty,” “a jury of peers,” and “the right to hear accusers” are safeguards to protect us, with the roots of these concepts stemming from the very inception of this country. Those of us who have never had a brush with the criminal court system can be lulled into a sense that it is functioning well. Sure, we all know that money can be important in getting a good defense, but the premise is that the innocent are rarely faced with having to defend themselves (an ironic reversal of our supposed “innocent until proven guilty” principle). Yet, for those of us with that perception, there has been a shocking embarrassment of wrongful convictions and of abuses and misuses of science that have come to light recently. Most of the information about wrongful convictions and the causes thereof has been revealed by the Innocence Project, but there are many other sources as well.

In the last decade or so:

- 1) Over 258 people in US prisons have been released after new evidence showed that they could not have committed the crime (more than 80 of these were from death row, of the approximately 6,000 people sent to death row since 1976). The most common cause of mis-conviction was mistaken identification by eyewitnesses.
- 2) Hair-matching methods, often used in court to establish that a single hair came from a particular person, have been declared nonsense. In 2003, Canada not only abandoned use of (non-DNA) hair matching methods, but began reviewing convictions that used hair matching, to see if the convictions should be overturned..
- 3) Fingerprints fell from grace. Long considered the icon of personal identification, fingerprint experts were finally subjected to proficiency tests in the mid-late 1990s and found to make false matches 10%-20% of the time.
- 4) Polygraph tests, widely used by the government though no longer allowed in many courts, were declared nonsense by a panel of the National Research Council
- 5) In a flurry of public attention in 2003, the Houston Crime lab lost its accreditation for DNA typing because of sloppy procedure

The list goes on, but these are some of the major ones. We owe many of these new revelations to DNA typing, because DNA typing exposed many of the mis-convictions, and also because it set high standards for the use of scientific data in court. As noted in the preceding (DNA) chapter, DNA typing was initially pursued enthusiastically by prosecutions but challenged by a group of scientists who felt (with some justification) that it was not being used correctly. No one on either side of the argument in those early days seemed to foresee the huge impact DNA typing would have in exposing the long history of bad science used in criminal courts.

Before delving into specific examples, we can summarize the main problems in the context of our ideal data template:

1. **Failure to gather and analyze data blindly:** This is probably the most pernicious violation of ideal data. Prosecution agencies and their consultants (labs) know whether there is a prime suspect, who it is, and know the identities of the samples being analyzed. As evidence begins to form around a suspect, it allows the entire process to continually reinforce the apparent guilt of that suspect by biasing the gathering and selection of evidence toward that person and away from others. There are many documented cases of this bias and much of it stems from a lack of blind protocols. This bias is difficult to correct, however, because many aspects of prosecutorial duties cannot be done blindly.

Since the routine use of DNA testing was implemented, 25% of the time the prosecution's prime suspect has been cleared by DNA before trial. This means that the prosecution's initial stage of gathering evidence led them to the wrong person. Because of the biases built into the prosecutorial practices, many of these 25% would have been convicted if DNA typing had not been available.

2. **Bad standards:** These problems include (i) failures to conduct blind proficiency tests of the labs, and (ii) inadequate (sometimes non-existent) reference databases.
3. **Bad protocols:** In some cases, methods have been used widely despite the lack of protocols for analyzing the data. In other cases, protocols for gathering the data have been inadequate for protecting against human and technical error. but challenged by a group of scientists who felt (with some justification) that it was not being used correctly. No one on either side of the argument in those early days seemed to foresee the huge impact DNA typing would have in exposing the long history of bad science used in criminal courts.

# Ideal Identification Methods

Critical to many if not most criminal trial is some kind of (physical) evidence linking the suspect to a crime or crime scene. This evidence may consist of DNA, hair, eyewitness accounts, fingerprints, shoe prints, and so on. There are some features of any identification method that render it suitable for scientific inquiry:

Template of an ideal identification method:

FEATURE	WHY NEEDED	ERROR RED'N PRINCIPLE	FLAGS- INDICATORS OF ABSENCE
1) reference database	gives the population frequencies of the different characteristics, thus knowing the RMP	allows determination of sampling error when a match occurs	not mentioned or claims that match is unique
2) characteristics measured are discrete	it is clear whether a person has it or not, allowing consistent scoring	no RPA error	no description of specific characteristics
3) independent verification a) universal protocol b) characters permanent	someone else can challenge the conclusions	= replication, to detect many types of error	methods of one expert cannot be evaluated by another; no explicit protocol; characters being measured are not permanent
4) labs subjected to blind proficiency tests	provides assurance of the accuracy of the methods	= standards to estimate overall error rate	no error rate given; tests internal, not blind, undocumented

We now consider some specific examples of forensic methods that have been used to identify people over the last 50 years in U.S. courts. The following table summarizes how well they fit the ideal template (and whether the method has been discredited):

Summary of Identification methods and characteristics:

<b>METHOD</b>	<b>(1) REFERENCE DATABASE</b>	<b>(2) CHARACTERISTICS MEASURED ARE DISCRETE</b>	<b>(3) INDEPENDENT VERIFICATION POSSIBLE</b>	<b>(4) LABS SUBJECTED TO BLIND PROFICIENCY TESTS</b>	<b>DISCREDITED</b>
DNA	+	+	+	+	No
fingerprints pre-1990	+	-	-	-	Yes
fingerprints post 2000	+	+	+	+	No
hair matching					Yes
bite marks	-	-	-	-	In some cases
shoe print ID	-	-	-	-	Yes
bullet lead	?	-	-	-	Yes
dog sniffing	-	-	-	-	Yes?
eyewitness	-	-	-	-	

# Detailed Discussion of Methods

## **Hair Matching: Bad Protocols, Bad Standards**

Once the most trusted method of identification in forensics, fingerprint matching has been shown to have major problems.

The year 1911 first successful introduction of fingerprint evidence in US court, but not as the sole evidence. Thirty years later, a legal precedent was established for convictions based on fingerprint evidence alone. Although the uniqueness of a person's fingerprints was originally established for the complete set of fingerprints from all 10 digits, somewhere around this time or later, there was acceptance of the general assertion that a single fingerprint was also unique and could be used to establish identity.

Surprisingly, the main international association of fingerprint experts (which, incidentally, consisted mostly of US experts) resisted the establishment of criteria for demonstrating a match into the 1990s. That is, they refused to accept an analytical protocol for declaring a match. They instead proclaimed that each decision about a match was to be made on a case-by-case basis and should be left up to the expert reviewing the case. (There was disagreement about this point between the two main fingerprint organizations, and the British adopted a minimal set of criteria for declaring a match.)

In the years 1995-8, there were 4 voluntary proficiency tests offered to different fingerprint labs. These involved multiple fingerprint comparisons. Not all labs responded, but for those that did respond, the false positive error rate of labs was at least a few percent and was as high as 22 percent!

## **Hair Matching: Bad Protocols, Bad Standards**

Hair matching was put to rest in 2003, both in the U.S. and Canada. In hindsight, there were many major problems with it that should have kept it from ever seeing the light of day. Specifically, there were

1. no data banks for hairs
2. no way of coding hairs
3. no protocols for analysis.

It should thus not be surprising that a full 18 of 62 wrongful convictions listed by the Innocence Project involved hair matching. What is astonishing is that hair matching was used for so long: proficiency tests from the early 1970s had found error rates of 28%-68% when labs were asked to match hairs, and different labs made different mistakes (as expected if there is no uniform protocol for doing the matches).

### **Dog Sniffing Identification:**

Effectively a method lacking in protocols. There is no way to know what method a dog is using for odor identification and matching. Tests using trial dogs have found that dogs are not very good at matching odors from different parts of the same person (e.g., hand versus neck).

### **Polygraph: Bad Protocols, Bad Standards**

A report released 8 October, 2002 by the National Academy of Sciences described polygraph testing as little more than junk science. Although a 1988 federal law banned the use of such tests for employment screening in most private businesses, and polygraph data had been inadmissible in nearly all state courts, the method has been widely used in government agencies concerned with national security. There was a time when polygraph data were used in court, and the polygraph has been used for unofficial purposes in criminal investigations to help prosecutors decide who to rule out as suspects. Thus, the fact that it has been inadmissible in court has not prevented it from assuming an important role in criminal investigations.

### **Interviews With Suspects: Bad Protocols**

Interviews have commonly not been videotaped or transcribed, so accounts of what was said have been based on recollections; the conduct of interviews has also been variable. Nonetheless, law enforcement officials often use their “recollections” of what a suspect said during an interview. Claims of confessions or incriminating statements may thus have been in error. Information may also have been passed to the suspect that was then used to indicate intimate knowledge of the crime. The use of physical force during interviews was banned by the Supreme Court only by 1936.

## **Eyewitness Identification: Not Blind, Bad Protocols, Bad Standards.**

This is the most baffling of all evidence used in courts. Eyewitness identification of a suspect is the most powerful evidence there is for swaying a jury. And it is among the most fallible of all evidence: 52 of 62 wrongful convictions tabulated by the Innocence Project involved mistaken ID, the most common error attributed to a wrongful conviction.

It has been known for over a century that eyewitness accounts are less than perfect. A 1902 experiment conducted in class (involving a gun – not something we'd do now) revealed that the best witnesses were wrong on 26% of important details; the worst had an 80% error rate. In a more recent California State University experiment with a staged attack on the professor, only 40% of the class later identified the attacker, and 25% attributed the attack to a bystander.

Errors by eyewitness testimony in court have been documented for decades, and some of them are profound. There have been several cases in which half a dozen or more eyewitnesses identified the same person, and it was the wrong person. In many cases, there is not even a close resemblance between the right and the wrong person.

How can this happen? How can many people independently arrive at the same wrong identification?

Let's start with a single eyewitness. Some psychologists use a 3-sequence model of memory:

1. **acquisition** – the events are recorded in your brain
2. **retention** – the acquired events remaining in your brain are lost at some rate
3. **retrieval** – the events are recalled by you

The acquisition phase is known to be imperfect – you never record all aspects of an event. That is, your memory starts out like a photograph with gaps. The more distracted or stressed you are at the time, the less you acquire. Thus, a person being raped or facing a gun/knife will have a much faultier acquisition than a person facing a non-threatening situation.

The retention phase is also imperfect. However, not only can you lose the memory of an event you had once acquired, you can also add events that never happened. Your memory is dynamic, and you are constantly building it, often filling in the holes. This rebuilding of the memory is where problems arise with eyewitnesses. In particular, your memory is very prone to subtle suggestions which eventually bias what you remember. Here are two problems that confound witness identification.

- Subtle hints can influence a witness to choose a particular suspect, even if that suspect is not the right one. These hints can be as innocuous as the police merely asking the witness to “take a careful look at #3” for example. Anything that makes one suspect stand out from the others can be a subtle influence on the witness (the way a picture of a suspect is taken, what they are wearing, etc.). Bad protocols and an absence of blind testing contribute to this bias. (The absence of blind exists if the police are influencing the witness and know who is their preferred suspect). The witness should not experience any outside factors that will push their choice toward a particular suspect.

- Familiarity transfers from one setting to another. If an eyewitness has had some previous exposure to a suspect in a setting unrelated to the crime, it is common for that familiarity to be transferred over in the witness's mind to the new context. In one case, the witness mistakenly identified a man who lived on her block (but whom she had only seen at a distance on a couple of occasions). In an experiment listed above, 25% of witnesses chose the person who had been a bystander to the crime – no doubt because that person was familiar to them, even though not committed the act. Another effect of familiarity is also problematic. Once a witness has been exposed to a set of suspects, any subsequent exposure to one of the suspects reinforces the witness's memory of THAT suspect. For example, if a witness is shown two lineups (at different times), and one suspect is common to both lineups, that suspect is likely to be chosen because of the familiarity. This problem is difficult to eliminate completely, because the familiarity may have been obtained before the witness saw the crime, so police procedure could not, in that case, prevent it.

These are the problems that one witness can experience, even without knowing it. But how can several people all make the same mistake? When several people all make the same mistake, it is a clear indication that police protocols are bad – it generally means that the police are influencing witnesses, or witnesses are influencing each other.

---

## Lab Tests: Lack of Blind

Although some of the specific examples mentioned above stem from lab errors, there are generic problems that apply to some labs, independent of the method of analysis. Perhaps the most widespread problem is a lack of blind analysis. Knowing which samples belong to which people allows fabrication of evidence (of which there have been many cases (Fred Zain's string of fraud resulted in so many convictions that prosecutors sought him out, as did the news program "60 minutes" when he was exposed). For honest technicians, the absence of blind encourages honest mistakes and selective replication of results (you repeat the test if the results don't fit your preconceived ideas about guilt). And as revealed in the Castro Case (early DNA testing), the lack of blind causes people to over interpret results and make them fit preconceived notions.

As an example of the absence of blind analysis by labs, here are two letters sent from the Chicago Police Department to the FBI, requesting DNA typing. All names are omitted from our text; where names were included in the letters, a description is given in square brackets [].

Letter 1: From Chicago Police Crime Lab to F.B.I. DNA Laboratory Division, 10 August, 1989

Dear [name of Director of F.B.I. lab],

I am writing to request that DNA typing be performed on several items of serological evidence. The names of the people involved are: [name of female victim] F/W (the victim) and [name of male suspect] M/B (the suspect). The evidence I am sending you consists of the following:

- Blood standard from [name of victim]
- Blood standard from [name of suspect]
- Extract from swab
- Extract from victim's pants
- Extract from victim's bra

All three of these extracts were found to be semen/spermatozoa positive and the two extracts from the clothing were found to have ABO, PGM and PEP A activity consistent with that of the suspect. I am also enclosing a copy of my laboratory report stating these results.

The facts of the case are that on 25 May 1989, the victim was grabbed from behind, pulled into the woods and sexually assaulted. The victim never got a good look at her offender and therefore is not able to make a positive I.D. of the suspect. The suspect [name] had just been released from the ILLINOIS DEPARTMENT OF CORRECTIONS after serving time for the same type of crime in the same area. At this time the suspect has not been charged.

Thank you very much for your assistance in this matter. Please feel free to contact me if you need more information.

Letter 2: From Chief of Detective Division, Chicago Dept. of Police to F.B.I. DNA lab

Dear [name, Commanding Officer, F.B.I. DNA lab],

In early January, 1990, detectives assigned to the Chicago Police Department's Detective Division, Area Three Violent Crimes Unit were assigned to investigate the particularly brutal Aggravated Criminal Sexual Assault, Robbery and Kidnapping of one [name of victim], recorded under Chicago Police Department Records Division Number N-005025. On January 10, 1990, one [name of suspect] M/N/ 31 years, FBI [#], C.P.D. Record Number [#], was arrested and charged with this and other offenses.

Blood and saliva samples of the offender and victim were obtained and tendered to Technician [name of technician] of the C.P.D. Criminalistics Unit. A sexual assault kit (Vitullo Kit) was also completed and submitted for the victim.

The undersigned requests that the recovered specimens and evidence be evaluated and subjected to DNA comparison testing. Although the offender has been identified and charged, we feel this comparison would greatly enhance the prosecution of [name of suspect], who was arrested after a week long crime spree.

If any additional information is needed, kindly contact Detective [name], star [#], Area Three Violent Crimes Unit, 3900 South California, Chicago, Illinois 60632, Telephone #(312)-744-8280, or the office of the undersigned.

Sincerely,  
[name]

# Combining Different Sources of Error

A declared match between a suspect and a forensic sample may not be real for several reasons. First, the RMP (random match probability) indicates how possible it is that the match is coincidence. But the RMP calculation assumes that the suspect and sample do indeed match. The match could be erroneous because of any number of human and technical errors in the process of gathering, labeling, and testing the samples. There are thus several reasons that the sample may not have come from the suspect despite the declared match.

A lot of ink and words have been exchanged over the best way to calculate the RMP in DNA typing. For the most informative DNA typing method (STR), the RMPs are typically 1 in a million (much larger with mitochondrial DNA profiles and Y-STR profiles). With larger and larger reference databases, the uncertainty in those calculations has gone down. But the impressively low RMPs are not the whole story. Something realized over a decade ago by J. Koehler (then at U. Texas) is that labs make mistakes, and the lab error rate (LER) needs to be factored into the probability of an erroneous or spurious match.

From the forensic perspective, we want to know the chance that the suspect is not the source of the sample. For the match to be 'real,' the lab and forensic team must not have erred in processing AND the match must not be due to chance. In the simplest case (LER independent of RMP), the match is 'real' with probability.

$$(1.0 - \text{LER}) * (1 - \text{RMP}) = 1 - \text{LER} - \text{RMP} + \text{RMP} * \text{LER}. \text{ (match is real)}$$

If both LER and RMP are small (e.g., less than 0.1), the probability that the match is not real is very nearly

$$\text{LER} + \text{RMP} \text{ (match is not real)}$$

In other words, you add the different possible causes of a spurious match to get the overall rate that the match is not real.

The implications are profound when you consider the history of argument over RMP calculations. Most labs do not divulge their error rates (nor subject themselves to blind proficiency tests), but Koehler's estimate of error rates was around 2%. Some labs are certain to be better than others, but even if the error rate is 0.2%, this value is large enough to render the RMP meaningless in comparison, at least when the RMP is 1 in 10,000 or less. So the emphasis should be on lab error rates (and reducing them) rather than on vanishingly small RMPs.

# External Links

[Eyewitness Identification - Getting it Right](#)

[The True Story Behind "Conviction"](#)

[Innocence Project Event - The Wrongfully Convicted](#)

[Confession Contamination](#)

## CHAPTER 15: DATA PRESENTATION

# 15

EVIDENCE

The way data are presented can have a big influence on your interpretation.

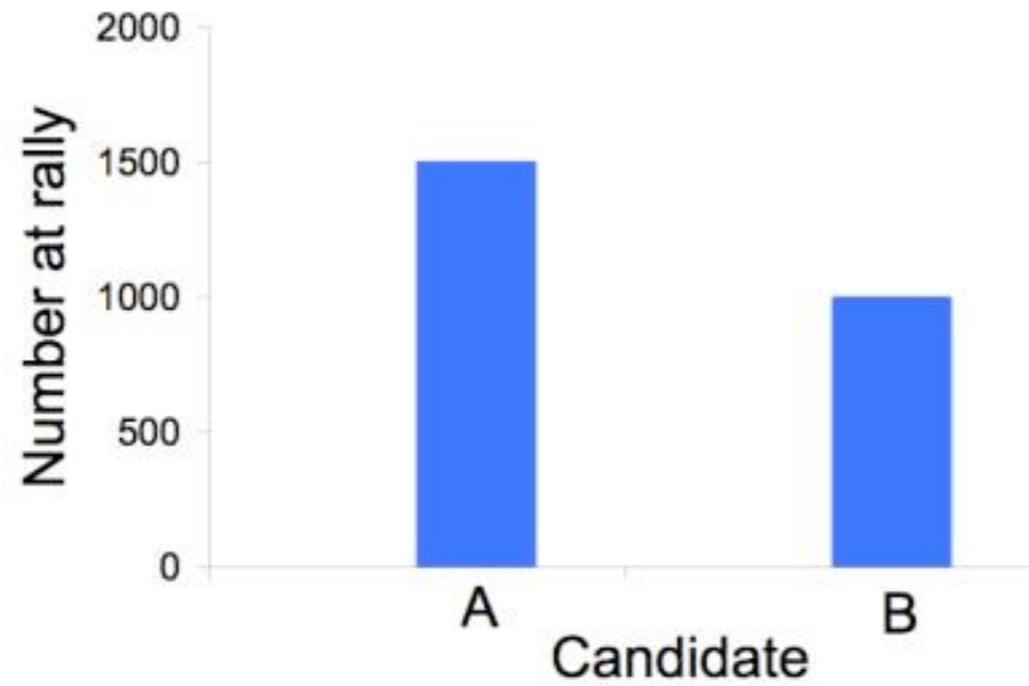
---

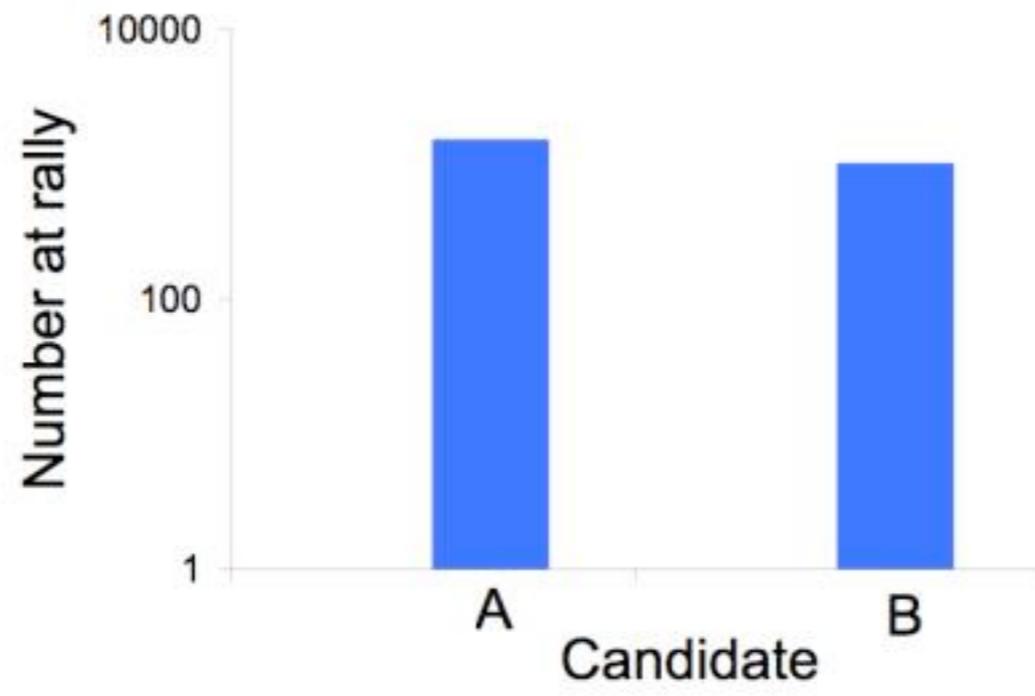
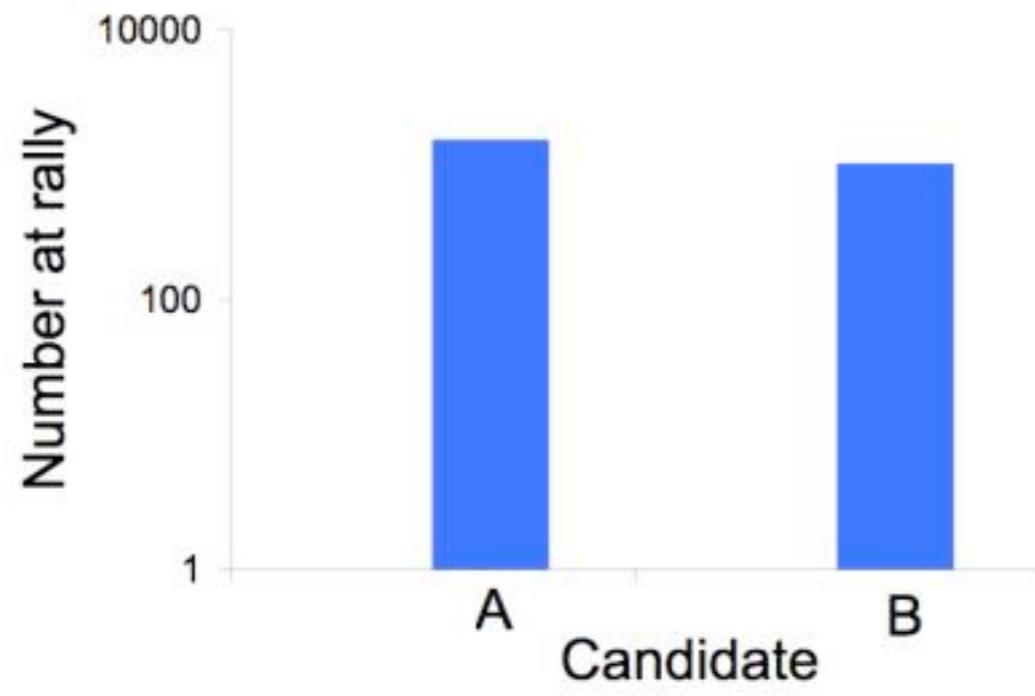
# Lots of Ways to Show Something

There are usually countless ways of presenting the same data. You are already familiar with the difference between raw numbers, tables, and graphs. Microsoft Excel gives you many choices of graph (chart) types, and if you have ever attempted to make a graph in Excel, you can probably appreciate the immense variety of ways there are to show something (what kind of chart, what colors to use, how wide to make lines and bars, where to put the text, ...), and how easy it is to make it more complicated than you want.

