Probability Theory: Background and Bayes Theorem

Psychology (Statistics) 484

Statistics, Ethics, and the Social and Behavioral Sciences

June 13, 2013

Beginning Quotations

Probability theory is nothing but common sense reduced to calculation.
— Pierre-Simon Laplace (*Essai Philosophique Sur Les Probabilités*, 1814)

Misunderstanding of probability may be the greatest of all impediments to scientific literacy.
— Stephen Jay Gould

— the case of Sally Clark, wrongly convicted in England of killing her two children; this miscarriage of justice was due to an inappropriate assumption of statistical independence and the commission of the "Prosecutor's Fallacy"

— breast cancer screening though mammograms; understanding Bayes' theorem, test sensitivity and specificity, prior probabilities, and the positive predictive value (for example, what is the probability of having breast cancer if the mammogram is "positive"?)

Required Reading:

SGEP (19–86) —
The (Mis)assignment of Probabilities
The Probabilistic Generalizations of Logical Fallacies Are No Longer Fallacies
Using Bayes' Rule to Assess the Consequences of Screening for Rare Events
Ethical issues in medical screening
Bayes' Rule and the Confusion of Conditional Probabilities
Bayes' Rule and the Importance of Base Rates
The (legal) status of the use of base rates
Forensic evidence generally

Popular Articles —
Trawling the Brain, Laura Sanders (*ScienceNews*, October 19, 2009)
The Cancer-Cluster Myth, Atul Gawande (*New Yorker*, October 4, 1999)
Duped, Margaret Talbot (*New Yorker*, July 2, 2007)

Suggested Reading:
Suggested Reading Relevant to the Whole Chapter
Suggested Reading on Probability Issues
Suggested Reading on Forensic Issues
Suggested Reading on Screening

Film:

Unfinished Business (58 minutes)

The Infamous Dreyfus Affair (50 minutes)

Probability: Event Independence

The formalism of thought offered by probability theory is one of the more useful portions of any beginning course in statistics in helping to promote ethical reasoning.

As typically presented, we speak of an event represented by a capital letter, say A, and the probability of the event as some number in the range from 0 to 1, 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 mutually exclusivity implies that both events cannot occur at the same time),
$P(A \text{ or } B) = P(A) + P(B)$.

As a final beginning definition, we say that two events are independent whenever the probability of the joint event, $P(A \text{ and } B)$, factors as the product of the individual probabilities, $P(A)P(B)$.

The Case of Sally Clark

The idea of statistical independence and the factoring of the joint event probability immediately provides a formal tool for understanding a number of historical miscarriages of justice.

In particular, if two events are not independent, then the joint probability cannot be generated by a simple product of the individual probabilities.

A recent example is the case of Sally Clark; she was convicted in England of 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.

People v. Collins

Several other examples of a misuse for the idea of statistical independence exist in the legal literature, such as the notorious 1968 jury trial in California, *People v. Collins*.

Here, the prosecutor suggested that the jury merely multiply several probabilities together, which he conveniently provided, to ascertain the guilt of the defendant.

In overturning the conviction, the Supreme Court of California criticized both the statistical reasoning and the framing of the decision for the jury:

We deal here with the novel question whether evidence of mathematical probability has been properly introduced and used by the prosecution in a criminal case. . . . Mathematics, a veritable sorcerer in our computerized society, while assisting the trier of fact in the search of truth, must not cast a spell over him. We conclude that on the record before us, defendant should not have had his guilt determined by the odds and that he is entitled to a new trial. We reverse the judgment.

Conditional Probability

The definition of conditional probability plays a central role in all our uses of probability theory;

in fact, many misapplications of statistical/probabilistic reasoning involve confusions of some sort regarding conditional probabilities.

Formally, the conditional probability of some event A given that B has already occurred, denoted $P(A|B)$, is defined as $P(A \text{ and } B)/P(B)$.

When A and B are independent,
$P(A|B) = P(A)P(B)/P(B) = P(A)$;

or in words, knowing that B has occurred does not alter the probability of A occurring.

If $P(A|B) > P(A)$, we will say that B is "facilitative" of A; when $P(A|B) < P(A)$, B is said to be "inhibitive" of A.

In any case, the size and sign of the difference between $P(A|B)$ and $P(A)$ is an obvious raw descriptive measure of how much the occurrence of B is associated with an increased or decreased probability of A, with a value of zero corresponding to statistical independence.

The Urn Model

One convenient device for interpreting probabilities and understanding how events can be "facilitative" or "inhibitive" is through the use of a simple 2×2 contingency table that cross-classifies a set of objects according to the events A and $\bar{A}$, and B and $\bar{B}$ (here, $\bar{A}$ and $\bar{B}$ represent the complements of A and B, which occur when the original events do not).

For example, suppose we have a collection of N balls placed in a container; each ball is labeled A or $\bar{A}$, and also B or $\bar{B}$, according to the notationally self-evident table of frequencies below:

	A	$\bar{A}$	Row Sums
B	N_{AB}	$N_{\bar{A}B}$	N_B
$\bar{B}$	$N_{A\bar{B}}$	$N_{\bar{A}\bar{B}}$	$N_{\bar{B}}$
Column Sums	N_A	$N_{\bar{A}}$	N

The process we consider is one of picking a ball blindly from the container, where the balls are assumed to be mixed thoroughly, and noting the occurrence of the events A or $\bar{A}$ and B or $\bar{B}$.

Based on this physical idealization of such a selection process, it is intuitively reasonable to assign probabilities according to the proportion of balls in the container satisfying the attendant conditions.

An Example

As a numerical example of using a 2×2 contingency table to help explicate probabilistic reasoning, suppose we have an assumed population of 10,000, cross-classified according to the presence or absence of Colorectal Cancer (CC) [A: +CC; $\bar{A}$: −CC], and the status of a Fecal Occult Blood Test (FOBT) [B: +FOBT; $\bar{B}$: −FOBT].

Using data from Gerd Gigerenzer, *Calculated Risks* (2002, pp. 104–107), we have the following 2×2 table:

	+CC	−CC	Row Sums
+FOBT	15	299	314
−FOBT	15	9671	9686
Column Sums	30	9970	10,000

The probability $P(+CC \mid +FOBT)$ is simply $15/314 = .048$, based on the frequency value of 15 for the cell (+FOBT, +CC), and the +FOBT row sum of 314.

The marginal probability, P(+CC), is $30/10{,}000 = .003$, and thus, a positive FOBT is "facilitative" of a positive CC because .048 is greater than .003.

The size of the difference, $P(+CC \mid +FBOT) - P(+CC) = +.045$, may not be large in any absolute sense, but the change does represent a fifteenfold increase over the marginal probability of .003.

(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.)

Some Useful Probability Results

(1) For the complementary event, $\bar{A}$, which occurs when A does not, $P(\bar{A}) = 1 - P(A)$.

(2) For events A and B that are not necessarily mutually exclusive,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \ .$$

(3) The *rule of total probability*: Given a collection of mutually exclusive and exhaustive events, $B_1, \ldots, B_K$ (that is, all are pairwise mutually exclusive and their union gives the sure event),

$$P(A) = \sum_{k=1}^{K} P(A|B_k)P(B_k) \ .$$

(4) Bayes' theorem (or rule) for two events, A and B:

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

(5) Bonferroni inequality: for a collection of events, $A_1, \ldots, A_K$,

$$P(A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_K) \leq \sum_{k=1}^{K} P(A_k) .$$

(6) $P(A \text{ and } B) \leq P(A \text{ or } B) \leq P(A) + P(B)$.

In words, the first inequality results from the event "A and B" being wholly contained within the event "A or B" ("A or B" occurs when A or B or both occur); the second results from the Bonferroni inequality restricted to two events.

(7) $P(A \text{ and } B) \leq \min(P(A), P(B)) \leq P(A) \text{ or } \leq P(B)$.

In words, the first inequality results from the event "A and B" being wholly contained both within A and within B; the second inequalities are more generally appropriate—the minimum of any two numbers is always less than either of the two numbers.

The (Mis)assignment of Probabilities

Although the assignment of probabilities to events consistent with the disjoint rule may lead to an internally valid system mathematically, there is still no assurance that this assignment is "meaningful," or bears any empirical validity for observable long-run expected frequencies.

