

A condensed summary of probability*

Greg Kochanski

February 28, 2006

1 Probabilities are Logic

Formally, a probability is a bundle of four things: a universe, a number, a constraint, along with some operations on these things.

- The universe is the set of all possible results. Sometimes, the universe can be big (like all of English), or even infinite. Sometimes the universe can be small, like when you are flipping a coin: then, it just contains “heads” and “tails”.
- The number is an extension of a Boolean truth value. Zero corresponds to false, one corresponds to true, and everything in between corresponds to degrees of uncertainty.
- The constraint matches the logical law of the excluded middle: the probabilities of all the alternatives sum to one.
- The arithmetical operations correspond to the standard logical operations of Boolean logic.

If you use the right arithmetic and meet the other three requirements, you have a probability, and you can reproduce all results of Boolean logic by using probabilities of zero or one. Consequently, *probabilities form a system of logic that is a generalisation of Boolean logic.*

2 What to Ignore

- There are two alternative interpretations of probabilities, which are called “Subjective” and “Frequentist.” Ignore the difference: they are not mutually exclusive and one gets the same answers either way.
- People can be divided into “Bayesian” or “non-Bayesian”, which refers to their attitudes toward a subjective interpretation of probability and their willingness to use Bayes’ Theorem. Don’t bother dividing: it’s a psychological divide, not a mathematical one.

*This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/2.5/> or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA. This work is available under <http://kochanski.org/gpk/teaching/04010xford>.

- Some people make a big deal about the distinction between “probabilities” and “conditional probabilities.” Ignore that, too. You can think of probabilities are conditional probabilities; but sometimes we’re too lazy to say what the conditions are; especially when the conditions are unimportant. That’s fine, so long as there is no confusion.

3 What is a Probability?

Any of the items below are perfectly reasonable ways to interpret a probability.

- It is a truth value that allows you to be uncertain, if you wish.
- It describes how much you know about a situation.
- It counts how big a specified subset of possible results is, relative to the size of the universe¹.
- It is an estimate of how often something happens under specified conditions.

4 What isn’t a probability?

Probabilities are not concrete things you can touch. They are almost Platonic ideals, properties of a model that you can never measure directly or exactly, but can only estimate from the data.

For instance, there is a very close relationship between probabilities and the frequency at which events happen, but they are not quite the same thing. They become numerically identical only if you count all events in your universe; until then, you always have a certain amount of sampling error, and your observed frequency will be above or below the probability, depending on whether you were lucky or not when you chose your sample.

4.1 The Probability of What?

We will talk about the probability that some event named E will occur under some conditions (called C). This is written as $P(E|C)$. The conditions can also include a statement of our knowledge², K : this can be written as $P(E|C, K)$. One reads it as “P of E, given C and K.” The comma between conditions is really an **and** operation. Thus, $P(X|C, K)$ is the probability of observing X when you know both C and K to be true. Sometimes if C and K are obvious or irrelevant, you can write just $P(E)$.

Note that E , C , and K are all just statements about the universe. Conditions can be treated as events, and one can calculate $P(C)$ or $P(K)$ just as sensibly as one can calculate $P(E)$. Likewise, conditions can be thought of as selecting the universe that you care about.

¹ The size of the universe matters. For instance, I have just flipped a coin 23 times, and found empirically that the the universe of possible results is (**heads, tails, rolls under my desk**). In this universe, if $P(\text{heads}) = 50\%$, then $P(\text{tails})$ must be less than 50%, because the coin will occasionally roll away and get lost. If the universe were only (**heads, tails**), then $P(\text{tails})$ would be exactly 50%.

² Our knowledge can matter for subjective probabilities. It is quite reasonable to say “I have a 50% chance of getting the job.” even if the decision may already have been made. The people making the decision may know, but you don’t, not yet. So, $P(\text{get job}|\text{Their Knowledge}) = 1$ (or zero), but $P(\text{get job}|\text{Your Knowledge}) = 0.5$.

Note also that it is quite valid to read $P(E|C)$ as the probability that statement E is true, given that statement C is true. One does not need to imagine a discrete event in order to have a probability.

Finally, one can read $P(E|C)$ as comparing³ the size of set E to the size of set C . For instance, if E is the set of all people who can speak Swedish and C is the set of all people, then $P(E|C)$ is just the fraction of all people that can speak Swedish, which is the same as the probability that a randomly selected person speaks Swedish.

³ Mathematicians speak of a “probability measure” and treat probabilities as measurements of the size of sets or events.

4.2 Examples

An Event Name	Another Reasonable Name	Description
E	next="e"	The next letter is "e".
Z	final=/z/	The final sound in the word is /z/.
F	Has-fricative	The word is pronounced with a fricative.
D	$\tau < 0.180$ s	The duration of the final syllable is less than 180 ms.
I	failed	My Internet connection will transfer less than 1 MB in the next 30 minutes.
S	is_speaker	The next person that George will pass on the street is before a group of more than 10 people.
G	age=40	George is 40 years old.
T	isverb(W_{12})	The twelfth lexical item of <i>Jabberwocky</i> acts as a verb.
W	isverb(W_{200})	The two-hundredth lexical item of <i>War and Peace</i> is a verb.
C	X had Fifi	Person X had a dog named "Fifi."
P	next=MD or last=electrician	The next person I meet will be a doctor, or the next person I meet was an electrician.
L	HTH	The next time I spin a 20p coin, it will come up heads. The following spin will be tails and the following spin will be heads again.
Ω		A special event that always happens.
\emptyset		A special event that never happens.

An event could be nearly anything:

(The event names are arbitrary. They are just provided for convenience and to show reasonable naming practise.)

Likewise, the conditions can be nearly anything imaginable:

Condition Name	Another Condition Name	Description
t	red table	I am sitting at a table with a red tablecloth.
m	male	X is male.
r	railings	The word is “railings.”
n	pay NTL	I pay NTL for broadband service.
o	Oxford	George is in Oxford.
j	–Remember	You cannot remember the text of Jabberwocky.
p	Author=Plato	War and Peace was written by Plato.
e	error	A disfluency will occur on the next word.

Now, you can write the probability of any event under any conditions, such as:

Symbol	Description
$P(F r)$	The probability that “railings” contains a fricative.
$P(F r, t)$	The probability that “railings” contains a fricative, given that I am sitting at a table with a red tablecloth. (Normally, one would expect that $P(F r, t) = P(F, r)$.)
$P(Z r)$	The probability that the word “railings” ends with a /z/ sound. (Note that if we are talking about an experiment, $P(Z r) < 1$, because some people will occasionally end it with a /zs/ sound.)
$P(C m)$	The probability that a male person had a dog named Fifi.
$P(X \text{ had Fifi} \mid X \text{ is male})$	Same as above: The probability that X had a dog named Fifi, given that X is male. This number is very close to the fraction of all male humans who had a dog named Fifi.
$P(C)$	If this appears alongside $P(C m)$, it probably means “The probability that a person had a dog named Fifi, assuming that I do not know the sex of that person.”
$P(W p)$	The probability that the 200 th lexical item if “War and Peace” acts as a verb, given that “War and Peace” was written by Plato. ⁴

⁴ $P(W|p)$ is undefined, since p is false. In other words, it can be assigned any value between zero and one, and it will never affect a calculation of the probability of any possible event.

Of course, in reality, there are many conditions for each probability, most of which are irrelevant:

$P(\text{heads}|\text{coin} = 20p,$
coin is spun, temperature = 20C,
time = 20:15:32, earthquake = no,
tablecloth colour = red,
Greg has never read “Jabberwocky” in Russian, ...).

In the above probability, only a few of the conditions are important at all: coin = 20p, coin is spun, and earthquake = no would affect the probability somewhat if they were changed to coin = \$0.01, coin is flipped⁵, and earthquake = yes. The rest are irrelevant, and one normally doesn't bother writing them.

5 Equations that connect Probabilities to Logic

Because probabilities can reproduce logic, we can pair up any logical operation with the equivalent arithmetic operation on probabilities.

We'll write $P(A|X)$ for the probability of event A given condition X . In these equations, X is any arbitrary condition. We'll write \cup for logical “inclusive or” or a set union operation. We'll write \cap for logical “and” or a set intersection operation. We'll write \neg for “not”, and \implies for “implies”.

⁵ Flipping coins gives you very near a 50-50 chance of heads or tails, while spinning a coin gives you a different probability of heads for each coin. US pennies, when spun, give heads about 1/3 of the time. British 20p coins give heads about 60% of the time. British pennies are pretty close to 50%.

Operation	Probability of result	Shorthand	Formula for probability of result.
True	$P(\text{True} X)$	$P(\Omega X)$	1.
False	$P(\text{False} X)$	$P(\emptyset X)$	0.
Not	$P(\text{not } A X)$	$P(\neg A X)$	$1 - P(A X)$.
And	$P(A \text{ and } B X)$	$P(A \cap B X)$	$P(A B, X) \cdot P(B X)$.
Or	$P(A \text{ or } B X)$	$P(A \cup B X)$	$P(A X) + P(B X) - P(A \cap B X)$.
Tautology	$P(X \Omega)$		1.
Tautology	$P(X X)$		1.
Tautology	$P(X X, Y, Z, \dots)$		1.
Tautology	$P(\text{not } X X)$	$P(\neg X X)$	0.
Tautology	$P(\text{not } X X, Y, Z, \dots)$	$P(\neg X X, Y, Z, \dots)$	0.
Implication	$P(A B, X)$ if B implies A	$P(A B, X)$ if $B \implies A$	1.
And	$P(A \text{ and } B X)$	$P(A \cap B X)$	$P(A X)$ if $P(B X) = 1$.
And	$P(A \text{ and } X X)$	$P(A \cap X X)$	$P(B X)$.
And	$P(A \text{ and } B X)$	$P(A \cap B X)$	0 if $P(A) = 0$ or $P(B) = 0$.
And	$P(A \text{ and } A X)$	$P(A \cap A X)$	$P(A X)$.
Exclusive Or	$P(A \text{ xor } B X)$		$P(A X) + P(B X) - 2 \cdot P(A \cap B X)$.

Now, let's explain some of those lines one-by-one.

True: By definition, the probability that something true is true is one.

False: By definition, the probability of something false being true is zero. The null event which never happens is written as \emptyset .

$P(\neg A|X)$: (See Figure 1.) By the definition of a probability, the probabilities of all the alternatives add up to 1. Now, since we have only two alternatives (either A is true or $\neg A$ is true) then we have

$$P(A|X) + P(\neg A|X) = 1. \quad (1)$$

A bit of algebra then tells us that

$$P(\neg A|X) = 1 - P(A|X). \quad (2)$$

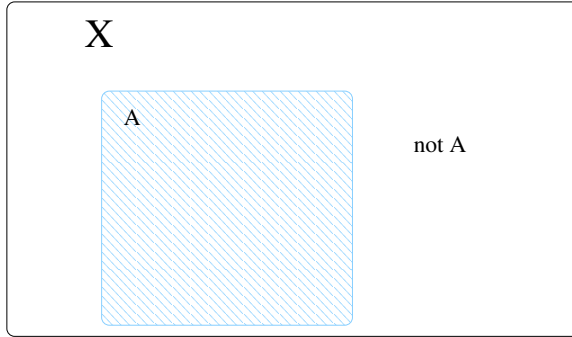


Figure 1: Logical **not** operation.

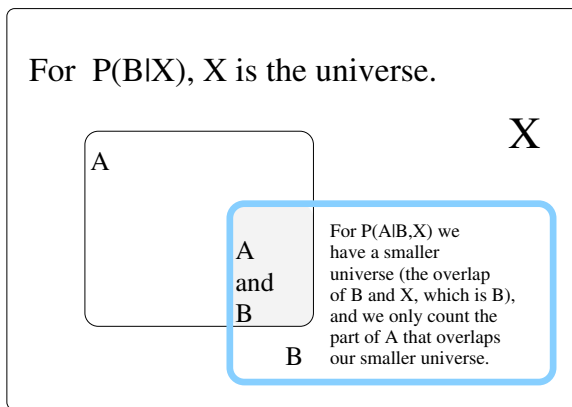


Figure 2: Logical **and** operation.

$P(A \cap B|X) = P(A|B, X) \cdot P(B|X)$: (See Figure 2.) This one is the most general form of Boolean **and**. It is also the only way to change a condition into an event or *vice versa*.

One can construct a plausibility argument for it by looking at the cases where everything is certainly true or certainly false so that both $P(B|X)$ and $P(A|B, X)$ are either zero or one. Then, the result of the multiplication is zero unless both $P(B|X)$ **and** $P(A|B, X)$ are one: *i.e.* both **B and A** are true.

Note that you can swap A for B , and the equation is still valid.

Note also that you cannot calculate $P(A \cap B|X)$ from $P(A|X)$ and $P(B|X)$. You need $P(A|B, X)$, instead. For example, $P(A|B, X)$ tells you how often you have wine (A) with lunch (B) in Oxford (X), while $P(A|X)$ answers a more general question: “How often do you have wine in Oxford?” without specifying the time. If you have wine only with dinner, those two probabilities can be entirely different.

$P(A \cup B|X) = P(A|X) + P(B|X) - P(A \cap B|X)$: (See Figure 5.) This equation for **inclusive or**

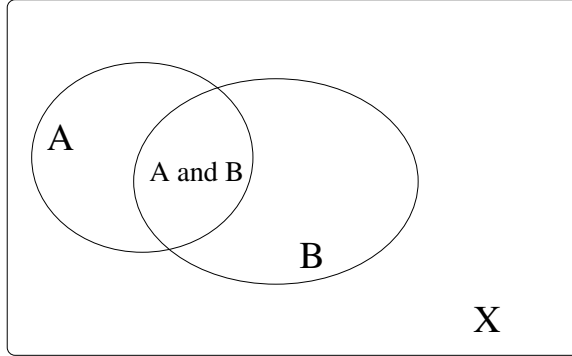


Figure 3: Logical **or** operation. If you add up the areas (i.e. probabilities) for A and B , you find that the overlap area is counted twice. The equation for logical **or** then subtracts off this overlap region to compensate for the double counting.

can easily be tested for the Boolean cases. For instance, $P(A|X) = 0$ and $P(B|X) = 0$, then A and B are both false, and so is $A \cap B$, so the probabilities on the right hand side are all zero.

To take another example, if A is false, but B is true, then we have $P(A|X) = 0$, $P(B|X) = 1$, and $P(A \cap B|X) = 0$, so we get $P(A \cup B) = 1$.

Finally, if A and B are both true, we have $P(A|X) = 1$ and $P(B|X) = 1$, so that $P(A \cap B|X) = 1$, and the result is $P(A \cup B|X) = 1$, as it should be.

$P(X|X, Y, Z, \dots)$: This is the probability that X is true, given that X and Y and Z are true. This reduces to $P(X|X)$, which is one.

$P(\neg X|X)$: This is the probability that X is false, given that X is true. Obviously X will never be true and false simultaneously, so this probability is zero.

$P(A|B, X)$ **if** $B \implies A$: We are given that B is true, and since $B \implies A$, then A must be true. The probability is therefore one.

$P(A \cap B|X)$ **if** $P(B|X) = 1$: Well, if $P(B|X) = 1$, then B is true. If B is true, then $A \cap B = A$. The result is then just $P(A|X)$. In English, one can say that if B is always true, then A **and** B is the same as A . This also works if you interchange A and B , of course.

$P(A \cap B|X)$ **if** $P(A) = 0$: Since $P(A) = 0$, A is false. Therefore, A **and** B is false, so the answer is zero. The same logic holds if $P(B) = 0$.

5.1 Changing Universes

When you change conditions, you are changing the set of possible outcomes. People talk about “changing universes” or “choosing a different universe.”⁶

⁶ Unfortunately, one can really only shrink the universe of probabilities. If it were possible to change it, rather than taking a subset of our normal universe, doubtless conditions like Z (Earth takes 25 hours to rotate so you can

For example, start with a big universe of coin flip possibilities: $\Omega = (\mathbf{heads, tails, rolls\ under\ the\ desk, coin\ stolen\ in\ mid-air, coin\ gets\ stuck\ in\ ceiling\ tiles})$.

We might have $P(\mathbf{heads}) = 0.48$, $P(\mathbf{tails}) = 0.48$, $P(\mathbf{desk}) = 0.03$, $P(\mathbf{stolen}) = 0.005$, and $P(\mathbf{ceiling}) = 0.005$.

Then, let's shrink the universe. Let's say you only count those flips where you catch the coin. Then, we're left with only two possible results in our reduced universe: $(\mathbf{heads, tails})$, and we write their probabilities like this: $P(\mathbf{heads|caught}) = 0.50$, $P(\mathbf{tails|caught}) = 0.50$.⁷ The “|caught” part just means that we are talking about a reduced-size universe.

The equation for logical **and** that we've seen above is the mathematical rule for changing universes. Start from the general equation $P(A \cap B|X) = P(A|B, X) \cdot P(B|X)$ and ignore X for simplicity. We get this:

$$P(A \cap B) = P(A|B) \cdot P(B). \tag{3}$$

A bit of algebra re-arranges it into

$$P(A|B) = P(X \cap B)/P(B). \tag{4}$$

In English, this says that you can compute the probability of A in a smaller universe (the universe where all events have property B) by taking all the events which have both properties A and B , and dividing by the size of the smaller universe⁸. In the example at the top of the section, it is this division by $P(B) = 0.96$ that converts $P(A) = 0.48$ in the large universe into $P(A|B) = 0.50$ in the smaller universe where B is true.

6 Dependence and Independence

Two events are dependent, or correlated, if knowing one helps you to predict the other. For instance, events Z and F defined above (§4.2) are dependent: if the final sound of a word is /z/, then that word certainly contains a fricative. Thus, knowledge of event Z helps you predict F . One can express this probabilistically as $P(F|Z) \neq P(F|\neg Z)$. Specifically, $P(F|Z) = 1$ and $P(F|\neg Z) < 1$.

The opposite of dependence is independence. Two events are *independent* if knowledge of one does not help you to predict the other. This is an important concept, because 1) a lot of events are independent of each other, and 2) if events are independent, calculations get much easier.

Proving that two events are independent of each other takes statistical tests, but often we can assume that events are independent if there is no shared underlying cause. For instance, event L (HTH, above), can be assumed to be independent of event C (Fifi, above), because there is no plausible mechanism by which one's history of dog ownership can affect a coin flip. Thus, we expect that $P(L|C) = P(L|\neg C)$.

If two events are independent, then we have a new equation we can use:

$$P(A \cap B|X) = P(A|X) \cdot P(B|X).$$

This simplifies the equations for both **and** and **or**. The equation for **and** no longer needs the $P(A|B, X)$ number which can be hard to deduce, and the equation for **or** simplifies to

$$P(A \cup B|X) = P(A|X) + P(B|X) - P(A|X) \cdot P(B|X).$$

sleep late every day) would be quite popular.

⁷ Note that the probability for $P(\mathbf{tails|caught})$ has gone up from 0.48 to 0.50? Can you explain why?

⁸ This is the rule that most texts call the definition of a conditional probability.

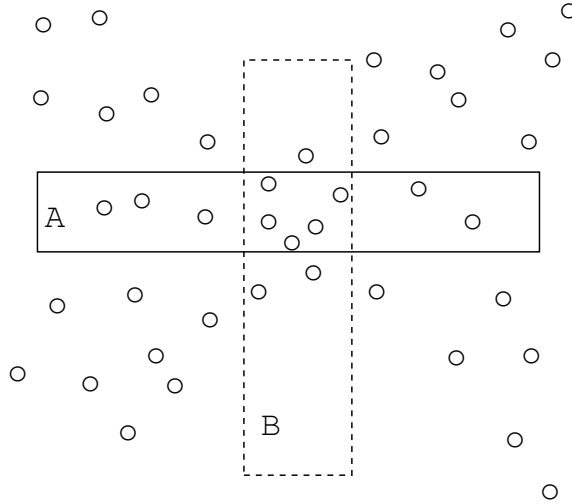


Figure 4: Dependent Probabilities. If you stand in a circle on B , it is likely that you are also on A : knowledge of one event helps you to predict the likely outcome of the other event. (Section 6.)

In practise, it is useful to say that things are almost independent, like the weather in Barcelona and San Francisco. While there is a causal connection between the weather in the two cities (air from one city will eventually blow over the other city), the weather has been modified so much in the interim that knowing one gives very little advantage in predicting the other.

Figure 6 shows an example of two dependent probabilities, A and B . These probabilities correspond to the events “a dot lands on stripe A,” and “a dot lands on stripe B,” respectively. We can see that A and B are dependent because the probability distribution in the vertical direction is different if you look on B vs. if you look off B . Specifically, $P(A|B) > P(A|\neg B)$: if you are looking at the vertical stripe (B), the probability distribution is narrow, and many of the points are on stripe A . Thus, $P(A|B)$ is large.

Conversely, if you are looking off stripe B , the probability distribution is wider, and points are much less likely to land on the A stripe. Thus, $P(A|\neg B)$ is small.

7 Elementary Events

One very useful way to think of probabilities is to break them up into *elementary events*. Elementary events are small, non-overlapping events from which you can conveniently assemble the events you care about. We’ll call them E_i for i from 1 to however many you need.

Because the elementary events do not overlap, $E_i \cap E_j$ is false, so long as $i \neq j$. That means you can add probabilities very simply: $P(E_i \text{ or } E_j) = P(E_i) + P(E_j)$, since the $P(E_i \cap E_j)$ term is zero.

You should define the elementary events so that the events you care about are just the union of a few E_j . The boundaries of the elementary events can be found by laying out the boundaries of all the events you care about, and using that collection of boundaries to chop the universe up into

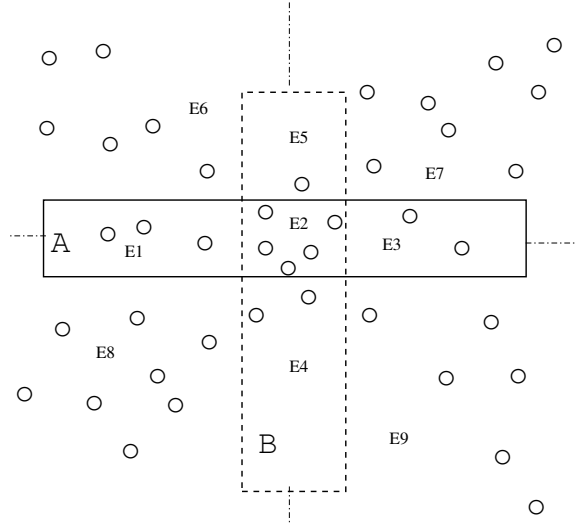


Figure 5: Elementary Events E1–E9. Elementary events are non-overlapping regions; thus no two elementary events are ever true at the same time. You can compute any probability by simply adding up the right set of elementary events.

elementary events. In Figure 7, for instance, $A = E_1 \cup E_2 \cup E_3$, $\neg A = E_6 \cup E_5 \cup E_7 \cup E_8 \cup E_4 \cup E_9$, $B = E_4 \cup E_2 \cup E_5$, and $\neg B = E_6 \cup E_1 \cup E_8 \cup E_9 \cup E_3 \cup E_7$. Consequently, $P(A) = P(E_1) + P(E_2) + P(E_3)$ and so forth.

Using elementary events also makes the connection to Boolean algebra more obvious. For instance, you can get

- $P(A)$ by adding up the probabilities for elementary events that are inside A .
- $P(\neg A)$ by adding up the probabilities for elementary events that are **not** inside A .
- $P(A \cup B)$ by adding up the probabilities for elementary events that are in either A or B .
- $P(A \cap B)$ by adding up the probabilities for elementary events that are in both A and B .

If you are dealing with coin flips, the elementary events are combinations of flips: HHH , HTH , THH , \dots If, for example, you want to know the probability of getting two heads, you add up the probability of the relevant elementary events: $P(\text{two heads}) = P(HHT) + P(HTH) + P(THH)$. These events cover all the possible ways of getting two heads out of three flips.

Elementary events are often the least confusing way to look at probabilities.

8 Recommended Reading

See “The Cartoon Guide to Statistics” Gonick and Smith [1994] for a good informal account, and “A First Course in Probability” Ross [1984] for a more complete, formal account. “Bayesean Data

Analysis” Gelman et al. [1995] is a rather dense book, going into details of many specific analysis procedures. A better book, if you want an understanding of the probabilities and statistics more than specific procedures is “Statistical Theory” Lindgren [1976], which has good sections on the background of probabilities [Lindgren, 1976, 1–36], [Lindgren, 1976, 36–49] (see also [Gelman et al., 1995, 12–18], [Gelman et al., 1995, 18–25], [Gelman et al., 1995, 25–27]).

Thanks to Peet Morris for suggestions.

References

- