A Brief Primer on Applied Probabilistic Reasoning

Ehsan Bokhari and Lawrence Hubert

University of Illinois at Urbana-Champaign

Author Note

Ehsan Bokhari and Lawrence Hubert, Department of Psychology, University of Illinois at Urbana-Champaign.

Ehsan Bokhari is now with the Los Angeles Dodgers, 1000 Elysian Way, Los Angeles, CA, 90012.

Correspondence concerning this article should be addressed to Lawrence Hubert, Department of Psychology, 603 East Daniel Street, The University of Illinois, Champaign, Illinois, 61820. E-mail: lhubert@illinois.edu

Abstract

This primer is intended as an informal introduction to several central ideas in probabilistic reasoning. The first topic introduced concerns Sally Clark, who was convicted in England of killing her two children, partially on the basis of an inappropriate assumption of statistical independence. Next, the infamous O.J. Simpson murder trial is recalled along with defense lawyer Johnnie Cochran's famous dictum: "if it doesn't fit, you must acquit." This last statement is reinterpreted probabilistically and then used to introduce the two key probabilistic reasoning concepts of an event being either facilitative or inhibitive of another. Based on these two notions of facilitation and inhibition, a number of topic areas are then reviewed in turn: probabilistic reasoning based on data organized in the form of 2×2 contingency tables; the Charles Peirce idea of abductive reasoning; Bayes' theorem and diagnostic testing; the fallacy of the transposed conditional; how to interpret probability and risk and deal generally with probabilistic causation; where the numbers might come from that are referred to as probabilities and what they may signify; the misunderstandings that can arise from relying on nontransparent odds ratios rather than on relative risks; and finally, how probabilistic causation has been dealt with successfully in a federal program to compensate workers exposed to ionizing radiation and other toxic materials through their involvement with the United States' nuclear weapons industry. The last section of this brief primer discusses a set of eleven additional instructional modules that cover a variety of (other) probabilistic reasoning topics. These modules are available through a web location given in this last section.

Keywords: abductive reasoning, Bayes theorem, fallacy of the transposed conditional, probabilistic reasoning, probability of causation

A Brief Primer on Applied Probabilistic Reasoning

Contents

Introduction

The formalism of thought offered by probability theory is one of the more useful portions of any beginning course in statistics in helping promote quantitative literacy. As typically presented, we speak of an event represented by a capital letter, say A, and the probability of the event occurring as some number in the range from zero to one, written as $P(A)$. The value of 0 is assigned to the "impossible" event that can never occur; 1 is assigned to the "sure" event that will always occur. The driving condition for the complete edifice of all probability theory is one single postulate: for two mutually exclusive events, A and B (where mutual exclusivity implies that both events cannot occur at the same time), the probability that A or B occurs is the sum of the separate probabilities associated with the events A and B: $P(A \text{ or } B) = P(A) + P(B)$. As a final beginning definition, we say that two events are independent whenever the probability of occurrence for the joint event, A and B, factors as the product of the individual probabilities: $P(A \text{ and } B) = P(A)P(B)$. Intuitively, two events are independent if knowing that one event has already occurred doesn't alter an assessment of the probability of the other event occurring.[1]

The idea of statistical independence and the factoring of the joint event probability immediately provides a formal tool for understanding several historical miscarriages of justice; it also provides a good introductory illustration for the general importance of correct probabilistic reasoning. Specifically, if two events are not independent, then the joint probability cannot be generated by a simple product of the individual probabilities. A fairly recent and well-known judicial example involving probabilistic (mis)reasoning and the (mis)carriage of justice, is the case of Sally Clark; she was convicted in England of

[1] In somewhat more formal notation, it is common to represent the event "A or B" with the notation $A \cup B$, where "$\cup$" is a set union symbol called "cup." The event "A and B" is typically denoted by $A \cap B$, where "$\cap$" is a set intersection symbol called "cap." When A and B are mutually exclusive, they cannot occur simultaneously; this is denoted by $A \cap B = \emptyset$, the impossible event (using the "empty set" symbol $\emptyset$). Thus, when $A \cap B = \emptyset$, $P(A \cup B) = P(A) + P(B)$; also, as a definition, A and B are independent if and only if $P(A \cap B) = P(A)P(B)$.

killing her two children, partially on the basis of an inappropriate assumption of statistical independence. The purveyor of statistical misinformation in this case was Sir Roy Meadow, famous for Meadow's Law: " 'One sudden infant death is a tragedy, two is suspicious, and three is murder until proved otherwise' is a crude aphorism but a sensible working rule for anyone encountering these tragedies." We quote part of a news release from the Royal Statistical Society (October 23, 2001):

The Royal Statistical Society today issued a statement, prompted by issues raised by the Sally Clark case, expressing its concern at the misuse of statistics in the courts.

In the recent highly-publicised case of *R v. Sally Clark*, a medical expert witness drew on published studies to obtain a figure for the frequency of sudden infant death syndrome (SIDS, or 'cot death') in families having some of the characteristics of the defendant's family. He went on to square this figure to obtain a value of 1 in 73 million for the frequency of two cases of SIDS in such a family.

This approach is, in general, statistically invalid. It would only be valid if SIDS cases arose independently within families, an assumption that would need to be justified empirically. Not only was no such empirical justification provided in the case, but there are very strong a priori reasons for supposing that the assumption will be false. There may well be unknown genetic or environmental factors that predispose families to SIDS, so that a second case within the family becomes much more likely.

The well-publicised figure of 1 in 73 million thus has no statistical basis. Its use cannot reasonably be justified as a 'ballpark' figure because the error involved is likely to be very large, and in one particular direction. The true frequency of families with two cases of SIDS may be very much less incriminating than the figure presented to the jury at trial.

The Sally Clark case will be revisited in a later section as an example of committing the "prosecutor's fallacy." It was this last probabilistic confusion that lead directly to her conviction and imprisonment.

Some Initial Basics: The O.J. Simpson Case and the Legend of Cinderella

The most publicized criminal trial in American history was arguably the O.J. Simpson murder case held throughout much of 1995 in Superior Court in Los Angeles County, California. The former football star and actor, O.J. Simpson, was tried on two counts of murder after the death in June of 1994 of his ex-wife, Nicole Brown Simpson, and

a waiter, Ronald Goldman. Simpson was acquitted controversially after a televised trial lasting more than eight months.

Simpson's high-profile defense team, led by Johnnie Cochran, included such illuminaries as F. Lee Bailey, Alan Dershowitz, and Barry Scheck and Peter Neufeld of the Innocence Project. Viewers of the widely televised trial might remember Simpson not being able to fit easily into the blood-splattered leather glove that was found at the crime scene and which was supposedly used in the commission of the murders. For those who may have missed this high theater, there is a YouTube video that replays the glove-trying-on part of the trial; just "google": OJ Simpson Gloves & Murder Trial Footage

This incident of the gloves not fitting allowed Johnnie Cochran in his closing remarks to issue one of the great lines of 20th century jurisprudence: "if it doesn't fit, you must acquit." The question of interest here in this initial module on probabilistic reasoning is whether one can also turn this statement around to read: "if it fits, you must convict." But before we tackle this explicitly, let's step back and introduce a small bit of formalism in how to deal probabilistically with phrases such as "if p is true, then q is true," where p and q are stand-in symbols for two (arbitrary) propositions.

Rephrasing in the language of events occurring or not occurring, suppose we have the following correspondences:

glove fits: event A occurs

glove doesn't fit: event $\bar{A}$ (the negation of A) occurs

jury convicts: event B occurs

jury acquits: event $\bar{B}$ (the negation of B) occurs

Johnnie Cochran's quip of "if it doesn't fit, you must acquit" gets rephrased as "if $\bar{A}$ occurs, then $\bar{B}$ occurs." Or stated in the notation of conditional probabilities, $P(\bar{B}|\bar{A}) = 1.0$; that is, the probability that $\bar{B}$ occurs "given that" $\bar{A}$ has occurred is 1.0 (where this latter phrase of "given that" is represented by the short vertical line "|"); in words, we have "a sure thing."

Although many observers of the O.J. Simpson trial might not ascribe to the absolute nature of the Johnnie Cochran statement implied by $P(\bar{B}|\bar{A})$ being 1.0, most would likely agree to the following modification: $P(\bar{B}|\bar{A}) > P(\bar{B})$. Here, the occurrence of $\bar{A}$ (the glove not fitting) should increase the likelihood of acquittal to somewhere above the original (or marginal or prior) value of $P(\bar{B})$; there is, however, no specification as to how big an increase there should be other than it being short of the value 1.0 representing "a sure thing."

To give a descriptive term for the situation where $P(\bar{B}|\bar{A}) > P(\bar{B})$, we will say in a non-causal descriptive manner that $\bar{A}$ is "facilitative" of $\bar{B}$ (that is, there is an increase in the probability of $\bar{B}$ occurring over its marginal value of $P(\bar{B})$). When the inequality is in the opposite direction, and $P(\bar{B}|\bar{A}) < P(\bar{B})$, we say, again in a non-causal descriptive sense, that $\bar{A}$ is "inhibitive" of $\bar{B}$ (that is, there is a decrease in the probability of $\bar{B}$ occurring over its marginal value of $P(\bar{B})$).

Based on the rules of probability, the one phrase of $\bar{A}$ being facilitative of $\bar{B}$, $P(\bar{B}|\bar{A}) > P(\bar{B})$, leads inevitably to a myriad of other such statements: $\bar{B}$ is facilitative of $\bar{A}$ and inhibitive of A; $\bar{A}$ is facilitative of $\bar{B}$ and inhibitive of B; B is facilitative of A and inhibitive of $\bar{A}$; A is facilitative of B and inhibitive of $\bar{B}$.[2]

[2]A number of alternative words or phrases could be used in place of the terms "facilitative" and "inhibitive." For instance, one early usage of the phrases "favorable to" (for "facilitative") and "unfavorable to" (for "inhibitive") along with demonstrations for all the implications just summarized, is in Kai-Lai Chung's "On Mutually Favorable Events" (*Annals of Mathematical Statistics*, *13*, 1942, 338–349). In an evidentiary context, such as that developed in detail by Schum (1994), the phrase "positively (or favorably) relevant" could stand for "facilitative," and the phrase "negatively (or unfavorably) relevant" could substitute for "inhibitive." Rule 401 in the *Federal Rules of Evidence* (FRE) defines evidence relevance as follows:

Evidence is relevant if

(a) it has any tendency to make a fact more or less probable than it would be without the evidence; and

(b) the fact is of consequence in determining the action.

Nevertheless, as discussed in the next section, just because evidence may be relevant doesn't automatically then make it admissible under FRE Rule 403.

To give one example of these latter statements, consider "A being facilitative of B." In a formula, this says that $P(B|A) > P(B)$, or in words, the probability of a conviction (B) given that the glove fits (A) is increased over the marginal or prior probability of a conviction. Most would likely agree that this is a reasonable statement; the point being made here is that once we agree that $\bar{A}$ is facilitative of $\bar{B}$, we must also agree to statements such as A being facilitative of B.

Although this introductory section is intended primarily to be just that, introductory, the reader may be interested to see in a more formal way how the steps would proceed from $P(\bar{B}|\bar{A}) > P(\bar{B})$ to, say, $P(A|B) > P(A)$. In the series of expressions below, the symbol $\Leftrightarrow$ means an "equivalent statement" and $\cap$ stands for "and." Also, there are repetitive uses of two main ideas — (1) the definition of a conditional probability; for example,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \ ;$$

(2) the notion that the probability of an event is 1.0 minus the probability of the negation of the event; for example,

$$P(\bar{A}) = 1 - P(A) \text{ and } P(\bar{B}|\bar{A}) = 1 - P(B|\bar{A}) \ .$$

$$P(\bar{B}|\bar{A}) > P(\bar{B})$$

$$\Leftrightarrow \ 1 - P(B|\bar{A}) > 1 - P(B)$$

$$\Leftrightarrow \ 1 - \frac{P(B \cap \bar{A})}{P(\bar{A})} > 1 - P(B)$$

$$\Leftrightarrow \ \frac{P(B \cap \bar{A})}{P(B)} < P(\bar{A})$$

$$\Leftrightarrow \ P(\bar{A}|B) < 1 - P(A)$$

$$\Leftrightarrow \; 1 - P(A|B) < 1 - P(A)$$

$$\Leftrightarrow \; P(A|B) > P(A)$$

In the language of probability theory, and as noted above in the Sally Clark case, two events A and B are said to be "independent" if the probability of the joint event, $A \cap B$, factors into the two (marginal) probabilities of A and of B. Or, to restate the formal definition: the events A and B are independent if and only if

$$P(A \cap B) = P(A)P(B)$$

Using the definition of a conditional probability, A and B are then independent if and only if

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

or

$$P(B|A) = \frac{P(A)P(B)}{P(A)} = P(B)$$

In words, A and B are independent if and only if A (or B) is neither inhibitive nor facilitative of B (or A).

To give another illustration of how probabilistic reasoning might reasonably operate, we go back to the legend of Cinderella and make one slightly risque modification. On her hurried way out of the castle just before midnight, Cinderella drops the one glass slipper (but say, she holds on to the other one) and loses all of her fitted clothes and jewelry including tiara, bra, panties, and so on. When the Prince sets off to find Cinderella, the following events are of interest:

slipper fits: event A occurs

slipper doesn't fit: event $\bar{A}$ (the negation of A) occurs

person is Cinderella: event B occurs

person is not Cinderella: event $\bar{B}$ (the negation of B) occurs

As in the Johnnie Cochran context, the occurrence of the event A (that the slipper fits) increases the likelihood that event B occurs (that the person is Cinderella) over the prior probability of this particular individual being Cinderella. But in our risque version of the legend, the Prince also has an array of fitted jewelry and clothes that also could be tried on sequentially, with each fitting item being itself facilitative of the event B of being Cinderella. Although one may never reach a "sure thing" and have Cinderella identified "beyond a shadow of a doubt," the sequential weight-of-the-evidence may lead to something at least "beyond a reasonable doubt"; or stated in other words, the cumulative probability of the event B (of being Cinderella) increases steadily with each newly fitting item.

The Cinderella saga we have laid out may be akin to what occurs in criminal cases where a conviction is obtained when the weight-of-the-evidence has reached a standard of "beyond a reasonable doubt." The tougher standard of "beyond a shadow of a doubt" may be attainable only when there is a proverbial "smoking gun." In Cinderella's case, this "smoking gun" might amount to producing the exact matching slipper that she held onto that night. For the O.J. Simpson case, it's unclear whether there could have ever been a "smoking gun" produced; even the available DNA evidence was discounted because of possible police tampering. If the blood-soaked Bruno Magli shoes had ever been found and if they had fit O.J. Simpson perfectly, then maybe — but then again, maybe not.