There seems to be a never-ending string of misunderstandings in the way probabilities can be generated that are either blatantly wrong, or more subtly incorrect, irrespective of the internally consistent system they might lead to.

One inappropriate way of generating probabilities is to compute the likelihood of some joint occurrence after some of the outcomes are already known.

For example, there is the story about the statistician who takes a bomb aboard a plane, reasoning that if the probability of one bomb on board is small, the probability of two is infinitesimal.

Or, during World War I, soldiers were actively encouraged to use fresh shell holes as shelter because it was very unlikely for two shells to hit the same spot during the same day.

Whenever coincidences are culled or "hot spots" identified from a search of available information, the probabilities that are then regenerated for these situations may not be valid.

There are several ways of saying this:

when some set of observations is the source of an initial suspicion, those same observations should not be used in a calculation that then tests the validity of the suspicion.

In Bayesian terms, you should not obtain the posterior probabilities from the same information that gave you the prior probabilities.

Alternatively said, it makes no sense to do formal hypothesis assessment by finding estimated probabilities when the data themselves have suggested the hypothesis in the first place.

Some cross-validation strategy is necessary; for example, collecting independent data.

Generally, when some process of search or optimization has been used to identify an unusual situation (for example, when a "good" regression equation is found through a step-wise procedure;

when data are "mined" and unusual patterns identified;

when DNA databases are searched for "cold-hits" against evidence left at a crime scene;

when geographic "hot spots" are identified for, say, some particularly unusual cancer;

or when the whole human genome is searched for clues to common diseases),

the same methods for assigning probabilities before the particular situation was identified are generally no longer appropriate after the fact.

A second general area of inappropriate probability assessment concerns the model postulated to aggregate probabilities over several events.

When wrong models are used to generate probabilities, the resulting values may have little to do with empirical reality.

For instance, in throwing dice and counting the sum of spots that result, it is not true that each of the integers from two through twelve is equally likely.

The model of what is equally likely may be reasonable at a different level (for example, pairs of integers appearing on the two dice), but not at all aggregated levels.

Random Match Probability

Flawed calculations of probability can have dire consequences within our legal systems, as the case of Sally Clark and related others make clear.

One broad and current area of possible misunderstanding of probabilities is in the context of DNA evidence (which is exacerbated in the older and more fallible system of identification through fingerprints).

In the use of DNA evidence (and with fingerprints), one must be concerned with the Random Match Probability (RMP):

the likelihood that a randomly selected unrelated person from the population would match a given DNA profile.

Again, the use of independence in RMP estimation is questionable;

also, how does the RMP relate to, and is it relevant for, "cold-hit" searches in DNA databases.

In a confirmatory identification case, a suspect is first identified by non-DNA evidence; DNA evidence is then used to corroborate traditional police investigation.

In a "cold-hit" framework, the suspect is first identified by a search of DNA databases;

the DNA evidence is thus used to identify the suspect as perpetrator, to the exclusion of others, directly from the outset (this is akin to shooting an arrow into a tree and then drawing a target around it).

Here, traditional police work is no longer the focus.

The Probabilistic Generalizations of Logical Fallacies Are No Longer Fallacies

In beginning statistics we commonly introduce some simple logical considerations early on that revolve around the usual "if p, then q" statements, where p and q are two propositions.

As an 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.

Probabilistic Generalizations

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, $P(B|A) > P(B)$).

Both of the latter two probability statements can be algebraically shown true using the simple 2×2 cross-classification frequency table and the equivalences among frequency sums given earlier:

(original statement) $P(A|B) > P(A) \Leftrightarrow N_{AB}/N_B > N_A/N \Leftrightarrow$

(denying the antecedent)
$P(\bar{A}|\bar{B}) > P(\bar{A}) \Leftrightarrow N_{\bar{A}\bar{B}}/N_{\bar{B}} > N_{\bar{A}}/N \Leftrightarrow$

(affirming the consequent)
$P(B|A) > P(B) \Leftrightarrow N_{AB}/N_A > N_B/N \Leftrightarrow$

(contrapositive) $P(\bar{B}|\bar{A}) > P(\bar{B}) \Leftrightarrow N_{\bar{A}\bar{B}}/N_{\bar{A}} > N_{\bar{B}}/N$

Another way of understanding these results is to note that the original statement of $P(A|B) > P(A)$ is equivalent to $N_{AB} > N_A N_B / N$.

Or in the usual terminology of a 2×2 contingency table, the frequency in the cell labeled (A, B) is greater than the typical expected value constructed under independence of the attributes based on the row total, N_B, times the column total, N_A, divided by the grand total, N.

The other probability results follow from the observation that with fixed marginal frequencies, a 2×2 contingency table has only one degree of freedom.

These results derived from the original of B being facilitative for A, $P(A|B) > P(A)$, could have been restated as $\bar{B}$ being inhibitive of A, or as $\bar{A}$ being inhibitive of B.

Abductive (Probabilistic) Reasoning

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," we may be more prone to conjecture the presence of "p."

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 a variant 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.

Using Bayes' Rule to Assess the Consequences of Screening for Rare Events

Bayes' theorem or rule was given earlier in a form appropriate for two events, A and B.

It allows the computation of one conditional probability, $P(A|B)$, from two other conditional probabilities, $P(B|A)$ and $P(B|\bar{A})$, and the prior probability for the event A, $P(A)$.

A general example might help show the importance of Bayes' rule in assessing the value of screening for the occurrence of rare events:

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], or fetal hemoglobin).

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.

Natural Frequencies

Gigerenzer and colleagues 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.

Ethical Issues in Medical Screening

Premarital screening

Prenatal screening

Costs of screening

Informed consent and screening

The (social) pressure to screen

Bayes' Rule and the Confusion of Conditional Probabilities

One way of rewriting Bayes' rule is to use a ratio of probabilities, $P(A)/P(B)$, to relate the two conditional probabilities of interest, $P(B|A)$ (test sensitivity) and $P(A|B)$ (positive predictive value):

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

With this rewriting, it is obvious that $P(A|B)$ and $P(B|A)$ will be equal only when the prior probabilities, $P(A)$ and $P(B)$, are the same.

Yet, this confusion error is so common in the forensic literature that it is given the special name of the "Prosecutor's Fallacy."

In the behavioral sciences, the "Prosecutor's Fallacy" is sometimes referred to as the "Fallacy of the Transposed Conditional" or the "Inversion Fallacy."

In the context of statistical inference, it appears when the probability of seeing a particular data result conditional on the null hypothesis being true, $P(\mathrm{data} \mid H_o)$, is confused with $P(H_o \mid \mathrm{data})$;

that is, the probability that the null hypothesis is true given that a particular data result has occurred.

Sally Clark Revisted

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.

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.

Sally Clark's conviction was overturned in 2003, and she was released from prison.

Sally Clark died of acute alcohol poisoning in her home four years later in 2007, at the age of 42.

Roy Meadow (1933–) is still an active British pediatrician.

He rose to fame for his 1977 academic article in the *Lancet* on Munchausen Syndrome by Proxy (MSbP); he is the person who coined the name.

He has spent his whole career crusading and testifying against parents, especially mothers, who supposedly wilfully harmed or killed their children.

We quote from Lord Howe, the opposition spokesman for health, speaking in the House of Lords on MSbP (February 2003):

. . . [O]ne of the most pernicious and ill-founded theories to have gained currency in childcare and social services over the past 10 to 15 years. The theory states that there are parents who induce or fabricate illnesses in their children in order to gain attention for themselves. The name given to it is Münchausen's syndrome by proxy, or factitious or induced illness—FII, as it is now known. It is a theory without science. There is no body of peer-reviewed research to underpin MSBP or FII. It rests instead on the assertions of its inventor and on a handful of case histories. When challenged to produce his research papers to justify his original findings, the inventor of MSBP stated, if you please, that he had destroyed them.

Dreyfus Affair

A much earlier and historically important *fin de siecle* case, is that of Alfred Dreyfus, the much maligned French Jew, and captain in the military, who was falsely imprisoned for espionage.

In this case, the nefarious statistician was Alphonse Bertillon, who through a very convoluted argument reported a small probability that Dreyfus was "innocent."

This meretricious probability had no justifiable mathematical basis and was generated from culling coincidences involving a document, the handwritten *bordereau* (without signature) announcing the transmission of French military information.

Dreyfus was accused and convicted of penning this document and passing it to the (German) enemy.

The "Prosecutor's Fallacy" was more or less invoked to ensure a conviction based on the fallacious small probability given by Bertillon.

In addition to Émile Zola's well-known article, *J'accuse* . . . !, in the newspaper *L'Aurore* on January 13, 1898, it is interesting to note that turn-of-the-century well-known statisticians and probabilists from the French Academy of Sciences (among them Henri Poincaré) demolished Bertillon's probabilistic arguments, and insisted that any use of such evidence needs to proceed in a fully Bayesian manner, much like our present understanding of evidence in current forensic science and the proper place of probabilistic argumentation.

The O.J. Simpson Murder Trial

Any number of conditional probability confusions can arise in important contexts and possibly when least expected.

A famous instance of such a confusion was in the O.J. Simpson case, where one conditional probability, say, $P(A|B)$, was equated with another, $P(A|B \text{ and } D)$.

We quote the clear explanation of this obfuscation by Krämer and Gigerenzer (2005):

Here is a more recent example from the U.S., where likewise $P(A|B)$ is confused with $P(A|B \text{ and } D)$. This time the confusion is spread by Alan Dershowitz, a renowned Harvard Law professor who advised the O.J. Simpson defense team. The prosecution had argued that Simpson's history of spousal abuse reflected a motive to kill, advancing the premise that "a slap is a prelude to homicide." Dershowitz, however, called this argument "a show of weakness" and said: "We knew that we could prove, if we had to, that an infinitesimal percentage—certainly fewer than 1 of 2,500—of men who slap or beat their domestic partners go on to murder them."

Thus, he argued that the probability of the event K that a husband killed his wife if he battered her was small, $P(K|\text{battered}) = 1/2{,}500$. The relevant probability, however, is not this one, as Dershowitz would have us believe. Instead, the relevant probability is that of a man murdering his partner given that he battered her and that she was murdered, $P(K|\text{battered and murdered})$. This probability is about 8/9. It must of course not be confused with the probability that O.J. Simpson is guilty; a jury must take into account much more evidence than battering. But it shows that battering is a fairly good predictor of guilt for murder, contrary to Dershowitz's assertions.

Defendant's Fallacy

A specious argument on the part of the defense is the "Defendant's Fallacy."

Suppose that for an accused individual who is innocent, there is a one-in-a-million chance of a match (such as for DNA, blood, or fiber).

In an area of, say, 10 million people, the number of matches expected is 10 even if everyone tested is innocent.

The Defendant's Fallacy would be to say that because 10 matches are expected in a city of 10 million, the probability that the accused is innocent is $9/10$.

Because this latter probability is so high, the evidence of a match for the accused cannot be used to indicate a finding of guilt, and therefore, the evidence of a match should be excluded.

O.J. Simpson Again

A version of this fallacy appeared (yet again) in the O.J. Simpson murder trial; we give a short excerpt about the Defendant's Fallacy that is embedded in the Wikipedia article on the Prosecutor's Fallacy:

A version of this fallacy arose in the context of the O.J. Simpson murder trial where the prosecution gave evidence that blood from the crime scene matched Simpson with characteristics shared by 1 in 400 people. The defense retorted that a football stadium could be filled full of people from Los Angeles who also fit the grouping characteristics of the blood sample, and therefore the evidence was useless.

The first part of the defenses' argument that there are several other people that fit the blood grouping's characteristics is true, but what is important is that few of those people were related to the case, and even fewer had any motivation for committing the crime. Therefore, the defenses' claim that the evidence is useless is untrue.

Bayes' Rule and the Importance of Baserates

In the formulation of Bayes' rule, the two prior probabilities, $P(A)$ and $P(B)$, are also known as "baserates";

that is, in the absence of other information, how often do the events A and B occur.

