1. Probability. Let lowercase letters such as *a*, *b*, etc. denote either statements (which might be true or false) or events (which might occur or not). We use $P(a)$ to denote the probability that *a* will occur (if *a* is an event) or the probability that *a* is true (if *a* is a statement). The probability of any event (or statement) *a* lies between 0 and 1 if *a* is uncertain. $P(a)=1$ if *a* is certainly true, and $P(a)=0$ if it is certainly false. In symbolic notation,

   $$0 \le P(a) \le 1$$

2. If *a* and *b* are independent (that is, the truth or falsity of one does not affect the probability of the other), then

   $$P(a \text{ and } b) = P(a) \cdot P(b) \quad \text{and} \quad P(a \text{ or } b) = P(a) + P(b) - P(a) \cdot P(b)$$

3. If *a* and *b* are mutually exclusive (that is, if one is true then the other must be false), then

   $$P(a \text{ and } b) = 0 \quad \text{and} \quad P(a \text{ or } b) = P(a) + P(b)$$

4. If we use ~*a* to denote the negation of *a* (that is, ~*a* is true when *a* is false, and vice versa), then the probability of ~*a* is

   $$P(\sim a) = 1 - P(a)$$

   (since *a* and ~*a* are mutually exclusive, and therefore $P(a \text{ or } \sim a)=P(a)+P(\sim a)=1$).

5. If *a* and *b* are equivalent (that is, if one is true if and only if the other is true, which we denote *a*<=>*b*), then they have the same probability:

   $$\text{if } a \Leftrightarrow b \text{ then } P(a) = P(b)$$

   The converse is not true; statements with the same probability need not be equivalent.

6. If *a* implies *b* (denoted *a*=>*b*), then the probability of *b* is at least as great as the probability of *a*:

   $$\text{if } a \Rightarrow b \text{ then } P(b) \ge P(a)$$

   (since *b* occurs every time *a* occurs, and *b* might also occur at some other times as well). When *b* occurs if and only if *a* occurs, $P(b)=P(a)$ as in (5) above; otherwise $P(b)>P(a)$. Note that the fact that *a* implies *b* does not exclude the possibility that other things may also imply *b*.

7. Conditional probability. If *a* depends on *b* (that is, the probability of *a* is affected by whether *b* is true or false), we use $P(a \mid b)$ to indicate the probability of *a*, given *b* (that is, the probability of *a* when *b* is true). This is also referred to as the probability of *a* conditioned on *b*, and can be expressed as

   $$P(a \mid b) = \frac{P(a \text{ and } b)}{P(b)}$$

   This can be rewritten in more intuitively obvious form as

   $$P(a \text{ and } b) = P(a \mid b) \cdot P(b)$$

   In other words, the probability that both *a* and *b* are true is simply the probability that *b* is true, times the probability that if *b* is true, *a* is also true. The strongest possible conditioning occurs when *b* implies *a*, whence $P(a \mid b)=1$ and $P(a \text{ and } b)=P(b)$. Conversely, when *a* and *b* are independent, $P(a \mid b)=P(a)$, whence $P(a \text{ and } b)=P(a)P(b)$, as in (2) above.

8. Because $P(a \text{ and } b)$ equals $P(b \text{ and } a)$, then from (7),

   $$P(a \mid b) \cdot P(b) = P(b \mid a) \cdot P(a)$$

   Note that although $P(a \text{ and } b)$ and $P(a \text{ or } b)$ both commute (that is, the order of a and b can be reversed without changing *P*), this is not true for $P(a \mid b)$. That is, $P(a \mid b)$ is not equal to $P(b \mid a)$, except in the special case where $P(a)=P(b)$. Here's a trivial example. If someone is named "Jean", then the probability is high that this person is female (except in France), but if someone is female, the probability is not high that she is named "Jean". In such a simple example, the fallacy is easy to see. But the fact remains that some of the most common, and insidious, errors in probabilistic reasoning consist of mistaking $P(a \mid b)$ for $P(b \mid a)$. A simple example: say a hypothesis *h* strongly suggests that an observable event *e* should occur (i.e. $P(e \mid h)$ is high). It does not follow that observing the event *e* provides strong evidence for *h*, (i.e. $P(h \mid e)$ need not be high) because *e* might be probable whether or not *h* is true.