Alternative Approaches to Probabilistic Reasoning

The approach taken thus far to the basics of applied probabilistic reasoning has been rather simple. Given events A and $\bar{A}$ and B and $\bar{B}$, the discussion has been framed merely as one event being facilitative or inhibitive (or neither) of another event, and without any particular causal language imposed. As might be expected, all this can be made more complicated. For example, we begin by introducing likelihood ratios. If an event A is facilitative of B, then $P(B|A) > P(B)$; but, also, $\bar{A}$ must then be inhibitive of B or

$P(B|\bar{A}) < P(B)$. The fraction, $\frac{P(B|A)}{P(B|\bar{A})}$, is called a likelihood ratio, and must in this case be greater than 1.0 because of the inequality $P(B|A) > P(B) > P(B|\bar{A})$.

In a later section the all-powerful Bayes' theorem will be introduced and discussed in some detail. Although we won't do so in that section, an alternative version of Bayes' theorem could be given in the form of posterior odds being equal to a likelihood ratio times the prior odds. Remembering that odds are defined by the ratio of the probability of an event to the probability of the complement, the formal statement would be:

$$(\frac{P(A|B)}{P(\bar{A}|B)}) = (\frac{P(B|A)}{P(B|\bar{A})}) \times (\frac{P(A)}{P(\bar{A})})$$

or (posterior odds of A) = (likelihood)×(prior odds of A). In words, when A is facilitative of B, and thus, the likelihood, $\frac{P(B|A)}{P(B|\bar{A})}$, is greater than 1.0, and given that the event B occurs, the posterior odds of A occurring is greater than the prior odds of A occurring. In an analogous manner, we could also derive the form

$$(\frac{P(B|A)}{P(\bar{B}|A)}) = (\frac{P(A|B)}{P(A|\bar{B})}) \times (\frac{P(B)}{P(\bar{B})})$$

And, again in words, given that the event A occurs, the posterior odds of B occurring is greater than the prior odds of B occurring.

These two equivalent forms of Bayes' theorem appear regularly in the judgment and decision making literature whenever the discussion turns to the reliability (or unreliability, as the case might be) of eyewitness testimony. To illustrate this numerically with a rather well-known type of example, we paraphrase a presentation from Devlin and Lorden (2007, pp. 83–85) involving taxi cabs:

A certain town has two taxi companies, Blue Cabs and Black Cabs, having, respectively, 15 and 75 taxis. One night when all the town's 90 taxis were on the streets, a hit-and-run accident occurred involving a taxi. A witness sees the accident and claims a blue taxi was responsible. At the request of the police, the witness underwent a vision test with conditions similar to those on the night in question, indicating the witness could

successfully identify the taxi color 4 times out of 5. So, the question: which company is the more likely to have been involved in the accident?

If we let B be the event that the witness says the hit-and-run taxi is blue, and A the event that the true culprit taxi is blue, the following probabilities hold: $P(A) = 15/90$; $P(\bar{A}) = 75/90$; $P(B|A) = 4/5$; and $P(B|\bar{A}) = 1/5$. Thus, the posterior odds are 4 to 5 that the true taxi was blue: $[P(A|B)/P(\bar{A}|B)] = [(15/90)/(75/90)][(4/5)/(1/5)] \approx 4$ to 5. In other words, the probability that the culprit taxi is blue is $4/9 \approx 44\%$. We note that this latter value is much smaller than the probability (of $4/5 = 80\%$) that the eyewitness could correctly identify a blue taxi when presented with one. This effect is due to the prior odds ratio reflecting the prevalence of black rather than blue taxis on the street.

Another approach to certain aspects of probabilistic reasoning that is in contrast to inductive generalization (which argues from particular cases to a generalization) is through a "statistical syllogism" which argues from a generalization that is true for the most part to a particular case. As an example, consider the following three statements:

1) Almost all people are taller than 26 inches

2) Larry is a person

3) Therefore, Larry is almost certainly taller than 26 inches

Statement 1) (the major premise) is a generalization and the argument tries to elicit a conclusion from it. In contrast to a deductive syllogism, the premise merely supports the conclusion rather than strictly implying it. So, it is possible for the premise to be true and the conclusion false but that is not very likely.

One particular statistical syllogism justifies the common understanding of a confidence interval as containing the true value of the parameter in question with a high degree of certainty. When we teach beginning statistics with an eye toward preciseness, the confidence interval discussion might be given as follows: (1) "if this particular confidence interval construction method were repeated for multiple samples, the collection of all such random intervals would encompass the true population parameter, say, 95% of the time";

(2) "this is one such constructed interval"; (3) "it is very likely that this interval contains the true population value."

The use of statistical syllogisms must obviously be done with care so that we don't inappropriately judge individuals only as members of some group or category and ignore those characteristics that might "set them apart." For instance, the *Federal Rules of Evidence*, Rule 403, implicitly excludes the use of base rates that would be more prejudicial than probative (that is, having value as legal proof). Examples of such exclusions abound but generally involve some judgment as to which types of demographic groups commit which crimes and which ones don't. Rule 403 follows:

Rule 403. Exclusion of Relevant Evidence on Grounds of Prejudice, Confusion, or Waste of Time: Although relevant, evidence may be excluded if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury, or by considerations of undue delay, waste of time, or needless presentation of cumulative evidence.

Particularly egregious violations of Rule 403 have been ongoing for some time in Texas capital murder cases. As one recent and pernicious example, a psychologist, Walter Quijano, has regularly testified that because a defendant is black, there is an increased probability of future violence; or in our event language, the event of being black is facilitative of the occurrence of a future event (or act) of violence. We give a redaction of the majority opinion (at the web site listed in the footnote) in of the recent Supreme Court case of *Duane Edward Buck v. Rick Taylor* (2011), where the defendant, Duane Edward Buck, was attempting to avoid the imposition of a death penalty sentence.[3] Buck's lawyers argued that the death penalty should be lifted because Quijano stated at Buck's trial that because he was black, there was an increased probability he would engage in future acts of violence. In Texas capital murder cases, such predictions of future dangerous behavior are needed to have a death penalty imposed. The Supreme Court refused to hear the case (that is, to grant what is called *certiorari*), not because Buck didn't have a case of prejudicial racial evidence being introduced (in violation of Rule 403), but because,

[3]http://cda.psych.uiuc.edu/buck_v_thaler.pdf

incredibly, Quijano was a witness for the defense (that is, for Buck).

Data in the Form of a 2×2 Contingency Table

The introduction just given to some elementary ideas in probabilistic reasoning was phrased in terms of events $\bar{A}$ and A and their relation to the events $\bar{B}$ and B. This discussion can be extended to frequency distributions, and particularly to those defined by cross-classifications of individuals according to the events A and $\bar{A}$, and B and $\bar{B}$. Organizing the available data in the form of 2×2 tables helps facilitate the use of several different approaches to the interpretation of the data – much like Sherlock Holmes looking at all the data details, and then making conjectures that could then be verified (or not).

To give a numerical illustration that will be carried through for awhile, we adopt data provided by Gerd Gigerenzer, *Calculated Risks* (2002, pp. 104–107) on a putative group of 10,000 individuals cross-classified as to whether a Fecal Occult Blood Test (FOBT) is positive [B: +FOBT] or negative [$\bar{B}$: −FOBT], and the presence of Colorectal Cancer [A: +CC] or its absence [$\bar{A}$: −CC]:

	+CC: A	−CC: $\bar{A}$	Row Sums
+FOBT: B	15	299	314
−FOBT: $\bar{B}$	15	9671	9686
Column Sums	30	9970	10,000

Corresponding to any frequency distribution of cases, there are probability distributions generated when single cases are selected from the total group of cases at random and with replacement. Based on the frequency distribution given above (in the form of what is called a 2×2 contingency table), we have the following probabilities, where the triple lines, "$\equiv$", are meant to indicate a notion of "defined as":

	+CC: A	−CC: $\bar{A}$	Row Sums
+FOBT: B	$P(A \cap B)$ $\equiv \frac{15}{10{,}000}$	$P(\bar{A} \cap B)$ $\equiv \frac{299}{10{,}000}$	$P(B)$ $\equiv \frac{314}{10{,}000}$
−FOBT: $\bar{B}$	$P(A \cap \bar{B})$ $\equiv \frac{15}{10{,}000}$	$P(\bar{A} \cap \bar{B})$ $\equiv \frac{9671}{10{,}000}$	$P(\bar{B})$ $\equiv \frac{9686}{10{,}000}$
Column Sums	$P(A)$ $\equiv \frac{30}{10{,}000}$	$P(\bar{A})$ $\equiv \frac{9970}{10{,}000}$	

The specification of theoretical probability distributions by the idea of randomly sampling with replacement from the available cases (or to use a German word, a "Gedankenexperiment" where we just "think about it") might also go under the short-hand title of an "urn model": 10,000 balls labeled by A or $\bar{A}$ and B or $\bar{B}$ according to the frequencies from the Gigerenzer table are put into a (big) urn; we repeatedly sample with replacement from the mixed-up balls in the urn to generate a sample from the underlying theoretical distribution just defined in the 2×2 table of probabilities given above.[4]

When discussing data descriptively, we will naturally slip into the language of probabilities and justify this by an appeal to the urn model. So, if one asks whether B is facilitative of A (that is, whether testing positive in a Fecal Occult Blood Test is facilitative of having Colorectal Cancer), the question can be restated as follows: is $P(A|B) > P(A)$? The answer is "yes," because

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{15/10,000}{314/10,000} > P(A) = \frac{30}{10,000}$$

$$\Leftrightarrow \frac{15}{314} > \frac{30}{10,000} \Leftrightarrow .048 > .003$$

The size of the difference, $P(+CC|+FOBT) - P(+CC) = +.045$, may not be large in any

[4]Another way to characterize this type of Gedankenexperiment is through a resampling method called the "bootstrap"; see, for example, Bradley Efron and Robert Tibshirani (1994), *An Introduction to the Bootstrap* (Boca Raton, FL., Chapman & Hall/CRC Press); also, see Module 12 and its brief discussion of the bootstrap.

absolute sense, but the change does represent a fifteenfold increase over the marginal probability of .003 for $P(+CC)$. But note that if you have a positive FOBT, over 95% ($= \frac{299}{314}$) of the time you don't have cancer; that is, there are 95% false positives.

Unfortunately, dismal results such as these appear regularly. Even though an event may be facilitative or inhibitive of another, this can be a very weak condition by itself. The degree of facilitation or inhibition may be so weak in absolute terms that reliance on it is mistaken both practically and ethically.

We will come back in a later section to the use of a cross-classified 2×2 contingency table (for A and $\bar{A}$ and B and $\bar{B}$) to explain several concepts and anomalies in diagnostic testing and related areas. Even though we will speak in terms of probabilities and conditional probabilities, these are typically obtained from frequencies and an underlying urn model. This general way of developing the descriptive statistics (or what we might label as descriptive probabilities) is referred to as using "natural frequencies" by Gigerenzer and others.

Abductive Reasoning

An alternative strategy to explain what it means for certain events to be inhibitive or facilitative for other events is through the idea of abductive reasoning or inference introduced by Charles Peirce in the late 19th century (for an extensive discussion of Peirce's life and work, consult the Wikipedia entry for Charles Sanders Peirce). We begin by providing Peirce's beanbag analogy to distinguish between the three reasoning modes of deduction, induction, and abduction:[5]

Deduction

(Step 1) Rule: All the beans from this bag are white

(Step 2) Case: These beans are from this bag

[5]These distinctions between modes of reasoning made by examples such as this one are available in much of Peirce's writing; a particularly accessible source for this particular bean-bag illustration is C.S. Peirce, "Deduction, Induction, and Hypothesis" (*Popular Science Monthly*, *13*, 1878, 470–482).

Therefore,

(Step 3) Result: These beans are white

Induction

(Step 1) Case: These beans are from this bag

(Step 2) Result: These beans are white

Therefore,

(Step 3) Rule: All the beans from this bag are white

Abduction

(Step 1) Rule: All the beans from this bag are white

(Step 2) Result: These beans are white

Therefore,

(Step 3) Case: These beans are from this bag

Abduction is a form of logical inference that goes from an observation to a hypothesis that accounts for the observation and which explains the relevant evidence. Peirce first introduced the term "abduction" as "guessing" and said that to abduce a hypothetical explanation, say a: these beans are from this bag, from an observed circumstance, say b: these beans are white, is to surmise that "a" may be true because then "b" would be a matter of course. Thus, to abduce "a" from "b" involves determining that "a" is sufficient (or nearly sufficient) for "b" to be true, but not necessary for "b" to be true.

As another example, suppose we observe that the lawn is wet. If it had rained last night, it would be unsurprising that the lawn is wet; therefore, by abductive reasoning the possibility that it rained last night is reasonable. Or, stated in our language of events being facilitative, the event of the lawn being wet (event A) is facilitative of it raining last night (event B): $P(B|A) > P(B)$. Obviously, abducing rain last night from the evidence of a wet lawn could lead to a false conclusion – even in the absence of rain, some other process such as dew or automatic lawn sprinklers may have resulted in the wet lawn.

The idea of abductive reasoning is somewhat counter to how we introduce logical

considerations in our beginning statistics courses that revolve around the usual "if p, then q" statements, where p and q are two propositions. To give a simple example, we might let p be "the animal is a Yellow Labrador Retriever," and q, "the animal is in the order *Carnivora*." Continuing, we note that if the statement "if p, then q" is true (which it is), then logically, so must be the contrapositive of "if not q, then not p"; that is, if "the animal is not in the order *Carnivora*," then "the animal is not a Yellow Labrador Retriever." However, there are two fallacies awaiting the unsuspecting:

denying the antecedent: if not p, then not q (if "the animal is not a Yellow Labrador Retriever," then "the animal is not in the order *Carnivora*");

affirming the consequent: if q, then p (if "the animal is in the order *Carnivora*," then "the animal is a Yellow Labrador Retriever").

Also, when we consider definitions given in the form of "p if and only if q," (for example, "the animal is a domesticated dog" if and only if "the animal is a member of the subspecies *Canis lupus familiaris*"), or equivalently, "p is necessary and sufficient for q," these separate into two parts:

"if p, then q" (that is, p is a sufficient condition for q);

"if q, then p" (that is, p is a necessary condition for q).

So, for definitions, the two fallacies are not present.