Baserates are obviously important in the conversion of $P(B|A)$ into $P(A|B)$, but as shown by Tversky and Kahneman, and others, baserates are routinely ignored when using various reasoning heuristics.

An example is given below on the importance of baserates in eyewitness identification.

The example will be made-up for clarity, but the principle it illustrates has far-reaching real-world implications.

It will be phrased in the language of "odds," so we first digress slightly to introduce that language.

Bayes' Rule in Terms of Odds

Given an event, A, the odds in favor of the event occurring (in relation to $\bar{A}$, the event not occurring), is the ratio

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

Thus, if $P(A) = 6/7$, the odds in favor of A is $(6/7)/(1/7) = (6/1)$, which is read as 6 to 1, meaning that A occurs 6 times for every single time $\bar{A}$ occurs.

Bayes' rule can be restated in terms of odds:

$$O_{odds}(A|B) = O_{odds}(A) \times \Lambda(A|B) \ ,$$

where $\Lambda(A|B)$ is the likelihood ratio:

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

$O_{odds}(A|B)$ is the posterior odds of A given B:

$$O_{odds}(A|B) = \frac{P(A|B)}{P(\bar{A}|B)} \ ;$$

and $O_{odds}(A)$, the prior odds of A by itself:

$$O_{odds}(A) = \frac{P(A)}{P(\bar{A})} \ .$$

The Taxicab Example

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:

$O_{odds}(A|B) = [(15/90)/(75/90)][(4/5)/(1/5)] \approx 4 \text{ to } 5$.

In other words, the probability that the culprit taxi is blue is $4/9 \approx 44\%$.

We note that this latter value is much different from 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.

Eyewitness Evidence Generally

Another arena in which Bayes' theorem has a role is in assessing and quantifying in a realistic way the probative (that is, legal-proof) value of eyewitness testimony.

The faith the legal system has historically placed in eyewitnesses has been shaken by the advent of forensic DNA testing.

In the majority of the numerous DNA exonerations occurring over the last twenty years, mistaken eyewitness identifications have been involved.

Polygraph Examination

A third area in which Bayesian notions are crucial to an understanding of what is possible, is in polygraph examinations and the quality of information that they can or cannot provide.

Again, what appears to happen is that people want desperately to believe in some rational mechanism for detecting liars and cheats, and thereby increase one's sense of security and control.

So, irrespective of the statistical evidence marshalled, and probably because nothing else is really offered to provide even an illusion of control in identifying prevarication, lie detector tests still get done, and a lot of them.

An illuminating tale is Fienberg and Stern's, "In Search of the Magic Lasso: The Truth About the Polygraph," (2005) and the work of the National Research Council Committee to Review the Scientific Evidence on the Polygraph (2003).

Blood Alcohol Content

We mention one last topic where a knowledge of Bayes' rule might help in arguing within another arena of forensic evidence: the assessment of blood alcohol content (BAC).

The United States Supreme Court heard arguments in January of 2010 (Briscoe v. Virginia, 2010) about crime analysts being required to make court appearances, and to (presumably) testify about the evidence and its reliability that they present now only in written form.

The case was spurred in part by a California woman convicted of vehicular manslaughter with a supposed blood alcohol level two hours after the accident above the legal limit of .08.

The woman denied being drunk but did admit to taking two shots of tequila (with Sprite chasers)

There are several statistically related questions pertaining to the use of a dichotomous standard for BAC (usually, .08) as a definitive indication of impairment and, presumably, of criminal liability when someone is injured in an accident.

Intuitively, it would seem that the same level of BAC might lead to different levels of impairment conditional on individual characteristics.

Also, was this value set based on scientifically credible data?

A variety of different BAC tests could be used (for example, urine, blood, saliva, breath, hair);

thus, there are all the possible interchangeability and differential reliability issues that this multiplicity implies.

A knowledge of Bayes' theorem and the way in which sensitivity, specificity, the positive predictive value, and the prior probability all operate together may at times be helpful to you or to others in mitigating the effects that a single test may have on one's assessment of culpability.

There are many instances where the error rates associated with an instrument are discounted, and it is implicity assumed that an "observed value" is the "true value."

The example of blood alcohol level just discussed seems to be, on the face of it, a particularly egregious example.