Data presentation is not just about making something pretty. Choice of one presentation format over another can influence how people respond to or evaluate the data. You will encounter examples of this many times in your lives (and no doubt already have), as the method of data presentation is often a deliberate attempt to influence you. Furthermore, how you interpret the data could make a difference to you. People who were around in the 1960s may remember a Peace Corps commercial with a glass that was half full of milk. The ad pointed out that you could interpret the glass as ‘half full’ or ‘half empty.’ A current but disturbing example of this same issue is seen in the CDC (Center for Disease Control) web page on condom efficacy against sexually transmitted disease. In the past, the web page emphasized that proper condom use had been shown to greatly reduce the transmission of HIV. Now, perhaps under pressure from an Executive Administration that wants to discourage promiscuity, the web page emphasizes that HIV transmission can occur even when condoms are properly used. Both views are correct – condoms have been shown to drastically cut HIV transmission, but some transmissions have occurred among ‘consistent’ condom users.

The scale of numbers used to illustrate a point can have a big effect on perception, as in the figure below. All three figures show the same invented data about the number of attendants at a rally for candidate A (1500) and for candidate B (1000). The most legitimate representation is in the top figure. The bottom figure appears to inflate the excess for candidate 2 by starting the vertical axis at 500 instead of 0. The middle figure appears to deflate the excess by using a log scale.





# How to market a new drug or medical procedure: relative versus absolute benefits

The marketing of new drugs and medical procedures (mammography, certain cancer treatments) has reached unprecedented levels. It is not uncommon to hear that a cholesterol-lowering drug (known as a statin) reduced mortality by 20%, or that some other medical offering improves outcomes by an appreciable percentage.

A percentage is a relative measure. The alternative is an absolute measure.

To fully appreciate the difference, let's invent some numbers and present them both ways. Suppose we tell you that a new drug – which may have some side effects – reduces death from heart attacks by 50% for people in your parents' age group over 5 years. How strongly would you encourage your parents to go on this drug? Many of you would no doubt be enthusiastic. It sounds too good to be true.

Now let's present the same data as absolute rates. For 1000 people in your parents' age group that do not take the drug, there are 4 heart attack deaths every 5 years; if they are on the drug, the number of heart attack deaths would be 2. Thus, the drug will save one heart attack death in 5 years for every 500 people taking it. Is your level of enthusiasm still the same? How about if we told you that the number of total deaths was not statistically different between those taking the drug and those not taking it?

Most medical data are now given to the public in relative numbers. It may be no coincidence that those data presentations often ultimately come from those who benefit from having you buy the drug. The number '50%' is intrinsically impressive, because it makes us automatically think of 'half.' But in many cases the 50% applies to only a very small fraction of the total. It's not the same as telling you that 50% of you will improve your grades if you buy my course packet. It's more like telling you that, for those of you in the lowest 2% of the class, 50% of you will improve your grades by purchasing my packet.

# Diagnostic tests: don't be alarmed (or overconfident) at initial test results

There is now an impressive and bewildering array of tests that can be run on us: many types of cancer screening, STD tests, blood cell counts, and drug tests; there are also lots of enzymes that can be measured as some indicator of metabolic health. How common are mistakes? Or more importantly, if you get an unexpected result (e.g., a positive HIV test), what is the chance that it is wrong? A test result that is erroneously positive is known as a false positive.

The way that numbers are typically presented by medical practitioners is so confusing that almost no one seems to know how to answer this question, even the practitioners themselves (a book 'Calculated Risks' by G. Gigerenzer makes this point in detail). For instance, consider the following numbers for a fictitious viral infection

1%	fraction of population infected
99%	fraction of population not infected
90%	how often an infected person tests positive
2%	how often an uninfected person tests positive

Can you convert these numbers into the probability that a positive test is in error? Not many people can. It goes like this:

$$\begin{aligned}
 \text{Prob. of false positive} &= (\text{all positive tests from uninfected people}) / (\text{all positive tests}) \\
 &= (0.99 \times .02) / (0.99 \times .02 + 0.01 \times 0.9) \\
 &= .06875 \approx 2/3
 \end{aligned}$$

Simple, eh?

It is possible to present the data in a more intuitive way. The probabilities presented above can be given as numbers of people in different categories. For example, if we imagine testing 100 people, then:

- 1 is infected and has a 90% chance of testing positive
- the other 99 are not infected, but nearly 2 of them will test positive
- It is now fairly simple to see that, out of 100 people, there will be nearly 3 positive tests, and 2 of them will be from uninfected people.

Studies have been done in which the two methods of data presentation were given to different subjects. The ones who were taught the second method had a far better chance of making the correct calculation than were the subjects taught the first method.

Why might this matter? Medical personnel are the ones handing out test results. Most of them are apparently unaware of how to calculate these probabilities. They have unknowingly told people with positive HIV tests that their test results indicate infection with virtual certainty. People have been known to kill themselves or engage in risky behavior upon receiving positive results, and in some cases at least, the results were erroneous. Had they realized how likely it is that the results were wrong, a great deal of their angst could have been avoided.

The point here is that there is a simple way to present the data that can be grasped by many people. In contrast to the use of relative versus absolute numbers, here the misunderstanding created by presenting numbers the 'hard' way seems to benefit no one.

# The Scale Used to Provide Choices

An interesting phenomenon is observed in the way choices are provided to people who are uncertain about the answer. If asked for their best guess as to the probability of some rare event about which they are not sure, people tend to choose from the middle of the options. Thus the questionnaire itself can influence the response.

Suppose you are asked for your guess as to the probability of contracting HIV in a single, unprotected sexual encounter with someone who is infected.

*On one form the choices presented are:*

< 0.01%	0.01%	0.1%	1%	10%	> 10%
---------	-------	------	----	-----	-------

*And on another form the options are:*

< 1%	1%	10%	20%	40%	> 40%
------	----	-----	-----	-----	-------

People will tend to choose higher probabilities in the second set. This makes no sense of course, if a person actually knows what the probability is. When they don't, they apparently tend to use the range given as a guide. Someone can, of course, use this knowledge to influence the outcome of a survey. One way to overcome this problem is to ask the respondents to provide their own estimate, without giving them multiple choices.

# Summaries Versus the Raw Data

As consumers of information, we are exposed to many levels of simplification. If we desperately wanted to know what a study found, we should obtain the raw data and analyze it ourselves. This option is rarely pursued, except by researchers with a strong interest in the study. The next level of getting close to the actual data is to read the original publication. Yet even the published study omits many details: it will provide statistical summaries of the data, and some data may be omitted entirely. Much of the driving force behind this initial compaction of data comes from the journals themselves. Prestigious journals have limited space, and a paper that presents the raw data is usually too long to get published. The authors of a study are almost always told to shorten their document.

Beyond this first round of simplification (in the original paper itself), important studies are often written up in other sources – newspapers, magazines, and the web. In some cases, especially with medicine, professional societies and journals may commission a formal summary and overview of the study. This next level affords a striking opportunity to compress information and even to distort it, such as overselling the study and claiming that its results apply more broadly than is indicated by the data. This form of exaggeration has been documented in summaries of clinical trials: the actual study finds a modest benefit of a drug in a narrow set of patients (e.g., men aged 50-60), and the summary claimed a benefit to a much wider audience (e.g., men aged 40+ or men AND women).

We all have limited time. Thus it is not surprising that summaries of studies are read far more widely than is the study itself. It is thus important to know whether a summary accurately reflects the original study, but this knowledge is often difficult to obtain without effort. Two books have documented several such abuses in medicine, and the overselling of drug benefits may stem from financial ties between the authors of the summaries and the companies whose drug was tested. (See ‘Overdosed America’ by J. Abramson and ‘Should I be Tested for Cancer’ by H. G. Welch.)

# External Links

[Eyewitness Identification - Getting it Right](#)

[The True Story Behind "Conviction"](#)

[Innocence Project Event - The Wrongfully Convicted](#)

[Confession Contamination](#)

## CHAPTER 16: IS SCIENCE LOGICAL?

# 16

INTERPRETATION AND CONCLUSIONS

An earlier chapter revealed that all models are false. This chapter reveals another blemish on the face of science -- how we decide the fate of models is arbitrary.

# Introduction

Once the data have been gathered according to the ideal data template, a challenging phase of scientific inquiry is faced next: what do the data mean? That is, what models can be rejected? The fact that data have been gathered to ensure accuracy does not guarantee that they will be particularly useful for the goals of the study. They need to be sufficiently accurate, but they also need to address the models at hand. Even assuming that the data DO address the models at hand, how do we decide when to abandon one model and move on to a new one? A surprising feature of the scientific method is that this aspect is arbitrary -- not everyone uses the same criteria, and the criteria of one person change from situation to situation. Thus, two objective scientists can evaluate the same data yet come away supporting different models.

# The Language of Evaluation

No one can prove that a model is correct, but we nonetheless want to use good models and avoid bad ones. Yet there are many different degrees of accepting/rejecting a model. A modest terminology surrounds the evaluation of models. The most extreme evaluations are

**refute:** the data are not compatible with a model and force us to reject it

**support:** the data are not only compatible with a model but refute many of the alternatives and lead us to think that it is possibly useful.

A model cannot be supported unless the data would (had they turned out different in certain ways) have refuted the model. That is, "support" means that the model could have failed the test but didn't. Refuting a model is an absolute classification -- there is no returning to reconsider a refuted model (for those data). Supporting a model, however, is a reversible designation -- additional data may ultimately refute it.

A lesser degree of compatibility between data and a model is

**consistent:** the data don't refute the model

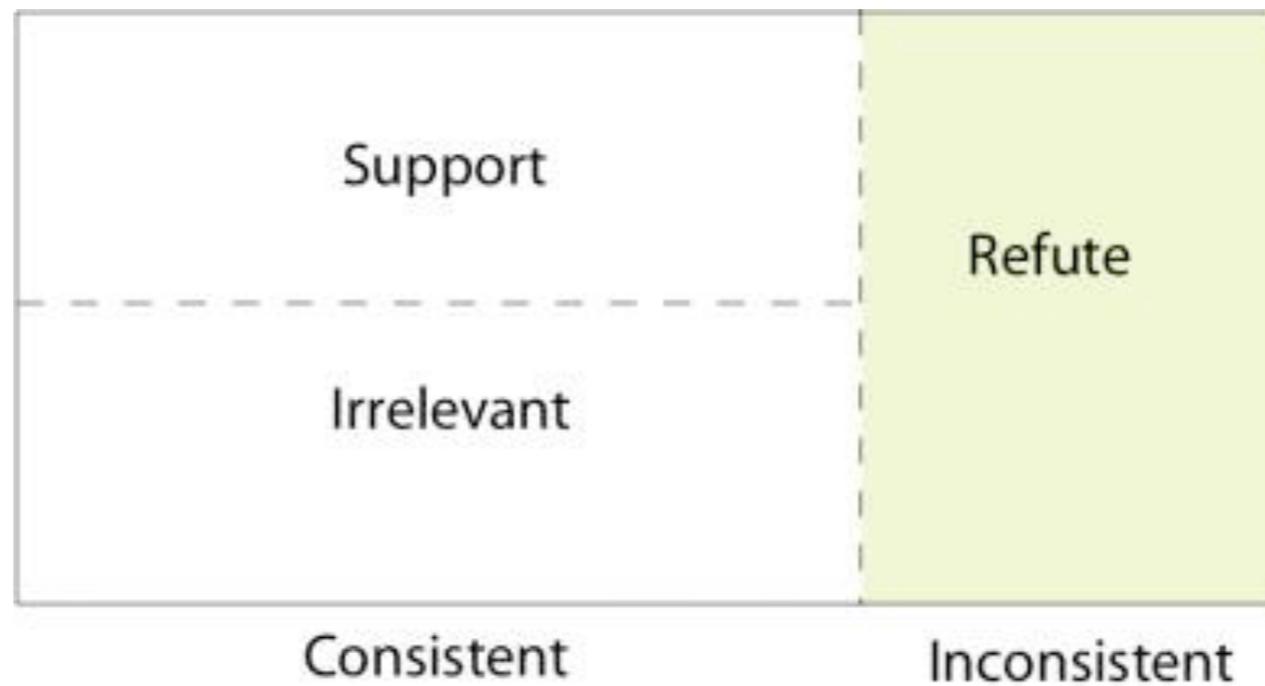
Data that support a model are consistent with it, but data may also be consistent without giving much confidence in it.

At the furthest extreme, data may be consistent with a model but be

**irrelevant:** the data do not address the model in any way that could have caused us to reject it

A simple picture representing the relationships of these different concepts is shown below. Support is surrounded with a dashed line because it is a fuzzy concept in some cases.

A simple picture representing the relationships of these different concepts is shown below. Support is surrounded with a dashed line because it is a fuzzy concept in some cases.



# The Right Stuff -- Which Models Do You Start With?

The notion of progress in the scientific process involves rejecting models and replacing them with new models. How rapidly this progression occurs depends both on goals and on the models chosen at each step. Obviously, if one is lucky enough to start with a model that is close to the "truth" (as concerns the goal), then little progress will occur, because there is simply little room for improvement. Alternatively, starting with a bad model may ensure lots of "progress" because there is so much room for improvement.

There are different philosophies about what kinds of models to choose initially. One approach is the null model approach. A null model is a default model -- one chosen just to have an obvious starting point. A null model might be one that we think is commonly true, or might be a model that we don't think is true but we use anyway, merely to demonstrate that we can reject something. For example, in choosing a model for the probability of Heads in a coin flip, most of us suspect that this probability is close to  $1/2$ , so we would start with  $P=1/2$  as a null model. Or if we were investigating whether alcohol impairs coordination, most of us realize that it does, but we would nonetheless start with a null model that alcohol does not impair coordination, just to show that this simple idea is wrong. There are thus several different reasons for starting with null models, and the choice of the which null model to use will depend on those reasons.

In some cases, people start with one or more specific models. These may or may not be contrasted with a null model. For example, if a particular theory proposed that two thirds of all cancers caused by electromagnetic fields should be leukemia, then we would want to test that model of  $2/3$  specifically (and we might not even know what an appropriate null model should be). Or someone might propose that a particular bacterium is the cause for "sick building syndrome" (the phenomenon in which certain buildings make lots of the inhabitants feel sick), and we would test that model specifically by looking for that microbe in the ventilation ducts of SBS buildings.

The choice of models for testing is thus arbitrary to a large degree.

# No data can reject all alternative models (you can't prove a negative)

There is no limit to how many models are relevant toward a particular goal. Typically, only a few models are considered in a test, but there are countless others that might be considered. For any goal, there will be infinitely many possible models that are relevant. For example, in the simple case of the probability of Heads in a coin flip, we can choose as a model any single value from the infinite set of values between 0 and 1; there is not only an infinity of such models, but the infinity is so "large" that it is considered uncountable. But we could also choose an interval of values in this range -- the probability of Heads lies between 0.467 and 0.522. We could even choose refined models that offered details about a succession of coin flips ("2 Heads will be followed by 1 Tail" and so on). With models for the effect of radiation on cancer, there are infinitely many models which assume the relation is a straight line, infinitely many assuming a curved relationship, and so on.

In testing a model, therefore, the best we can hope for is to reject some of the models. Invariably, no matter how good of a test we conduct, there will be countless others remaining after the test. Since it is impossible even to list all of the models, so the results of a test are usually stated in terms of just the few models considered up front (which may be as few as one model -- the null model).

This inability to reject all possible alternatives is the main reason we can never prove that a model is correct -- there are always many models consistent with any set of results, so we have no basis for claiming that a particular one is correct. Thus, in a coin flip, we can never prove that the probability of Heads is exactly  $1/2$ , because no matter how many times the coin is flipped, there will always be a range of values consistent with the data. There is an infinite number of values within that range, any of which could be true. A special case of this is the statement that we "cannot prove a negative," which is to say that we cannot prove that a phenomenon absolutely fails to exist. In testing astrology predictions, we can never prove that there is NOTHING to them, because there will always be a range of outcomes consistent with any test, and that range will include at least a tiny bit of nonrandom prediction. In testing whether sugar in a diet influences cancer rates, we can never prove that sugar has no effect, because the data will always be consistent with a range of cancer levels. Hence a reason to rely on null models.

# How Disagreement Can Persist After a Test

One would think that objective people should be able to achieve consensus with one another once the relevant models have been tested. Yet differences of opinion abound in science. These differences stem from the points described above. First, not everyone starts with the same set of models. Some people want desperately to think that trace amounts of pesticides in food are harmful; others want to think that the traces are harmless. Any particular study may fail to discriminate between two alternative models, and the proponents of each model will feel accordingly bolstered each time that their model survives a test. So a test that fails to resolve between two models can actually increase the acrimony in a debate. Furthermore, in the case of trace pesticide levels, there will always be some low level of pesticide that cannot be shown to cause harm (even if it does), simply because of intrinsic limitations of the scientific method (see the subsequent chapter "Intrinsic Difficulties" in Section V).

# Criteria for rejection

Each of us personally makes many decisions daily about what to believe or accept and what to reject. The sales pitch promising high returns on your investment is obviously to be questioned. We are used to campaign promises being forgotten on the night of the election. But if our physician tells us something, or we read about a government report of a decrease in AIDS deaths, we are inclined to believe it. (In contrast, the public has come to mistrust many government statistics, especially rosy forecasts about the economy, and war casualty reports during wartime.) This is true for all of us -- we trust some sources more than others and accept some things at face value. But for something like the result of a research study or a government-approved release, somewhere back along the information hierarchy, someone has made a decision about what is true enough to be accepted and what is not. That is, someone has made a decision to accept some models and reject others.

## **Statistics:**

The most common protocol for making acceptance/rejection decisions about a model is statistics. In some cases, results are so clear that the accepted and rejected models are obvious. But far more commonly, mathematical rigor is required to make these decisions. For example, if you want to know if a drug is helping to cure ulcers, and the tests show that 15% are cured with the drug versus 12% cured without the drug, any benefit of the drug won't be obvious. Statistical tests are mathematical tools that tell us how often a set of data is expected by chance under a particular model. (A statistical model is actually a mathematical model built on the assumptions of the abstract model we are testing, so it involves layers upon layers of models.) If the results would be expected under the model infrequently, we reject it; otherwise we accept it (which doesn't mean that it has been "proven"). Ironically, we know in advance that the statistical model is false. The question is, however, whether it can be refuted.

Scientists have agreed by "convention" what criteria to use in rejecting/accepting models. Commonly, if a set of observations (data) would be expected to occur under a particular model only 1/20 times or less often (5%), we reject the model. What this means is that, if the model is true, we will make a mistake in rejecting it 1 in 20 times. So "rejection" is imperfect. Because scientists often test many things, and they don't like to be wrong about it, they are sometimes conservative and don't get excited about a rejection unless the data would be expected less than 1 in 100 times under the model.

These criteria for rejection and acceptance are arbitrary. Yet science is often portrayed as objective and absolute. Furthermore, scientists often have difficulty relating to the public willingness to accept many things for which there is little support, when in fact, their own criteria for acceptance are subjective. There is nothing magic about using a 5% criterion for rejection. As an institution, science is fairly unwilling to adopt new ideas and abandon old ones (reluctant to reject the "null" model) -- the burden of proof is on the challenger, so to speak. But there are many facets of life for which we don't need to be so discriminating and thus don't need to wait until the 5% threshold is reached. We can be willing to try a cheap, new over-the-counter medicine without 95% confidence in its worth because the cost of being wrong is slight and the benefit is great. Conversely, when it comes to designing airline safety, we want to be extremely cautious and conservative about trying new things -- we will not tolerate a 1 in a million increased risk of a crash. Many people play the lottery occasionally; the chance of winning is infinitesimal, so it is a poor investment. Yet, the cost of a few tickets is trivial, and the hope of winning is entertainment value that it can actually make sense for people to play. After all, we pay \$6 to see a movie and have no chance of recovering any money. The criteria for acceptance of a model, at least in the short run, thus depend on the cost of being wrong. Where these costs are small, we can afford to set less stringent standards than where the costs are high.

### ***Repeated Successes:***

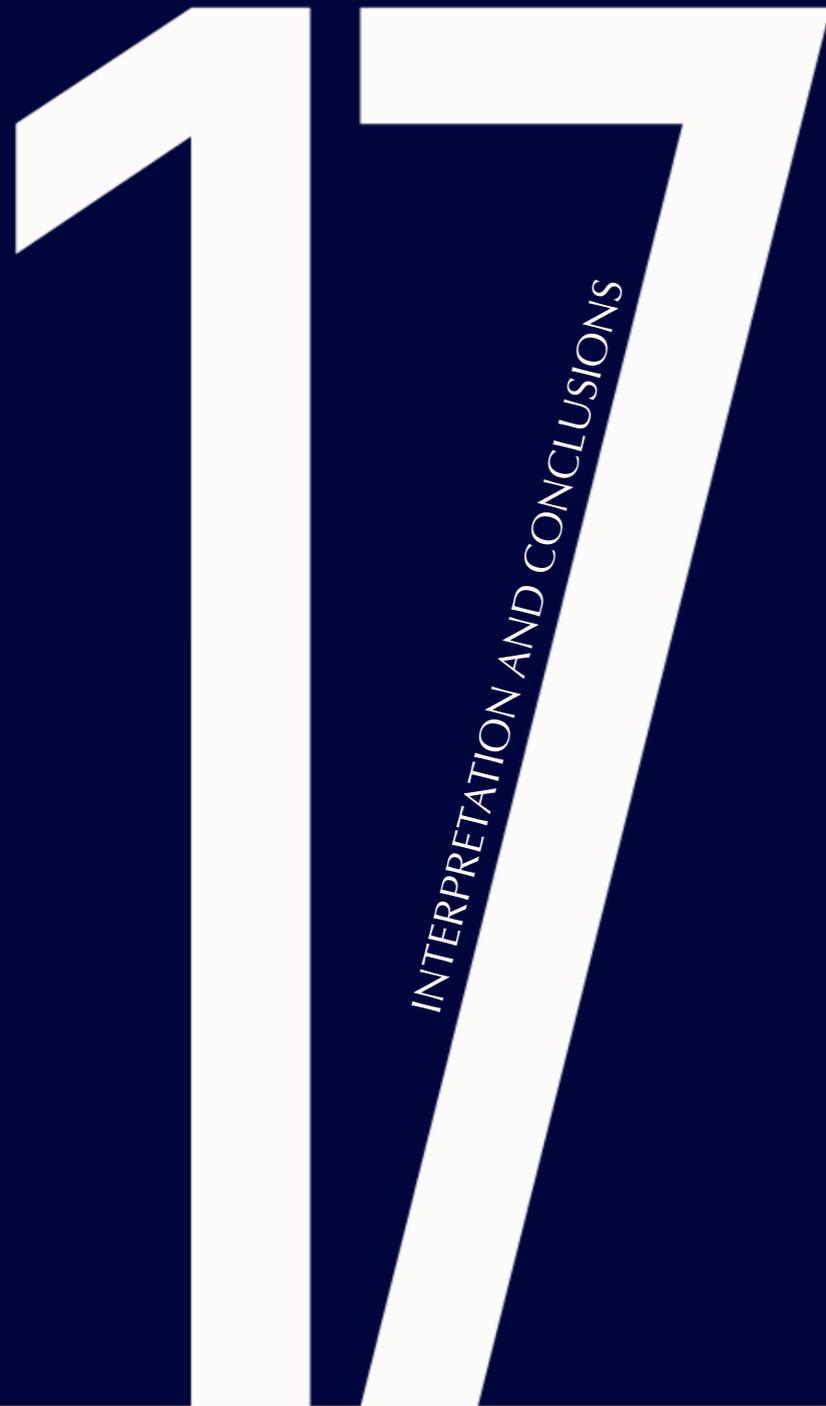
Statistical tests are substitutes for what really counts -- whether something works time and again. We no longer need a statistical model to convince ourselves that the Sabin polio vaccine works, because it has been tried with success on billions of people. The major theories in physics, chemistry, and biology (including evolution) have held up to many different tests. Each time we conduct a specific test, we may use statistics to tell us if the results for that study are significant, but in the long run a model had better hold up time and again, or we will abandon it.