In a probabilistic context, we reinterpret the phrase "if p, then q" as B being facilitative of A; that is, $P(A|B) > P(A)$, where p is identified with B and q with A. With such a probabilistic reinterpretation, we no longer have the fallacies of denying the antecedent (that is, $P(\bar{A}|\bar{B}) > P(\bar{A})$), or of affirming the consequent (that is, $P(B|A) > P(B)$); all of these are now necessary implications of the first statement that B is facilitative of A.

In reasoning logically about some situation, it would be rare to have a context that would be so cut and dried as to lend itself to the simple logic of "if p, then q," and where we could look for the attendant fallacies to refute some causal claim. More likely, we are

given problems characterized by fallible data, and subject to other types of probabilistic processes. For example, even though someone may have a genetic marker that has a greater presence in individuals who have developed some disease (for example, breast cancer and a mutation in the BRAC1 gene), it is not typically an unadulterated causal necessity. In other words, it is not true that "if you have the marker, then you must get the disease." In fact, many of these situations might be best reasoned through using our simple 2×2 tables; A and $\bar{A}$ denote the presence/absence of the marker; B and $\bar{B}$ denote the presence/absence of the disease. Assuming A is facilitative of B, we could go on to ask about the strength of the facilitation by looking at, say, the difference, $P(B|A) - P(B)$, or possibly, the ratio, $P(B|A)/P(B)$.

As developed in detail by Schum (1994) and others, such probability differences and ratios (as well as various other transformations) are considered important in defining what might be called measures of "inferential force." Our discussion will be confined to these kinds of simple differences and ratios and to rather uncomplicated statements about their relative sizes. In the context of genetics, for example, the conditional probability, $P(A|B)$, is typically reported by itself; this is called "penetrance" – the probability of disease occurrence given the presence of the marker. A fairly recent and high profile instance of the BRAC1 mutation being assessed as strongly facilitative of breast cancer (that is, having high "penetrance") was for the actress Angelina Jolie, who opted for a prophylactic double mastectomy to reduce her chances of contracting breast cancer. A few excerpts follow from her Op-Ed article, "My Medical Choice," that appeared in the *New York Times* (May 14, 2013):

> My mother fought cancer for almost a decade and died at 56. She held out long enough to meet the first of her grandchildren and to hold them in her arms. But my other children will never have the chance to know her and experience how loving and gracious she was.
>
> We often speak of "Mommy's mommy," and I find myself trying to explain the illness that took her away from us. They have asked if the same could happen to me. I have always told them not to worry, but the truth is I carry a "faulty" gene, BRCA1, which sharply increases my risk of developing breast cancer

and ovarian cancer.

My doctors estimated that I had an 87 percent risk of breast cancer and a 50 percent risk of ovarian cancer, although the risk is different in the case of each woman.

Only a fraction of breast cancers result from an inherited gene mutation. Those with a defect in BRCA1 have a 65 percent risk of getting it, on average.

Once I knew that this was my reality, I decided to be proactive and to minimize the risk as much I could. I made a decision to have a preventive double mastectomy. I started with the breasts, as my risk of breast cancer is higher than my risk of ovarian cancer, and the surgery is more complex.

...

I wanted to write this to tell other women that the decision to have a mastectomy was not easy. But it is one I am very happy that I made. My chances of developing breast cancer have dropped from 87 percent to under 5 percent. I can tell my children that they don't need to fear they will lose me to breast cancer.

The idea of arguing probabilistic causation is, in effect, the notion of one event being facilitative or inhibitive of another. If a collection of "q" conditions is observed that would be the consequence of a single "p," one may be more prone to conjecture the presence of "p," much like we could do in the Cinderella example. Although this process may seem like merely affirming the consequent, in a probabilistic context this could be referred to as "inference to the best explanation," or as we have noted above, an interpretation of the Charles Peirce notion of abductive reasoning. In any case, with a probabilistic reinterpretation, the assumed fallacies of logic may not be such. Moreover, most uses of information in contexts that are legal (forensic) or medical (through screening), or that might, for example, involve academic or workplace selection, need to be assessed probabilistically.[6]

[6]The Angelina Jolie decision to have a preventive double mastectomy based on her high probability of eventually contracting breast cancer seems a most rational choice. Other forms of prophylactic breast removal, however, are more controversial when based on only a small probability of cancer arising (or being lethal) in an otherwise healthy breast. As a case in point, Peggy Orenstein in an article for the *New York Times* (July 26, 2014), entitled "The Wrong Approach to Breast Cancer," relates her own story about a cancer recurrence in a breast that had undergone an earlier lumpectomy and radiation in 1997 and that now

Bayes' Rule (Theorem)

One of the most celebrated mathematical results in all of probability theory is called Bayes' theorem (or Bayes' rule or Bayes' law). Its modern formulation has been available since the 1812 Laplace publication, *Théorie analytique des probabilities*. In commenting on its importance, Sir Harold Jeffreys (1973, p. 31) noted that Bayes' theorem "is to the theory of probability what Pythagoras's theorem is to geometry." There are several ways to (re)state and extend Bayes' theorem but here we only need a form for the event pairs of A and $\bar{A}$, and B and $\bar{B}$, however the latter are defined.

To begin, note that

$$P(A|B) = P(A \cap B)/P(B)$$

and

$$P(B|A) = P(A \cap B)/P(A)$$

---

would have to be removed. The question was whether the otherwise healthy breast should also be removed at the same time, through a procedure called "contralateral prophylactic mastectomy" (CPM). The published evidence for undergoing a CPM shows virtually no survival benefit from the procedure; but still, the use of CPM is mushrooming. As Orenstein notes, there is a "need to recognize the power of 'anticipated regret': how people imagine they'd feel if their illness returned and they had not done 'everything' to fight it when they'd had the chance. Patients will go to extremes to restore peace of mind, even undergoing surgery that, paradoxically, won't change the medical basis for their fear." In a letter to the editor of the *New York Times* (July 31, 2014), Noreen Sugrue states the point particularly well that small or large probabilities are not the sole (or even the major) determinant of personal medical choice:

When a woman is given a diagnosis of cancer, the choices she makes about treatment are based on a number of risk assessments and subjective probabilities. But perhaps most important, those decisions are made so that the woman can find some peace of mind and move on with her life.

Any model of decision-making has two sets of inputs: probabilities of outcomes and preferences, goals or desires. Divergent choices can be made on the same factual basis expressed in the probabilities that a woman assigns to various outcomes.

The decision to have a mastectomy or a lumpectomy, or remove a seemingly healthy breast, should be a woman's choice without others second-guessing that a wrong decision was made.

These two statements directly lead to

$$P(A|B)P(B) = P(B|A)P(A)$$

and the simplest form of Bayes' theorem:

$$P(A|B) = P(B|A)(\frac{P(A)}{P(B)})$$

Thus, if we wish to connect the two conditional probabilities $P(A|B)$ and $P(B|A)$, the latter must be multiplied by the ratio of the marginal (or prior) probabilities, $\frac{P(A)}{P(B)}$. Noting that $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$, the simplest form of Bayes' theorem can be rewritten in a less simple but more common form of

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$

Bayes' theorem assumes great importance in assessing the value of diagnostic screening for the occurrence of rare events. For now we merely give a generic version of a diagnostic testing context and introduce some associated terms. Two introductory numerical examples are then given: one is for breast cancer screening through mammography; the second involves bipolar disorder screening through the Mood Disorders Questionnaire.

Suppose we have a test that assesses some relatively rare occurrence (for example, disease, ability, talent, terrorism propensity, drug or steroid usage, antibody presence, being a liar [where the test is a polygraph]). Let B be the event that the test says the person has "it," whatever that may be; A is the event that the person really does have "it." Two "reliabilities" are needed:

(a) the probability, $P(B|A)$, that the test is positive if the person has "it"; this is referred to as the *sensitivity* of the test;

(b) the probability, $P(\bar{B}|\bar{A})$, that the test is negative if the person doesn't have "it"; this is the *specificity* of the test. The conditional probability used in the denominator of Bayes' rule, $P(B|\bar{A})$, is merely $1 - P(\bar{B}|\bar{A})$, and is the probability of a "false positive."

The quantity of prime interest, the *positive predictive value* (PPV), is the probability that a person has "it" given that the test says so, $P(A|B)$, and is obtainable from Bayes' rule using the specificity, sensitivity, and prior probability, $P(A)$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + (1 - P(\bar{B}|\bar{A}))(1 - P(A))} \ .$$

To understand how well the test does, the facilitative effect of B on A needs interpretation; that is, a comparison of $P(A|B)$ to $P(A)$, plus an absolute assessment of the size of $P(A|B)$ by itself. Here, the situation is usually dismal whenever $P(A)$ is small (such as when screening for a relatively rare occurrence), and the sensitivity and specificity are not perfect. Although $P(A|B)$ will generally be greater than $P(A)$, and thus B facilitative of A, the absolute size of $P(A|B)$ is commonly so small that the value of the screening may be questionable.[7]

As will be discussed in greater detail in Module 4 on diagnostic testing, there is some debate as to how a diagnostic test should be evaluated; for example, are test sensitivity and specificity paramount or should our emphasis instead be on the positive and negative predictive values? Our view at this point is to argue that sensitivity and specificity, being properties of the test itself and obtained on persons known to have or not to have the condition in question, would be of primary interest when deciding whether to use the test. But once the diagnostic test results are available, and irrespective of whether they are positive or negative, sensitivity and specificity are no longer relevant. For clinical or other applied uses, the main issue is to determine whether the subject in question has the condition given the observed test results, and this is measured by the positive and negative predictive values. In other words, if we knew the status of the subject so sensitivity and

[7] As the example of Angelina Jolie illustrates, the absolute size of $P(A|B)$ may be large enough to generate decisive action. When B is the event of a BRAC1 mutation and A the event of contracting breast cancer, the probability $P(A|B)$ was estimated at .87 for the actress; also, when A is the event of contracting ovarian cancer, the probability estimate for $P(A|B)$ drops to .50; but this is still a "as likely as not" assessment of chances for ovarian cancer at some point in her life.

specificity would be relevant, it is unnecessary to perform the diagnostic test in the first place. In short, test sensitivity and specificity are important in initial test selection; the positive and negative predictive values are then most relevant for actual test usage.

As noted above, our first numerical example considers the efficacy of mammograms in detecting breast cancer. In the United States, about 180,000 women are found to have breast cancer each year from among the 33.5 million women who annually have a mammogram. Thus, the probability of a tumor is about 180,000/33,500,000 = .0054. Mammograms are no more than 90% accurate, implying that

$P(\text{positive mammogram} \mid \text{tumor}) = .90;$

$P(\text{negative mammogram} \mid \text{no tumor}) = .90.$

Because we do not know whether a tumor is present, all we know is whether the test is positive, Bayes' theorem must be used to calculate the probability we really care about, the positive predictive value (PPV). All the pieces are available to use Bayes' theorem to calculate the PPV of the test to be .047:

$$P(\text{tumor} \mid \text{positive mammogram}) =$$

$$\frac{.90(.0054)}{.90(.0054) + .10(.9946)} = .047,$$

which is obviously greater than the prior probability of .0054, but still very small in magnitude; again, as in the Fecal Occult Blood Test example, more than 95% of the positive tests that arise turn out to be incorrect.

Gigerenzer and colleagues (see Gigerenzer et al., 2007) have argued for the importance of understanding the PPV of a test, but suggest the use of "natural frequencies" and a simple 2×2 table of the type presented earlier, rather than actual probabilities substituted into Bayes' rule. Based on an assumed population of 10,000, the prior probability of A, plus the sensitivity and specificity values, we have the following 2×2 table:

	tumor	no tumor	Row Sums
+ mammogram	49	995	1044
− mammogram	5	8951	8956
Column Sums	54	9946	10,000

The PPV is then simply 49/1044 = .047, using the frequency value of 49 for the cell (+ mammogram, tumor) and the + mammogram row sum of 1044.

The second example is from clinical psychology and uses data from the Mark Zimmerman et al. article entitled "Performance of the Mood Disorders Questionnaire in a Psychiatric Outpatient Setting." (*Bipolar Disorders*, 2009, *11*, 759–765). Part of the abstract from this article follows:

Objectives: The Mood Disorders Questionnaire (MDQ) has been the most widely studied screening questionnaire for bipolar disorder, though few studies have examined its performance in a heterogeneous sample of psychiatric outpatients. In the present report from the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project, we examined the operating characteristics of the MDQ in a large sample of psychiatric outpatients presenting for treatment.

Methods: A total of 534 psychiatric outpatients were interviewed with the Structured Clinical Interview for DSM-IV and asked to complete the MDQ. Missing data on the MDQ reduced the number of patients to 480, 10.4% ($n = 52$) of whom were diagnosed with bipolar disorder.

Results: Based on the scoring guidelines recommended by the developers of the MDQ, the sensitivity of the scale was only 63.5% for the entire group of bipolar patients. The specificity of the scale was 84.8%, and the positive and negative predictive values were 33.7% and 95.0%, respectively ...

Conclusions: In a large sample of psychiatric outpatients, we found that the MDQ, when scored according to the developers' recommendations, had inadequate sensitivity as a screening measure ... These results raise questions regarding the MDQ's utility in routine clinical practice.

The data reported in the article can be given in the form of the following 2×2 contingency table. The row attribute is a classification by the Mood Disorder Questionnaire (MDQ); the column attribute is a clinical classification according to a Structured Clinical Interview for DSM Disorders (SCID), which is the supposed "gold standard" for bipolar disorder diagnosis.

	SCID:Bi	SCID:NBi	Row Sums
MDQ:Bi	33	65	98
MDQ:NBi	19	363	382
Column Sums	52	428	480

As given in the abstract, various MDQ test characteristics can be computed from the frequencies given in the table:

sensitivity = .635 (= 33/52);

specificity = .848 (= 363/428);

positive predictive value = .337 (= 33/98);

negative predictive value = .950 (= 363/382).

Several comments are in order about these rather dismal values. First, the diagnostic accuracy of the MDQ (the proportion of correct diagnoses) is 82.5% (= (33 + 363)/480), but this value is less than simple prediction by the base rates which would consistently predict someone to be "not bipolar" (these predictions would be correct 89.2% of the time (= 428/480)). Second, the event (B) of receiving an MDQ diagnosis of "bipolar" is facilitative of an SCID diagnosis of "bipolar" (event A); that is, $P(A|B)$ (= 33/98 = .337) > $P(A)$ (= 52/480 = .108). But because the PPV of .337 is below 1/2, a person testing "bipolar" with the MDQ is more likely than not to be assessed as "not bipolar" with the supposedly more accurate SCID. In fact, 2 out of 3 diagnoses of "bipolar" with the MDQ are incorrect. This is a clear indictment of the MDQ as a reasonable screening device for the diagnosis of being bipolar.[8]

[8]These kinds of anomalous situations where prediction by base rates outperforms prediction by a diagnostic test and where positive predictive values are less than one-half, are discussed in greater detail in Module 4 on Probabilistic Reasoning and Diagnostic Testing.

Beware the Fallacy of the Transposed Conditional

As noted at the start of the last section, the simplest form of Bayes' theorem relates $P(A|B)$ and $P(B|A)$ by multiplying this later conditional probability by the ratio of the prior probabilities:

$$P(A|B) = P(B|A)(\frac{P(A)}{P(B)})$$

Given this form of Bayes' theorem, it is clear that for $P(A|B)$ and $P(B|A)$ to be equal, the two prior probabilities, $P(A)$ and $P(B)$, must first be equal. When $P(A)$ and $P(B)$ are not equal and then to assert equality for $P(A|B)$ and $P(B|A)$, is to commit the "fallacy of the transposed conditional," the "inverse fallacy," or in a legal context, the "prosecutor's fallacy." We give four examples where the fallacy of the transposed conditional can be seen at work: (1) in the (mis-)interpretation of what a p-value signifies in statistics; (2) returning to the Sally Clark case that opened this primer, her ultimate conviction is partly attributable to the operation of the "prosecutor's fallacy"; (3) in deciding when to be screened for colon cancer by a colonoscopy rather than by a simpler and less invasive sigmoidoscopy; (4) the confusion between test sensitivity (specificity) and the positive (negative) predictive value.

(1) In teaching beginning statistics, it is common to define a "p-value" somewhat as follows: assuming that a given null hypothesis, H_o, is true, the p-value is the probability of seeing a result as or more extreme than what was actually observed. It is not the probability that the null hypothesis is true given what was actually observed. The latter is an example of the fallacy of the transposed conditional. Explicitly, the probability of seeing a particular data result conditional on the null hypothesis being true, $P(\text{data} \mid H_o)$, is confused with $P(H_o \mid \text{data})$, the probability that the null hypothesis is true given that a particular data result has occurred.

(2) For our second example, we return to the Sally Clark conviction where the invalidly constructed probability of 1 in 73 million was used to successfully argue for Sally Clark's guilt. Let A be the event of innocence and B the event of two "cot deaths" within

the same family. The invalid probability of 1 in 73 million was considered to be for $P(B|A)$; a simple equating with $P(A|B)$, the probability of innocence given the two cot deaths, led directly to Sally Clark's conviction.[9]

We continue with the Royal Statistical Society news release:

Aside from its invalidity, figures such as the 1 in 73 million are very easily misinterpreted. Some press reports at the time stated that this was the chance that the deaths of Sally Clark's two children were accidental. This (mis-)interpretation is a serious error of logic known as the Prosecutor's Fallacy.

The Court of Appeal has recognised these dangers (*R v. Deen* 1993, *R v. Doheny/Adams* 1996) in connection with probabilities used for DNA profile evidence, and has put in place clear guidelines for the presentation of such evidence. The dangers extend more widely, and there is a real possibility that without proper guidance, and well-informed presentation, frequency estimates presented in court could be misinterpreted by the jury in ways that are very prejudicial to defendants.

Society does not tolerate doctors making serious clinical errors because it is widely understood that such errors could mean the difference between life and death. The case of *R v. Sally Clark* is one example of a medical expert witness making a serious statistical error, one which may have had a profound effect on the outcome of the case.

Although many scientists have some familiarity with statistical methods, statistics remains a specialised area. The Society urges the Courts to ensure that statistical evidence is presented only by appropriately qualified statistical experts, as would be the case for any other form of expert evidence.

(3) The third example is inspired by Edward Beltrami's book, *Mathematical Models for Society and Biology* (Academic Press; 2013), and in particular, its Chapter 5 on "A Bayesian Take on Colorectal Screening ..."[10] We begin with several selective quotations

[9]The exact same circumstances can occur in the (mis)use of DNA evidence. Here, the event B is the existence of a "match" between a suspect's DNA and what was found, say, at the crime scene; the event A is again one of innocence. The value for $P(B|A)$ is the probability of a DNA match given that the person is innocent. Commission of the "prosecutor's fallacy" would reverse the conditioning and say that this latter probability is actually for $P(A|B)$, the probability of innocence given that a match occurs.

[10]As readers, you may wonder where we have our minds, given that an earlier numerical example used a Fecal Occult Blood Test to check for Colorectal cancer, but trust us, this is a very informative example of how the fallacy of the transposed conditional plays an important role in fostering misunderstanding within "evidence-based-medicine" and how the media then perpetuates the interpretative error.

from an article in the *New York Times* by Denise Grady (July 20, 2000), "More Extensive Test Needed For Colon Cancer, Studies Say":

> The test most commonly recommended to screen healthy adults for colorectal cancer misses too many precancerous growths and should be replaced by a more extensive procedure that examines the entire colon, doctors are reporting today.
>
> ...
>
> The more common test, sigmoidoscopy, reaches only about two feet into the colon and is generally used to screen people 50 and older with an average risk of colon cancer. The more thorough procedure, colonoscopy, probes the full length of the colon, 4 to 5 feet, and is usually reserved for people with a higher risk, like those with blood in their stool, a history of intestinal polyps or a family history of colon cancer.
>
> ...
>
> Sigmoidoscopy, which is cheaper and easier to perform, has been used for screening on the optimistic theory that if no abnormalities were seen in the lower colon, none were likely to be found higher up.
>
> But that theory is contradicted by two studies being published today in *The New England Journal of Medicine*, which included a total of more than 5,000 healthy people screened by colonoscopy. One study, which involved more than 3,000 patients, is the largest study to date of the procedure. Both studies show that it is not safe to assume that the upper colon is healthy just because the lower third looks normal. The studies found that half the patients who had precancerous lesions in the upper colon had nothing abnormal lower down. If those patients had had only sigmoidoscopy, they would have mistakenly been given a clean bill of health and left with dangerous, undetected growths high in the colon.

Based on the two studies mentioned by Denise Grady and the analyses done by Beltrami, we have approximately the following conditional probabilities involving the two events U: there are advanced upper colon lesions, and L: there are no lower colon polyps: $P(U|L) \approx .02$ and $P(L|U) \approx .50$. A doctor wishing to convince a patient to do the full colonoscopy might well quote the second statistic, $P(L|U)$, and say "50% of all upper colon cancerous polyps would be missed if only the sigmoidoscopy were done." Although this statement is true, it might not be as convincing to undergo the much more invasive colonoscopy compared to a sigmoidoscopy if the first statistic, $P(U|L)$, were then quoted: "there is a very small probability of 2% of the upper colon showing cancerous lesions if the sigmoidoscopy shows no lower colon polyps." Confusing the 2% in this last statement with

the larger 50% amounts to the commission of the transposition fallacy.

(4) The last example deals with the generic diagnostic testing context where B is the event of testing "positive" and A is the event that the person really is "positive." Equating sensitivity and the positive predictive value requires $P(A|B)$ to be equal to $P(B|A)$; or in words, the probability of having "it" given that the test is positive must be the same as the test being positive if the person really does have it. As our example on breast cancer screening illustrates, if the base rate for having cancer is small (as it is here: $P(A) = .0054$), and differs from the probability of a positive test (as it does here: $P(B) = .90(.0054) + .10(.9946) = .1044$), the positive predictive value can be very disappointing ($P(A|B) = .047$; so there are about 95% false positives), and nowhere near the assumed test sensitivity ($P(B|A) = .90$).

Probability of Causation

In mass (toxic) tort cases (such as for asbestos, breast implants, and Agent Orange), there is a need to establish, in a legally acceptable fashion, some notion of causation. First, there is a concept of *general causation* concerned with whether an agent can increase the incidence of disease in a group; because of individual variation, a toxic agent will not generally cause disease in every exposed individual. *Specific causation* deals with an individual's disease being attributable to exposure from an agent.

The establishment of general causation (and a necessary requirement for establishing specific causation) typically relies on a *cohort study*. This is a method of epidemiologic study where groups of individuals are identified who have been or in the future may be differentially exposed to agent(s) hypothesized to influence the probability of occurrence of a disease or other outcome. The groups are observed to assess whether the exposed group is more likely to develop disease.

One common way to organize data from a cohort study is through a simple 2×2 contingency table, similar in form to those seen earlier in this introductory discussion:

	Disease	No Disease	Row Sums
Exposed	N_{11}	N_{12}	N_{1+}
Not Exposed	N_{21}	N_{22}	N_{2+}

Here, N_{11}, N_{12}, N_{21}, and N_{22} are the cell frequencies; N_{1+} and N_{2+} are the row frequencies. Conceptually, these data are considered generated from two (statistically independent) binomial distributions for the "Exposed" and "Not Exposed" conditions. If we let p_E and p_{NE} denote the two underlying probabilities of getting the disease for particular cases within the conditions, respectively, the ratio $\frac{p_E}{p_{NE}}$ is referred to as the relative risk (RR), and may be estimated with the data as follows:

estimated relative risk $= \widehat{\text{RR}} = \frac{\hat{p}_E}{\hat{p}_{NE}} = \frac{N_{11}/N_{1+}}{N_{21}/N_{2+}}$.

A measure commonly referred to in tort litigations is attributable risk (AR), defined as

AR $= \frac{p_E - p_{NE}}{p_E}$, and estimated by

$\widehat{\text{AR}} = \frac{\hat{p}_E - \hat{p}_{NE}}{\hat{p}_E} = 1 - \frac{1}{\widehat{\text{RR}}}$.

Attributable risk, also known as the "attributable proportion of risk" or the "etiologic fraction," represents the amount of disease among exposed individuals assignable to the exposure. It measures the maximum proportion of the disease attributable to exposure from an agent, and consequently, the maximum proportion of disease that could be potentially prevented by blocking the exposure's effect or eliminating the exposure itself. If the association is causal, AR is the proportion of disease in an exposed population that might be caused by the agent, and therefore, that might be prevented by eliminating exposure to the agent.

The common legal standard used to argue for both specific and general causation is an RR of 2.0, or an AR of 50%. At this level, it is "as likely as not" that exposure "caused" the disease (or "as likely to be true as not," or from English law, "the balance of the probabilities"). Obviously, one can never be absolutely certain that a particular agent was "the" cause of a disease in any particular individual, but to allow an idea of "probabilistic

causation" or "attributable risk" to enter into legal arguments provides a justifiable basis for compensation. It has now become routine to do this in the courts.

Besides toxic tort cases, genetics is an area where the idea of attributable risk is continually discussed in informed media outlets such as the *New York Times*. The "penetrance" of a particular genetic anomaly or mutation was briefly explained earlier in the context of Angelina Jolie's decision to undergo a preventive mastectomy. But there now seems to be a stream of genetic studies reported on regularly where an informed understanding of attributable and relative risk would be of benefit for our own personal medical decision making. To give one such example, again in the context of genetics and contracting breast cancer, we have the recent article in the *New York Times* by Nicholas Bakalar (August 6, 2014), entitled "Study Shows Third Gene as Indicator for Breast Cancer." Several paragraphs of this piece are given below that emphasize attributable and relative risk in some detail:

> Mutations in a gene called PALB2 raise the risk of breast cancer in women by almost as much as mutations in BRCA1 and BRCA2, the infamous genes implicated in most inherited cases of the disease, a team of researchers reported Wednesday.
>
> ...
>
> Over all, the researchers found, a PALB2 mutation carrier had a 35 percent chance of developing cancer by age 70. By comparison, women with BRCA1 mutations have a 50 percent to 70 percent chance of developing breast cancer by that age, and those with BRCA2 have a 40 percent to 60 percent chance.
>
> The lifetime risk for breast cancer in the general population is about 12 percent.
>
> The breast cancer risk for women younger than 40 with PALB2 mutation was eight to nine times as high as that of the general population. The risk was six to eight times as high among women 40 to 60 with these mutations, and five times as high among women older than 60.

The Energy Employees Occupational Illness Compensation Program (EEOICP)

David Michaels is the current Assistant Secretary of Labor for the Occupational Safety and Health Administration (OSHA); he was nominated by President Obama and unanimously confirmed by the U.S. Senate in 2009 – quite a feat in an era of Congressional

gridlock. During the Clinton administration, Michaels served as the United States Department of Energy's Assistant Secretary for Environment, Safety, and Health (1998–2001), where he developed the initiative to compensate workers in the nuclear weapons industry who developed cancer or lung disease as a consequence of exposure to radiation, beryllium, and other toxic hazards. The initiative resulted in the program that entitles this section, and which has provided some ten billion dollars in benefits since its inception in 2001. David Michaels, an epidemiologist on leave from George Washington University School of Public Health and Health Services, is also the author of the well-received book, *Doubt is Their Product: How industry's assault on science threatens your health* (2008; Oxford).

The EEOICP was signed into law on December 7, 2000 by President Clinton, along with Executive Order 13179 reproduced below:

Since World War II, hundreds of thousands of men and women have served their Nation in building its nuclear defense. In the course of their work, they overcame previously unimagined scientific and technical challenges. Thousands of these courageous Americans, however, paid a high price for their service, developing disabling or fatal illnesses as a result of exposure to beryllium, ionizing radiation, and other hazards unique to nuclear weapons production and testing. Too often, these workers were neither adequately protected from, nor informed of, the occupational hazards to which they were exposed.

Existing workers' compensation programs have failed to provide for the needs of these workers and their families. Federal workers' compensation programs have generally not included these workers. Further, because of long latency periods, the uniqueness of the hazards to which they were exposed, and inadequate exposure data, many of these individuals have been unable to obtain State workers' compensation benefits. This problem has been exacerbated by the past policy of the Department of Energy (DOE) and its predecessors of encouraging and assisting DOE contractors in opposing the claims of workers who sought those benefits. This policy has recently been reversed.

While the Nation can never fully repay these workers or their families, they deserve recognition and compensation for their sacrifices. Since the Administration's historic announcement in July 1999 that it intended to compensate DOE nuclear weapons workers who suffered occupational illnesses as a result of exposure to the unique hazards in building the Nation's nuclear defense, it has been the policy of this Administration to support fair and timely compensation for these workers and their survivors. The Federal

Government should provide necessary information and otherwise help employees of the DOE or its contractors determine if their illnesses are associated with conditions of their nuclear weapons-related work; it should provide workers and their survivors with all pertinent and available information necessary for evaluating and processing claims; and it should ensure that this program minimizes the administrative burden on workers and their survivors, and respects their dignity and privacy. This order sets out agency responsibilities to accomplish these goals, building on the Administration's articulated principles and the framework set forth in the Energy Employees Occupational Illness Compensation Program Act of 2000. The Departments of Labor, Health and Human Services, and Energy shall be responsible for developing and implementing actions under the Act to compensate these workers and their families in a manner that is compassionate, fair, and timely. Other Federal agencies, as appropriate, shall assist in this effort.

The EEOICP is one of the most successful and well-administered Federal compensation programs. It has its own non-profit advocacy group called "Cold War Patriots" (submotto: We did our part to keep America Free!); the web site for this organization is:

www.coldwarpatriots.org

This advocacy group provides informational meetings and help for those who might be eligible under the program. Below is part of an ad that appeared in the *New Mexican* (Santa Fe, New Mexico; June, 2014) announcing informational meetings in Penasco, Los Alamos, and Espanola:

Attention Former LANL (Los Alamos National Lab), Sandia Labs, and Uranium Workers:

— Join us for an important town hall meeting

— Learn if you qualify for benefits up to $400,000 through the Energy Employees Occupational Illness Compensation Program Act (EEOICPA)

— Learn about no-cost medical benefit options

— Learn how to apply for consequential medical conditions and for impairment re-evaluation for approved conditions

The EEOICP represents an implementation of the "as likely as not standard" for attributing possible causation (and compensation), and has gone to great technical levels

(and which should keep many biostatisticians gainfully employed for years to come). The web site given in the footnote provides an extensive excerpt from the *Federal Register* concerning the Department of Health and Human Services and its *Guidelines for Determining the Probability of Causation and Methods for Radiation Dose Reconstruction Under the [Energy] Employees Occupational Illness Compensation Program Act of 2000*.[11] This material should give a good sense of how the modeling principles of probability and statistics are leading to ethically defensible compensation models; here, the models used are for all those exposed to ionizing radiation through an involvement with the United States' nuclear weapons industry.[12]

The Interpretation of Probability and Risk

The Association for Psychological Science publishes a series of timely monographs on *Psychological Science in the Public Interest*. One recent issue was from Gerd Gigerenzer

[11] http://cda.psych.uiuc.edu/eeoicpa.pdf

[12] Several points need emphasize about the *Federal Register* excerpt: (1) the calculation of a "probability of causation" is much more sophisticated (and fine-grained) than one based on a simple aggregate 2×2 contingency table where attributable risk (AR) is just calculated from the explicit cell frequencies. Statistical models (of what are commonly referred to as the generalized linear model variety) are being used to estimate the AR tailored to an individual's specific circumstances—type of cancer, type of exposure, other individual characteristics; (2) all the models are now implemented (interactively through a graphical user interface) within the Interactive RadioEpidemiological Program (IREP), making obsolete the very cumbersome charts and tables previously used; also, IREP allows a continual updating to the model estimation process when new data become available; (3) it is not just a point estimate for the probability of causation that is used to determine compensation, but rather the upper limit for a 99% confidence interval; this obviously gives a great "benefit of the doubt" to an individual seeking compensation for a presumably radiation-induced disease; (4) as another "benefit of the doubt" calculation, if there are two or more primary cancers, the probability of causation reported will be the probability that at least one of the cancers was caused by the radiation. Generally, this will result in a larger estimate for the probability of causation, and thus to a greater likelihood of compensation; (5) when cancers are identified from secondary sites and the primary site is unknown, the final assignment of the primary cancer site will be the one resulting in the highest estimate for the probability of causation.

and colleagues, entitled "Helping Doctors and Patients Make Sense of Health Statistics" (Gigerenzer et al., 2007). It discusses aspects of statistical literacy as it concerns health, both our own individually as well as societal health policy more generally. Some parts of being statistically literate may be fairly obvious; we know that just making up data, or suppressing information even of supposed outliers without comment, is unethical. The topics touched upon by Gigerenzer et al. (2007), however, are more subtle. If an overall admonition is needed, it is that context is always important, and the way data and information are presented is absolutely crucial to an ability to reason appropriately and act accordingly. We review several of the major issues raised by Gigerenzer et al. in the discussion to follow.

We begin with a quotation from Rudy Guiliani from a New Hampshire radio advertisement that aired on October 29, 2007, during his run for the Republican presidential nomination:

> I had prostate cancer, five, six years ago. My chances of surviving prostate cancer and thank God I was cured of it—in the United States, 82 percent. My chances of surviving prostate cancer in England, only 44 percent under socialized medicine.

Not only did Guiliani not receive the Republican presidential nomination, he was just plain wrong on survival chances for prostate cancer. The problem is a confusion between survival and mortality rates. Basically, higher survival rates with cancer screening do not imply longer life.

To give a more detailed explanation, we define a five-year survival rate and an annual mortality rate:

five-year survival rate = (number of diagnosed patients alive after five years)/(number of diagnosed patients);

annual mortality rate = (number of people who die from a disease over one year)/(number in the group).

The inflation of a five-year survival rate is caused by a *lead-time bias*, where the time of diagnosis is advanced (through screening) even if the time of death is not changed.

Moreover, such screening, particularly for cancers such as prostate, leads to an *overdiagnosis bias*, the detection of a pseudodisease that will never progress to cause symptoms in a patient's lifetime. Besides inflating five-year survival statistics over mortality rates, overdiagnosis leads more sinisterly to overtreatment that does more harm than good (for example, incontinence, impotence, and other health-related problems).

Screening does not "prevent cancer," and early detection does not prevent the risk of getting cancer. One can only hope that cancer is caught, either by screening or other symptoms, at an early enough stage to help. It is also relevant to remember that more invasive treatments are not automatically more effective. A recent and informative summary of the dismal state and circumstances surrounding cancer screening generally, appeared in the *New York Times* as a "page one and above the fold" article by Natasha Singer (July 16, 2009), "In Push for Cancer Screening, Limited Benefits."

A major area of concern in the clarity of reporting health statistics is in how the data are framed as relative risk reduction or as absolute risk reduction, with the former usually seeming much more important than the latter. We give examples that present the same information:

Relative risk reduction: If you have this test every two years, your chance of dying from the disease will be reduced by about one third over the next ten years.

Absolute risk reduction: If you have this test every two years, your chance of dying from the disease will be reduced from 3 in 1000 to 2 in 1000, over the next ten years.[13]

[13]In informed media outlets such as the *New York Times*, the distinction between relative and absolute risk reduction is generally highlighted whenever there is also a downside to the medical procedure being reported. An example of this caution is present in the article by Tara Parker-Pope (August 6, 2014), entitled "Prostate Cancer Screening Still Not Recommended for All." The article gives a lifetime risk of dying of prostate cancer of 3 percent and a drop to 2.4 percent under a PSA testing regime. Although the absolute risk reduction of .6 percent does represent a 21 percent lower relative risk of dying, it is highly questionable whether this drop is worth the over-diagnosis and over-treatment that it requires. A few paragraphs from the article follow:

A major European study has shown that blood test screening for prostate cancer saves lives, but doubts

A useful variant on absolute risk reduction is given by its reciprocal, the *number needed to treat* (NNT); if 1000 people have this test every two years, one person will be saved from dying from the disease every ten years. (Numerically, the NNT is just the reciprocal of the absolute risk reduction, or in this case, $1/(.003 - .002) = 1/.001 = 1000$.)[14]

Because bigger numbers garner better headlines and more media attention, it is expected that relative rather than absolute risks are the norm. It is especially disconcerting, however, to have potential benefits (of drugs, screening, treatments, and the like) given in relative terms, but harm in absolute terms that is typically much smaller numerically. The latter has been referred to as "mismatched framing" by Gigerenzer and colleagues.

The issues involved in presenting two probabilities or proportions either as an absolute difference or relatively as a ratio reappears continually when there is a need to assess and report magnitudes. For example, in the Fecal Occult Blood Test illustration, the absolute difference between $P(+CC \mid + FOBT)$ and $P(+CC)$ was a small value of +.045 (but still would be one way of stating the degree of facilitation of $+FOBT$ on $+CC$). As a ratio, however, with respect to the prior probability of .003 for $P(+CC)$, this absolute difference does represent a fifteen-fold change. So, a relative measure again appears much

remain about whether the benefit is large enough to offset the harms caused by unnecessary biopsies and treatments that can render men incontinent and impotent.

The study, published Wednesday in *The Lancet*, found that midlife screening with the prostate-specific antigen, or PSA, screening test lowers a man's risk of dying of the disease by 21 percent. The relative benefit sounds sizable, but it is not particularly meaningful to the average middle-age man, whose risk of dying of prostate cancer without screening is about 3 percent. Based on the benefit shown in the study, routine PSA testing would lower his lifetime cancer risk to about 2.4 percent.

[14]In addition to the use of relative and absolute risk, or the number needed to treat, a fourth way of presenting benefit would be as an increase in life expectancy. For example, one might say that women who participate in screening from the ages of 50 to 69 increase their life expectancy by an average of 12 days. This is misleading in terms of a benefit to any one particular individual; it is much more of an all-or-nothing situation, like a lottery. Nobody who plays a lottery gains the expected payout; you either win it all or not.

more impressive than an absolute difference. All of this, by the way, is in the context of a very large false positive rate of over 95%. The exact same story is told in the illustration for breast cancer screening with mammography: we have small absolute differences, large relative ratios, and dismal performances as to the occurrence of false positives.

An ethical presentation of information avoids nontransparent framing of information, whether intentional or unintentional. Intentional efforts to manipulate or persuade people are particularly destructive, and unethical by definition. As Tversky and Kahneman have noted many times (for example, 1981), framing effects and context have major influences on a person's decision processes. Whenever possible, give measures that have operational meanings with respect to the sample at hand (for example, the Goodman–Kruskal gamma, the median or the mode, the interquartile range) and avoid measures that do not, such as the odds ratio (discussed in greater detail in a subsection to follow).

Where Do the Numbers Come From that Might Be Referred to as Probabilities and What Do They Signify

The introductory section of this module introduced the notion of probability as some number between zero and one that could be attached to an event, say A. The numbers so assigned were assumed implicitly to satisfy the usual axioms of probability theory, particularly the additive property for mutually exclusive events. The presentation here will be brief and necessarily basic, and follows a few simple distinctions that might be made in a beginning statistics class. For a much more thorough introduction to the topic of where probabilities come from that touches on a several hundred-year history, the reader is referred to the on-line *Stanford Encyclopedia of Philosophy* and the article "Interpretations of Probability."[15]

[15]However it is done, in assigning numbers to the occurrence of an event or to the truth of some statement, it may at times be of value to separate out the actual numerical estimate from one's confidence in it. For example, in throwing two dice a probability of 6/36 might be assigned for getting a seven and with complete confidence in that assignment. A weatherman, however, may be very confident about a forecast of 30 percent

A distinction can be made between two broad categories of methods for assigning probabilities: objective and subjective. An objective method is based on some point-at-able that might, for example, be a randomly controlled physical system or a database containing a variety of objects and attributes that could be queried to estimate a set of desired relative frequencies (or probabilities under an implicit urn model). A subjective method for assigning probabilities would typically be subject-specific, and based on some collection of presented evidence or result from aggregating the assessments of others. We begin with objective assignment methods.

As noted in greater detail in Module 5 on Probabilistic Reasoning in the Service of Gambling, the foundations of a modern theory of probability were laid down by Blaise Pascal and Pierre de Fermat in the early 17th century in response to a request from the Chevalier de Méré. This has become known as the "classical definition" of probability, and rests on the idea that some random process produces outcomes that can be considered equally-likely. We give a quotation from Pierre-Simon Laplace (*A Philosophical Essay on Probabilities*, 1814):

The theory of chance consists in reducing all the events of the same kind to a certain

---

rain for tomorrow but may be very uncertain when that same forecast of 30 percent rain is made for a week from now. In considering the range of reasonable forecasts for tomorrow, there may be some small uncertainty around the 30% estimate; a week from now, however, and even though the same estimate of 30% might be made, there is a much larger range of uncertainty around that same value – or as the old adage goes: there is many a slip 'twixt the cup and the lip.

At times there may be some equivocation between what is a probability estimate and what is one's confidence in that estimate. For example, an eyewitness who picks a particular subject out of a line-up has, in effect, assigned a probability of 1.0 to the event that the person is the "perp"; if pressed, additional phrases such as "yup, he's the one" or "I could see pretty good in the moonlight" express various levels of confidence in that initial assessment. In the evaluation discussed later as to whether Osama bin Laden would be at the compound in Pakistan when the Navy SEALs arrived, the values given in this particular vignette may reflect more confidence than (degree-of-belief) probability for the statement "bin Laden is at Abad Abad all the time."

number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible. As a simple example involving the random process of tossing two dice, the 36 pairs of integers from 1 to 6 could be considered equally likely. Out of the 36 pairs, 6 have the sum of spots equal to 7 ((6,1), (1,6), (2,5), (5,2), (4,3), (3,4)); thus, the probability of the event of rolling a 7 is 6/36.

A second objective method of assigning (approximate) probabilities is through observed relative frequencies. Here, there are n trials in which the event, say, A, could have occurred, and on n_A of these, event A did occur. Thus, $P(A) \approx n_A/n$, which is assumed to converge to the true probability as n goes to infinity (that is, "in the long run"). The type of repeated sampling using an urn model and obtaining the relative frequencies for the events of interest would be another example of this type of assignment.

A third form of objective probability assignment might be called the "database strategy." For example, suppose we have a multi-way contingency table that cross-classifies women according to various personal attributes, and as to whether they have the BRAC1 mutation and have contracted breast cancer (over their lifetimes). Entering the multi-way table with the personal attributes of Angelina Jolie and her positive status on the BRAC1 mutation, the proportion of women with these same characteristics that contracted breast cancer would be a probability estimate for her contracting breast cancer (that is, the value of .87 given in an earlier section).

The NPR News program, *All Things Considered*, ran a series of five programs on *Risk and Reason* in the Summer of 2014 that all dealt with probability in some way. One discussed weather forecasting: "Pop Quiz: 20 Percent Chance of Rain. Do You Need An Umbrella?" The issue discussed was how people generally interpreted the value of 20

percent. The best answer was to consider a database assignment: "it will rain on 20 percent of the days like tomorrow." Or, from a followup clarification (July 23, 2014):

> Many listeners and readers felt a concise explanation of "a 20 percent chance of rain" was missing from this story about weather forecasts and probability, so we followed up with two meteorologists.
>
> Will it rain or not? How you interpret the forecast could mean the difference between getting soaked or staying safe.
>
> From meteorologist Eli Jacks, of the National Oceanic and Atmospheric Administration's National Weather Service:
>
> "There's a 20 percent chance that at least one-hundredth of an inch of rain – and we call that measurable amounts of rain – will fall at any specific point in a forecast area."
>
> And from Jason Samenow, chief meteorologist with *The Washington Post*'s Capital Weather Gang:
>
> "It simply means for any locations for which the 20 percent chance of rain applies, measurable rain (more than a trace) would be expected to fall in two of every 10 weather situations like it."

The last form of objective probability assignment might be called "algorithmic" and would be represented by Nate Silver's *FiveThirtyEight* Blog that nicely predicted the outcome of the 2012 Presidential Election. The method used by Nate Silver involves aggregating information about some subject (for example, impending Senate races) across available polls that are weighted in various ways according to assumed biases – a nascent Bayesian approach some might say. Such aggregated probability estimates are another "objective" means for assigning probabilities that appears to work remarkably well.[16]

Now that Nate Silver has moved on to ESPN, the *New York Times* has a new statistical Blog called *The Upshot*, edited by David Leonhardt. As an example of the type of probability assessments they now give, we extract a first few paragraphs from an article about the 2014 midterm elections by Amanda Cox and Josh Katz, "Republicans' Senate Chances Rise Slightly to 60 Percent" (July 27, 2014):

> For the last month, we've been adding one or two polls a day to *The Upshot*'s Senate forecasting model. Today, we update all 36 races, based on estimates from a YouGov online panel that covers every

[16]Nate Silver's success has raised the self-esteem of all statisticians. The Associated Press states that "Nate Silver had made statistics sexy again"; Bloomberg Businessweek comments that "Nate Silver-led statistics men crush pundits in election."

congressional and governor's race across the country.

The panel, asked for its preferences in collaboration with CBS and *The New York Times*, is unusual in its scope: It comprises more than 66,000 people living in states with a Senate race this year. YouGov, a polling firm, also conducted online interviews with roughly half of the panelists in previous years, allowing it to know how they voted in the past.

With the addition of the YouGov estimates to our model, the overall outlook for the Senate remains roughly the same. The Republicans appear to have a slight advantage, with the most likely outcome being a Republican gain of six seats, the minimum they need to finish with a 51-to-49-seat majority. But we, like many other forecasters, would not be surprised by a gain of anywhere from four to eight seats.

Summing up the possible outcomes, our model gives the Republicans a 60 percent chance of taking control, up from 54 percent on April 1.

Polls are only one part of the model. (And we adjust polls from partisan firms according to our best estimates of how Republican- or Democratic-leaning the pollster has been this cycle.) The model also includes the candidates' political experience, fund-raising, a state's past election results and national polling.

The relative weight of these factors depends on the number and the quality of the polls in each state, as well as how useful each factor was in predicting past Senate elections. Currently, polls make up about 80 percent of the forecast in the most competitive races.

A subjective assignment of probability typically takes the form of an individual, or a group of individuals such as a jury or committee, receiving information (or hearing evidence) about the potential occurrence of some event or the truth of some statement. This latter assessment might be given in terms of explicit numbers that signify probability, or alternatively, in words that might suggest a (numerical) range of possibilities; for example, an assessment of a "preponderance of the evidence" suggests some numerical value greater than one-half. A particularly good example of the issue of using words versus numbers to refer to subjective (or "evidentiary") probabilities was in the second program in the NPR series on *Risk and Reason* that aired on July 23, 2014, entitled "In Facing National Security Dilemmas, CIA Puts Probabilities Into Words." This program dealt with the circumstantial evidence of Osama bin Laden's whereabouts, how the evidence was assessed by individuals in the CIA, and eventually on the assessment of President Obama

before he made the decision for the Navy SEALs to invade the compound in Pakistan. We give a few paragraphs from the show's transcript (with Host, Robert Siegel) that concern using numbers and/or words to indicate the likeliness of bin Laden being in the Abad Abad compound:[17]

KENNETH POLLACK: There was a real injunction that no one should ever use numbers to explain probability.

SIEGEL: That's Kenneth Pollack who used to be a military analyst at the CIA, where he now teaches intelligence analysis. Pollack says CIA analysts are told if you are asked what the chances of something happening, use words.

POLLACK: Almost certainly or highly likely or likely or very unlikely.

SIEGEL: What's the problem with numbers?

POLLACK: Assigning numerical probability suggests a much greater degree of certainty than you ever want to convey to a policymaker. What we are doing is inherently difficult. Some might even say it's impossible. We're trying to project the future. And, you know, saying to someone that there's a 67 percent chance that this is going to happen, that sounds really precise. And that makes it seem like we really know what's going to happen. And the truth is that we really don't.

SIEGEL: So Ken Pollack was surprised by the accounts of one especially high-profile event, in which CIA analysts and others in the intelligence agencies used numbers, very specific numbers, to express probabilities. Let's go back to May 1, 2011.

...

PRESIDENT BARACK OBAMA: Good evening. Tonight, I can report to the American people and to the world that the United States has conducted an operation that killed Osama bin Laden, the leader of al-Qaida.

SIEGEL: According to writers who investigated the decision to send Navy SEALs to the compound in Abad Abad, Pakistan, the estimates of certainty that bin Laden was the man they'd spotted in the compound covered a range. Both Peter Bergen and Mark Bowden wrote separately that the lead analyst at the CIA put his confidence level at 90 percent or 95 percent. The deputy director of the CIA was at 60

[17]We need to remember that any number attached to the event that bin Laden is at the Abad Abad compound is a statement about "degree-of-belief." It does not refer to any repeatable event. The latter would require something like the following fanciful situation: bin Laden comes and goes from the compound more or less randomly each day; the probability that bin Laden happens to be at the compound when the Navy SEALs show up is based on some estimate of time that he is at home.

percent. Other analysts, they say, settled on a much lower number – 40 percent. A week after the raid, President Obama went on "60 Minutes" and acknowledged that it'd been a tough decision, because the evidence was circumstantial.

(SOUNDBITE OF TV SHOW, "60 MINUTES")

OBAMA: At the end of the day, this was still a 55-45 situation. I mean, we could not say definitively that bin Laden was there.

SIEGEL: Other accounts have Obama concluding it was basically 50-50 – a coin flip. Reading these accounts, Jeff Friedman wondered about how the intelligence analysts had presented their views and whether they could've done it better. Friedman researches national security decision-making as a postdoctoral fellow at Dartmouth. He's the co-author of a paper examining the decision to stage the raid on Abad Abad. And he says this – people's beliefs about how probable something is are subjective and how much we trust those people is also subjective. But Jeff Friedman argues you can still picture the subjective judgments effectively, especially if you remember where people making those judgments are coming from, as in the bin Laden case.

JEFF FRIEDMAN: The low estimate of 30 or 40 percent likelihood that bin Laden was at Abad Abad was issued by a CIA red team. And the red team, which is a common institutional practice, is to be skeptical on purpose, right? They were meant to poke holes in the intelligence. So they, of course, came out the lowest estimate – 30 or 40 percent. The deputy director of Intelligence, Michael Morell, says that he assessed that there was a 60 percent chance that bin Laden would be at Abad Abad. And he tells President Obama explicitly that he's lowballing that assessment a bit because he remembers this assessment of Iraq's weapons of mass destruction. And he knows how easy it is to sort of connect these dots and to be overoptimistic in Intelligence. What does President Obama do, in the end? He says it's a coin flip. We're going with 50-50. So he ends up implicitly giving the most weight to the estimates at the bottom that are the least credible.

In making legal and quasi-legal assessments of guilt or innocence, or to justify certain police actions, it is most common to use verbal phrases to denote levels of evidence (or proof). Some of these are indicated in the list given below along with suggested approximate numerical probabilities that might reflect the strength of evidence or degree of belief. Much of this is discussed in greater detail in Module 9 on Probability and Litigation, particularly in the work of Federal Judge Jack Weinstein and his opinion in the

case of Daniel Fatico redacted in that module.[18]

"Stop-and-Frisk" level:

reasonable suspicion (≈ .20)

"Grand Jury Indictment" level:

probable cause (≈ .40)

"Burden of Proof" levels:

1) preponderance of the evidence (≈ .50+)

2) clear and convincing evidence (≈ .70)

3) clear, unequivocal, and convincing evidence (≈ .80)

4) beyond a reasonable doubt (≈ .95)

"Sure Event" level:

beyond a shadow of a doubt (= 1.0)

Whether actual numerical values are assigned to these phrases that refer to the level of evidence or proof, the induced ordinal ranking that these numbers imply is relevant to our everyday legal discourse. For example in a *New York Times* article by Richard Pérez-Peña (July 2, 2014), entitled "Harvard to Bring on Specialists to Examine Sexual Assault Claims," we have the informative paragraph about the level of evidence now needed to assert sexual harassment and/or assault:

> Under pressure from the Obama administration, many colleges have shifted from a "clear and convincing" standard for finding that a person committed an offense to a looser "preponderance of the evidence" standard. In its new policy, Harvard has also adopted the less stringent standard, which some civil libertarians say tilts the scales too steeply against the accused.

[18]Module 9 on Probability and Litigation also discusses the case "In re As.H" (2004) where the quantification of levels of proof was at issue. The dissenting Associate Judge Farrel noted pointedly: "I believe that the entire effort to quantify the standard of proof beyond a reasonable doubt is a search for fool's gold." There will always be uncertainty as to what numerical probability values should be attached to verbal phrases. But generally if we remember that any mapping needs enough "wiggle room" to be viable, the type of assignment given here should help provide a common frame of reference for those needing to make these evidentiary determinations.

A particularly fertile area where the assessment of probability plays a crucial and varied role is in sports. Baseball, in particular, is a game completely dominated by probabilities, and which dictate how the game is played down to a fine-detailed level. We have the on-going struggles between the batter and the pitcher/catcher combination as to what pitch should be thrown in any given situation;[19] in soccer, there is a related penalty

[19]In major league baseball, the catcher is responsible for selecting the pitch to be thrown by the pitcher; the catcher tries to do pitch selection in such a way that the batter is maximally uncertain as to what pitch will be thrown. (For example, a catcher who wishes for a fast ball, puts down one finger; two fingers (the deuce) indicates a curve ball.) Batters in the major leagues are generally so good that any pitch could be hit if they knew for sure what was coming. In the movie, *Bull Durham*, a wizened old-time catcher, Crash Davis (played by Kevin Costner) is given the task of educating an immature young pitcher with a great arm, Ebby LaLoosh (played by Tim Robbins). Two quotes are given below that show the type of education he receives:

Crash Davis: This son of a bitch is throwing a two-hit shutout. He's shaking me off. You believe that shit? Charlie, here comes the deuce. And when you speak of me, speak well.

...

[Crash calls for a curve ball, Ebby shakes off the pitch twice]

Crash Davis: [stands up] Hey! *Hey*!

[walks to meet Ebby at the mound]

Crash Davis: Why are you shaking me off?

Ebby Calvin LaLoosh: [Gets in Crash's face] I want to give him the heat and announce my presence with authority!

Crash Davis: Announce your fucking presence with authority? This guy is a first ball, fast ball hitter!

Ebby Calvin LaLoosh: Well he hasn't seen my heat!

Crash Davis: [pauses] All right meat, show him your heat.

[Walks back towards the batter's box]

Crash Davis: [to the batter] Fast ball.

...

Ebby Calvin LaLoosh: [pause] God, that sucker teed off on that like he knew I was gonna throw a fastball!

Crash Davis: He did know.

Ebby Calvin LaLoosh: How?

Crash Davis: I told him.

kick confrontation between where the kicker aims the ball in the goal and the direction of goalie movement; in football, there is the long history of what happens on fourth-down plays; in basketball, there are various probability assessments of scoring potential depending on where the ball is on the court and who has it. We even have some quasi-legal verbal descriptions entering the area of sports. For example, in baseball there are now standards for changing a call originally made on the field:

> To change a reviewable call, the Replay Official must determine that there is clear and convincing evidence to change the original call that was made on the field of play. In other words, the original decision of the Umpire shall stand unchanged unless the evidence obtained by the Replay Official leads him to definitively conclude that the call on the field was incorrect.

The Odds Ratio: A Statistic that Only a Statistician's Mother Could Love

As noted in the introduction, it is common in teaching beginning statistics to introduce the terminology of probability by saying that an event, A, occurs with probability, $P(A)$, with the latter represented by a number between zero and one. An alternative way of stating this fact is to say that the "odds" of A occurring is a ratio, $P(A)/(1-P(A)) = P(A)/P(\bar{A})$; that is, the probability of the event A occurring to the event not occurring (or equivalently, to $\bar{A}$ occurring). So, if $P(A) = 2/5$, then the odds of A occurring is $(2/5)(3/5)$ or $(2/3)$, which is read as "2 to 3." Another interpretation is to note that there are $2 + 3 = 5$ chances for A to occur; and that A occurs in 2 out of the 5 for a probability of $2/5$ $(= P(A))$

Now, consider another event B with $P(B) = 4/5$. Here, the odds of B occurring is $P(B)/(1-P(B)) = (4/5)(1/5) = 4/1$, or "4 to 1". When we take the ratio of the odds of B occurring to the odds of A occurring (that is, $(4/1)(2/3)$), the value of 6 is obtained. In words, the odds of B occurring is six times greater than the odds of A occurring. But the real question should be one of how this odds ratio relates to a relative risk of B to A given by $P(B)/P(A) = (4/5)/(2/5) = 2$. Generally, the odds ratio will be larger than the relative

risk; moreover, the odds ratio, because it is such a nontransparent statistic, is consistently (mis)identified in the literature as a relative risk statistic.

To indicate the widespread confusion that exists between relative risk and the odds ratio, the abstract of an article is given below that appeared in *Obstetrics & Gynecology* (2001, *98*, 685–688), entitled "An Odd Measure of Risk: Use and Misuse of the Odds Ratio" (William L. Holcomb, Tinnakorn Chaiworapongsa, Douglas A. Luke, & Kevin D. Burgdorf):

OBJECTIVE: To determine how often the odds ratio, as used in clinical research of obstetrics and gynecology, differs substantially from the risk ratio estimate and to assess whether the difference in these measures leads to misinterpretation of research results.

METHODS: Articles from 1998 through 1999 in *Obstetrics & Gynecology* and the *American Journal of Obstetrics and Gynecology* were searched for the term "odds ratio." The key odds ratio in each article was identified, and, when possible, an estimated risk ratio was calculated. The odds ratios and the estimated risk ratios were compared quantitatively and graphically.