But there are other tests that could be usefully approached with an understanding of Bayes' rule, such as drug/steroid/human growth hormone use in athletes, blood doping in bicycle racers, polygraph tests for spying/white collar crime, fingerprint or eyewitness (mis)identification, or laser gun usage for speeding tickets.

We are not saying that a savvy statistician armed with a knowledge of how Bayes' theorem works can "beat the rap," but it couldn't hurt.

Anytime a judgment is based on a single fallible instrument, the value of the positive predictive value assumes a great importance in establishing guilt or innocence

The (Legal) Status of the Use of Baserates

The Gileadites seized the fords of the Jordan before the Ephraimites arrived. And when any Ephraimite who escaped said, "Let me cross over," the men of Gilead would say to him, "Are you an Ephraimite?" If he said, "No," then they would say to him, "Then say, 'Shibboleth'!" And he would say, "Sibboleth," for he could not pronounce it right. Then they would take him and kill him at the fords of the Jordan. There fell at that time forty-two thousand Ephraimites.
— Judges 12:5-6

Shibboleth: This word comes directly from the Old Testament Biblical quotation (Judges 12:5-6) regarding the Gileadites and Ephraimites.

It refers to any distinguishing practice, usually one of language, associated with social or regional origin that identifies its speaker as being a member of a group.

Criminal trials: In the *Federal Rules of Evidence*, Rule 403 implicitly excludes the use of baserates 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: Federal Rules of Evidence

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.

Racial profiling: Although the Arizona governor vehemently denies the label of racial profiling attached to its Senate Bill 1070, her argument comes down to officers knowing an illegal alien when they see one, and this will never depend on racial profiling because that, she says, "is illegal."

How an assessment of "reasonable suspicion" would be made is left to the discretion of the officers—possibly a shibboleth will be used, such as speaking perfect English without an accent.

Or as the then governor of the state adjoining Arizona (Arnold Schwarzenegger) said: "I was also going to go and give a speech in Arizona but with my accent, I was afraid they were going to deport me back to Austria."

The Fourteenth Amendment

Section 1. All persons born or naturalized in the United States, and subject to the jurisdiction thereof, are citizens of the United States and of the State wherein they reside. No State shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any State deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws.

Although the "due process" and "equal protection" clauses seem rather definitive, the United States judicial system has found ways to circumvent their application when it was viewed necessary.

One example discussed later is the Supreme Court decision in *McCleskey v. Kemp* (1987) on racial disparities in the imposition of the death penalty (in Georgia).

But probably the most blatant disregard of "equal protection" was the Japanese-American internment and relocation of about 110,000 individuals living along the United States Pacific coast in the 1940s.

These "War Relocation Camps" were authorized by President Roosevelt on February 19, 1942, with the infamous *Executive Order 9066*.

The Fourth Amendment

The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

Various interpretations of the Fourth Amendment have been made through many Supreme Court opinions.

We mention two here that are directly relevant to the issue of law-enforcement application of baserates, and for (racial) profiling:

Terry v. Ohio (1968) and *Whren v. United States* (1996).

The Wikipedia summaries are given in both cases in the readings.

Government institution protections: Although government institutions should protect rights guaranteed by the Constitution, there have been many historical failures.

Many of these (unethical) intrusions are statistical at their core, where data are collected on individuals who may be under surveillance only for having unpopular views.

To give a particularly salient and egregious example involving the FBI, J. Edgar Hoover, Japanese-American internment, and related topics, we redact (in your readings) the Wikipedia entry on the Custodial Detention Index used by the FBI from the 1930s to the 1970s (with various renamed successor indices, such as Rabble-Rouser, Agitator, Security, Communist, Administrative).

USA PATRIOT Act: The attitude present during World War II that resident Japanese-Americans had a proclivity for espionage has now changed after September 11, 2001, to that of Middle Eastern men having a proclivity for committing terrorist acts.

The acronym of being arrested because of a DWB ("driving while black") has now been altered to FWM ("flying while Muslim").

Section 412 of the *USA PATRIOT Act* allows the United States Attorney General to detain aliens for up to seven days without bringing charges when the detainees are certified as threats to national security.

The grounds for detention are the same "reasonable suspicion" standard of *Terry v. Ohio* (1968).

The Attorney General certification must state that there are "reasonable grounds to believe" the detainee will commit espionage or sabotage, commit terrorist acts, try to overthrow the government, or otherwise behave in a way that would endanger national security.

After seven days, the detention may continue if the alien is charged with a crime or violation of visa conditions. When circumstances prohibit the repatriation of a person for an immigration offense, the detention may continue indefinitely if recertified by the attorney general every six months.

Under the *USA PATRIOT Act*, a person confined for a violation of conditions of United States entry but who cannot be deported to the country of origin, may be indefinitely confined without criminal charges ever being filed.

Eyewitness reliability and false confessions: Several troublesome forensic areas exist in which baserates can come into nefarious play.

One is in eyewitness testimony and how baserates are crucial to assessing the reliability of a witness's identification.

The criminal case reported later of "In Re As.H (2004)" illustrates this point well, particularly as it deals with cross-racial identification, memory lapses, how lineups are done, and so forth.

Also, we have the earlier taxicab anecdote of this section.

One possibly unexpected use that we turn to next involves baserate considerations in "false confessions."

False confessions appear more frequently than we might expect and also in some very high profile cases.

The most sensationally reported example may be the Central Park jogger incident of 1989, in which five African and Hispanic Americans all falsely confessed.

Reid Technique

To put this issue of false confession into a Bayesian framework, our main interest is in the term, $P(\text{guilty} \mid \text{confess})$.

Based on Bayes' rule this probability can be written as

$$\frac{P(\text{confess} \mid \text{guilty})P(\text{guilty})}{P(\text{confess} \mid \text{guilty})P(\text{guilty}) + P(\text{confess}|\text{notguilty})P(\text{notguilty})}$$

The most common interrogation strategy taught to police officers is the 9-step Reid Technique.

The proponents of the Reid Technique hold two beliefs:

that $P(\text{confess} \mid \text{not guilty})$ is zero, and that they never interrogate innocent people, so the prior probability, $P(\text{guilty})$, is 1.0.

Given these assumptions, it follows that if a confession is given, the party must be guilty.

There is no room for error in the Reid system; also, training in the Reid system does not increase accuracy of an initial prior assessment of guilt but it does greatly increase confidence in that estimate.

We thus have a new wording for an old adage: "never in error and never in doubt."

People have a naive faith in the power of their own innocence to set them free.

They maintain a belief in a just world where people get what they deserve and deserve what they get.

People are generally under an illusion of transparency where they overestimate the extent that others can see their true thoughts.

When in doubt, just remember the simple words—"I want a lawyer." (Or, in the idiom of the *Law & Order* series on TV, always remember to "lawyer-up.")

If an interrogation proceeds (against our recommendation), it is a guilt-presumptive process that unfolds (it is assumed from the outset that $P(\text{guilty})$ is 1.0).

Forensic Evidence Generally

Most of us learn about forensic evidence and how it is used in criminal cases through shows such as *Law & Order*.

Rarely, if ever, do we learn about evidence fallibility and whether it can be evaluated through the various concepts introduced to this point, such as baserates, sensitivity, specificity, prosecutor or defendant fallacy, or the positive predictive value.

Contrary to what we may come to believe, evidence based on things such as bite marks, fibers, and voice prints are very dubious.

National Academy Report

Because of the rather dismal state of forensic science in general, Congress in 2005 authorized "the National Academy of Sciences to conduct a study on forensic science, as described in the Senate report."

The Senate Report states in part:

"While a great deal of analysis exists of the requirements in the discipline of DNA, there exists little to no analysis of the remaining needs of the community outside of the area of DNA. Therefore . . . the Committee directs the Attorney General to provide [funds] to the National Academy of Sciences to create an independent Forensic Science Committee. This Committee shall include members of the forensics community representing operational crime laboratories, medical examiners, and coroners; legal experts; and other scientists as determined appropriate."

The results of this National Research Council (NRC) study appeared in book form in 2009 from the National Academies Press: *Strengthening Forensic Science in the United States: A Path Forward*.