---

# Unscrupulous Exploitation and the Limits of Evaluation

Unfortunately, however, we can't wait for dozens of trials on everything, and we must rely on statistics and other short-term evaluation methods to show us what to accept. This reliance on short-term evaluations provides a loop-hole that can be exploited to gain acceptance of a model that should be rejected. Businesses can run "scams" (legal or illegal) that take full advantage of the time lag between initial marketing of a product and widespread evaluation of its success -- with good or lucky marketing (based on hearsay, for example), a product can sell millions before it is shown to be useless. In the meantime, the consumer wastes money and time. The market of homeopathic "medicines" ("natural remedies") is full of products with suggested benefits for which there is no reliable evidence; the FDA ensures that these products are not advertised with claims of health benefits, but many counter-culture (and even mainstream) magazines provide articles touting various homeopathies. For products that do seek FDA approval, careful selection of statistical tests can obscure mildly negative effects of a health product, and careful design of clinical trials can avoid certain types of outcomes that would be detrimental to gaining approval of the product (the FDA estimates that only 1 in 5 newly approved drugs constitute real advances). For a product used by only a small percentage of the population, it may be impractical or impossible to accumulate enough data to provide long term evaluations of beneficial and detrimental effects.

# CHAPTER 17: UNCERTAINTY AND RANDOM: WHEN IS CONCLUSION JUSTIFIED?



Deduction – the use of facts to reach a conclusion – seems straightforward and beyond reproach. The reality is that uncertainty underlies every step in deductive inference. The uncertainty applies at many levels.

# Introduction

Any fan of Conan Doyle's Sherlock Holmes has no doubt marveled at the stunning deductive powers of the mythical detective: Holmes's glance at a suspect's shoes evokes a proclamation that the gray smudge is of a clay found only in a particular quarry outside of Dover, that the person's wrinkled clothes indicates a recent train ride that day, and before you know it, Holmes has produced a dazzling chain of connected facts that accounts for the suspect's whereabouts for the previous 24hr. It sounds so logical and flawless.

The problem is that those deductions never recognize uncertainty. There is some chance that the gray smudge on the person's shoes is not clay from Dover, but is pigeon poop, paint, or any of 100 other materials; the wrinkling of clothes might stem from being worn two days in a row, and so on.

In applying the scientific method to reach a conclusion, we want to acknowledge the uncertainty. Ideally, we hope to reduce the major sources of uncertainty, but in any case, we should not do what Holmes does – we should not regard our conclusion as fact.

# The Roots of Uncertainty

Consider the conclusion that levels of a particular protein hormone in the body (leptin) determine the body mass index (BMI): whether the person is thin, of normal weight, or obese. The initial studies of leptin were based on mice, and indeed, the biotech company Genentech spent several hundred million dollars acquiring the rights to use the leptin gene therapeutically. The data we might imagine using to reach this conclusion could include:

leptin levels and BMI in a mouse strain

leptin levels and BMI in a sample of humans

Suppose we find that BMI and leptin show a trend in both mice and humans. Where is the uncertainty in concluding that leptin is the basis of BMI? Here are some issues to consider:

1. inappropriate model - mice may not be a good model of humans
2. bad protocol – the measured leptin levels may be inaccurate, so the patterns are not real.
3. bias - the sample of people used may not be representative of most humans (e.g., perhaps they were all middle-aged, white males)
4. insufficient replication - the number of people in the sample may be small, so that any pattern may have arisen by chance. We use statistics to decide this possibility, and the matter is addressed below under ‘random.’
5. correlations - the leptin-BMI pattern may be real but leptin is not ‘causal.’ We deal with this

Thus any conclusion about leptin and BMI must acknowledge and address these and other sources of uncertainty. Initially, it may not be possible to quantify the uncertainty or even to decide that leptin is ‘probably’ a major determinant of BMI. As more data are obtained, the role of leptin on BMI should be increasingly resolved.

In general, uncertainty underlies the models used and data quality in many ways.

# Random

Randomness is a form of uncertainty that we often attempt to quantify. When you play cards or roll dice, the so-called games of chance, you are knowingly allowing randomness to have a big influence on your short-term fate. Of course, randomness is what makes those games interesting and puts everyone on a somewhat equal basis for winning. Not all variation is due to chance – when you step on the gas pedal to make the car go faster, you are creating non-random variation in your speed. Random is specifically reserved to explain why we get different outcomes (= variation) when trying to keep everything the same, as with a coin flip. When it comes to the scientific method, we are mainly interested in whether some observed variation is due to chance or something else (e.g., is the accident rate of drivers talking on cell phones higher than that of drivers not on cell phones).

# Not All Randomness is the Same

Randomness comes in different flavors. A coin flip represents one type of random – two possible outcomes with equal probability. (A die is a similar type of random but with 6 possible outcomes.) Random variation may instead fit a bell curve, as if we were considering how much your daily weight differed from its monthly average: most of the daily differences would be small, but a few might be large. Yet another type of randomness describes how many condoms are expected to fail in a batch of 1000.

---

# Statistics: Testing Models

Most people have heard of statistics, and we mentioned it in a previous chapter. This mathematical discipline should probably be considered a top-ten phobia for most college students, but it is unfortunately useful in the scientific method. The principle behind most statistical tests is simple, however. A statistical test merely compares a particular model of randomness with some data. When a null model is rejected, it means that the data are NOT compatible with that particular brand of randomness. In essence, a statistical test is a substitute for replication, but instead of replicating the data, the test replicates the model of randomness to see often the random process fits the real data.

# Wierdnesses of Random

Some properties of randomness are intuitive, but others are not. Some of the interesting properties of randomness can be explained without any use of mathematics. It can be useful to be aware of them, so you do not get ‘fooled’ by randomness. There is in fact a book with that title (‘Fooled by Randomness’) that explains how many seemingly significant events in our lives and in the stock market are due merely to chance, and the demise of many investment analysts has resulted from their failure to appreciate the prevalence of randomness in their early success.

## **Runs and excesses**

If you flip a coin (randomly), you expect a Head half the time on average. Sampling error will cause deviation from exactly 50%, but as the number of flips gets really large, the proportion of heads will get closer and closer to  $1/2$ .

You can ask a different question, however. At any step in the sequence of coin flips, you will have either an excess of heads overall, an excess of tails, or have exactly 50% of each. If you have observed more heads than tails, for example, how likely is it that the number of tails will ‘catch up’ so that you then have as many or more tails than heads? From the fact that the observed proportion of heads gets closer and closer to 0.5 as more flips are done, it might seem that an excess of heads (or tails) will not last long. In fact, the opposite is true. As the number of flips increases, an excess tends to persist. From a gambler’s point of view, the fact that ‘he’ is losing does not mean that ‘he’ is ever likely to catch up, even if the game is fair and the odds of winning each hand are 50%. The longer the game goes on, there is less and less chance of ever breaking even.

A 'run' is a succession of wins with no losses (or a succession of losses with no wins). In athletics, runs can occur in a team's wins and losses or in a player's hits/baskets. There is a tendency to think that a player is 'hot' during a succession of good plays but is cold in a succession of misses. To describe a player is hot means, of course, that we don't think the string of good plays is due to chance, but instead stems from their being really good at those times. Yet when hot and cold strings have been analyzed statistically, they are usually consistent with random (like a coin flip, but one in which the odds of success differ from 50%).

### **Rare Encounters:**

We know that the chance two unrelated people have the same birthday is approximately 1 in 365 (slightly less due to leap year and seasonal trends in birth rates). We might thus imagine that the probability of finding two people with the same birthday is small even when we consider a group of people. This intuition is wrong (again). In a group of 23 people, the chance that at least 2 of them share a birthday is approximately 1/2.

The reason for this paradox is that there are many different pairs of individuals to consider in a group of 23 (253 pairs to be exact), although not all pairs are 'independent' of the others.

There are many 'birthday problem' events in our lives. As you get older and have more experiences, there will be accidental meetings of people from your past and other coincidences that seem to improbable to arise from chance. However, when you average over the countless opportunities that you and others have for those rare events, it is not surprising that they happen occasionally.

A related phenomenon concerns the improbability of events in our lives. We often marvel at unique events and assume that something so unusual could not happen by chance. Yet our lives are a constant string of statistically improbable events. When you consider the identities of each card in a poker hand, each hand is just as improbable as every other hand. In fact, the probability of getting a royal flush is higher than the probability of getting the specific hand you were dealt; it's just that the vast majority of poker hands are worthless in terms of winning the game.

## Scams:

An apparently common scam in investment circles exploits randomness. It works like this. The scammer sends out monthly predictions about the stock market to 4096 potential clients. In the first month, half the clients receive a prediction that the market will go up, half receive the opposite prediction. At the end of the month, only half the predictions were correct (neglecting the possibility of no change). The scammer then sends out 'predictions' to the 2048 people who received correct predictions for the first month; once again, half of them receive predictions of an increase in the market, half receive predictions of a decrease. At the end of the second month, there are 1024 people who have received 2, consecutive correct predictions. Furthermore, if the scammer is clever, most of these prospective clients will not know the others who have been sent letters, so they will be unaware that half the letters sent out have made incorrect predictions. By continuing this methodology, after 5 months the scammer will be guaranteed of having 128 clients who have received 5 consecutive, correct predictions. If even a modest fraction of them are impressed, they may be prepared to invest heavily in the scammer's fund, with absolutely no assurance that it does any better than random.

A somewhat similar, though more legitimate process occurs with investment companies. Big companies have lots of funds (separate investment accounts). Even if most of the funds lose money, some – by pure chance – will do well in the short term. Thus a company can always point to funds with a good track record as worth of investment, even though they are no better on average than the others.

## CHAPTER 18: CORRELATIONS ARE HARD TO INTERPRET

# 18

INTERPRETATION AND CONCLUSIONS

In his essay *The Danger of Lying in Bed*, Mark Twain made folly of people who bought travel insurance. He pointed out that far more people died in bed than on public transportation, so the REAL danger came from lying down.

# Introduction

In most scientific inquiries, we seek the cause of something. We want to know what causes cancer, what drugs cause us to recover from disease or to feel less pain, what cultural practices cause environmental problems, what business practices lead to (cause) increased profits, what kind of sales pitch increases sales, what kind of resume is the most effective in getting a job, and so on. In these cases, we are testing causal models. Not everything we've discussed so far requires evaluation of a causal model: in DNA and drug testing, we are merely trying to measure properties of an individual (a drug level, a DNA bar code). But these exceptions notwithstanding, the most common kind of evaluation everyone encounters is testing of a causal model. "What can we change in our lives or our world to cause a certain outcome?" is the essence of what we want in a causal model.

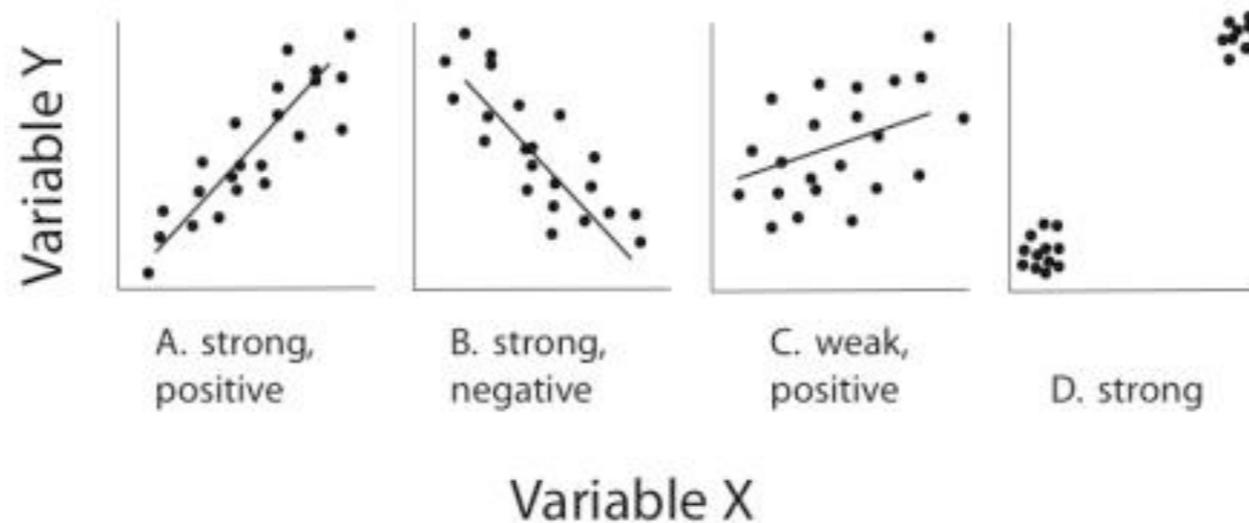
Causal models are typically evaluated, at least initially, with data that describe an association or correlation between variables. If smoking causes lung cancer, then cancer rates should be higher (associated) with smokers. If some patterns of investment lead to higher profits, then companies which practice those kinds of investment ought to be associated with greater returns to their investors. If alcohol causes reckless driving, then a higher rate of accidents should be associated with drunk driving. The catch is this. Although a causal relationship between 2 sets of data leads to an association between them (drinking and driver accidents), an association may occur even when there is no causation. How then do we decide if the causal model is supported or refuted? This chapter is about associations among variables -- correlations -- and how and when we can tease out causation.

Recall from an earlier chapter that epidemiologists in Britain noted a higher incidence of cancers in young people living near nuclear power plants than in the population at large. These data pointed to a possible environmental hazard of the nuclear power plants -- perhaps the power plants were causing excess cancers. However, the fact that excess cancers were also found in proposed sites that still lacked power plants suggested that the power plants were not the cause of excess cancers. This example is typical of the problems that often arise from a failure to appreciate the limitations of correlations.

# What Are Correlations?

Correlations are associations between variables. The first question to answer in understanding a correlation is therefore "What are variables?" Variables are things we measure that can differ from one observation to the next, such as height, weight, behavior, fat intake, life-span, grade-point average, and income. With these variables we can easily assign a number to represent the value of the variable. Perhaps less obviously, we can also treat sex (gender), country of origin, and political preference as variables, even though we don't know how to assign a number to represent each category. In general, a variable is a measure of something that can take on more than one value. It is somewhat arbitrary how we define a variable, but in general, you must be able to put the different values a variable can take onto a single axis of a graph. If you are wondering whether something you have defined is a variable and it would require two axes, then you are likely dealing with a couple of variables combined.

When an association exists between two variables, it means that the average value of one variable changes as we change the value of the other variable (Fig. 18.1). A correlation is the simplest type of association -- linear. When a correlation is weak (e.g., Model C), it means that the average value of one variable changes only slightly (only occasionally) in response to changes in the other variable. In some cases, the correlation may be positive (Models A, C), or it may be negative (Model B). If the points in such a graph pretty much fall inside a circle or horizontal ellipse such that the "trend-line" through them is horizontal, then a correlation does not exist (the same as a zero or no correlation). When either or both variables cannot be assigned numbers (e.g., political party or country of origin), a correlation may still exist but we no longer apply the terms positive and negative (e.g., Model D, depending on the nature of the variables). Since a correlation is an association among variables, a correlation cannot exist (is not defined) with just one variable; "undefined" is not the same as a zero correlation or no correlation. A graph of points with only one variable would have all points on a perfectly horizontal line or a perfectly vertical line (with no scatter around the line).



### **Different kinds of correlations:**

The horizontal axis represents one variable (X) and the vertical axis represents a different variable (Y), with values of X and Y increasing according to the distance from the origin. Models A, B & C show correlations for continuous variables which can take on a range of values (e.g., height, weight), whereas Model D reveals a correlation for discrete variables (variable X might be gender, variable Y presence or absence of the Y chromosome). Model A reveals a strong positive correlation, Model B a strong negative correlation, and Model C a weak positive correlation. The correlation in Model D would be regarded as positive if values could be assigned to X and Y, but if values cannot be assigned (e.g., gender and presence of Y chromosome), we would not refer to the correlation as being positive or negative.

Correlations are common in Business . Businesses often obtain large quantities of correlational data as they go about their activities (Table 18.1). An insurance company in the course of doing business obtains data about which types of customers are more often involved in accidents. These data are purely observational - the company can't force a 68 year old grandmother to drive a pickup if she doesn't want to. The data consist of driver age, sex, make and model of car, zip code, street address and so forth. In addition, the company knows how many and the type of accidents for each customer. These correlations are clearly quite useful in predicting what customers will have more accidents.

Correlated variables having substantial impact on profits and losses or on the efficiency of government operations:

<b>INSURANCE</b>
accident rate vs. age, and sex of driver, make and year of car hurricane frequency vs. city death rate vs. age and sex
<b>FINANCE</b>
personal loan default rate vs. age, gender, and income of borrower corporation bond default rate vs. Moody's rating of the bond
<b>RETAIL SALES</b>
total sales vs. day of the week customer's name vs. product brand, amount, and dollar value of items sold
<b>TRANSPORTATION AND COMMUNICATION</b>
volume of mail vs. city express mail deliveries vs. zip code profits vs. route
<b>MANUFACTURING</b>
steel mill profits vs. type of order and type of ingots and coke used
<b>GOVERNMENT</b>
parolee recidivism rate vs. age, sex, family status, crime committed

Correlations are used to manipulate us. Most advertisements, sales pitches, and political speeches invoke correlations to influence our behavior. A company tends to display its product in favorable settings to build an imaginary correlation between its product and the desirable surroundings (e.g., beer commercials using attractive members of the opposite sex, 4WD autos being pictured with a backdrop of remote, montane scenery). Negative campaigning usually involves describing some unfavorable outcome that occurred during an opponent's tenure in office to develop a correlation in the viewer's mind between the candidate and bad consequences of their election to office.

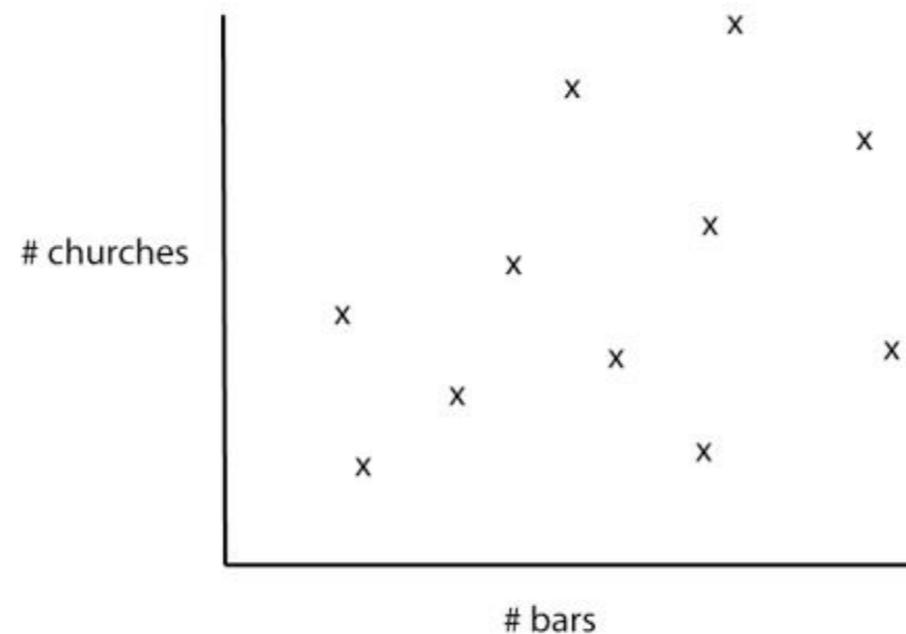
The reason that correlations are used so often in commercials is that they work-- people make the causal extrapolation from correlations. We tend to blame our current president for many social problems, even though the president has little control over many of them. In a well-known but unfortunate psychological experiment of some decades ago, a child was encouraged to develop a close attachment to a white rat, whereupon the experimenters intentionally frightened the child with the rat. Thereafter, the child avoided white objects -- a rather surprising correlate of the rat. Other studies have shown that people respond differently to an item of clothing according to what they are told about an imaginary person who wore it: the response is more favorable if the supposed previous wearer is famous than if the person is infamous. The information thus established a correlation between the clothing and a desirable or undesirable person, and the subjects mentally extrapolated that correlation to some kind of causation of good or bad from wearing the object. And some of our responses to correlations are very powerful. The experience of getting overly drunk on one kind of alcoholic beverage is often enough to cause a person to avoid that beverage years into the future but not to avoid other kinds of alcoholic beverages.

Negative uses. A more negative context for the application of correlation to influence behavior is the practice known as character assassination. A person can be denigrated in one aspect of their life by identifying an unfavorable characteristic in some other (and perhaps trivial) aspect of their life. We automatically extrapolate the negative correlation to them as a whole.

# The Problem With Correlations: Hidden Variables

The problem that underlies evaluation of correlations is extremely common in science. We observe an association, or correlation, between two or more variables. In the nuclear power plant example, there is a correlation between residential proximity to a nuclear plant and cancer, because people near power plants are more likely to get cancer than those who live away from power plants. And we try to infer the causation from that correlation (does the plant actually cause cancer?). Time and again, science has learned the hard way that we cannot infer causation from correlation: **correlation does not imply causation.**

What does this mean? Say that you observe a correlation between smoking and lung cancer. To infer that smoking CAUSES lung cancer, you would argue that people should stop smoking to lower their lung cancer rates. If smoking does not cause lung cancer, however, then stopping smoking would actually have no effect on lung cancer rates (we are very confident, however, that smoking causes lung cancer). How can a correlation not reflect causation? Consider a plot of the number of churches in a town (city) and the number of bars in a town:



This is drawn so that there tend to be more churches than bars in a town, but as the number of churches increases, so does the number of bars on average. Although these data were made up for illustration, the correlation is almost certainly true. To argue causation from these data, we would either have to say that churches cause people to drink more (whether intentionally or unintentionally), or argue that lots of drinkers in a town causes more churches to be built (e.g., churches move in where there are sinners). Furthermore, causation would suggest either that banning bars would reduce the number of churches in the town, or that the way to cut down on the number of bars was to close down churches (depending on which way the causation went). In reality, the correlation is due to a hidden variable -- population size. That is, larger towns have more demand for churches and for bars, as well as other social institutions.