9. If $a_1, a_2, a_3, \ldots a_n$ are mutually exclusive (that is, no two can be true simultaneously) and jointly exhaustive (that is, at least one must be true), then

   $$P(a_1) + P(a_2) + P(a_3) + \ldots + P(a_n) = 1$$

   (A set of statements that are mutually exclusive and jointly exhaustive is sometimes termed a "partition", meaning something that partitions the realm of the possible into discrete propositions a_1, a_2, etc.)

10. If $a_1, a_2, a_3, \ldots a_n$ are mutually exclusive and jointly exhaustive, then for any proposition *b*,

    $$P(b) = P(b \text{ and } a_1) + P(b \text{ and } a_2) + \ldots + P(b \text{ and } a_n)$$

    since if *b* occurs at all it must occur with one of the *a*'s (since one of the *a*'s must occur). This is sometimes termed the "theorem of total probability".

11. In the same vein, if a_1, a_2, a_3, ...a_n are mutually exclusive and jointly exhaustive, then for any proposition b,

$$P(b) = P(b \mid a_1) \cdot P(a_1) + P(b \mid a_2) \cdot P(a_2) + ... + P(b \mid a_n) \cdot P(a_n)$$

This is the basis for "fault tree analysis" and other forms of risk analysis that quantify the likelihood of an outcome (e.g., a nuclear reactor meltdown) that can occur through several different causal pathways.

12. If an hypothesis h implies some observable evidence e, and if h is possible, and if e is not inevitable, (that is, h=>e, $P(h)$>0, and $P(e)$<1), then $P(h \mid e)$, the probability that the hypothesis is true, given that the evidence is observed, is greater than $P(h)$, the probability that h is true in the absence of any information about whether or not e is observed,

$$P(h \mid e) > P(h)$$

In other words, *if h predicts e, then the occurence of e raises the likelihood that h is true.* Although h and e could be any two statements, and need not be an hypothesis and evidence for it, they are used in this example because it is such an important part of intuitive scientific inference. Where does (12) come from? Directly from (8), as follows: if $P(h \mid e)P(e)=P(e \mid h)P(h)$, and $P(e \mid h)=1$ (because h implies e), then $P(h \mid e)=P(h)/P(e)$, which must be greater than $P(h)$ because $P(e)$<1.

13. If an hypothesis h indicates that some observable evidence e is probable, with a probability of $P(e \mid h)$, then the probability that the hypothesis is true if e is indeed observed, $P(h \mid e)$, depends on $P(e \mid h)$, $P(h)$, and $P(e)$, as follows:

$$P(h \mid e) = \frac{P(e \mid h) \cdot P(h)}{P(e)}$$

where, as in (12), $P(h)$ is the probability that h is true in the absence of any information about whether or not e is observed, and $P(e)$ is the probability that e would be observed whether or not h was true. This is Bayes' Theorem, which forms the basis for Bayesian inference, in conjunction with the two variants given below. It can be derived directly by rearranging the terms in (8).

Bayes' Theorem quantifies how strongly experimental observations confirm or undermine a hypothesis. If h greatly increases the probability of e (that is, $P(e \mid h)$ is large compared to $P(e)$), then observing e greatly increases the likelihood that h is true (that is, $P(h \mid e)$ will be much larger that $P(h)$. Conversely, if h strongly suggests that e should not occur (that is, $P(e \mid h)$ is small), then observing e provides strong disconfirmation for h (that is, $P(h \mid e)$ will be small).