RESULTS: Of 151 studies using odds ratios, 107 were suitable to estimate a risk ratio. The difference between the odds ratio and the estimated risk ratio was greater than 20% in 47 (44%) of these articles. An odds ratio appears to magnify an effect compared with a risk ratio. In 39 (26%) articles the odds ratio was interpreted as a risk ratio without explicit justification.

CONCLUSION: The odds ratio is frequently used, and often misinterpreted, in the current literature of obstetrics and gynecology.

In general, an odds ratio is a reasonable approximation to relative risk only when the disease frequency is small (for example, in our "exposed versus not exposed" by "disease versus no disease" contingency table). Otherwise, the odds ratio can be a serious overestimate. About the only place that odds ratios may have a justifiable place is in what are called case-control studies designed for the assessment of rare events (for example, when dealing with diseases that have very low frequencies). In these cases the distinction in risk assessment produced by interpreting an odds ratio as a relative risk may be negligible. In all other instances, however, odds ratios should be avoided.

A final cautionary tale illustrates the damage that can be done when the media picks

up on a story and confuses an odds ratio with relative risk. We give a short article that appeared in the *New York Times* (February 25, 1999), entitled "Doctor Bias May Affect Heart Care, Study Finds":

> Unconscious prejudices among doctors may help explain why women and blacks complaining of chest pain are less likely than men and whites to receive the best cardiac testing, a study in today's issue of *The New England Journal of Medicine* suggests.
>
> A new study of 720 physicians found that with all symptoms being equal, doctors were 60 percent as likely to order cardiac catheterization for women and blacks as for men and whites. For black women, the doctors were 40 percent as likely to order catheterization, considered the gold standard diagnostic test for heart disease.
>
> "Most likely this is an underestimate of what's occurring," Dr. Kevin Schulman of Georgetown University Medical Center said, because the doctors knew their decisions were being recorded, but not why.

Sometime later (August 17, 1999), the *Times* published the following "Correction":

> A brief report by The Associated Press on Feb. 25 about a study of bias in heart care cited a statistic incorrectly. The study, published in *The New England Journal of Medicine*, showed that doctors were 7 percent less likely to order cardiac catheterization tests for female or black patients than for male or white patients – not 40 percent less likely. The error is discussed in the current issue of the journal. Editors of the journal told the A.P. that they "take responsibility for the media's overinterpretation" of the study, which used an unusual statistical method.

The "unusual statistical method" referred to is the use of odds ratios. The article in *The New England Journal of Medicine* (1999, *341*, 279–283) that critiqued the Schulman et al. piece was entitled "Misunderstandings About the Effect of Race and Sex on Physicians' Referrals for Cardiac Catheterizations" (Lisa M. Schwartz, Steven Woloshin, & H. Gilbert Welch). The abstract and two explanatory paragraphs from this later article follow:

> In the February 25 issue of the *Journal*, Schulman et al. claimed that the "race and sex of a patient independently influence how physicians manage chest pain." Their study received extensive coverage in the news media. It was reported in most major newspapers and was a feature story on ABC's *Nightline*, with Surgeon General David Satcher providing commentary. Unfortunately, in each case, the results were overstated. We explore what went wrong and suggest ways to improve the communication of data to the public. Our purpose is not to deny the occurrence of racial or sex bias, rather to emphasize the importance of presenting information accurately.

...

The use of odds ratios is unfortunate. Few people think in terms of odds or encounter them in daily life. Perhaps for this reason, many people tend to equate odds with probability (the most familiar way to characterize chance) and thus to equate odds ratios with risk ratios. The quotations noted in Table 2 [given in the article] suggest that the major newspapers, *Nightline*, and even the surgeon general did just that in characterizing the results of the study by Schulman et al. When the outcome of interest is uncommon (i.e., it occurs less than 10 percent of the time), such confusion makes little difference, since odds ratios and risk ratios are approximately equal. When the outcome is more common, however, the odds ratio increasingly overstates the risk ratio.

Because the study by Schulman et al. involved a very common event (84.7 percent of blacks and 90.6 percent of whites were referred for catheterization), the overstatement in this case was extreme. The reported odds ratio of 0.6 actually corresponds to a risk ratio of 0.93 (i.e., 84.7 percent divided by 90.6 percent). Inappropriately equating odds ratios with risk ratios led to the mistaken impression that blacks had a 40 percent lower probability of referral than whites, whereas in fact, the probability of referral for blacks was 7 percent lower. In this case, the failure to distinguish between odds ratios and risk ratios had profound consequences for how the magnitude of the difference in referral rates for blacks and whites (or women and men) was portrayed. Regardless of the magnitude, however, the comparison itself was misleading.

Probabilistic Reasoning and the Prediction of Human Behavior

The assignment of a (degree-of-belief) probability to an event is a form of prediction where a numerical assessment is given to the likelihood that some event will occur. As noted in the case of Angelina Jolie, the probability of her contracting breast cancer because of the BRAC1 mutation was set at .87. Presumably, this later value is an empirically generated estimate based on the specific "at risk" group(s) to which she belongs. Although medical databases may be extensive enough to arrive at these kinds of decisive assignments, events that involve human behavior are generally more difficult to predict and thus to assign reasonable numerical values to various event occurrences that could be ethically justified. Module 2, for example, is devoted to clinical (that is, expert judgement) and actuarial (that is, statistical) predictions of dangerous behavior; such prediction is of

interest for various legal purposes such as civil commitment, or the granting of parole or bail. As will be shown in that module, and no matter how much society would wish it to be otherwise, we don't do very well in predicting dangerous behavior – or in the vernacular, we generally "suck" at behavioral prediction irrespective of whether it is done clinically or actuarially. This unfortunate fact remains true in the face of all the "risk assessment" instruments offered and touted in the literature.

Although there is now ample evidence that the reliable prediction of human behavior that might be of interest to the criminal justice system is extremely difficult (if not damn near impossible), "hope springs eternal in the human breast." There is now the push for evidence-based sentencing (EBS) that depending on the prediction of a future recidivism might change an individual's length of sentence. The reason given for this push is an analogy to the build-up of the Oakland Athletics baseball team in the early 2000s; here, the argument goes something like the following: "well, if Billy Beane can get a great team with predictive analytics, we obviously can do the same in the criminal justice context." For a particular uniformed (by any data) TED talk on this fraught analogy, see Anne Milgram's "Why Smart Statistics Are the Key to Fighting Crime" (filmed October 2013). A recent and highly informative *Stanford Law Review* (2014, *66*, 803–872) article by Sonja Starr entitled "Evidence-Based Sentencing and the Scientific Rationalization of Discrimination," discusses in some detail the constitutional issues involved in EBS. We give part of an Attorney General Eric Holder speech (delivered at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference; August 1, 2014) that issues appropriate cautions about EBS:

> It's increasingly clear that, in the context of directing law enforcement resources and improving reentry programs, intensive analysis and data-driven solutions can help us achieve significant successes while reducing costs. But particularly when it comes to front-end applications – such as sentencing decisions, where a handful of states are now attempting to employ this methodology – we need to be sure the use of aggregate data analysis won't have unintended consequences.
>
> Here in Pennsylvania and elsewhere, legislators have introduced the concept of "risk assessments"

that seek to assign a probability to an individual's likelihood of committing future crimes and, based on those risk assessments, make sentencing determinations. Although these measures were crafted with the best of intentions, I am concerned that they may inadvertently undermine our efforts to ensure individualized and equal justice. By basing sentencing decisions on static factors and immutable characteristics – like the defendant's education level, socioeconomic background, or neighborhood – they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.

Criminal sentences must be based on the facts, the law, the actual crimes committed, the circumstances surrounding each individual case, and the defendant's history of criminal conduct. They should not be based on unchangeable factors that a person cannot control, or on the possibility of a future crime that has not taken place. Equal justice can only mean individualized justice, with charges, convictions, and sentences befitting the conduct of each defendant and the particular crime he or she commits. And that's why, this week, the Justice Department is taking the important step of urging the Sentencing Commission to study the use of data-driven analysis in front-end sentencing – and to issue policy recommendations based on this careful, independent analysis.

There are two general approaches to the prediction of human behavior. One is through the use of data that pertains to only one specific individual such as age, previous criminal history, and mental status. The second concerns what particular groups a person might belong to, such as having the BRAC1 genetic mutation, race, sex, and ethnicity. In legal contexts, the prediction of a specific person's behavior through individual variables like past criminal behavior is typically permissible; but when prediction is made based on the group(s) one is in, such as race or gender, that usage is usually unconstitutional (see the earlier Federal Rules of Evidence and the distinction between evidence that may be relevant but inadmissible under Rule 403).

There are other methods of prediction that even if not inadmissible in a court of law, should nevertheless be excluded. One good example would be the labeling done by so-called (clinical) experts that by itself supposedly predicts behavior reliably. There is the notorious example of James Grigson ("Dr. Death") discussed in Module 2 who justified imposing a death sentence under Texas law by simply assigning the label of "sociopath" to a defendant; in Grigson's view this meant that a perfect prediction of violent behavior was

possible, and thus, the defendant should be executed. A second current example involves evidence-based-sentencing which contends that we can obviously predict recidivism extremely well because of *Moneyball* – this is sophistry at best.

Besides the pernicious assignment of a label such as "sociopath" to a single individual (which at one time in Texas allowed that individual's execution to proceed), a group of criminologists and sociologists in the 1980s engaged in the faulty labeling of a large swath of upcoming teenagers as "superpredators" without any credible evidence whatsoever. The latter assertion that superpredators were about to emerge, led many states to enact laws permitting the incarceration of children to life without parole. We give parts of an article by Gail Garinger from the *New York Times* (March 14, 2012), entitled "Juveniles Don't Deserve Life Sentences," which provides some of the background for these misguided "get tough on crime" efforts:

> In the late 1980s, a small but influential group of criminologists predicted a coming wave of violent juvenile crime: "superpredators," as young as 11, committing crimes in "wolf packs." Politicians soon responded to those fears, and to concerns about the perceived inadequacies of state juvenile justice systems, by lowering the age at which children could be transferred to adult courts. The concern was that offenders prosecuted as juveniles would have to be released at age 18 or 21.
>
> At the same time, "tough on crime" rhetoric led some states to enact laws making it easier to impose life without parole sentences on adults. The unintended consequence of these laws was that children as young as 13 and 14 who were charged as adults became subject to life without parole sentences.
>
> Nationwide, 79 young adolescents have been sentenced to die in prison – a sentence not imposed on children anywhere else in the world. These children were told that they could never change and that no one cared what became of them. They were denied access to education and rehabilitation programs and left without help or hope.
>
> But the prediction of a generation of superpredators never came to pass. Beginning in the mid-1990s, violent juvenile crime declined, and it has continued to decline through the present day. The laws that were passed to deal with them, however, continue to exist. This month, the United States Supreme Court will hear oral arguments in two cases, *Jackson v. Hobbs* and *Miller v. Alabama*, which will decide whether children can be sentenced to life without parole after being convicted of homicide.
>
> The court has already struck down the death penalty for juveniles and life without parole for young

offenders convicted in nonhomicide cases. The rationale for these earlier decisions is simple and equally applicable to the cases to be heard: Young people are biologically different from adults. Brain imaging studies reveal that the regions of the adolescent brain responsible for controlling thoughts, actions and emotions are not fully developed. They cannot be held to the same standards when they commit terrible wrongs.

Homicide is the worst crime, but in striking down the juvenile death penalty in 2005, the Supreme Court recognized that even in the most serious murder cases, "juvenile offenders cannot with reliability be classified among the worst offenders": they are less mature, more vulnerable to peer pressure, cannot escape from dangerous environments, and their characters are still in formation. And because they remain unformed, it is impossible to assume that they will always present an unacceptable risk to public safety.

The most disturbing part of the superpredator myth is that it presupposed that certain children were hopelessly defective, perhaps genetically so. Today, few believe that criminal genes are inherited, except in the sense that parental abuse and negative home lives can leave children with little hope and limited choices.

As a former juvenile court judge, I have seen firsthand the enormous capacity of children to change and turn themselves around. The same malleability that makes them vulnerable to peer pressure also makes them promising candidates for rehabilitation.

An overwhelming majority of young offenders grow out of crime. But it is impossible at the time of sentencing for mental health professionals to predict which youngsters will fall within that majority and grow up to be productive, law-abiding citizens and which will fall into the small minority that continue to commit crimes. For this reason, the court has previously recognized that children should not be condemned to die in prison without being given a "meaningful opportunity to obtain release based on demonstrated maturity and rehabilitation."

The criminologists who promoted the superpredator theory have acknowledged that their prediction never came to pass, repudiated the theory and expressed regret. They have joined several dozen other criminologists in an *amicus* brief to the court asking it to strike down life without parole sentences for children convicted of murder. I urge the justices to apply the logic and the wisdom of their earlier decisions and affirm that the best time to decide whether someone should spend his entire life in prison is when he has grown to be an adult, not when he is still a child.

The cases mentioned in the article, *Jackson v. Hobbs* and *Miller v. Alabama*, were decided in favor of the juveniles, Jackson and Miller, in 2012. The 5 to 4 judgement of the Court, delivered by Justice Elena Kagan, follows:

The two 14-year-old offenders in these cases were convicted of murder and sentenced to life imprisonment without the possibility of parole. In neither case did the sentencing authority have any discretion to impose a different punishment. State law mandated that each juvenile die in prison even if a judge or jury would have thought that his youth and its attendant characteristics, along with the nature of his crime, made a lesser sentence (for example, life with the possibility of parole) more appropriate. Such a scheme prevents those meting out punishment from considering a juvenile's "lessened culpability" and greater "capacity for change, and runs afoul of our cases' requirement of individualized sentencing for defendants facing the most serious penalties. We therefore hold that mandatory life without parole for those under the age of 18 at the time of their crimes violates the Eighth Amendment's prohibition on "cruel and unusual punishments."

When behavioral prediction for relatively rare events is attempted actuarially, the facilitative effect of the available evidence is typically small, and sometimes painfully so. We never reach the type of situation wished for by Johnnie Cochran where the occurrence of some condition makes another event a "sure thing." As discussed in Module 4 on diagnostic testing, given the usual type of individual level information available, we rarely can outperform a simple prediction rule using base rates in terms of the number of correct predictions; for example, a simple base rate rule might merely assert that everyone will be non-violent.

Even though experts might be tasked with the prediction of rare events, there also seems to be the unreal expectation that this should be done perfectly, irrespective of the available evidence. The reasoning goes that if Billy Beane can do it for the Oakland Athletics, it also must be possible to do close-to-perfect prognostications throughout the criminal justice system. As an extreme case of such twisted reasoning, there is the Italian judge who convicted seven seismologists of manslaughter when they failed to predict or give a warning for a specific earthquake that occurred on April 6, 2009. Several paragraphs about this incident are given below taken from an article by Florin Diacu in the *New York*

Times (October 26, 2012) entitled "Is Failure to Predict a Crime?":

> I learned with disbelief on Monday about the decision of an Italian judge to convict seven scientific experts of manslaughter and to sentence them to six years in prison for failing to give warning before the April 2009 earthquake that killed 309 people, injured an additional 1,500 or so and left more than 65,000 people homeless in and around the city of L'Aquila in central Italy.
>
> By this distorted logic, surgeons who warn a patient that there's a small chance of dying during surgery should be put in prison if the patient does, in fact, die. Imagine the consequences for the health system. The effect on other fields would be just as devastating. In response to the verdict, some Italian scientists have already resigned from key public safety positions. Unless this shortsighted verdict is overturned by wiser minds, it will be very harmful in the long run.
>
> In L'Aquila, the scientists presented a risk assessment in late March 2009 after small seismic events made the public anxious. They found that a major quake was unlikely. Certainly, the timing of the scientists' statements played against them. On April 6, a 6.3-magnitude earthquake devastated the area, where earthquakes had been recorded since 1315. And L'Aquila is built on the bed of a dry lake, so the soil tends to amplify the motions of the ground. These facts, however, do not alter the truth of the scientists' claim that earthquakes are extremely rare there. One of the most important ones took place back in 1703.

We might end this section on predicting human behavior with a clever twist on Reinhold Niebuhr's Serenity Prayer given by Nate Silver in his well-received book, *The Signal and the Noise*:

> Prediction is difficult for us for the same reason that it is so important: it is where objective and subjective reality intersect. Distinguishing the signal from the noise requires both scientific knowledge and self-knowledge: the serenity to accept the things we cannot predict, the courage to predict the things we can, and the wisdom to know the difference. (p. 453)

Where to Go From Here

The short introduction to applied probabilistic reasoning given by this primer had to be necessarily selective in the topics presented. To make up for this brevity, more extensive coverage is available through a series of instructional modules that cover specific topic areas in probabilistic reasoning. The general web site directory containing these modules has the following address:

http://cda.psych.uiuc.edu/applied_probabilistic_reasoning

Several of the modules include material adapted and rewritten from *A Statistical Guide for the Ethically Perplexed* (Lawrence Hubert and Howard Wainer, 2013); other modules have been newly constructed for this site (for example, Module 4 on diagnostic testing). The twelve modules are listed below along with short abstracts that give the topic(s) covered by the particular module (all except for the first module which is just this current primer):

Applied Probabilistic Reasoning: A *Vade Mecum* to Accompany a First Course in Statistics

Module 1: A Brief Primer on Applied Probabilistic Reasoning

Module 2: The (Un)reliability of Clinical and Actuarial Predictions of Dangerous Behavior

The prediction of dangerous and/or violent behavior is important to the conduct of the United States justice system in making decisions about restrictions of personal freedom such as preventive detention, forensic commitment, or parole. This module discusses behavioral prediction both through clinical judgement as well as actuarial assessment. The general conclusion drawn is that for both clinical and actuarial prediction of dangerous behavior, we are far from a level of accuracy that could justify routine use. To support this later negative assessment, two topic areas are discussed at some length: 1) the MacArthur Study of Mental Disorder and Violence, including the actuarial instrument developed as part of this project (the Classification of Violence Risk (COVR)), along with all the data collected that helped develop the instrument; 2) the Supreme Court case of *Barefoot v. Estelle* (1983) and the American Psychiatric Association "friend of the court" brief on the (in)accuracy of clinical prediction for the commission of future violence. An elegant Justice Blackmun dissent is given in its entirety that contradicts the majority decision that held: There is no merit to petitioner's argument that psychiatrists, individually and as a group, are incompetent to predict with an acceptable degree of reliability that a particular criminal will commit other crimes in the future, and so represent a danger to the community.

Module 3: The Analysis of $2 \times 2 \times 2$ (Multiway) Contingency Tables: Explaining Simpson's

Paradox and Demonstrating Racial Bias in the Imposition of the Death Penalty

This module discusses the two major topics of Simpson's paradox and the Supreme Court decision in *McCleskey v. Kemp* (1987). Simpson's paradox is ubiquitous in the misinterpretation of data; it is said to be present whenever a relationship that appears to exist at an aggregated level disappears or reverses when disaggregated and viewed within levels. A common mechanism for displaying data that manifests such a reversal phenomenon is through a multiway contingency table, often of the $2 \times 2 \times 2$ variety. For example, much of the evidence discussed in *McCleskey v. Kemp* was cross-categorized by three dichotomous variables: race of the victim (black or white), race of the defendant (black or white), and whether the death penalty was imposed (yes or no). Despite incontrovertible evidence that the race of the victim plays a significant role in whether the death penalty is imposed, the holding in *McClesky v. Kemp* was as follows: Despite statistical evidence of a profound racial disparity in application of the death penalty, such evidence is insufficient to invalidate defendant's death sentence.

Module 4: Probabilistic Reasoning and Diagnostic Testing

Two main questions are discussed that relate to diagnostic testing. First, when does prediction using simple base rate information outperform prediction with an actual diagnostic test?; and second, how should the performance of a diagnostic test be evaluated in general? Module 2 on the (un)reliability of clinical and actuarial prediction introduced the Meehl and Rosen (1955) notion of "clinical efficiency," which is a phrase applied to a diagnostic test when it outperforms base rate predictions. In the first section to follow, three equivalent conditions are given for when "clinical efficiency" holds; these conditions are attributed to Meehl and Rosen (1955), Dawes (1962), and Bokhari and Hubert (2015). The second main section of this module introduces the Receiver Operating Characteristic (ROC) curve, and contrasts the use of a common measure of test performance, the "area under the curve" (AUC), with possibly more appropriate performance measures that take base rates into consideration. A final section of the module discusses several issues that

must be faced when implementing screening programs: the evidence for the (in)effectiveness of cancer screening for breast (through mammography) and prostate (through the prostate-specific antigen (PSA) test); premarital screening debacles; prenatal screening; the cost of screening versus effectiveness; the ineffectiveness of airport behavioral detection programs implemented by the Transportation Security Administration (TSA); informed consent and screening; the social pressure to screen.

Module 5: Probabilistic Reasoning in the Service of Gambling

Probabilistic reasoning is applied to several topics in gambling. We begin with the Chevalier de Méré asking the mathematician Blaise Pascal in the early 17th century for help with his gambling interests. Pascal in a series of letters with another mathematician, Pierre de Fermat, laid out what was to be the foundations for a modern theory of probability. Some of this formalization is briefly reviewed; also, to give several numerical examples, the Pascal-Fermat framework is applied to the type of gambles the Chevelier engaged in. Several other gambling related topics are discussed at some length: spread betting, parimutuel betting, and the psychological considerations behind gambling studied by Tversky, Kahneman, and others concerned with the the psychology of choice and decision making.

Module 6: Probabilistic Reasoning Through the Basic Sampling Model

One mechanism for assisting in various tasks encountered in probabilistic reasoning is to adopt a simple sampling model. A population of interest is first posited, characterized by some random variable, say X. This random variable has a population distribution (often assumed to be normal), characterized by (unknown) parameters. The sampling model posits n independent observations on X, denoted by $X_1, \ldots, X_n$, and which constitutes the sample. Various functions of the sample can then be constructed (that is, various statistics can be computed such as the sample mean and sample variance); in turn, statistics have their own sampling distributions. The general problem of statistical inference is to ask what sample statistics tell us about their population counterparts; for

example, how can we construct a confidence interval for a population parameter such as the population mean from the sampling distribution for the sample mean.

Under the framework of a basic sampling model, a number of topics are discussed: confidence interval construction for a population mean where the length of the interval is determined by the square root of the sample size; the Central Limit theorem and the Law of Large Numbers; the influence that sample size and variability have on our probabilistic reasoning skills; the massive fraud case involving the Dutch social psychologist, Diederik Stapel, and the role that lack of variability played in his exposure; the ubiquitous phenomenon of regression toward the mean and the importance it has for many of our probabilistic misunderstandings; how reliability corrections can be incorporated into prediction; the dichotomy and controversy encountered every ten years about complete enumeration versus sampling (to correct for, say, an undercount) in the United States Census.

Module 7: Probabilistic (Mis)Reasoning and Related Confusions

The introductory module started with the well-known case of Sally Clark and how a misunderstanding about probabilistic independence helped lead to her wrongful imprisonment for killing her two children. The present module will provide more examples of mistaken probabilistic reasoning, with many involving misinterpretations of conditional probability. We will revisit the O.J. Simpson criminal case where his defense team took advantage of what is termed the "Defendant's Fallacy," as well as some specious reasoning about conditional probability (perpetrated by Alan Dershowitz). Several additional high-profile legal cases will be mentioned that were mishandled because of the prosecutor's fallacy, much like that of Sally Clark. One is recent – the Dutch nurse, Lucia de Berk, was accused of multiple deaths at the hospitals she worked at in the Netherlands; another is much older and involves the turn-of-the-century (the late 1800s, that is) case of Alfred Dreyfus, the much maligned French Jew who was falsely imprisoned for espionage.

Module 8: Probabilistic Reasoning, Forensic Evidence, and the Relevance of Base Rates

The topics developed in this module have at least a tacit connection to Bayes' theorem, and specifically to how base rates operate formally in the use of Bayes' theorem as well as more informally for several legally-related contexts. A number of topic areas are pursued: the general unreliability of eyewitness identification and testimony; polygraph testing; the assessment of blood alcohol level; the legal status and use of base rates; racial and ethnic profiling; false confessions; police interrogations; and the overall dismal state of the forensic "sciences."

An earlier Module 4 discussed the relevance of base rates in the evaluation of diagnostic tests and did so in several important contexts. One involved the Meehl and Rosen (1955) notion of "clinical efficiency" where prediction with a diagnostic test could be shown to outperform prediction using simple base rates. A second was a critique of the area under a Receiver Operating Characteristic curve (the AUC) as the sole mechanism for evaluating how well a particular diagnostic test performs; in general, the AUC is independent of base rates and fails to assess how well a diagnostic instrument does in specific populations that have relatively low base rates for the characteristic to be detected. When base rates are equal, test sensitivity and the positive predictive value (PPV) are equal (and so are the negative predictive value (NPV) and test specificity). Because of these equivalences, simple functions of the PPV and NPV make sense in communicating just how well or how badly a diagnostic instrument performs.

Module 9: Probability and Litigation

This module explores the connection between statements that involve probabilities and those phrases used for evidentiary purposes in the courts. We begin with Jack Weinstein, a federal judge in the Eastern District of New York, and his views on the place that probability has in litigation. Jack Weinstein may be the only federal judge ever to publish an article in a major statistics journal; his primary interests center around subjective probability and how these relate, among others, to the four levels of a "legal burden of proof": preponderance of the evidence; clear and convincing evidence; clear,

unequivocal, and convincing evidence; and proof beyond a reasonable doubt. The broad topic area of probability scales and rulers is discussed in relation to several more specific subtopics: Jeremy Bentham and his suggestion of a "persuasion thermometer"; some of Jack Weinstein's legal rulings where probabilistic assessments were made: the cases of Vincent Gigante, Agent Orange, and Daniel Fatico. An appendix gives a redacted Weinstein opinion in this later Fatico case. Two other appendices are also given: the text of Maimonides' 290th Negative Commandment, and a District of Columbia Court of Appeals opinion "In re As.H" (2004) that dealt with the assignment of subjective probabilities and various attendant verbal phrases in eyewitness testimony.

Module 10: Sleuthing with Probability and Statistics

Statistical sleuthing is concerned with the use of various probabilistic and statistical tools and methods to help explain or "tell the story" about some given situation. In this type of statistical detective work, a variety of probability distributions can prove useful as models for a given underlying process. These distributions include the Bernoulli, binomial, normal, Poisson (especially for spatial randomness and the assessment of "Poisson clumping"). Other elucidating probabilistic topics introduced include Benford's Law, the "birthday probability model," survival analysis and Kaplan-Meier curves, the Monty Hall problem, and what is called the "secretary problem" (or more pretentiously, the "theory of optimal stopping"). An amusing instance of the latter secretary problem is given as a *Car Talk* Puzzler called the "Three Slips of Paper"; a full listing of the script from the NPR show is included that aired on February 12, 2011.

Module 11: Cross-validation and the Control of Error Rates

This module emphasizes what might be termed "the practice of safe statistics." The discussion is split into three parts: (1) the importance of cross-validation for any statistical method that relies on an optimization process based on a given data set (or sample); (2) the need to exert control on overall error rates when carrying out multiple testing, even when that testing is done only implicitly; (3) in the context of "big data" and associated

methods for "data mining," the necessity of some mechanism for ensuring the replicability of "found results."

Module 12: An Olio of Topics in Applied Probabilistic Reasoning

The last module is a collection of topics in applied probabilistic reasoning that were all too small to command their own separate modules. Topics include: 1) the randomized response method as a way of asking sensitive questions and hopefully receiving truthful answers; 2) the use of surrogate end points (or proxies) in the study of some phenomenon where the connections to "real" outcomes of interest (for example, to mortality) are indirect and probabilistically linked (for example, to lowered cholesterol levels); 3) the comparison between a normative theory of choice and decision making derived from probability theory and actual human performance; 4) permutation tests and statistical inference derived directly from how a randomized controlled study was conducted. As an oddity that can occur for this type of statistical inference procedure, the famous 1954 Salk polio vaccine trials are discussed. Also, three brief subsections are given that summarize the jackknife, the bootstrap, and permutation tests involving correlational measures. This latter material is provided in an abbreviated form suitable for slide presentation in class, and where further explanatory detail would be given by an instructor.

References

Devlin, K., & Lorden, G. (2007). *The Numbers Behind NUMB3RS: Solving Crime with Mathematics*. New York: Penguin Group.

Gigerenzer, G. (2002). *Calculated Risks: How to Know When Numbers Deceive You*. New York: Simon & Schuster.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 55–96.

Jeffreys, Harold (1973). *Scientific Inference* (3rd ed.). New York: Cambridge University

Press.

Schum, David A. (1994). *The Evidential Foundations of Probabilistic Reasoning*. New York: Wiley.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.