To reiterate the theme of this chapter, the major difficulty with all correlations is that there are many models consistent with any correlation: the correlation between two variables may be caused by a third, fourth, or dozens of variables other than the two being compared. Thus we are left with countless alternative models in addition to the obvious ones. For example, we initially think that the correlation between cancer and residence near a power plant shows that nuclear power plants cause cancer. Then we learn that another factor, site of the power plant, may be important. It appears that the important factor is not the power plant itself, but rather some characteristic of sites chosen for power plants (one obvious possibility is that nuclear power plants are situated in low income areas that have higher cancer rates than suffered by the general population). That is, there are correlations between all sorts of other variables besides just residence and cancer.

There are many issues in society that hinge on correlations (Table 18.2). In some cases, a correlation may identify a causal relationship, such as health defects being caused by environmental toxins. Yet because the correlational data don't reject countless alternatives models, no action is taken to correct the problem. In other cases, a correlation may be assumed to reflect the cause when it does not.

**Public policy issues that involve understanding the cause of a correlation:**

<b>ISSUE</b>	<b>POSSIBLE CAUSATION</b>
High cancer incidence near industrial sites, toxic waste dumps, nuclear power plants.	If the increased cancer rate is actually caused by the hazard, there would be compelling motivation for taking action. But it is often difficult to rule out the alternative explanation that those living near the hazard have different diets or for other reasons are more susceptible to cancer than the general population.
Racial differences in standardized test scores.	There are two opposing positions in this acrimonious debate: i) a person's race, per se, causes them to have low test scores, or ii) minorities often have low incomes, and it is income rather than race that determines test score. The first explanation states that a person is born with a certain intellectual ability, the second states that they acquire it.

---

# Correlations Complicate Studying Diet and Heart Disease

The medical news over the last decade or so has been obsessed with the relationship between diet and heart disease. (Heart disease is chiefly the build-up of deposits inside blood vessels, hardening the arteries and enabling the vessels to rupture and clog.) A report that dietary fiber lowered heart attack risk led to an avalanche of pills and breakfast cereals high in fiber. More recently, a trendy topic has been iron levels in the blood. It is not clear what to make of these reports, but we can be confident that associations between diet and heart disease will continue to be the subject of studies for decades to come. However, let's consider the problems such studies pose.

Your diet consists of literally hundreds of correlated components. For example, people who eat a lot of meat also tend to eat a lot of fat, and people that eat lots of vitamin C tend to also eat much fiber. These, and numerous similar correlations, create huge problems in determining what diet you should eat to avoid heart disease. A study that found an correlation between heart disease and fat, for example, would be hard to interpret because we would not know if it was the fat, per se, or the meat that was the problem. The problem in this example is not as great as it is in other cases, because we can actually conduct experiments with human diets to explore causal relationships. But even in these experiments, it is difficult to control and randomize all relevant factors.

# Do Electromagnetic Fields Cause Cancer?

Beginning in the 1960's and the 1970's, evidence arose that intense electromagnetic fields (EMFs) could influence behavior and physiology. No study was particularly conclusive. In all cases, the fields were intense and effects seemed reversible, and the concern was neither about cancer nor about effects from fields of low intensity such as those in the typical neighborhood. But in 1979, epidemiologists Nancy Wertheimer & Ed Leeper reported that childhood leukemia rate in Denver was higher for dwellings "near" a transformer than for dwellings away from a transformer. The result was incredible because it suggested that many of us are exposed to a cancer risk in our own dwellings. There have been at least 6 attempts to repeat Wertheimer and Leeper's epidemiological correlations, and the overall trend continues to be born out (with some inconsistencies); studies appear maybe a couple of times a year now. Overall, it appears that there is a slightly elevated risk of leukemia associated with living near high current transformers and the wires that emanate from them (the risk factor is 1-2). The baseline rate for childhood leukemia is about 1/20,000, so the EMF risk raises it to 1/10,000.

Once public awareness had been elevated by these original studies, there was a plethora of anecdotal and post-hoc observations that highlighted incidents in which EMFs might be causing harm. The news was filled with a cluster of miscarriages in women working at CRT's (cathode ray tubes – the computer monitors in the days before flat screens), the news carried stories of people with cell phones who got brain cancer, and so on. A study of a NY telephone company made an attempt to determine if there was a correlation between cancer and occupations which had varying exposure to EMFs, in the hope of showing that more cancers were found with higher doses (Table 1).

Cancer incidence vs. exposure to electromagnetic fields:

<b>OCCUPATION</b>	<b>RELATIVE EXPOSURE</b>	<b>CANCER</b>
cable splicers	highest	2X overall cancer rate
central office	next highest	3X prostate cancer; 2X oral; some male breast cancer
other	lowest	nothing of particular note

The occupations were ranked according to exposure and the cancer incidence showed some hint of a dose-response. However, this was the only study (of many) showing a possible dose response effect, and even in this case, the results present a heterogeneous array of cancers.

## Reasons for Being Skeptical

In determining whether electromagnetic fields might cause cancer, it is reasonable to compare EMFs to a form of radiation that does cause cancer – ionizing radiation. To compare ionizing radiation with household EMFs, one needs to consider the energy and intensity of EMFs. Electromagnetic fields from alternating current are low in energy. The energy of electromagnetic radiation increases with the frequency of radiation. Alternating current cycles at 60 times/second, so its frequency is 60 cycles/second. Visible light has a frequency of  $10^{14}$  cycles/second, UV light has a frequency of  $10^{15}$ - $10^{16}$  cycles/second, and X-rays have a frequency of  $10^{16}$ - $10^{20}$  cycles/second (gamma and cosmic rays have even higher frequencies). So the electromagnetic fields (EMFs) of alternating current have only a trivial level of energy compared to the mildest form of ionizing radiation that can cause cancer -- UV.

In addition to the energy of EM radiation, one needs to consider the intensity. Intensity is the amount of radiation per unit time. For example, a light bulb emits more intensely when it is bright than when it is dim, even though the energy level of individual photons is the same. So even though EMF from alternating current might be too low in energy to produce mutations, high intensity fields might have some biological effects. Here again, however, there would seem to be little reason for concern. Field intensity falls rapidly with distance, so even though the field intensity of various household appliances is high at the source (e.g., the motor in a hair dryer), the field is quite small only a few inches away. And intensities experienced in the household are small relative to the Earth's magnetic field and to the electrical fields generated by our own cells. The only possible cause for concern, therefore is that the man-made fields oscillate, whereas the cells' electrical fields and Earth's magnetic fields do not.

Oscillating magnetic fields do have biological effects – they generate currents in body tissue that are easily measured. However, normal muscle activity also generates currents as well (with no known function). On the whole, this EMF effect on bodies and tissues is not large compared to normal levels. However, there is general ignorance about these effects, so any conclusions are tentative.

## Where Things Stand Now – no cause for concern

A report by the National Academy of Sciences in 1997 (Possible Health Effects of Exposure to Residential Electric and Magnetic Fields, <http://www.nap.edu/openbook/0309054478/html> ) summarized the then present status of residential EMFs and cancer:

1. There remains a statistically significant correlation between childhood leukemia and the wire code of a house (mostly based on the distance between the house and high current power lines). The highest-code houses have about a 1.5 risk factor (50% increase). There is no significant correlation for other childhood cancers or for any adult cancer.
2. There is no correlation between EMFs measured inside the households and childhood leukemia (measured after leukemia was diagnosed). The cause of the correlation in (1) remains unknown.
3. In vitro effects (cell culture) reveal abnormalities only at EMF doses 1,000-100,000 times greater than typical residential exposures. These effects on cells do not include genetic damage.
4. Exposure of lab animals to EMFs has not shown any consistent pattern with cancer, even at high EMF doses. Some behavior responses are seen at high doses, and there is an intriguing result that animals exposed to both a known carcinogen and intense EMF show increased breast cancer levels.

As it stands, there is no reason to be concerned about residential EMF levels. As is true in all scientific matters, our current conclusions may change as new evidence comes in. But there is already compelling evidence that any cancer-causing effect of EMFs is not very large.

# Why Do We Bother With Correlations At All?

Given the problems with interpreting correlational data, one might reasonably ask: why do we bother with them at all if it is a causal relationship that we seek? Why not just gather data that could provide a more definite answer, or otherwise just ignore correlations? The reason is pragmatism. Correlational data are usually relatively easy and inexpensive to obtain, at least in comparison to experimental data. Also, many cause-effect relationships are so subtle that we often first learn of them through correlations detected in observational data. That is, they are often useful.

## CHAPTER 19: CONTROLS

# 19

INTERPRETATION AND CONCLUSIONS

One reprieve from the correlation-does-not-imply-causation difficulty is to seek out data that avoid certain problems with interpretation. Controls are essential to all evaluation of causal models, and better controls can bypass some of the problems in evaluating correlations.

# Introduction

If drunk driving is the cause of increased accident rates, then we should observe a higher rate of accidents when a driver is drunk than when sober. More generally, a causal model works on the simple principle that a substance or event (X) causes something else to happen (Y). If the model is correct, we should observe that Y occurs in the presence of X, but that Y doesn't occur (as often) in the absence of X.

What underlies this reasoning is a comparison:

- Y is observed along with X (more accidents with drunk driving)
- Y is not observed in the absence of X (fewer accidents with sober driving)



In order to evaluate a causal model, therefore, the data must address both sides of this comparison -- if we know only the accident rate for drunk driving and not for sober driving, we can't say whether drinking raises or lowers the accident rate. We need data for the baseline accident rate of sober driving, for comparison to the accident rate of drunk driving.

These baseline data are known as a control. A control serves as a reference point for the study, i.e., a point of comparison. (In the Ideal Data section we introduced the idea of a standard or control. Here we extend the concept of a control for evaluating or interpreting a model.) The following table lists control groups for various kinds of studies:

<b>MODEL</b>	<b>CONTROL GROUP</b>	<b>TREATMENT GROUP</b>
Smoking causes cancer	non-smokers	smokers
Smoking causes cancer	people who smoke less	people who smoke more
Aggressive questioning by a lawyer is more effective than passive questioning	outcome from passive questioning	outcome from aggressive questioning
Coca-Cola tastes different than Pepsi	People's responses to the taste of Pepsi	People's responses to the taste of Coca-Cola
	People's responses to the taste of Coca-Cola	People's responses to the taste of Pepsi
A new advertisement causes an increase in sales	sales rates under the old ad	sales rates under the new ad

In some cases, there is no clear boundary for the control group, but controls are nonetheless present. For example, if our model is that increased smoking results in increased cancer rates, then a control is present whenever people with different smoking levels are included. There is no cutoff at which we say people are definitely in or out of the control group, but controls are nonetheless included by virtue of the comparison between different levels of smoking. In other cases, we can say that a control is present, but there is no clear group which can be called "control" instead of "treatment" (as in the Pepsi example in the above table).

The control is possibly the most vital design feature in studies testing causal models (or other models which make a comparison). If the control group is chosen poorly, then no amount of ideal data can salvage the study. It is relatively simple to decide whether an appropriate control (or comparison) is present in a study: Merely list how each group is treated, and list the observations that are made systematically for all groups. A causal model can be evaluated with a set of data only if

- (i) the data measure the relevant characteristics described by the model, and
- (ii) the difference in treatment between the groups matches the comparison given in the model.

As an example, consider the model: aspirin lowers cancer rates. Any study testing this model would need to measure cancer rates in different groups of people, and the groups must differ in their exposures to aspirin. However, if one group received aspirin plus a low-fat diet, and another group received a high-fat diet without aspirin, the groups differ in more than just dose of aspirin. The data generated from the study would lack an adequate control, because the data could just as easily be argued to test a model of the cancer-causing effect of high-fat diets.

The problem with correlational data is that one often does not know how many factors differ between the main group (treatment group) and the control group.

---

# Eliminating Factors with a Control

The purpose of a control is to eliminate unwanted factors that could possibly explain a difference between groups also differing in the factor of interest. If we want to know whether smoking increases lung cancer rates, we don't want our smoking group to be uranium miners with dusty lungs and our non-smoking (control) group to be Himalayan monks, because any difference in lung cancer rates might be due to other differences in environment instead of the difference in smoking. We therefore want a control group to eliminate as many factors as possible other than smoking/non-smoking. What we are controlling for in the control group is not smoking (which is the "treatment" or main factor). Rather we are trying to control for or eliminate the myriad of other factors that we don't want to interfere with our assessment of what smoking does (we mean the same thing by "control for" as we do by "eliminate" or "match" a factor).

By "controlling for" or "eliminating" or "matching" a factor with a control group, we mean merely that the factor is (on average) the same between the treatment and control groups. That is, the control group attempts to be the same as the treatment group except for the treatment factor. Thus, if our smoking group consists of (smoking) uranium miners, our control group should likewise consist of non-smoking uranium miners.

A factor can be controlled for if:

- A. it is absent in the treatment and control groups,
- B. it applies to everyone in the control and treatment groups, or
- C. is present in only some members of each group but is present to the same degree between control and treatment groups.

Control groups that match the treatment group in every possible way other than the treatment thus eliminate all possible unwanted factors. But typically (except in the best experiments), it is not possible to obtain such a perfect match between treatment and controls.

# Better Controls

From the point we just made, not all controls are equally good, even if they are all considered adequate. Consider the British nuclear power plant example from Chapter 2, in which higher cancer rates were observed in people living near nuclear power plants than in the population at large. Residents living near the power plants are the "treatment" group (exposed to the possible environmental hazard). Controls are thus people not living near the nuclear power plant. These controls could, in principle, be comprised of

**control 1:** all other people living in Britain

**control 2:** people living at environmentally similar sites as the power plant locations but lacking a nuclear power plant

**control 3:** people living at sites of the power plants after the plant was built but before any radioactive material was brought in.

While all of these groups would be considered acceptable controls, some seem better than others. Why? The reason is that some control groups match the treatment group for more factors than others and thereby enable us to reject more alternative models than others. Here are 3 models that we might consider for the elevated cancer rates:

**model (a):** The radioactivity from nuclear power plants causes cancer. Cancer rates will thus increase after the plant is built and comes on line because only then is radioactivity present. The factor of interest is radioactivity.

**model (b):** Nuclear power plants do not cause cancer but are built in areas of poor environmental quality which cause the elevated cancer rates. The correlation between cancer rate and area of residence (near or far from a nuclear power plant) stems from a correlation between environmental quality and sites chosen for nuclear power plants. The factor of interest is thus environmental quality.

**model (c):** Once a site is targeted for nuclear plant development, land values decrease, and the people who move in have cultural practices that predispose them to higher cancer rates. There is thus a hidden correlation between social culture and area of residence. The factor of interest is thus social culture.

These three models collectively propose three different factors as the cause of elevated cancer rates: radiation, environmental quality, and culture. The latter two models suppose correlations among hidden variables and can be ruled out if control groups are appropriately matched with residents near nuclear power plants.

The first control group (1) does not eliminate either of the factors in (b) and (c) and thus would not allow us to distinguish among any of these models - low cancer rates in the control would be consistent with all 3 models. Control (2) matches both groups for the environmental factor and could allow us to reject model (b): if cancer rates were lower in environmentally-similar sites lacking nuclear plants then we would reject the idea that environmental quality was the cause of cancer. Control (3) eliminates both the environmental quality and cultural factors and thus could allow us to reject models (b) and (c): if cancer rates were low immediately before the plant started running but increased later, we could reject all models in which cancer rates were high before the plant opened [of which (b) and (c) are examples]. We thus say that control (3) is better than (2) because it matches more factors, and both are better than (1), again because they match control and treatment groups for more factors than (1).

The way to assess the quality of a control group is thus to consider the possible factors causing cancer (i.e., the different causal models) and to compare the control groups with each other and with the treatment group to see if some are superior to others. That is, which factors are matched between control and treatment groups; if they are matched, we say they are eliminated:

GROUP	FACTOR			
	POWER PLANT (RADIATION)	BRITAIN	ENVIRONMENTAL QUALITY OF POWER PLANT SITES	social culture similar to residents near power plants
Treatment	+	+	+	+
Control 1	-	+	-	-
Control 2	-	+	+	-
Control 3	-	+	+	+

*The third control is thus matched to the treatment group better than the other two controls.*

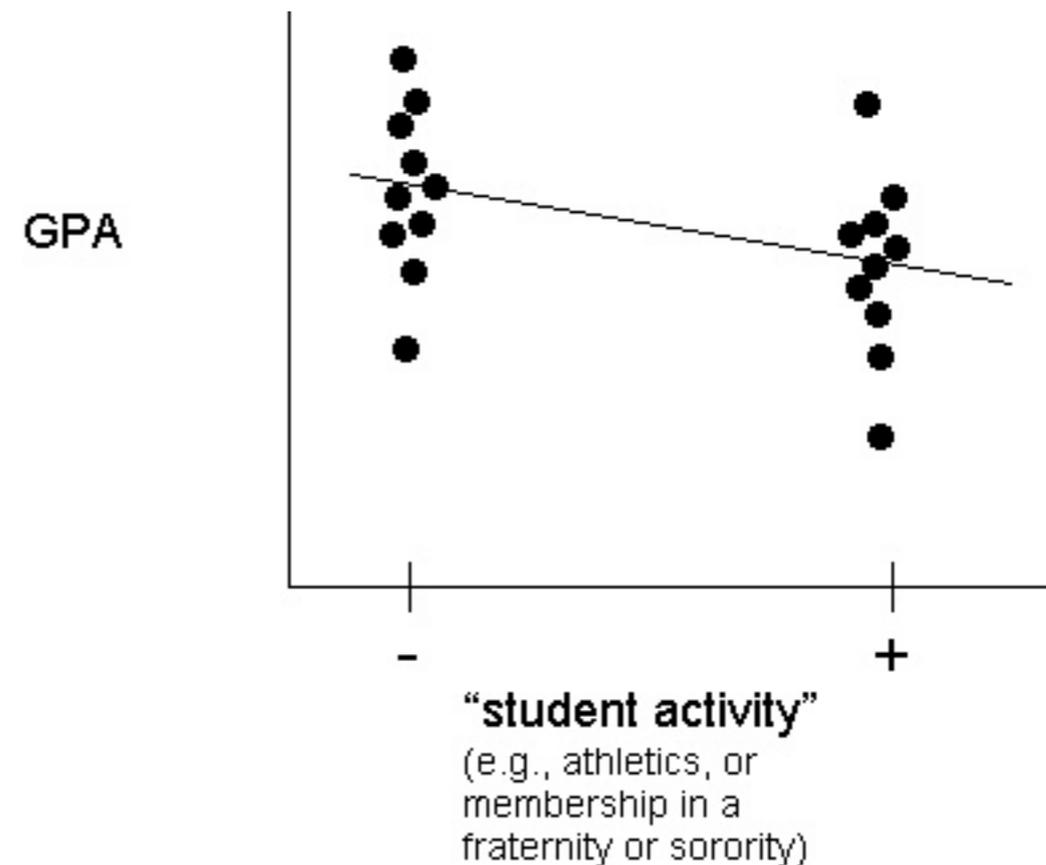
From this illustration, we can also see that the number of alternative models is virtually unlimited. But for each model, we can imagine a control group that would allow us to distinguish it from many of the important alternatives. By and large, we want controls that enable us to reject many of the alternative models. In this nuclear power plant example, we are chiefly interested in whether the radiation causes cancer, so we would want control groups that allow us to reject lots of alternatives to this possibility. Clearly, however, we can never find the perfect control group for all alternative models.

# A Second (Hypothetical) Example: GPA and Social Activity

Suppose for the sake of illustration, we observed a negative correlation between a student's GPA and the university-related social activities of a student:

There would be many possible causal models to explain this correlation:

1. activities limit time for studying, and studying causes better grades
2. the more "social" students are less prepared or able academically
3. the more "social" students set higher personal goals and take harder courses, which is the cause of poorer grades
4. students adopt activities in response to poor grades early in their college career



And so on...

As in the above example with nuclear power plants, many of these alternative models suppose that there are additional factors (variables) underlying this correlation and that one of those hidden variables is the cause of the correlation. To eliminate a factor, the correlation between GPA and "activity" would have to remain negative even when the control group was "matched" with the main group so that the hidden factor was the same in both:

<b>FACTOR</b>	<b>CONTROL GROUP THAT WOULD ELIMINATE THE FACTOR</b>
study time	students w/o "activity" who study as much (little) as students with the "activity"
academic preparation and ability	students w/o "activity" that had similar high school grades and SAT scores as students with "activity"
course difficulty	students w/o "activity" taking same courses as students with the "activity"
early grades	students w/o "activity" with similar first-year grades to those with the "activity"

Of course, if the original correlation disappeared when the control group was matched with the main group for a particular factor, we would then suspect that the actual cause of the correlation was that controlled factor.

In some cases, this approach of controlling for factors one-by-one is all that can be done. But there is no end to the number of such factors that can be considered, so this approach is limited.

# Diet and Heart Disease

Rates of heart disease are higher in the U.S. than in Japan (and many other countries). Two possible reasons for this difference are (i) genetics, and (ii) culture. That is, genetic differences between U.S. citizens and Japanese citizens could result in U.S. citizens being more prone to heart disease. Alternatively, the cultures are different enough and heart disease is so influenced by culture (diet), that the difference could be mostly cultural.

The most basic control is the comparison of heart disease rates between Japan and the U.S. A better control is to use heart disease rates in people of Japanese descent living in the U.S., so that culture is somewhat equalized between the two groups of different genetic backgrounds. Or we could compare Americans living in Japan with the Japanese in Japan. When the control group is taken from Japanese living in the U.S., the difference in heart disease largely disappears. (Japanese living in Hawaii are intermediate.) So this better control enables us to reject an important alternative model.

---

# Calculating an Expectation

A control serves merely to show us of an expected result in the absence of a particular treatment (a baseline, as we have said). There are times when a control can be calculated, without gathering data. For example, we may easily calculate the odds of winning a lottery, of obtaining any particular combination of numbers when rolling dice, and in other games of chance. These calculations can be very helpful in a variety of other circumstances as well. For example, people often marvel at the occurrence of seemingly rare and improbable events (e.g., having a "premonition"). Calculations can show us just why these individually improbable occurrences should happen, without invoking anything mysterious.

## CHAPTER 20: PRISONERS OF SILENCE

# 20

INTERPRETATION AND CONCLUSIONS

Thousands of children in the U.S. (and in other countries) are born with a severe mental handicap known as autism. No doubt, the classification of autism includes many different specific disabilities, but autistic people are unable to communicate. As a consequence, no one really knows if they are able to learn.

---

# Facilitated Communication

In the last decade or so, a revolutionary new method of dealing with autistic children was invented. First used in Australia, it was rapidly adopted by many institutions in the U.S. It is known as Facilitated Communication (FC, for short), and the principle idea behind it is that autistic children can learn but just cannot communicate what has been learned. FC is a tool to help the person communicate.