14. If several alternative hypotheses, h_1, h_2, ... h_n are mutually exclusive and jointly exhaustive, then the probability that any one of them, say h_k, is true is:

$$P(h_k \mid e) = \frac{P(e \mid h_k) \cdot P(h_k)}{\sum_{i=1}^{n} P(e \mid h_i) \cdot P(h_i)}$$

This is simply a generalization of (13), derived by substituting (11) for $P(e)$ in (13). The variables are likewise generalized from (13): $P(h_i)$ (or $P(h_k)$)is the likelihood that h_i (or h_k) is true, before one knows whether or not the observable evidence e is true, $P(e \mid h_i)$ (or $P(e \mid h_k)$) is the probability that e will be observed if h_i (or h_k)is true, and $P(h_k \mid e)$ is the likelihood that h_k is true, if e is indeed observed. This is known as the second form of Bayes's Theorem, but was actually proposed by Laplace.

This shows that observing e can provide strong evidence for h_k (that is, $P(h_k \mid e)$ can be large), but only if e is unlikely under any of the other hypotheses (that is, $P(e \mid h_i)$ is small for $h_i \neq h_k$).

15. Where just one hypothesis is being tested, (14) can be reduced to only two alternatives: the hypothesis h, and all other possibilities (which imply that h is false, indicated by ~h). Then (14) becomes (after several terms are rearranged) the third form of Bayes' Theorem:

$$P(h \mid e) = \frac{P(h)}{P(h) + \frac{P(e \mid \sim h)}{P(e \mid h)} P(\sim h)} = \frac{P(h)}{P(h) + \frac{P(e \mid \sim h)}{P(e \mid h)} [1 - P(h)]}$$

where $P(h)$ is the "prior" likelihood that h is true (that is, the likelihood in the absence of any evidence whether e is true or not), $P(h \mid e)$ is the "posterior" likelihood that h is true (that is, after e has indeed been observed), and $P(e \mid h)$ and $P(e \mid \sim h)$ are the probabilities that e would be observed if the hypothesis were true and if it were false.

The expression above shows that the posterior likelihood, $P(h \mid e)$, depends on only two factors: the prior assessment of likelihood, $P(h)$, and the *likelihood ratio* $P(e \mid \sim h) / P(e \mid h)$. Note that if e is likely if h is true, but very unlikely if h is false, the posterior likelihood $P(h \mid e)$ can approach 1 after e is observed, regardless of what the prior likelihood $P(h)$ is. That means that the evidence e could even convince a skeptic (someone who believed $P(h)$ was low) that h is indeed true, but it can only do so if e would be sufficiently unlikely if h were false.

Conversely, if e is likely when h is false, but very unlikely when h is true, the posterior likelihood $P(h \mid e)$ can approach 0 after e is observed, regardless of whether $P(h)$ was assumed to be high or low. That means that the evidence e could even convince a fanatical believer (someone who believed $P(h)$ was high) that h is indeed false, but it can only do so if e would be very unlikely if h were true.

Adapted in part from Chapter 2 of C. Howson and P. Urbach, *Scientific Reasoning: The Bayesian Approach*, Open Court Publishing, La Salle, Illinois, 1993.

Bayesian inference I:
John Doe and the AIDS test

Assume: the HIV test has a false positive rate of 5%
and a false negative rate of 0%

Data point: John Doe tests positive

Question: what is the probability that John Doe has HIV?

Wrong answer: 95%

Right answer via classical statistical inference:
"***Don't ask that question!***" (***Facts*** are not probabilistic!)

Right answer from Bayesian statistics:
It depends on the rate of HIV in the group John Doe came from!

Illustration:

Assume HIV rate in this group is 1%.
Test 10000 individuals.

Results:

100 true positives (1% of 10000)
495 false positives (5% of 9900)

Chance John Doe has HIV is:

= (100 true positives)/(100 true + 495 false positives)
= 100/595 ≈ 17% (***not*** 95%!)

Illustration 2:

What if the HIV rate in John Doe's group were 0.1%?
Test 10000 individuals.

Results:

10 true positives (0.1% of 10000)
499.5 false positives (5% of 9990)

Chance John Doe has HIV is:

= (10 true positives)/(10 true + 499.5 false positives)
= 10/509.5 ≈ 2% (***not*** 95%!)