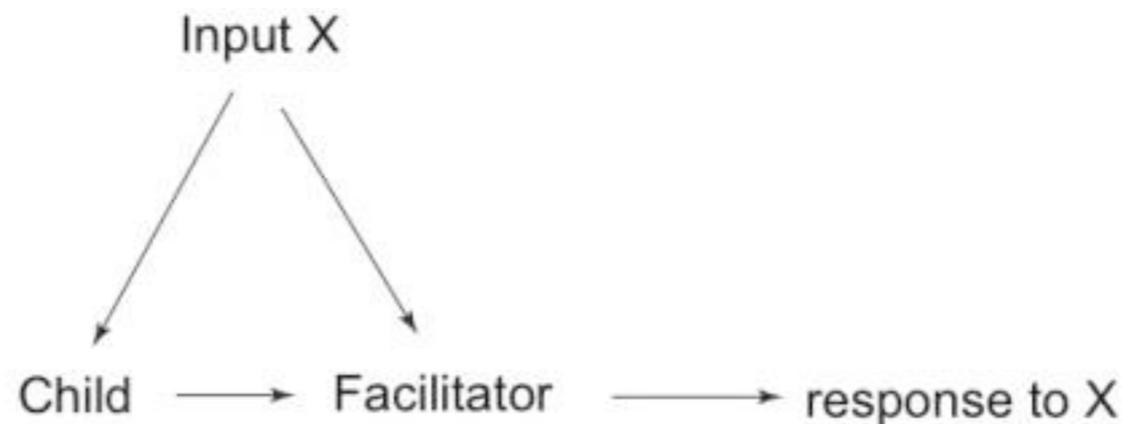
FC works like this. An autistic person is placed in front of a small keyboard. Next to this person sits an adult (the facilitator), who merely holds the autistic person's arm above the keyboard, so that a finger hangs down and can hit keys one at a time. The keyboard is connected to a printer that prints the letters being typed. In theory, these letters (words) represent the autistic person's thoughts.

When first introduced, FC seemed the most remarkable technology in the history of autism. Autistic people changed from being perceived as severely retarded to being perceived as of normal intelligence, merely needing help to communicate. Social workers were delighted, as were the parents of these children.

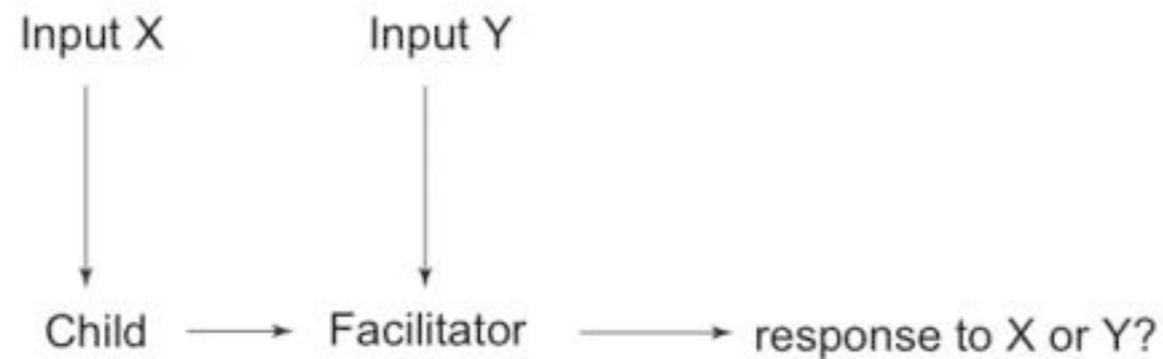
Everything with FC seemed fine. Not everyone was convinced that it really worked, but it seemed harmless at best. Then, however, things took a turn for the worse. Some of the FC transcripts graphically indicated sexual abuse. When these transcripts appeared, children were yanked from their parents and put into foster care. Families were being separated by the state because of the typed words. FC was no longer harmless. If the typed words were true, FC was helping to free kids from desperate situations. If the typed words were not true, however, FC was destroying families.

How was one to decide if the typed words were true? A smart lawyer brought into one of the abuse cases devised a clever test. His idea was to decide WHO was authoring the words -- the autistic child or the facilitator. This test was not quite the same as deciding if the typed words were true, because even if they came from the child, they might not be true. But if the words were not coming from the child, then they most certainly were not true.

The problem in deciding whether the words came from the child was that FC was usually done in a setting in which both the child and the facilitator were given the same information; the facilitator knew what answer was expected, so could possibly be influencing the answer in subtle ways. In essence, there was a correlation present -- both the facilitator and child were being exposed to the same environment.



The test to decide who was authoring the words was simple in hindsight: the child and facilitator were shown different objects and the team was then asked to describe what they saw. It would become immediately clear who was typing the words.



**FC was a fraud:**

*The test proved decisive: the descriptions were always of what the facilitator saw.*

The kind of test used to debunk FC is known as an experiment. In the absence of the experiment, we had no way of knowing whether the words came from the facilitator or child. The experiment manipulated the normal mode of FC in which both the facilitator and child were given the same information to a mode in which they were shown different information. By deviating from this "natural order" of things in a specific way, it was easy to find out that the child was not choosing the keys.

## CHAPTER 21: EXPERIMENTS MAKE THE BEST CONTROLS

# 21

INTERPRETATION AND CONCLUSIONS

Experimentation is just trial and error. But good experiments offer the only trustworthy method of resolving causal models.

# Experiments

What are experiments?	They are planned manipulations, or deliberate changes in the natural associations of objects, undertaken to observe the outcome.
Why use experiments?	To get rid of uncontrolled factors; they help destroy unwanted associations in the data. With randomization, they give the best possible controls.
When to use experiments?	Experiments are appropriate when testing causal models but may not always be feasible or ethical.
What are consequences of omitting experiments?	Unwanted, confounding factors are likely to be present in the data. The controls will not be entirely free of problems in the absence of experiments. Experiments don't guarantee easily-interpreted data, however.

Experiments are usually done to test causal models -- a manipulation to observe whether something changes as a consequence. Consider the testing of a cancer drug. The model tested is: the drug does (or does not) decrease the cancer rate of patients -- a causal model. In testing this model, an experiment is a study in which some patients are given the medicine to see whether it reduces their cancer rates. Evaluation of this model from such an experiment obviously requires some type of control -- baseline -- for the expected rate in the absence of the drug, but the manipulation of giving the drug to test whether the drug affects cancer rates is an experiment. In contrast, some manipulations are done merely to obtain measurements -- inflating a condom to its breaking point is a manipulation, but it is being done to test condom failure rates rather than testing whether condom inflation causes a particular outcome. So condom testing is not considered an experiment to test a causal model. In general, if the intent is only to gather information about the natural situation (unmanipulated), it is not an experiment.

Back to our anti-cancer drug example. The alternative to an experiment is to study cancer in people who took (or didn't take) the drug for their own reasons, such as would be accomplished by merely putting the drug on the market and seeing whether those peoples' cancer improved. Studies made in the absence of experimental manipulations are often referred to as epidemiological or correlational studies. In such an epidemiological study, there are many reasons why cancer rates might differ between the two groups of people, none of them having to do with the medicine. For example, those people most conscientious about avoiding cancer (watching their diet and the like) might be the ones most prone to take the medicine, hence we would observe a lower cancer rate in the "drugged" group even if the medicine was ineffective.

A more obvious example of the value of experiments involves studying the effects of alcohol on social behavior. Letting people choose whether to drink alcohol at a party and observing behavior as a function of alcohol consumption leads to the obvious problem that, even in the absence of alcohol, those who choose to drink are likely to behave differently than those who abstain. In both cases, we would want studies in which people were deliberately assigned to a treatment or control group, rather than giving them their choice. Other examples of experiments are described in the following table. From these examples, it is clear that many kinds of experiments are unethical or impractical when dealing with humans.

AN EXPERIMENT	EPIDEMIOLOGICAL OBSERVATIONS
HIV-infected people are assigned to two groups, one group being given a drug and the other a placebo	Patients who chose to and can afford to take the drug are monitored in comparison to those who don't take the drug.
People are assigned to different "sex" groups to test condom breakage	People are surveyed for their experiences with condom breakage
People are assigned different occupations for 5 years to monitor cancer rates	Cancer rates are monitored in people who have pursued different careers
You feed your children sugar at predetermined times to observe changes in their behavior	You observe changes in your children's behavior according to whether they chose to eat sugar or not

When experiments are performed, individuals are often assigned to different groups randomly, and it may seem that randomization implies experimentation. It is possible, however, to select subsets of data randomly in epidemiological studies (in which no manipulation of nature is performed), so these two features are distinct. But, as we will note below, randomization is virtually an essential component of experimental studies if the study is to be maximally useful.

### **Fortuitous experiments:**

Although deliberate, planned experiments offer the best source of data, there is a category of manipulation that falls between them and pure epidemiological observations. These intermediates are human-caused changes in the natural order, but without intent to test a particular model. The subjects of fortuitous experiments are subsequently observed for the effects of this manipulation.

There is a famous set of fortuitous experiments in the people exposed to large doses of radiation: the Japanese survivors of atomic bombs, U.S. soldiers involved in field tests of atomic weapons, certain occupational exposures, and medical exposures. In all cases, there was no intent to study the long-term effects of radiation on cancer, so these manipulations do not constitute proper experiments to study these models. Yet these exposures are far above any exposures that would occur through natural variation in background radiation levels, so we can think of them as a kind of experiment (deliberate modifications of the natural order) but without intent to study cancer. In the case of U.S. soldiers, the distinction becomes even more subtle, because these soldiers were sent in to detonation sites partly to observe the short-term effects of radiation (and the soldiers wore film badges to monitor their exposures), but the study was not planned to look at cancer. Observations from fortuitous experiments tend to be uncommon, probably because we don't often accidentally create a situation that is so easily regarded as a manipulation to test a model.

### **Experiments, Randomization, and Controls:**

The main benefit of a well-designed experiment is to produce a good control. By randomizing assignments between treatment and control groups, virtually all confounding factors are eliminated, and any final difference between control and treatment group should be due to the treatment. Randomization is the only way to ensure this elimination of confounding factors, but various pseudo-random methods may work just as well (e.g., assigning people by the second letter of their last name). Randomization must be performed over the relevant factors and involve a large enough sample size for the data to be interpretable, however.

# Dissecting Correlations Into Manageable Pieces

As we have indicated, a correlation may or may not be due to causation. Experimental manipulation provides one way of discovering causation, but when a full-blown experimental test is not feasible, there is another way that helps resolve causation. This method merely involves delving deeper into the basis of the correlations to trace causal chains. That is, correlations are usually detected at a superficial level, and merely looking at the underlying mechanisms producing the correlation may indicate whether the correlation is causal. In the nuclear power plant example, one approach might have been to assess whether levels of radiation were indeed higher around the plants than away from the plants. The difficulty here, of course, is that a nuclear power plant might be a cause of cancer for other reasons, and the measurement of radiation levels would not detect other causes of cancer.

The approach of partitioning a correlation into component mechanisms has worked in some cases, one of them being the correlation between diet and heart disease. People with high fat diets die of heart disease more often than do people with low fat diets. Yet these data do not demonstrate that high fat diets actually cause heart disease. It could be, for example, that people with high fat diets also have diets high in protein, and that protein, not fat, causes heart disease.

Additional data show that high fat diets really do cause heart disease. Experiments and other evidence show that each step in the causal chain is likely correct:

*high fat diets high* → *blood cholesterol hardening* → *of the arteries* → *heart disease*

Scientists have done experiments showing that when dietary fat increases, so does blood cholesterol. These experiments did not show directly that high-fat diets cause heart disease, but they did contribute to showing that mechanisms are present that can cause the correlation. Each step in the above chain has similar evidence supporting it.

The method of investigating the cause of correlations can be applied in many other situations. Some years ago consumer interest groups observed a correlation between Pintos and death rates in car accidents: people riding in Pintos were burned more frequently than were people involved in accidents with other types of cars. In this case it was shown that the problem was the Pinto itself, rather than the possibility that Pinto drivers were somehow the cause. The finding that the gas tank in Pintos often exploded after a rear-end collision was the critical evidence establishing the cause of the high death-by-fire rates suffered by Pinto drivers.

## CHAPTER 22: SOME PROBLEMS ARE INTRINSICALLY DIFFICULT



The scientific method leads to faster progress on some problems than others. Like children's puzzles and games, scientific problems vary in tractability from easy to difficult. Tic-tac-toe is easy to master. Jig saw puzzles require more patience, but with persistence can be solved by nearly anyone. Rubik's cube is mind-boggling.

# Introduction

In the last half century we have built bombs that could catastrophically alter our climate for centuries to come, yet we remain surprisingly inept at predicting the weather even a week in advance. Although we have eliminated smallpox, progress on curing cancer has been painfully slow. We can describe with some accuracy the flight of a baseball, but are inept at predicting the outcome of sporting events.

There are good reasons why ignorance persists about the weather, sports contests, and many other phenomena in spite of continual assault by the scientific approach. The very nature of these problems makes them difficult. Of course the rate of progress is in part controlled by factors other than the difficulty of the problem, including the amount of money spent, and the number of people working on the problem. But these social factors will not concern us. Here we will consider four factors that make some scientific questions intrinsically difficult:

1. time lags
2. rarity
3. interactions
4. the difficulties of using human subjects in an experiment

# Time Lags Slow Progress

For some problems, the data needed to test models can only be gathered slowly. Consider the procedure we use in adjusting the temperature of a shower and the dial of a radio. Both involve a simple application of the scientific method: We start at some initial setting, evaluate the setting, readjust the setting, evaluate the new setting, and repeat the process until the desired goal is achieved.

Because we obtain the needed data faster, progress in setting the radio dial occurs more quickly than progress on adjusting the shower temperature. The difference is due to time lags. Shower handles are typically several feet removed from the shower head, and it may take up to a minute before an adjustment at the handle translates into a change in the temperature of water on our skin. By contrast, turning the radio dial results in a nearly instantaneous change in the radio's output. Because gathering data about shower temperature involves a longer time lag than gathering data about a radio frequency, it takes longer to adjust a shower.

Time lags abound in our world. Some spectacular and renowned cases involve the orbits of heavenly bodies. Those of us who did not observe Haley's comet this century are not likely to have another chance, because the time lag is 77 years. No one knew to expect the Hale-Bopp comet in 1997 because it had not been observed for nearly 2000 years. We all fear overexposure to radiation because it increases the risk of cancer. However, the onset of cancer typically follows exposure to radiation by many years (5 years for leukemia and 20 years for other cancers). In economics, the Federal Reserve Bank Board (the "Fed") attempts to influence the U.S. economy by adjusting interest rates; the effect of a change in interest rates takes months to be translated into an impact on the economy. And couples wishing to become parents must wait at least 9 months if they do the job by the classical method, and often much longer if they wish to adopt.

Not only do time lags increase the cycle time of the scientific method, but they also are sometimes so long as to escape detection altogether. The first point is evident from our shower analogy. The longer the delay between faucet adjustment and temperature change, the longer it takes to find an acceptable setting. It may take the same number of adjustments to adjust both the shower and radio dial, but it simply takes more total time when the time lag is long.

The second point --- the possible failure to recognize a time lag --- is more subtle and more sinister. If a time lag is extraordinarily long, we may be unable even to determine that it is present. For example, if the cancer rates we experience today are determined by chemical exposures to our grandfathers when they were 10, it would be nearly impossible to discover the effect.

### Examples of time lags:

LONG TIME LAGS	
Greenhouse warming	There is a lag of decades between industrial activities that increase atmospheric carbon dioxide, and any resulting change in climate.
Weight loss diets	Only after weeks or months on a diet do you achieve significant weight loss.
Diet and heart disease	Many people eat a high fat diet for decades before suffering a heart attack.
SHORT TIME LAGS	
Computers	You can obtain data from many computers almost instantaneously; that is, with essentially no time lag. The computer responds almost instantly when you type in a command.
Steering a ship	There is an obvious time lag between turning the steering wheel of an ocean liner and the actual change in direction of the ship. Pilots also face time lags in landings and take-offs because the momentum of the plane does not change quickly in response to the controls.

---

## Social Problems Created by Long Time Lags

Consider the difficulties posed by time lags in drug manufacturing. If you are testing a new drug, how long should the participants in the study be followed before you can conclude that the drug is safe? Is one year sufficient? Five years? It strains the limits of credulity to imagine how our drug-based health treatment would be affected if trials needed to be followed even for 10 years before a product could be approved. New companies would have to find sources of revenue for at least 10 years before they could begin marketing their first product. The shelves are full of drugs that would not be available under such rules, and of course, improvements in those drugs would be even longer in coming.

The U.S. drug marketplace has witnessed such a problem. From the late 1940's until 1970, the drug DES (diethylstilbestrol) was administered to many women in early pregnancy to suppress miscarriage. It was only later discovered that its use results in an increased cancer rate in their offspring --- 20 to 40 years after exposure to the drug. If the drug had caused cancer immediately (e.g., in the pregnant women or their newborn), then it would have been contraindicated long before 1970, and many fewer people would have developed DES-caused cancer. Similarly, a 1993 trial of a hepatitis B drug killed 5 of the 15 volunteers. Part of the reason so many died was that the lethal effect of the drug was somewhat delayed --- a time lag had not been anticipated.

There is no definitive solution to this dilemma. Countless drugs that we are taking now could be having a delayed effect. Some compromise must be struck between the conflicting goals of adequate testing to ensure safety and maximizing the number of effective drugs available. Actions that the government takes to increase the safety of drugs available (by requiring tests run for longer periods of time), will often prevent or delay some safe and effective drugs from coming to market, because the drugs can't be sold while experiments determining their safety are undertaken. Moreover, regardless of the number of tests undertaken, there is no way to be absolutely certain that a given drug is safe.

Another set of problems arises because long time lags make it difficult to determine who is to blame for poor performance. Is the current recession a consequence of the policies of the current president, or a predecessor? Are the company's earnings the first year after a new CEO is hired a consequence of his/her actions, or the actions of a predecessor? In both these cases, uncertainty about the duration of a time lag obscures the answer.

# Avoiding Time Lags

Time lags are common problems, and scientists have discovered ways to lessen their impact. A common approach is to study alternative models that incorporate a shorter time lag (see Table 22.2). The utility of an alternative model with short time lags depends on its similarity to the main model in question. Viruses (bacteriophages) and fruitflies yielded major insights to the study of human genetics because they have a vastly shorter generation time than do humans.

Perhaps the biggest difficulty is posed by unexpected time lags, as with DES-caused cancer. If you aren't expecting a time lag, there's not much you can do about it until you stumble on it.

## Models that reduce time lags:

Genetics of viruses. Their short life cycle led to rapid understanding of genetic principles that apply to nearly all life.

Rodents are used in cancer research because their short life, relative to humans, enables testing for otherwise long-term effects

Flow charts enable coordination of different dimensions of complicated construction and other social projects, so that excessive delays are avoided.

Political polls provide politicians with rapid feedback about public perception of their performance. The politician can change their positions and their behavior in response to the poll, so they don't have to wait for an election to discover their popularity.

Sneak previews enable marketing agencies to anticipate public reaction to a product before it is made widely available. Changes in packaging and marketing strategy can occur much more quickly if they only affect a small market. Once the bugs are ironed out in a small market, the resulting marketing strategy is then used nationally.

Early reviews and advance advertising. A company may speed public awareness of a product prior to or coincident with its availability in the marketplace.

# Rare events are difficult to

We have all experienced the frustration of a car, stereo, or other complicated machine failing us, only to face the embarrassment of the machine working perfectly when brought in for repair. It is usually easier to fix something that consistently fails than to fix an intermittent problem. A common solution is to simply ignore an intermittent problem until it worsens.

The difficulty of a scientific problem depends heavily on the frequency of the event being studied. Models of rare events improve only slowly. Inconsistent or uncertain results increase the number of observations that must be made -- the number of samples that must be taken -- before we can make progress. For example, it does not require too many coin tosses to realize that we are being cheated with a 2-headed coin. But to detect whether a casino's slot machine offered "fair odds" of a win, we might need to pull the lever thousands or hundreds of thousands of times. Thus, when the event we seek is extremely rare, the problem can become physically insurmountable.

There are many kinds of rare events that confront us, many of them undesirable (Table 22.3). Any one of these events is rare enough that we are likely to ignore its possibility, but there are so many rare event possibilities that they pose a threat collectively. And from a social policy perspective, an individually rare event can still mean thousands of cases in a population the size of the U.S.

**RARE EVENTS IN OUR PERSONAL LIVES**

Adverse reactions to common drugs and vaccines

Side effects of food additives

Transportation accidents

Equipment failures on airplanes and space shuttles

Cardiac arrest under anesthesia

Large liability awards against insurance companies

Floods, tornadoes, lightning, and hurricanes

Leukemia

Winning the lottery

# Measuring a Rare Event Can Involve Enormous Sample Sizes

Childhood leukemia is one of the few cancers that occurs at appreciable levels in children. The disease is fatal unless halted with an extremely radical and difficult treatment. The odds in the U.S. are that about 1 in every 20,000 children will develop leukemia before becoming an adult. This number is a baseline, or average rate. We would obviously like to reduce the number of cases below 1 in 20,000, but we also want to ensure against environmental changes that increase it.

Studies over the last 15 years have suggested that the childhood leukemia rate may nearly double due to exposure to intense electromagnetic fields --- the sort of everyday radiation emitted from electric appliances, power lines, and transformers atop telephone poles. Even though a doubling of this rate still means that each individual has an excellent chance of avoiding leukemia, the doubling would constitute a serious increase in the number of childhood leukemia cases in a country the size of the U.S.

With a rate of 1 in 20,000, we expect only 5 cases in 100,000, or 10 cases if the rate is doubled. Yet, if we indeed observed 5 cases out of 100,000 for one group and 10 out of 100,000 for another group, the difference between 5 and 10 is not large enough to convince us that pure chance isn't responsible for the discrepancy. Even larger numbers of individuals would need to be sampled. Herein lies the problem: a sample of 200,000 children is not adequate for detecting even a doubling of the leukemia rate. When considering that a variety of data must be collected on each child, the enormity and cost of the problem becomes staggering.

# A Dilemma for Business

Although we have illustrated how the difficulty in measuring rare events can harm ordinary citizens, these same problems also impact business in pursuit of their goals. Suppose that a product is tested with 1000 subjects and found to be satisfactory and safe. If it is hazardous to 1 out of 10,000 people, then even this extensive study is likely to miss the hazardous effect. Yet when the product is marketed, it will come in contact with possibly millions of people, and its drawbacks will become obvious from the hundreds of people who suffer from it. Liability costs for even a few of those afflicted could easily wipe out all profits. This problem applies to manufacturers of drugs and food additives, obviously enough, but also to manufacturers of fabrics, household chemicals, equipment, toys, and an innumerable list of other items with which physical accidents may occur.

---

# Sometimes it's Impossible to Obtain an Adequate Sample

The rapid urbanization of the last century notwithstanding, much of the U.S. is populated by small communities of a few hundred to a few thousand people. Importantly, an environmental hazard may increase the incidence of cancer, birth defect, or miscarriage yet the entire community may be so small that there is no statistical basis for demonstrating an ill effect of the hazard.

Furthermore, a corporation exposing a small village to a toxic chemical, for example, may be virtually immune from legal accountability (provided that people are not killed or hospitalized en masse), because too few cases will ever come to pass. Disputes between small communities and large corporations spraying herbicides have in fact occurred over this very point. Similar debates have arisen over whether the emissions from chemical manufacturers have increased the number of cases of anencephaly (babies born with essentially no brain) in small communities along the Texas-Mexico border.

Even if a suspected hazard such as a toxic waste site or gasoline tank farm occurs in a large city, there is no guarantee that enough people will be affected to produce convincing scientific evidence that the suspected hazard really is bad. If the toxic waste site only increases cancer rates in residents who live within several blocks of the site, then it is likely that only a very small number will contract cancer because of the toxic waste site. Even though the toxic waste site is in a large city, the scientific issues are very similar to those encountered in understanding environmental hazards in small communities.

## Related Problems: Dispersed Impacts and Events that aren't Replicated

There are some obvious generalizations and extensions of rare event problems. One is dispersed effects: a large number of people are affected, but they are not clustered in any obvious way. Dispersion is a common problem in the detection of infectious diseases and is an acknowledged problem in bioterrorism awareness. For any one type of food item, there are relatively few food processing centers in the country. For example, most lettuce used in restaurants and fast-food chains is chopped up in a few sites. Suppose one of those sites was contaminated with an infectious bacterium that caused 400 consumers to get sick: the effect would be a distributed outbreak of the illness, but only 2-4 per major city. If the sickness was nothing out of the ordinary (e.g., diarrhea, with recovery in 3 days), the contamination would go undetected. If all 400 illnesses happened in one city, it might well be detected. Indeed, the clustering of illnesses was critical to the detection of an *E. coli* O157 outbreak in Seattle a few years ago (known as the “Jack-in-the-Box” episode) – there had been a similar outbreak in Nevada years earlier that had gone unnoticed. Likewise, the discovery of hantavirus infections in the U.S. was accidental, and found only because of a geographic cluster of illnesses in the four-corners area.

A second difficulty in applying the scientific method is that some events cannot be replicated. Historical events are the most obvious, and some controversy and angst in our society revolves around past events that weren't sufficiently documented and don't have satisfactory answers (Kennedy assassination, supposed aliens near Roswell). Some types of large-scale events have the same problem – they can only happen once, because the whole population is affected. (The mass polio vaccination with the live Sabin vaccine around 1960 comes to mind, but this problem affects the implementation of many government programs. Likewise, the HIV epidemic is unique for the size of its impact on the world.) In many large-scale events, there will be components that are replicated (e.g., in the HIV epidemic, infections are replicated millions of times) but will also be components that are unique (e.g., world-wide economic impacts of the massive toll).

# Overcoming Rarity

If a large sample can't be obtained, there are several alternatives that enable us to side-step the problem posed by rare events. The general solution is to turn to alternative models that facilitate the observation of large numbers of cases.

## **Models of surrogates:**

Although cancer is a common affliction of humans, the development of specific cancers in response to specific factors is rare (e.g., the leukemia risk from increased levels of radiation is not very high). To assess cancer risk, some studies instead look at abnormalities other than cancer, such as chromosome aberrations in blood cells or precancerous growths, and other studies assess mutation rates in bacteria, which can be analyzed by the billions. These cancer surrogates are chosen because they are thought to accurately reflect the likelihood of developing cancer and because they occur at higher frequencies than the cancers themselves. In the same vein, one could use the near-collisions of aircraft to study the factors influencing actual collisions, which are themselves exceedingly rare.

### **Inflating the rates:**

The world is heterogeneous, and when science studies a rare phenomenon, there may be special circumstances in which the phenomenon is common or can be rendered common. To test equipment failure, it is often a simple matter to stress equipment under laboratory conditions to increase its rate of failure, thereby obtaining information about its failure under more normal circumstances. In medicine, rats are often subjected to extremely high doses of substances, to increase the frequency of any ill effects that might be felt by a tiny minority of human consumers. And medical models with inflated rates are not always rats. People who for whatever reason receive higher-than-normal doses of radiation, alcohol, and other drugs are sometimes studied specifically for the purpose of determining risks of lower doses. Airline cockpit simulators can mimic unusual combinations of events, thereby increasing a pilot's ability to survive adverse conditions.

### **Tracing causal chains:**

We have implicitly assumed in this chapter that in order to demonstrate that, say, a toxic waste dump causes cancer in nearby residents, you must establish a correlation, or association, between the waste dump and cancer. This assumption is not completely valid. If we can understand why a rare event is occurring, we may be able to draw reliable conclusions with even small sample sizes. In the beginning of the chapter, we pointed out that it would likely take many thousands of pulls of a slot machine lever to determine the odds of winning. But there is a more direct approach: Simply open the machine up, and look at how the odds have been set (However, we don't recommend trying this on a casino floor.) As we discussed in the chapter on correlation and causation, many problems can be attacked similarly. Examples include scientists identifying the particular genes that an environmental hazard causes to mutate, and secret service agents looking at the details of past presidential assassinations to understand the psychological profile of assassins and the circumstances in which a threat is likely to develop.

---

# Pitfalls of Complexity

The arch-villain Joker in the 1989 movie *Batman* devised a plan to poison the citizens of Gotham City. Rather than simply put a single poison into one product, Joker used a poison which required the combined effects of multiple ingredients. No single product was by itself toxic. Batman discovered the formula to Joker's toxic scheme, and the public was advised accordingly: "avoid the following combinations: deodorants with baby powder, hair spray, and lipstick."

The sinister dimension to Joker's plan is readily apparent to us, because we can all appreciate how difficult it would have been to discover that a combination of products was deadly. Several years ago, when a real villain was lacing bottles of Tylenol with cyanide in the U.S., the problem was simple enough to trace, because a single product was the source of the poison. But imagine the difficulty of tracing the problem if a combination of three products was toxic, and that each of these products by itself was innocuous. Likely, many people would die before anyone determined that a particular combination of products was fatal.

The phenomenon that underlies this example is an interaction among many factors: we cannot discern the whole from a sum of the parts. This is a problem because science typically functions in the same way that we construct a jigsaw puzzle. That is, although the problem involves many pieces and is overwhelmingly complex, progress is made one piece at a time, building on previous successes. Most improved models are relatively minor modifications of their predecessors. But suppose that the puzzle consisted of many pieces, each of which could fit with several other pieces, yet only one combination enabled all pieces to fit together. In this case, we could make many starts, only to find that they invariably led nowhere.

Interactions are ubiquitous in our lives at one level or another (Table 22.4). Many events from the non-scientific and non-industrial side of our lives involve interactions at one level or another: a joke without its punchline is not half funny.

**Interactions of common experience:**

<b>EXAMPLE</b>	<b>INGREDIENTS</b>	<b>RESULT</b>	<b>BASIS OF INTERACTION</b>
Flash powder	mixture of magnesium powder and potassium nitrate, plus energy	explosion	neither ingredient alone generates a reaction
Lethal gas	mixing household bleach and ammonia cleansers	chlorine gas	each cleanser is safe when used alone
Atomic bomb	critical mass of plutonium or uranium	chain reaction of atomic disintegrations	Half of a critical mass does not release half the energy of a critical mass
Cooking recipes	various spices and food items	prepared meals	eating the prepared meal has a greater appeal than eating each ingredient separately
Drug complications	different drugs designed for different purposes	drug-induced death or illness	when used separately, drugs produce positive health effects

### **Why Interactions are a Problem:**

The problem posed by interactions is due to an inability to extrapolate from one model to a new one. For someone cleaning around the house, it seems perfectly logical to mix different cleansers to reduce the number of times a surface needs to be cleaned. Indeed, many household products and over-the-counter drugs actively advertise a multiplicity of components --- the all-in-one principle. But occasionally, the combination of two or more safe ingredients holds a surprise, such as deadly chlorine gas.

The extension of this principle from ordinary problems to scientific ones is simple, as is a realization of the difficulty it poses. Time and again, science fails to give us advance warning of dangerous interactions, and people are injured or die before we are able to arrive at an adequate model to explain the phenomenon. For example, the deadly combination afforded by sedatives and alcohol was discovered by trial and error. The death of a few celebrities in the 1950's and 1960's made this interaction well known. The history of new discoveries in applied chemistry is replete with examples of botched protocols that led to completely unexpected results.

### **Avoiding the Pitfall:**

To a large extent, science is simply saddled with this problem. The recent Noble Prize awarded for the discovery of bizarre combinations of metals that have superconducting properties reflects the difficulty of such problems. Two kinds of approaches help overcome this general problem, but neither is a completely satisfactory solution: models of mechanisms (or, equivalently, causal chains), and models of single components.

*Tracing causal chains:* This is exactly the same principle we have already discussed in reference to rare events, and in the chapters on causation and correlation. The atomic bomb, for example, was not discovered by accident, rather it was predicted from knowledge about radioactive disintegration products and energies for specific Uranium and Plutonium isotopes. In this case, an explosive chain reaction results when the sum of many individual fissions reaches a critical threshold.

*Models of single components:* In many other cases, complex interactions can be anticipated by first looking at one or more of the ingredients separately. The driving force in gunpowder and flash powder is an oxidizing chemical. Although explosions will result from specific combinations of ingredients, the oxidizing agent is capable of sustaining combustion with a much wider range of ingredients, so it becomes a simple matter to explore different combinations to optimize the rate of combustion.

# Humans Make Difficult Experimental Subjects

There are many problems facing humans which could be ameliorated using the scientific method except that the "ideal" experiments cannot be conducted because they involve humans -- they are unethical, too expensive, or just impractical. Consider how you would react if you discovered that the government or your employer had exposed you to high doses of radiation without your knowledge, or had tested drugs on you without your approval. These kinds of manipulations are routinely performed on non-human organisms, but we do not permit them to be conducted on ourselves, even when such manipulations could be most useful in solving an important problem.

Second, some manipulations with humans that are not unethical are nonetheless not feasible. Studies requiring humans to voluntarily change their behavior (for example by adopting a particular diet) pose the obvious problem that the subjects may not comply with the regimen. If the manipulation calls for an extreme change in behavior for a long period, the experiment is probably not feasible.

Experimental studies of human behavior constitute a gray area in terms of ethics. Most of us would likely frown on an experiment that involved teaching children to fear common, harmless objects. And many people would object to being "experimented upon" without being informed of this. Yet that is essentially what advertisers and many other businesses do when they gather data on the effectiveness of different product promotion techniques. When an advertising agency generates two versions of the same ad and compares the sales they generate, it is performing an experiment on its customers. Some ads are designed to create an aversion against a competitor's product, in the same spirit as the psychologists who taught the little girl to fear white rats. Furthermore, the customers do not know they are involved in the experiment, and quite probably do not know that such experiments are regularly done. Businesses not only attempt to discover our preferences and dislikes, but they attempt to alter our behavior in ways that benefit them --- actually teaching us to enjoy their products and dislike others. The very basis of capitalism, by which some products survive and others fail, is itself an ongoing set of experiments in human behavior.

# CHAPTER 23: BIOLOGICAL CORRELATES OF BEING GAY: BIOLOGICAL DETERMINISM?

# 23

IMPEDIMENTS TO SCIENTIFIC PROGRESS

Scientists are currently uncertain as to whether homosexuality is primarily caused by environmental or genetic factors. The uncertainty arises because much of the available data involves correlations, human behavior is likely affected by interactions (complexity) and humans are VERY difficult research subjects when it comes to sex. But evidence is pointing to a combination of genetic and environmental factors influencing sexual preference..

# Introduction

Some of the most profound questions about humans address our behavior. From culture to crime, science to society, we often want to know whether our destiny and who we are is in our genes or a matter of choice and determined by our environment. Our parents went to great lengths to teach us and equip us for this world, but many of those efforts would be wasted if it was all in our genes. To some extent, genes and environment are inseparable. But when it comes to why some of us are scientists and others artists, or other differences in behavior, we can at least attempt to partition genetic effects from environmental ones. As an analogy, if we want a hunting dog to retrieve birds, there are some obvious breeds to choose and train, but many breeds will be worthless for this task no matter how much they are trained. Behavior of humans is more challenging to understand than that of other animals, but some of the same principles can be applied to both humans and other animals.

The question of whether a behavioral difference among people is due to genetic differences or environmental differences is known as the question of biological determinism, often phrased as “is it in our genes?” or as “nature vs. nurture.” Biological determinism is relevant to many issues of social relevance:

- crime
- obesity
- mental illness
- IQ
- addiction
- risk-taking

There are several reasons that answering the question of biological determinism is intrinsically difficult. One is that humans make difficult subjects, especially limiting the kinds of experiments that can be done. This limitation in turn means that we have to rely on correlations for much of our conclusions.. Finally, many of our behaviors are due to a combination of many factors, with possible interactions (complexity). Consequently, there has been little resolution of any of the nature-nurture debate.

Your generation is perhaps unaware of the dark side of biological determinism that was manifested in the past. The most notorious abuse was the eugenics program of Germany's Nazi regime in the 1930s and 1940s. The eugenics program was used to justify genocide of the Jews as well as the killing of gypsies and homosexuals and others deemed socially inferior. The U.S. was never that extreme, but throughout much of the 1900s, the eugenics movement led to sterilization (castration) of those deemed mentally inferior. The motivation for eugenics was to improve the gene pool, and castration was done to prevent the inferior individuals from having babies. Yet, if the basis of their 'inferior' minds was environmental – perhaps nothing more than a lack of education – then preventing them from having offspring did nothing to improve the gene pool.

In this chapter, the focus is on sexual preference. A person's physical sex depends on gonad type (testes, ovaries) as well as secondary sexual traits, such as genitalia (penis versus clitoris and vagina), breast development, facial hair, and so forth.. A person's sexual preference is measured by whether they prefer to have sex with someone of the same sex (= homosexual preference, known as 'gay' if male and 'lesbian' if female), or whether they prefer someone of the opposite sex (= heterosexual or straight). A separate but related behavior concerns gender identity, which is whether a person thinks of himself/herself as man or woman. Gender identity can be fully separated from sexual preference. Individuals get sex change operations because of issues with gender identity, but men who get changed into women will sometimes have sexual preferences for women (for example).

Being gay/lesbian has social consequences, especially to the individuals with the homosexual preferences. It is estimated that 2%-5% of men are gay, 1%-2% of women are lesbian, and these percentages appear to hold across cultures, as best one can tell. Despite the relative abundance of these behaviors, many states have passed referenda that disallow same-sex marriages. On a more individual level, 'gay-bashing' has led to many deaths and less brutal beatings, based entirely on widespread intolerance of homosexual preferences. Perhaps for all of these reasons, there have been many searches for correlates of gay/lesbian behavior that might provide some clues to what determines it. In the recent past, and thus likely still, it has been commonly thought that homosexual preference is a choice and even learned: a 1970 U.S. survey found that 43% thought that young gays learned their SP from older gays. Thus, if we can find anatomical or physiological correlates of being gay, we may at least settle the question of whether sexual preference is learned or a 'choice.'

Overall, there are a few patterns of association with homosexual preference, but they are weak (hence demonstrated only statistically) and prone to poor repeatability. This topic is so far a difficult one to research because of the lack of decisive patterns. Nonetheless, there is collective support for both environmental and genetic causes.

# Genetics

Several lines of evidence suggest a weak-moderate genetic component to sexual preference. A genetic basis is especially difficult to establish for human behavioral differences, both because we don't do experimental crosses with people and because there is so much parental influence on behavior that confounds environmental effects with genetic ones. One of the most useful comparisons therefore makes use of identical twins versus non-identical (fraternal) twins. Identical twins are genetically the same, so any difference between a twin pair must be non-genetic (environmental). Fraternal twins are genetically related but not identical. Both kinds of twins share the womb and are the same age, so they experience many environmental similarities that might be thought to affect behavior. As a consequence, if identical twins more often have the same type of behavior than fraternal twins, we suspect a partial genetic basis to the behavior. If the behavior was 100% due to genes, two identical twins should always have the same behavior. And if there is no genetic basis to the behavior, then identical twins should no more often be similar to each other than fraternal twins.

The twin data show that identical twins have about 50% concordance for sexual preference in some studies, 30% in others. Fraternal twins have nearly half this concordance. So these data suggest that there is a modest effect of genetics. Other data, using a combination of molecular techniques and pedigrees, suggest that an X-linked gene or region influences sexual preference, but that finding has not been confirmed in all careful studies.

---

## Miscellaneous Correlations

**Fraternal birth order:** The probability that a man has homosexual preference increases with the number of older brothers he has. Each older brother increases the odds by  $1/3 - 1/2$ . This effect cannot have a genetic basis. Speculations for this effect focus on the mother progressively building antibodies against an unknown male protein, more so with each son.

**Finger length ratio:** The ratio of the index finger to the 4th finger is higher in women than men. In people with homosexual preference, there is a tendency for the ratio to be lower than in heterosexuals of the same sex. By this criterion, homosexuality is associated with overmasculinization.

**Childhood gender non-conformity:** Children that fail to conform to standard childhood gender roles (such as 'tomboy' girls and effeminate boys) have a higher incidence of adult homosexuality than children that conform to standard gender roles. This kind of study is difficult to do properly (prospectively). The danger of doing this kind of study retrospectively, after sexual preference is already known, is that there will be a biased tendency to selectively recall instances of childhood behaviors that fit the adult outcome.

**Otoacoustic emissions (OAE):** Our ears actually make sounds, though they are too weak to hear by ear. They have a characteristic frequency, starting in early childhood. The right ear's OAE is different from the left ear's, and males differ from females. A UT researcher (Dennis McFadden) is finding that gay males have slightly different OAE frequencies than heterosexual men. The direction of the difference supports an overmasculinization of gay men.

# Neuroanatomy

There is a strong temptation to think or hope that a behavioral difference as strong as the difference between homosexual and heterosexual preferences will have a physical manifestation in the brain. There is an increasing number of techniques that can be performed non-invasively on a live person, but many of the most direct assays, and those that can be applied to microscopic regions of the brain, require actual brain material. There are obvious difficulties in obtaining sufficient material for those studies, and work repeating any findings work is rare. Studies have reported the following, but one should not consider any of the patterns as demonstrated beyond reasonable doubt.

Interstitial nucleus of the Anterior Hypothalamus #3. In 1991, Simon LeVay reported a search for sexual preference differences in the size (volume) of 4 brain nuclei in a brain region known as the anterior hypothalamus. Work on rodents had demonstrated that this brain region affected sexual behavior, and work on humans had already identified a male-female difference in tiny regions or 'nuclei' of the anterior hypothalamus. LeVay found a difference between heterosexual and gay men in one of these nuclei (#3); the size of INAH3 in gay men was similar to that of (heterosexual) females and smaller than that of heterosexual men.

Other studies have reported brain differences associated with sexual preference. The regions involved (in different studies) have been the suprachiasmatic nucleus, the mid-sagittal plane of the anterior commissure, and the isthmus of the corpus callosum.

As mentioned above, the biological basis of sexual preference in humans is a research area in which it will be some time before we have definitive answers. The problems are

1. humans are difficult subjects
  - a. they do not readily divulge these sexual preferences
  - b. the experiments done are limited
2. because of (1), we have only correlations to work with
3. possible complexity: there appear to be many factors influencing sexual preference, none of which are strong (we don't know about interactions, but they may exist).

## CHAPTER 24: CONFLICT AND THE CORRUPTION OF SCIENCE

# 24

*IMPEDIMENTS TO SCIENTIFIC PROGRESS*

The scientific method can be and often is deliberately frustrated by those who would not benefit from an improved model. Not everyone has the goal of objectivity.

# Goals: Seek Not the Truth

The previous chapters have assumed that the goal is to seek scientific "truth" about some problem, hence identifying useful models. For many people, real life goals are made up of many factors, not just to seek scientific truth. (Or, if someone's goal is to seek scientific truth, they may want to keep others from finding that truth.) These factors should be obvious to you, because you have experienced them and can relate to them.

# Factors Affecting a Person's

## **I. Material gain: money, time, objects, power.**

Much of what people do is for tangible, material wealth or resources that lead to wealth. Our economy is based on such a resource. Such resources are at the root of many goals, and one person's gain of the resource may require concealing the truth from others. As will be noted below, companies go out of their way to conceal defects in their products, because such "truths" would deter many buyers.

## **II. Emotions**

*(com)passion:*

People's anger, love, sympathy, and parenting instincts often mean that they are not interested in knowing the full truth. The video about facilitated communication showed parents who wanted to believe that their autistic child could communicate; lovers are often blind to the truth about their mates; parents do not want to know what their children really did, if it is bad.

*ego:*

A person's reputation may influence her/his views. People are especially averse to admitting their own errors and fallibility, so much of progress is blocked by people refusing to change their original claims. One of the most egregious examples of ego came out of the history of medicine. In the mid 1800s, a physician named Semmelweis discovered that physicians and medical students at a maternity ward in Vienna were causing the deaths of the women: the medical staff was moving from autopsy room to maternity wards without washing their hands, and the live patients were becoming infected. He did an experiment that showed a dramatic effect of washing hands (dropping the mortality rate from 10% to less than 2%), yet his publications and advice were rebuffed by his colleagues because they did not want to believe that they -- the DOCTORS -- were killing their patients.

### **III. Philosophy**

Religious beliefs have stood in the way of many scientific advances. But some people just want to believe (or disbelieve) in things for no obvious reason. For example, many people believe in psychic powers, even though science has universally failed to support the existence of such powers.

### **IV. Politics**

Political beliefs and perhaps even laws may stand in the way of scientific progress. Laws may protect an individual's rights so that a particular kind of study cannot be undertaken -- there are certain types of data that we cannot obtain without permission. A recent proposal has even been offered in the U.S. to ban the use of any data on humans that was obtained in violation of U.S. ethics standards, even when such data were obtained outside of U.S. soil and without U.S. funding. And political views may simply prevent dissemination of scientific results. A sordid and aging example was the Lysenko era in the Soviet Union, when genetics was essentially outlawed as a science, and geneticists were put in prison. The influence of Lysenko set Soviet biology and agriculture back several decades. Lysenko's views were received favorably by the Communist regime because the Communist ideology held that everyone was equal; genetics, instead, allowed intrinsic differences to exist among individuals. On a more innocuous level, politicians in Congress can have large effects on the funding of different projects, merely because of their power to make important decisions. There are instances in which a Congressman or Senator has been able to direct funding of projects that are immediately important to them rather than the population (e.g., research into Lyme disease). Likewise, factors that may impact a large number of people but do not impact a politician's constituency can get ignored (e.g., multi-drug resistant TB).

When various factors enter into goals, it is often or invariably the case that conflict arises. Different people are then typically basing their goals on different factors, hence their goals differ or conflict. When conflict is present, progress will often be impeded or fully prevented.

# Arenas of Conflict

Our society is a collection of individuals and corporations with differing goals (e.g., Table 17.1). When people have conflicting goals, they don't all necessarily benefit when a model is improved. We cannot begin to reveal the ubiquitous presence of conflicting goals. To do justice to this topic we would have to review much of what is known about the human race, including history (e.g. wars), sociology (e.g. class struggle), and economics (competition between firms for the same market). Wherever we turn, there are individuals or institutions with conflicting goals. These conflicts are widespread, and it is not surprising that they have had a huge impact on science. Some poor science is certainly a consequence of ignorance. And as discussed in the correlation chapter, some poor science is certainly a consequence of the cost and difficulty of doing the science correctly. But many of the abuses of science can be traced to conflicts: If two people have conflicting goals, then it is to the advantage of each person to slow the progress of the other person toward their goal.

## **Conflict and the potential for bias in science:**

### **Expert witnesses in criminal cases:**

Scientists that testify for the defense or prosecution are often paid for doing so. At the same time, they are sworn to give accurate testimony.

### **Medical conferences:**

Consumer Reports discussed how pharmaceutical manufacturers that sponsor conferences bias the presentations at these conferences by choosing the conference speakers from a list of scientists previously known to have a favorable view toward the company's products

### **Environmental consulting:**

The engineering firm that renders an opinion about the environmental effects of the latest subdivision faces a conflict of interest between doing an unbiased review and doing a review that will ensure additional business in the future.

### **Funding for big ticket projects:**

Scientists explaining to the public what the benefits of projects such as the space station and the human genome project are may have incentive to overstate their case. Research funding for their project may depend on the perception by the public that the project is important, as evidenced by the death of the superconducting supercollider.

### **Furthering political dogma:**

When scientists decide before gathering any data what the best solution to a problem is, the data and analyses may be selectively presented so as to support an a priori decision about what the researcher wants to show.

There are some generalities about where to expect conflict, however. Some major categories include:

1. The buyer-seller interaction: the seller is often interested in the buyer's money, not whether the buyer is getting what is wanted.
2. The criminal court system: the prosecution and defense have opposite goals by constitutional definition.
3. Politics: the opponents for a political office have opposite goals
4. Industry and government regulation: regulation is often a situation of conflict, since the government uses its regulations to impose burdens on corporations. In some cases, however, these regulations level the playing field among companies, rather than taking away profits.

This list is not exhaustive, but it gets at several of the main categories. Some of these are obvious. Below we expand on conflict in the legal system, because it offers an interesting contrast with scientific discovery in several ways. We end the chapter with a description of a subtle form of conflict that has not entered into any of the examples above: tragedy of the commons.

# Conflict over Science and the Legal System

In principle and in practice, the U.S. legal system is a triumph of individual rights. Yet its nature, and indeed its success, is founded foremost on conflict. The introduction of scientific evidence into the courts thus exposes science to many potential abuses.

Consider a criminal trial. The purpose of the trial is to decide on the guilt or innocence of the defendant. In a case involving scientific evidence, there are at least 4 parties involved, each with different goals:

<b>PARTY</b>	<b>GOAL</b>
A. Prosecution	Conviction
B. Defense	Acquittal
C. Jury	A fair presentation of evidence
D. Science Lab	future contracts, reputation for quality

There is an obvious conflict between (A) and (B). But conflict may also arise between (D) and (A) or (D) and (B), because the goal of the lab might be to (i) support whichever agency hired it, or (ii) present a fair appraisal of the evidence, which will put it in conflict with whichever side is hurt most by the evidence.

A fundamental difference between science and our legal system. Now it might seem that, despite the conflicting goals, there is little room to argue about scientific evidence. But therein lies perhaps the most basic difference between the legal system and the scientific method. With respect to the scientific method, there are three possible outcomes that might be reached about the defendant based on the evidence at hand:

**Definitely Not Guilty | Not clear | Definitely Guilty**

The approach in science is to seek more evidence (additional data) to narrow the middle category. If more experiments/evidence are not feasible, a scientist merely states the ambiguity. But the legal system does not readily allow for these three possibilities - there are only two allowed outcomes, guilty or not guilty. (A hung jury represents a type of indecision, but judges sometimes coerce hung juries into rendering a verdict by refusing to let them go, thereby refusing to accept this "middle ground".) The task of the defense is thus to compress the middle category by convincing the jury that this ambiguous evidence contributes to the suspect's innocence, whereas the task of the prosecution is to compress the middle category and convince the jury that this ambiguous evidence contributes to the suspect's guilt. Uncertainty at some level applies to virtually all scientific data (some more-so than others), creating a nearly universal problem for the fair evaluation of scientific evidence in courts.

# Tragedy of the Commons: Conflict Between the Social Welfare and Individual Benefit

Until now, we have discussed cases in which the conflicting goals are obvious -- prosecution versus the defense, consumer versus a company selling a product. There is a more subtle conflict that pervades perhaps all attempts at providing social order -- government, and collective decision-making. The "tragedy of the commons" is a metaphor (verbal model) for a conflict that arises when apportioning resources held by many among the owners. It applies especially to public resources, but also can apply on a smaller scale (e.g., to multiple owners of a business or other resource). More generally, it is a model of conflict between what's good for the individual and what's good for the group.

Imagine that we have a single big dispenser of liquid for a class of thirsty students. Everyone is given a big cup and told that there is enough drink for everyone to have  $\frac{2}{3}$  of their cup full. We let everyone line up and pour drink into their cups, but no one can see how much another person takes. What will happen??

No doubt, the drink will all be used up before the last people in line get their chance at it.

The phrase "tragedy of the commons" was coined several decades ago by Biologist Garrett Hardin to describe the basic conflict between individuals and groups of those same individuals. In colonial days, a "common" was a public grazing area. (The Boston Commons still exists, but it is now a park.) Residents of the community were told how many head of livestock were allowed on the commons, but there was no real way of policing these quotas. Invariably, the commons was overgrazed and became useless as a grazing area.

To appreciate the problem underlying the tragedy of the commons, consider a resource owned equally by each of 10 people. By rights, everyone owns 1/10 of it. Let's say that I am one of those owners. If I take 2/10 instead of my 1/10, I get double my share. The other 9 owners must divide the remaining 8/10, so on average, each of them gets 8/90, which is only a little bit less than 9/90 (1/10). So my gain is big, but their individual losses are small. Of course, if the other 9 also behave like I do, then a few owners reap most of the benefit of this "communal" resource, and most owners lose. The other owners will object to my taking more than my share (or anyone else doing the same, unless it is them personally), so the success of such a communal resource relies on keeping track of who takes more than their share. Only if there is some way of policing or enforcing the division of resources can this tragedy be avoided.

There are many social institutions that suffer from tragedy of the commons (see table below). The success of minimizing the effects of this conflict depends on both the individual benefit of acting selfishly and the mechanisms to reward compliance with social goals or to punish non-compliance.

TOPIC	PUBLIC BENEFIT	INDIVIDUAL COSTS AND BENEFITS
Taxes	schools, road, defense, social infrastructure	by avoiding taxes, you enjoy the public infrastructure and have extra dollars to spend (because of the strong individual benefit of not paying taxes, we need a powerful tax collection agency).
Clean Air Autos	clean air	by not maintaining a clean-air car, you avoid costly repairs but you breath the same air as everyone else.
Fisheries	food supply, jobs over many years	each boat makes more money as it catches more of the fish in the public fishing grounds, but the reduced fish populations affect all fishing boats equally
Grades	society is able to evaluate student training	higher grades benefit a student; one student's grades do not affect the average
Letters of Recommendation	society is able to evaluate the quality of a job candidate with accurate letters	letter writers often have agendas for the former associates, and an honest appraisal of the candidate may be in conflict with that agenda.
Traffic Laws	an orderly system of transportation is vital to our society	following traffic rules often increases our transit time.
Teaching	it is essential to train people for jobs in the economy	universities offer little reward for outstanding teaching but offer big rewards for outstanding research, and teaching detracts from research

# Vaccines and Public Health

Unquestionably, the biggest medical advances ever are microbe-fighting drugs and vaccines. Even into the 1940s, infectious diseases took a major toll on Americans (and they still are major health problems outside of Western countries). A stroll through a cemetery from the early part of this century will reveal that death then was not confined to the elderly, as it chiefly is now. Much of this early death was from scourges that we no longer fear or, in some cases, that we even lack any familiarity with: pneumonia, TB, diphtheria, tetanus, whooping cough. Other diseases, such as polio were more often crippling than lethal, but nonetheless caused annual epidemics (the President FDR had been crippled by polio).

To give a perspective, 38,000 cases of polio were reported in the U.S. in 1954. That was one of the first years of a trial vaccine for the disease, caused by a virus. Now, the number of annual cases in the U.S. is less than 10, and those are usually caused by the vaccine. Other diseases, such as diphtheria, whooping cough (pertussis), measles, have likewise all but disappeared from their former, frightening levels. These diseases in particular, have been reduced to a tiny fraction of their historical levels because of vaccines. Most vaccines consist of killed or otherwise subdued microbes or parts of microbes that are injected or eaten by us to trigger our immune system to combat the real microbe. (Vaccines differ from antibiotics: antibiotics are drugs that kill the microbe directly and only work on bacteria. Vaccines trick our bodies into killing the invading microbe, which may be a bacterium or virus; some vaccines merely help our immune system block a toxic chemical.)

Vaccines offer a remarkable level of protection. They are so successful, and their efficacy is so accepted by the medical establishment and public that many of them are required for admittance to public schools. However, vaccines carry a small risk of complication, and when administering them to millions of people, a handful of those people (in some cases, a "big" handful) will develop problems, or rarely, will die.

What are the reasons for being vaccinated? There are two, and they are very different.

1. **Personal protection:** Many people get vaccinated to avoid getting a disease. This is the obvious benefit.
2. **Group benefit or herd immunity:** When a large fraction of the population is vaccinated (over 50% in most cases), the risk [to an unvaccinated person](#) of getting infected goes down. In other words, if everyone else in the population but you is vaccinated, then your chance of getting infected is tiny or zero because there is no one around to infect you. (This argument applies to diseases that are transmitted from person-to-person but not to those acquired from animals or the soil -- such as tetanus.)

Why does society require vaccination, then? Why shouldn't we let vaccination be a matter of personal choice? If someone wants to neglect vaccination, they do so at their own risk. That seems fair and seems consistent with U.S attitudes about individualism, provided we make the vaccines affordable and available to everyone. (We also need to inform the entire population about which vaccines are available and what the risks are.)

There is a compelling argument for mandatory vaccination: [the herd immunity afforded by a vaccinated population protects some groups of people that cannot be immunized](#). One group is babies. Their immune system needs time to develop, and although they have protection from mom's antibodies for a few months, there is a window of vulnerability before vaccination in which they are susceptible. A second group is more ill-defined but very real: most vaccines do not afford 100% lifetime protection. Some people just don't ever develop a good immunity, and in others, the immunity decays with time. There is no easy way to identify these people (e.g., to give them booster inoculations). In addition, the elderly and people with compromised immune systems may have difficulty in being protected by vaccines. Babies and non-immunized adults who have received the vaccine thus benefit from living in a population that is fully vaccinated. The argument in favor of mandatory vaccination is thus that the entire population benefits when everyone gets the vaccine.

There is a movement afoot that opposes mandatory vaccination. What would happen if we did not require these vaccinations? The answer is pretty clear that many people would not get the vaccinations, and we would have recurrent epidemics of these old diseases. Some people would certainly continue to get vaccinated (for individual reasons), but others would not on the grounds that their chance of getting the disease was small enough that vaccination was not worth the risk. They, in fact, would benefit from the many people who did get vaccinated, but as a group, they would be the ones who fueled the epidemics. Many "innocent" people would die, as would many others who were just negligent about getting vaccinated. In fact, Texas had a measles epidemic in the 1980s that killed at least a couple adults. It was suggested that this epidemic resulted from a high influx of people from Central America, who had never been vaccinated, but the epidemic affected many U.S. citizens as well. A second case is an outbreak of whooping cough in Northern Idaho in about 1996. Here, at least one young child died. The title of an editorial in an Idaho paper at the time accused parents who did not vaccinate their children as killing the children of other parents.

The vaccination example illustrates the tragedy of the commons. People make decisions for individual reasons but not for the good of society. In this case, the tragedy is avoided (a group benefit is maintained) by requiring individual compliance. However, influenza kills tens of thousands of Americans each year, yet the vaccine for that disease is not required. Influenza differs from the "childhood" diseases for which we have mandatory vaccination in that the vaccine does not confer life-long immunity. In fact, the vaccine for influenza is changed every year. It may be this need for annual vaccine renewal that makes it impractical to require population compliance.

# CHAPTER 25: DELIBERATE BIAS: HOW CONFLICT CREATES BAD SCIENCE

# 25

IMPEDIMENTS TO SCIENTIFIC PROGRESS

When proper scientific procedure is undermined by conflicting goals so that it results in deception, we say it is biased. This form of bias is prevalent in advertising - companies universally advocate their products, emphasizing product assets while concealing product faults and concealing the advantages of competitor products. You don't expect an advertisement from a private company to offer a fair appraisal of the commodity. But bias even exists in the way that states promote their lotteries by advertising the number of winners and money "given away" without telling you the number of losers and money taken in.

# Introduction

Deliberate bias occurs in science as well as in business and politics. The potential for bias arises when a scientist has some goal other than (or in addition to) finding an accurate model of nature, such as increasing profits, furthering a political cause, or protecting funding. As an example, the environmental consulting firm that renders an opinion about the environmental effects of the latest subdivision may not give a completely accurate assessment. The future work that this firm receives from developers is likely to depend on what they say. If the developers don't like the assessment, they will likely find another environmental next time. Thus there is a conflict between obtaining an unbiased model, and making the largest profit possible. In these cases, the scientists are motivated to present biased arguments. There are many situations which present a conflict between scientific objectivity and some other goal.

# Drugs and Medicine

We take for granted an unlimited supply of medicinal drugs. If we get pneumonia, gonorrhea, HIV, or cancer, the drugs of choice are invariably available. They may not cure us, but whatever drugs we know about (that have been approved) are in abundant supply.

For the most part, this abundance of and reliance on drugs comes from public trust of health care. We don't imagine that our doctors try to prescribe us useless or unnecessary drugs (if anything, the patient often requests drugs when they are unnecessary). But in reality, many drugs ARE unnecessary, and some drugs are no better than cheaper alternatives. The Food and Drug Administration (FDA) is charged with approving new foods and drugs for the U.S. In 1992, it was approving about 20 new drugs a year but regarded only about 20% of those as true advances. So many drugs are no better than alternatives (many are obviously at least slightly worse than alternatives). And physicians often don't have the evidence to know which drugs are best.

The goals of consumers are in conflict with those of drug companies in some respects. The consumer wants drugs that are cost-effective safe with few side-effects. If two drugs are equally effective, we want the cheaper one. We may even not want the most effective drug if a cheaper one will do the trick. But the goals of any drug company are different:

- company sales, which may involve
- company reputation
- a successful treatment but not necessarily a cure
- low risk of liability claims (which may include drug safety and low side-effect)

It costs to hundreds of millions of dollars to get a drug approved by the FDA now. Much of the cost is in research and trials, but even FDA consideration itself costs millions. So it is not cheap. Most important is time, because the sooner a drug hits the market, the sooner the company reaps the benefits. So drug companies have strong incentives to market any product that is approved by the FDA -- once approved, the major costs of money (and time) have already been borne.

Of course, it does not behoove a company to market a harmful product -- liability costs can be quite high. But most products that pass all the hurdles of FDA approval can be regarded as harmless at worst. The drug company has a very strong incentive to market its approved products regardless of whether the consumer benefits or not. One of the most economically successful drugs ever was an ulcer medicine that reduced suffering. It did not cure ulcers but instead had to be taken as long as the patient had the ulcer. Research later found that most ulcers were caused by a bacterium and that treatment with antibiotics cured the ulcer. So the original ulcer treatment was based on a misunderstanding of the cause of ulcers.

There is no shortage of scientific method abuses in the drug industry. Publications of books like *Overdosed America* (J. Abramson 2004, HarperCollins Pub. Inc) and *Should I be Tested for Cancer?* (H. G. Welch 2004, U. California Press) provide a wealth of examples in which drug companies have gone to great lengths to subvert the scientific method in gaining FDA approval or to market drugs to physicians and the public. As an indication of the general problem of bias created by the conflict between corporate versus public goals, it was found in 2003 that commercially sponsored studies are 3.6-4 times more likely to favor the sponsor's product than studies without commercial funding. Also in 2003, another study found that, in high quality clinical trials, odds of recommending the new drug were 5 times greater in studies with commercial sponsorship than in those funded by non-profit organizations. Some flagrant examples follow.

**Broaden the market:** Something that has been shown to be effective in a narrow segment of the population gets recommended for a wider segment even when the evidence opposes its benefit to the wider segment or the benefit is at best extraordinarily expensive. Examples: defibrillators, statin drugs.

**Ignore alternatives:** Companies of all types do not benefit from sales of competing products. In medicine, it is often the case that new drugs/methods are no better than old ones, but the benefit of old methods/drugs are often neglected (when a company's patent for a product has expired, it has less incentive to defend that product against a newer alternative). Examples: the benefits of modest exercise and diet changes are often demonstrated to be more effective in reducing heart attack rates and even cancer rates than many drugs, but those alternatives are not mentioned to patients when being sold the drug. In a second example, a trial with Oxycontin (similar to heroin) used a control group with no pain killer. Not surprisingly, Oxycontin was found to be superior to no drug in reducing pain. Had the control group used an older painkiller, Oxycontin might not have been superior.

**Comparisons confounded by dose differences:** The antacid drug Nexium (whose patent was not expiring) was tested in trials against the chemically similar Prilosec (whose patent had expired). Rights to both drugs were owned by the same company, but with expiration of the Prilosec patent, sales of it would no longer be highly profitable. Nexium proved to be more effective in reducing acid reflux in this comparison trial. However, the dose of Nexium tested was twice that of Prilosec, even though Prilosec is commonly taken at the higher dose.

**Testing the wrong age group:** Many studies choose subjects who are younger than the expected users of the drug. For example, only 2.1% of all patients in studies of anti-inflammatory drugs were over 65, even though those over 65 are among the largest users of those drugs. Why? Drug side-effects are less likely to arise in younger patients. Examples: Aricept (for Alzheimer's) and cancer drugs.

**Selective data release:** Drug companies rely on researchers to conduct trials and, importantly, publish the results. Drug companies routinely restrict access to the data, so the researchers publishing the study don't see all the results, only the favorable ones.

**Ghostwriters:** The authors of a study are not necessarily the ones who write it. Companies will hire non-authors to write the first draft, which then gets passed off to the official authors for approval. Companies choose ghostwriters who know how to spin the study in the most favorable light.

Other tricks include:

- Drug companies have paid for university research on their products and have then blocked publication of unfavorable results and/or cut continued funding of the work when the results began to look bad.
- Pharmacy sales people routinely visit physicians at work, offering them free lunches, free samples of medicines, gifts, information to promote their products, and notepads and pens with company logos. (Next time you visit a physician, look around the inner rooms for evidence of company logos on charts, pens, pads, and so on.)
- To maintain their licenses, physicians are required to take courses in continuing medical education (CME). These courses are often sponsored and paid for by drug companies in exotic locations and with hand-picked speakers who provide favorable coverage of company products.
- Drug companies publish ads in medical journals that look and read like research articles. These ads promote products.

# DNA

A second, well-documented case in which conflict is manifested is over DNA typing. These examples may not reflect current debate over DNA technology, but one should use them to appreciate the strong potential for conflict over any scientific issue in the legal system.

The rush to implement DNA typing in the U.S. criminal system was done before guidelines were set for proper DNA typing procedures. Consequently, there were varying levels of uncertainty in the use of these methods by law enforcement agencies and commercial labs. They were also reluctant to admit the uncertainty. The manifestation of conflict over DNA evidence was thus heated and surfaced in the popular press on several occasions. We introduce this material in a prospective manner - by first imagining how the prosecution, defense, and forensic lab can be expected to behave to achieve their goals in ways that are contrary to fair scientific procedure. You have already been exposed to the nature of the models and data in DNA typing, so now the issue is how the different legal parties deal with the problems in evaluation and ideal data.

If a case that uses DNA typing has come to trial, we can assume that the DNA results support the prosecution's case. There are thus three parties in conflict:

Prosecution <<< (in conflict with) >>> Defense <<< (in conflict with) >>> DNA lab

We can assume that the DNA lab's results support the Prosecution's case, or there would not be a trial, so the conflict will lie between the Defense and the other two agencies. Now consider how this conflict might be manifested.

I) What might the prosecution do to improve the chances of reaching its goals?

1. eliminate test procedures that benefit the suspect (eliminate standards; build a case of circumstantial evidence to shield criticism of the DNA fingerprint evidence).
2. harass or impede prior witnesses for the defense who might testify in the future.
3. keep a list of sympathetic expert witnesses
4. Maintain positive relationships with labs that have contributed to convictions in the past.

II) With respect to the errors and uncertainties of DNA evidence in specific cases:

5. argue that inconsistencies are plausible artifacts
6. fail to question any outcome in its favor

III) How is the defense likely to behave?

1. keep a list of sympathetic expert witnesses
2. emphasize all inconsistencies in the DNA analyses as evidence of innocence
3. question all assumptions, original data, calculations

IV) How is the lab likely to behave?

1. Produce results that enhance the goals of their economic benefactor, including
2. establish a reputation for a lack of indecisiveness; overstate case
3. defend its initial conclusions

### **Evidence from DNA cases:**

The case histories available from the last 4-5 years of DNA forensics verify many of these expectations. In particular, DNA testing has omitted such basic elements as standards and blind procedures (I.1 above); the prosecution in the Castro case ignored inconsistencies in the evidence (I.5, I.6); the lab in the Castro case overstated the significance of a match and defended such practices as failing to include a male control for the male-specific probe (III.2, III.3). Defense and prosecution agencies definitely keep lists of sympathetic witnesses (I.3, II.1), and defense agencies indeed choose witnesses to challenge the nature of DNA evidence based on its (necessarily false) assumptions (II.3). And finally, harassment by the prosecution of experts who testify for the defense is well documented, both in the courtroom and outside (I.2). This harassment includes character assassination on the witness stand, implied threats to personal liberties for witnesses who were in the U.S. on visas, and contacts made to journal editors to prevent publication of papers submitted by the witness. Some of these cases have been described in the popular press, and others are known to us through contacts with our colleagues. These latter manifestations of conflict don't make sense in the context of a single trial, but they stem from the fact that networks exist in which prosecuting attorneys share information and parallel networks exist among defense attorneys. Expert witnesses are often used in multiple trials across the country, so any expert witness who is discouraged at the end of one trial will be less likely to participate in future trials, and conversely, an expert who does well in a trial may be more likely to participate in future trials.

The suggestions of harassment have even extended to scientists who merely publish criticisms of forensic applications of DNA typing. In the 1991-92 Christmas break, two papers were published in Science on opposites of the DNA fingerprinting conflict (Science 254: 1745-50, and 1735-39). At the same time, news items were also published in Science, Nature, The New York Times, and The Washington Post, in which the details of this conflict were aired in full detail. The authors of the paper opposing use of current DNA typing methods (Lewontin & Hartl) were phoned by a Justice Department official and allegedly threatened with retaliation (having their federal funding jeopardized); the official denied the threats but did not deny the phone call. The Lewontin-Hartl article had apparently been leaked in advance of publication, and an editor for Science contacted the head editor to have a rebuttal published. However, it turned out that the editor requesting the rebuttal owns a patent for DNA fingerprinting and stands to benefit financially by forensic use of DNA typing methods. The two authors chosen for the rebuttal had been funded by the FBI to work on DNA fingerprinting methods. So, there appears to have been some serious conflicts of interest at least on one side of the issue.

This treatment of conflict in DNA trials has omitted a 4th category of people whose goals may conflict with the agencies above: the expert witnesses themselves. The goals of expert witnesses may be varied, including monetary (the standard rate is \$1000/day in court), notoriety, and philosophical (e.g., some people volunteer to assist the defense on a no-cost basis, merely to ensure a fair trial).

# The Footprints of Bias: Generalities

The attempt to deceive in science may take many specific forms. At a general level, arguments may avoid the scientific method entirely, or they may instead appear to follow scientific procedure but violate one or more elements of the scientific method (models, data, evaluation, revision). The next few sections of this chapter describe different kinds and levels of possible bias at a level that transcends specific cases. These generalities are useful in that they enable you to detect bias without knowing the specifics of the situation.

The standard scientific approach to evaluating a model is to gather data. If you suspect bias (e.g., you doubt the claim of support for a model), the ideal approach is to simply gather the relevant data yourself and evaluate the claim. But this approach requires time that none of us have (we can't re-search everything). In many cases, blatant examples of bias can be detected by noting some simple features of a situation.

## **Look for Conflict of Interest:**

The first and easiest clue to assist you in anticipating deliberate bias is conflict of interest. If another party's goal differs from your goal, and your goal is to seek the scientific "truth", then there is a good chance that that party is biased -- just as you may be biased if YOUR goal differs from seeking scientific truth. Service on many Federal panels requires a declaration of all conflicts of interest in advance (and you will be excused from consideration where those conflicts lie). That is, the government avoids the mere appearance of bias based on the existence of conflict, without looking for evidence of actual bias. However, in our daily lives, we are confronted with conflict at every turn, and we can't simply avoid bias by avoiding interactions involving conflict (e.g., every time you make a purchase, there is a conflict of interest between you and the seller). Thus, being aware of conflict is a first step in avoiding bias, but you can also benefit by watching for a few symptoms of bias.

## Non-Scientific Arguments as Indicators of Bias

Sometimes, someone is so biased that they resort to lines of reasoning and argumentation that are clearly in violation of science. These cases are easy to expose, because they can be detected without even looking at data or analysis. And many of them are already familiar to you, as given in the following tables:

<b>ARGUMENTS IN VIOLATION OF THE SCIENTIFIC METHOD</b>	
Appeal to Authority	Appeal to authority is the defense of a model by indicating that the model is endorsed by someone well known (an authority). A model should stand on its own merits. The fact that a particular person supports the model is irrelevant, though the specifics of what they have to say may assist you in evaluating the model.
Character assassination of opponent	Character assassination is the attempt to discredit someone's character (e.g., point out that they associate with undesirable people, etc.). The character of somebody is irrelevant to the evidence they present that supports or refutes the model. We should evaluate the evidence, not the person presenting it.
Refusal to admit error and rationalize failures	Refusal to admit error is the refusal to specify the conditions under which a model should be rejected or the refusal to accept its refutation in the face of solid evidence against it. All models are false, and anyone who refuses to discuss how their model could be seriously in error is obscuring a fair appraisal of their model (or is using an unfalsifiable model)
Identify trivial flaws in an opponent's model	This violation refers to the practice of searching for unimportant details about a model that are false, and using those minor limitations as the basis for refuting the model. The fact that all models are false does not mean that all are useless. Yet it is a common trick of lawyers to harp endlessly on the fact that a particular model advocated by their opponent is not perfect and thus should be abandoned.

**ARGUMENTS IN VIOLATION OF THE SCIENTIFIC METHOD**

Defend an unfalsifiable model

A model must be falsifiable to be useful. "Falsifiable" merely means that it could be refuted if the data turn out to be a certain way. An unfalsifiable model is one that cannot be refuted no matter how the data turn out. Creationists, for example, adopt and then defend an unfalsifiable model. An unfalsifiable model is one that is framed so that we could never gather data to show it is wrong. By contrast, science is predicated on the assumption that all models will eventually be overturned.

Defend an unfalsifiable model

A special case of defending an unfalsifiable model, this one is subtle. It takes the form of insisting that a class of models is correct until all variations of them have been rejected. As an example, we might refuse to accept that the probability of Heads in a coin flip is 0.5 unless we reject all alternatives to 0.5. Whereas it is possible to refute that the probability of Heads in a coin flip is 1/2, it is impossible to refute that the probability of Heads is anything other than 1/2, because that would mean showing it is exactly 1/2. (It would take an infinite number of flips to reject everything other than 1/2.) This argument also takes the form of claiming that there is some truth to a model until it has been shown that there is nothing to it at all.

Require refutation of all alternatives

Refusal to admit error is the refusal to specify the conditions under which a model should be rejected or the refusal to accept its refutation in the face of solid evidence against it. All models are false, and anyone who refuses to discuss how their model could be seriously in error is obscuring a fair appraisal of their model (or is using an unfalsifiable model)

Scientific-sounding statements

Scientific "buzz words" are often effective tools in persuading a naive audience that a speaker is saying something profound.

**ARGUMENTS IN VIOLATION OF THE SCIENTIFIC METHOD**

Heresy does not imply correctness

Just as some people will believe a conspiracy theory about almost anything, so many of us are sympathetic to the independent thinker who goes against the consensus and discovers something totally new. Unfortunately, most people who defy the consensus are simply wrong. The fact that a person has an off-the-wall or anti-establishment theory to propose is not a reason to assume that they must be on the right track.

Build causation from correlation

It is an easy trick to describe a correlation that appears to defy a causal model, but as we know, correlations can be misleading.

Unexplained is not inexplicable & either-or arguments

A common ploy is to attack a theory (e.g., evolution) on the grounds that it doesn't explain some observations, hence that it must be wrong. (Of course, since all progress is built on model improvement, it is absurd to suggest that a currently accepted model should have to explain everything.) However, when such a tactic is combined with an either-or proposition -- that if theory X isn't right, theory Y MUST be -- it can be an effective way of erroneously convincing an audience to support a useless model.

Use anecdotes and post hoc observations

This category represents a non-systematic presentation of special cases made in defense of (rather than as a test of) a particular model. An anecdote is an isolated, often informal observation made without a systematic, thorough evaluation of the available evidence. As a selected observation, it is not necessarily representative of a systematic survey of the relevant observations. Post hoc observations are observations made after the fact, often to bolster a particular model. It is easy to select unrepresentative data that support almost any model.

Perhaps the most subtle but useful of these points is the refusal to admit error. In science, models are tested precisely because the scientist acknowledges that every model has imperfections which may warrant its abandonment. Someone who is trying to advocate a model may want to suppress all doubt about its imperfections and thus suggest that it can't be wrong. That attitude is a sure sign that the person is biased. Of course, in many cases you will already know that the person is biased (as with a car salesperson), and the best that you can hope for is to determine how much they deviate from objectivity.

# Introducing Bias Before the Study: Controlling the Null Model

As noted in the earlier chapter *Interpreting the data: is science logical?* many evaluations are based on a null model approach: the null model is accepted until proven wrong. To "control" the null model means to "choose" the null model. Choice of the null model can have a big effect on the outcome of even the most unbiased scientific evaluation for the simple reason that a null model is accepted until proven guilty. Any uncertainty or inadequacy in the data will thus rule in favor of the null model. By choosing the null model, therefore, many of the studies testing the model will "accept" it, not because the evidence for it is strong, but because the evidence against it is weak. As a consequence, the null model enjoys a protected status, and it is to anyone's advantage to choose which model is adopted as the null model. Choice of the null model in this sense does not even mean developing it or proposing/inventing it. Given a set of alternatives decided upon in advance, controlling the null model means simply the selection of which model from that set is adopted.

Consider the two alternative models that might be used in approving a new food additive for baby formula:

- a. food additives are considered harmful unless proven safe to some limit
- b. food additives are considered safe unless shown to be harmful

As the null model, (a) requires a rigorous demonstration of the safety of a food additive before it is approved. In contrast, (b) requires that an additive can be used until a harmful effect is demonstrated. As noted in the Data chapters, an enormous sample size might be required to demonstrate a mild harmful effect, so a harmful product could reach market much more easily under null model (b) than under (a).

Choice of the null model represents a powerful yet potentially subtle way in which an entire program of research can be biased. Every other aspect of design, models used, and evaluation could meet acceptable standards, yet choice of a null model favorable to one side in a conflict will bias many outcomes in favor of that side.

# Bias in Experimental Design and Conduct of a Study

The template for ideal data presented earlier is a strategy for producing data with a minimum of bias. But the template can be applied in many ways, and someone with a goal of biasing data can nonetheless adhere to this template and still generate biased data. Let's consider a pharmaceutical company testing the efficacy of a new drug. How many ways can we imagine that the data reported from such a study might be deliberately biased, when the trials are undertaken by the company that would profit from marketing the drug? The following table lists a few of the possibilities.

## Bogus Designs:

VIOLATION OF ACCEPTED PROCEDURE	IMPACT
Change design in mid-course	An investigator may terminate an experiment prematurely if it is producing unwanted results; if the experiment is never completed, it will not be reported.
Assay for a narrow spectrum of unlikely results	The public well being is many-faceted, and a product is unlikely to have a negative impact on more than a few facets. With advance knowledge of the likely negative effects (e.g., a drug causes brain cancer), a study can be designed to purposefully omit measuring those negative effects and focus on others (e.g., colon cancer). Were the subjects a fair sample of the relevant population? The medicine might be more effective on some age groups than others, so the study might be confined to the most responsive age groups (determined in preliminary trials). While the data would be accurate as reported, details of the age group might be omitted to encourage a broader interpretation of the results than is warranted. (One commonly observes a related phenomenon in car commercials – a company pointing out the few ways in which their product is superior to all others.)

VIOLATION OF ACCEPTED PROCEDURE	IMPACT
Protocol concealed	It is easy to write a protocol that conceals how the study was actually conducted in some important respects. For example, was a blind design really used? Although a blind design exists on paper, it is possible to let patients and staff know which patients belong to which groups. Indeed, patients can sometimes determine whether they are receiving the drug or placebo. Were the controls treated in exactly the same manner as the group receiving the medicine? It is possible to describe countless ways in which the control group and treatment group were treated similarly, yet to omit ways in which they were treated differently. The medicine might be given along with some other substance that can affect patient response, with this additional substance being omitted from the placebo.
Small samples	Science often assumes "innocent until proven guilty" in interpreting experiments designed to determine if a product is hazardous. Small samples increase the difficulty of demonstrating that a compound is hazardous, even when it really is.
Non-random assignments	Most studies, especially those of humans, begin with enough variation among subjects that random assignment to control or treatment groups is essential to eliminate a multitude of confounding factors. Clever non-random assignments could produce a strong bias in favor of either outcome.
Pseudo controls	There are many dimensions to the proper establishment of controls, including assignment of the control groups and subsequent treatment of the controls. It is possible to describe many ways in which a control is treated properly while omitting other ways in which a control is treated differently.

There are obviously additional ways to bias studies. For example, the description of drug company tactics at the beginning of this chapter listed several specific examples: **ignore alternatives**, **comparisons confounded by dose differences**, and **testing the wrong age group**. However, it is also important to recognize that bias can creep in *after* the work is done, as addressed next.

---

# After the Study: Biased Evaluation and Presentation of Results

Even when the raw data themselves were gathered with the utmost care, there is still great opportunity for bias. Bias can arise as easily during data analysis, synthesis and interpretation, as during data gathering. This idea is captured in the title of a book published some years ago, "How to Lie With Statistics." Two methods of biasing evaluation are (i) telling only part of the story, and (ii) searching for a statistical test to support a desired outcome. Again, these are a few of the major types of abuses. The description of drug company practices (above) gave some specific examples that are not encompassed by the two cases below: broaden the market, and using ghostwriters to put a favorable spin on the outcome.

## **Telling only part of the story:**

A powerful and seeming 'honest' method of deceit is to omit important details. That is, nothing you say is untrue, it's just that you avoid saying part of the truth. In science you can do the same thing – present only part of the results. For example, a drug that was tested might be free of effects on blood pressure but elevate cholesterol. Telling an audience that there was no effect on blood pressure would be accurate. It's just that they would like to know everything that was found. Of course, advertisers do this all the time – describe only the good aspects of their product. In a court case, the defense will only present the data that they have that tends to exonerate their client. So we kind of expect it. But in science, this is unacceptable. Furthermore, there is a long gradation of omission that runs from innocent to totally dishonest.

The most extreme and dishonest form of omission is throwing out results because they do not fit the goal. We often assume that a study reports all relevant results. But studies often have (valid) reasons for throwing out certain results. Throwing out results can also bias a study, however. If we flip a coin ten times, and we repeat this experiment enough times, we will eventually obtain 10 heads in some trials and ten tails in others. We might then report that a random coin flip test produced ten heads (or tails), even though the entire set of results produced an equal number of heads and tails - by failing to report some results, we have biased those that we do report. For example, a product test may have been repeated many times, with the ones finally published being limited to those favoring the product.

A less extreme but equally improper form of discarding results has commonly been used by drug companies. They will fund studies into the efficacy of one of the drugs in their research pipeline. In some cases, they have terminated studies where the early results look unfavorable; if the study is never finished, then it doesn't get published. Alternatively, they have blocked publication of completed studies that they funded, so that the results never see the light of day.

### **Lies with statistics:**

There are hundreds of ways to conduct statistical tests. Some study designs fit cleanly into standardized statistical procedures, but in many cases, unexpected results dictate that statistical tests be modified to suit the circumstances. Thus, any one data set may have dozens to hundreds of ways of being analyzed. In reporting the results, someone may bias the evaluation step by reporting only those tests favorable to a particular goal. We should point out that this practice offers a limited opportunity to bias an evaluation. If the data strongly support a particular result, it won't be easy to find an acceptable test which obscures that result. Of course, a biased person would merely avoid presenting those results or avoid presenting a test of that hypothesis. There are many ways to selectively present data, as are listed below and will be shown in class.

### **Ways to bias statistics and graphics:**

1. Avoid presentation of unfavorable results and tests
2. Group different categories of subjects to obscure results
3. Chose an appropriate scale to display results favorably
4. Transform the data before testing
5. Do a post-hoc analysis of the data to modify the analysis
6. Suppress or inflate uncertainty

# Minimizing the Abuses

Recognizing the possible abuses of science is the simplest and most effective way to avoid being subjected to them. Beyond this, we can think of no single simple rule to follow that will minimize the opportunity for someone to violate the spirit of science and present misleading results -- there are countless ways to bias data. One strategy to avoid bias is to require detailed, explicit protocols. Another is to have the data gathered by an individual or company lacking a vested interest in the outcome. But even with these policies, there is no guarantee that deliberate biases can be weeded out. The following table gives a few pointers.

## Ensuring legitimate science:

PROPERTY OF STUDY	IMPACT
Publish protocols in advance of the study	Prevents mid-course changes in response to results; enables requests for design modifications with little cost.
Publish the actual raw data	Enables an independent researcher to look objectively at the data, possibly uncovering any attempts to obfuscate certain results
Specify evaluation criteria before obtaining results	Minimizes after-the-fact interpretation of data.
Anticipate vested interests	Conclusions of individuals, corporations, and political bodies can be predicted with remarkable accuracy by knowing their financial interests and their political and ideological leanings. Understanding these data helps immensely in understanding how they may have used biased (but perhaps, well-intentioned) methods in arriving at conclusions.

## CHAPTER 26: OUR BRAINS INTRINSICALLY MISLEAD US

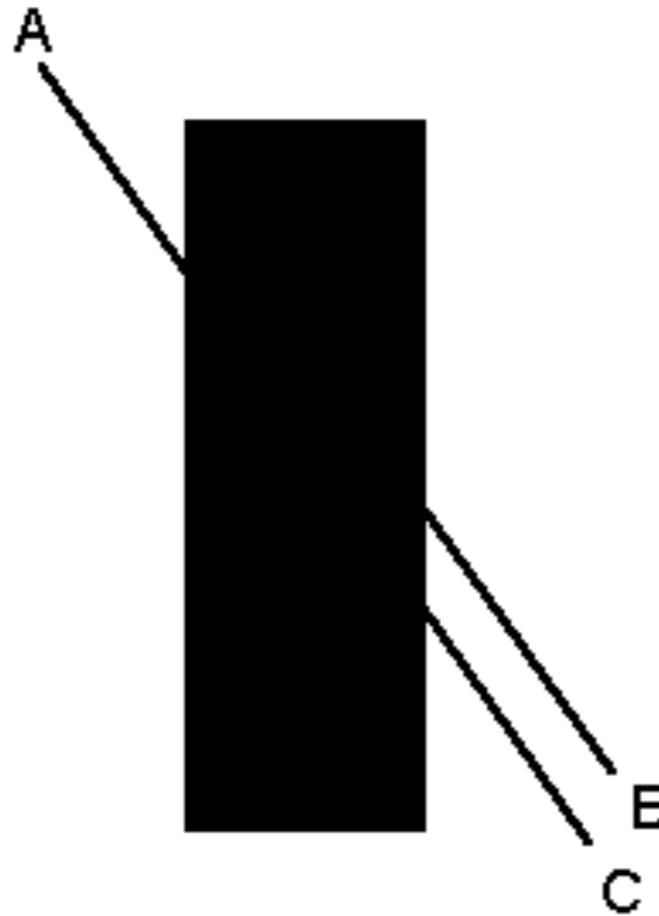
# 26

IMPEDIMENTS TO SCIENTIFIC PROGRESS

Our intuition and instincts can mislead us in some important ways unless we are alert to avoiding them.

# Introduction

Our brains are very useful organs in solving problems. Merely relying on intuition to solve a problem or make the right choice has some well-known limitations, however. Optical illusions and magic tricks are examples familiar to all of us in which appearances are deceiving. Consider the following picture:



The straight line appears to be AB, but it is really AC. There are many such optical illusions, and magic tricks often take advantage of them.

Magic tricks and optical illusions are examples in which we know our brains are fooled by appearances. Other ways in which we are deceived are more subtle. The following list offers some flaws in our natural instincts, which are discussed in turn.

# Behaviors That Can Lead Us Astray

1. We observe a correlation or association and automatically think causation
2. We respond unconsciously to many emotional, personal and environmental factors
3. We search for and remember confirmatory evidence of our beliefs rather than a test
4. Our memories are reconstructed over time
5. We prefer simplicity and certainty

## **(1) The power of association:**

This point goes back to the message of the correlation chapter, that correlation does not imply causation. Unfortunately, our minds are very easily influenced by the associations we observe. In an experiment done by psychologists approximately a decade ago, students were shown an item of clothing (e.g., jacket or sweater) and told to imagine that it had been owned and worn by someone famous or infamous. They were then asked to rank the desirability of wearing that item. The imagined previous ownership had a powerful effect. Items whose imagined previous owner was infamous were regarded as highly undesirable, whereas items imagined to be from the noble or famous were regarded as desirable. Of course, we know well that people collect items previously owned by the rich and famous. The fact that the desirability of an item is based on its history of previous ownership reflects just how strongly we respond to associations.

Advertisers use this principle effectively. Endorsements of a product by someone famous is now routine and begets the personality high dollars (e.g., Michael Jordan endorsing shoes, football players endorsing beer, Elizabeth Taylor endorsing her perfumes; the movie "Attack of the Killer Tomatoes" portrayed an ad in which Jesus Christ endorsed an electronics company). The practice works because of the association people make between the product and the personality, which may have nothing whatsoever to do with the quality of the product. Another advertising strategy which makes use of this principle is to advertise a product in an appealing setting - Coca Cola being consumed by exuberant, happy people, cars being driven on scenic roads or featured with people in remote settings. In the "old" days, cigarette commercials were allowed on television, and the Salem brand was typically advertised in a "springtime" setting with a young man and woman; Marlboro was portrayed in a remote Western setting by a man on horseback. In neither case did the commercial make any reference to the quality of the product. It was just a gimmick to build an association between the product and a situation that the viewer was likely to enjoy.

The fact that we respond to associations means simply that we need to be aware of this limitation and to separate our analysis of a phenomenon from the setting in which it is observed.

## **(2) Innate Responses:**

Closely related to the point above is the fact that we have many built-in behaviors. Situations that evoke fear and greed are very powerful at getting our attention. Sex (sex appeal) is another factor generating a strong response. Advertisers commonly use sex appeal to build an association with their product, as per (1) above. Greed is evoked by some ads as well, but fear is used much less, perhaps because fear is not a good emotion to associate with a product you want to sell. However, American Express travelers checks used the fear of vacationers being victimized by theft as a way to promote their product. Fear and greed are also two of the main driving forces behind most cons and scams. Most scams succeed because of greed, but once people get sucked in, the fear of exposure (or of notification to the IRS of illegal activities) is used to sustain the con. Many chain letters evoke both fear and greed in offering vast rewards for sustaining the chain (i.e., for sending money) while warning of the extraordinary bad luck that has befallen those who broke the chain.

Ideas that evoke these strong emotions can develop a contagious, social momentum that can have unfortunate consequences. Michael Shermer (1997, *Why people believe weird things*, W.H. Freeman and Co., NY) describes two instances of epidemics of false accusations: Medieval English Witch crazes (1560-1620), and much more recently, the "recovered memory" accounts of sexual abuse by parents (1992-1994). In the witch craze, the number of accusations arose from nothing to hundreds over a couple decades. The recovered memory epidemic was much more precipitous, accelerating from none to over 10,000 accusations in under 2 1/2 years. Then, as evidence to support the wild claims failed to surface, the accusations died out. Harmless examples of these epidemics of emotion-invoking tales nowadays are often described as "urban legends," because they evoke such strong emotions that people spread the story rapidly and aggressively. (One urban legend tells of a businessman who went into a hotel bar one night and awoke the next day in a bathtub, with both his kidneys surgically removed.) The point is that stories which evoke strong enough emotions and are told to enough people can create a hysteria that interferes with a careful evaluation of the evidence.

Control is another feature that is important to us. A situation that denies a person control over an outcome that affects them is more likely to be avoided than one that allows control. It is well known, for example, that people are willing to accept much higher levels of risk when they are given the choice than when not. For example, people are much more inclined to fear flying than driving, even though driving is much more dangerous. And people are very intolerant of low levels of pesticide residue in food, even though many people freely accept the risks associated with smoking and with drinking.

Built-in responses go far beyond these few cases just mentioned. People have favorite colors, they find certain shapes more appealing than others, and a person's "body language" and dress has a big influence on the responses of others. These factors can be very important when persuasion is an integral part of the goal. Lawyers and business people in particular must heed these factors, highlighting the simple fact that many people make decisions and choices in non-scientific ways.

### **3) Wanting to know we are right:**

One of the most damaging tendencies of ours is the search for confirmation rather than truth. If we just paid a large sum to purchase a car, we look for ways to convince ourselves that we made a good choice. We do not want to find out that it could have been purchased for less at another dealer, nor that some flaw in its design has just been discovered. Instead, we want to feel that we made the right choice.

This behavior can spill over and affect our ability to make good choices. We typically approach a situation with some initial preference, and we subconsciously bias our evaluation in favor of that preference. This problem is common even among scientists, because most scientists have strong preferences among the different models and theories they are testing. It is extremely easy to unconsciously bias a test by seeing only what they want to see. This human weakness motivates the use of double-blind designs.

#### **4) Rewriting the past:**

We all forget things, and we are aware that we can't remember some things. But for those things of which we have a clear memory, it seems that our memory should be trusted. Not so. Our memories of even recent events can be faulty, as demonstrated by the different responses of people who witnessed the same recent events. (This point was visited with experimental tests of eyewitness identifications of people.) But as the event falls further into our past, our memory of it becomes increasingly rebuilt. Psychologists have been very successful recently in showing how suggestible people are, by describing wholly fictitious "memories" to a person and then later discovering that the person now remembers the incident as if it were true. We thus need to be careful to document events when they happen.

#### **5) Skip the details:**

Much of how we respond to new information depends on our "world view" of things. Each of us has a mental model of how society, nature, and the universe works, and each new fact or observation is either accommodated into that view (perhaps changing it slightly) or discarded as unimportant or wrong. As we noted in the first chapter of this book and in the class response to the first-day questionnaire, people have very different world views about some things (e.g., the paranormal). However, some features are common to most people: they by-and-large prefer simple explanations of things and prefer apparent certainty. Observations and other models with both of these features are easier to accommodate in a world view than are those incorporating complexity and uncertainty. Even science operates this way, because simple models are preferred over complicated ones until the simple ones must be rejected. Thus, people and science are more accepting of explanations that appear simple and unambiguous.

## **Conclusions:**

This list of inherent limitations of the human mind is certainly (!) incomplete. The simple point of this chapter is to remind us that we each have many behaviors that interfere with our ability to objectively and rationally evaluate evidence. We can train ourselves to understand and avoid these pitfalls, but identifying them is the first step. Understanding these pitfalls also enables us to understand how other people will make mistakes in objective thinking. Con artists and magicians are masters of exploiting these weaknesses in people, but even aside from those extreme cases, we need to be alert to our tendencies to make poor choices.

# Blind Subjects

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

---

## Related Glossary Terms

Drag related terms here

---

**Index**

Find Term

**Chapter 11 - Blind Data**