Important lessons:

Test results alone are ***meaningless***.
Everything depends on likelihood that positive results are true positives rather than false positives.
This likelihood is ***unknowable*** without knowing the background risk of HIV in the population!

Illustration 3:

Repeat the HIV test on the 509.5 positives
(assume these results are *independent* of previous ones)

Results:

10 true positives (all of them)
24.975 false positives (5% of 499.5)

Chance John Doe has HIV is:

= (10 true positives)/(10 true + 24.975 false positives)
= 10/35 ≈ 30% (***not*** 95%!)

Illustration 4:

Repeat the HIV test *again* on the 35 positives
(assume these results are *independent* of previous ones)

Results:
10 true positives (all of them)
1.25 false positives (5% of 25)

Chance John Doe has HIV is:
= (10 true positives)/(10 true + 1.25 false positives)
= 10/11.25 ≈ 90% (***not*** 95%!)

Illustration 5:

Repeat the HIV test *yet again* on the 11.25 positives
(assume these results are *independent* of previous ones)

Results:
10 true positives (all of them)
0.0625 false positives (5% of 1.25)

Chance John Doe has HIV is:
= (10 true positives)/(10 true + 0.0625 false positives)
= 10/10.0625 ≈ 99.4% (***not*** 95%!)

Important lesson:

By generating enough information from testing, we can reach a clear result (almost) regardless of the background risk in the population.

Bayesian inference II:
The Reverend Thomas Bayes, FRS, and "Bayes' Theorem"

... which isn't a theorem, and wasn't derived by Bayes ...

T. Bayes, An Essay Towards Solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society of London*, 53, 370-418, 1763.
(Poses the problem but solves it only for one special case).

Pierre Simon, Marquis De Laplace, *Theorie Analytique des Probabilites*, 1812.
(States a general solution but doesn't prove it).

Problem: Given some set of observations, what is the probability that a given model or hypothesis is correct?

Define:

H	a hypothesis
d	some observable data
$P(d\|H)$	probability of observing d if H is true
$P(d\|\sim H)$	probability of observing d if H is false
$P(H\|d)$	confidence that H is true if d is observed

note! $\boldsymbol{P(d|H) \neq P(H|d)}$ **!**

Question:

What is the probability that H is true, and d is observed?

note $P(H \text{ and } d) = P(d \mid H)\, P(H)$

and $P(H \text{ and } d) = P(H \mid d)\, P(d)$

so $P(d \mid H)\, P(H) = P(H \mid d)\, P(d)$

SO!

$$P(H \mid d) = \frac{P(d \mid H)\, P(H)}{P(d)}$$

$$= \frac{P(d \mid H)\, P(H)}{P(d \mid H)\, P(H) + P(d \mid \sim H)\, P(\sim H)}$$

$$= \frac{\text{chance of } d \text{ occurring because } H \text{ is true}}{\text{chance of } d \text{ occurring whether or not } H \text{ is true}}$$

This is "Bayes' Theorem", or more humbly an "updating rule" which uses data d to update our prior confidence $P(H)$, yielding our posterior confidence $P(d \mid H)$.

For multiple hypotheses H_1, H_2, H_3...:

$$P(H_i \mid d) = \frac{P(d \mid H_i)\, P(H_i)}{P(d)}$$

$$= \frac{P(d \mid H_i)\, P(H_i)}{\sum_k P(d \mid H_k)\, P(H_k)}$$

$$= \frac{\text{chance of } d \text{ occurring because } H_i \text{ is true}}{\text{chance of } d \text{ occurring for any reason}}$$

I am
My Lord
Your Lordship's
most obedient
humble Servant
T. Bayes.

Bayesian inference III: How science really works

Two possibilities:

H some theory is true

$\sim H$ that theory is false

Two outcomes of a test of a theory:

d data are consistent with the theory

$\sim d$ data are inconsistent with the theory

Classical statistics can tell us:

$P(\sim d \mid \sim H) = 1-\beta$ = power
(likelihood of rejecting a theory that is false)

$P(d \mid \sim H) = \beta$ = false negative rate (Type II error rate)
(risk of accepting a theory that is actually false)

$P(\sim d \mid H) = \alpha$ = false positive rate (Type I error rate)
(risk of rejecting a theory that's actually true)

$P(d \mid H) = 1-\alpha$ = statistical significance
(likelihood that theory will be accepted if it's true)

(The semantics above are conventional if H is the *null* hypothesis)

But what we really want to know is:

$P(H \mid d)$ **(*If the data support the theory, what are the chances that the theory is actually true?*)**

$$P(H \mid d) = \frac{P(d \mid H)\, P(H)}{P(d)}$$

$$= \frac{P(d \mid H)\, P(H)}{P(d \mid H)\, P(H) + P(d \mid \sim H)\, P(\sim H)}$$

$$= \frac{\text{chance of supportive evidence because theory is true}}{\text{chance of supportive evidence whether or not theory is true}}$$

So post-test confidence in the theory can be high, but only if there was a low risk that the theory could spuriously pass the test

In the language of power and significance:

$$P(H \mid d) = \frac{(1-\alpha)\, P(H)}{(1-\alpha)\, P(H) + \beta\,(1-P(H))}$$

$$= \frac{1-\alpha}{1-\alpha-\beta+\frac{\beta}{P(H)}}$$

$$= \frac{1}{1+\frac{\beta}{1-\alpha}\left[\frac{1}{P(H)}+1\right]}$$

So if power is high enough (β is low enough), the test should convince a non-dogmatic skeptic that the theory is valid
(non-dogmatic means $P(H)>0$)

Bayesian inference IV: Hotspot detection

Two possibilities:

H a hotspot exists

$\sim H$ hotspot does not exist

Two outcomes of an attempt to detect the hotspot:

d hotspot detected

$\sim d$ hotspot not detected

Classical statistics can tell us:

$P(d \mid H) = 1-\beta$ = power
(probability that hotspot will be detected if it's there)

$P(\sim d \mid H) = \beta$ = false negative rate (Type II error rate)
(probability of missing hotspot even though it's there)

$P(d \mid \sim H) = \alpha$ = false positive rate (Type I error rate) (=0?)
(probability of falsely detecting a hotspot that's not there)

$P(\sim d \mid \sim H) = 1-\alpha$ = statistical significance (=1)
(likelihood that if there's no hotspot, none will be detected)

But what we really want to know is:

$P(H \mid \sim d)$
(*i.e., the risk that there's a hotspot there, even though we've failed to detect it*)

$$P(H \mid \sim d) = \frac{P(\sim d \mid H)\, P(H)}{P(\sim d)}$$
$$= \frac{P(\sim d \mid H)\, P(H)}{P(\sim d \mid H)\, P(H) + P(\sim d \mid \sim H)\, P(\sim H)}$$
$$= \frac{\text{chance of nondetection even though hotspot is there}}{\text{chance of nondetection whether or not hotspot is there}}$$

Assuming $P(\sim d \mid \sim H) = 1$ (we won't falsely detect a nonexistent hotspot),

$$P(H \mid \sim d) = \frac{P(\sim d \mid H)\, P(H)}{P(\sim d \mid H)\, P(H) + P(\sim H)}$$
$$= \frac{\beta\, P(H)}{\beta\, P(H) + 1 - P(H)}$$

If β is small (high power to detect hotspots if they exist)

$$P(H \mid \sim d) \approx \beta \frac{P(H)}{1 - P(H)}$$

i.e., if hotspot detection is reliable enough, failure to detect a hotspot should be convincing to a non-dogmatic skeptic

that is, if β is small enough, post-test confidence that a hotspot exists -- $P(H \mid \sim d)$ -- can be small
... even though pre-test confidence -- $P(H)$ -- was high
... as long as $P(H)<1$ (non-dogmatic prior)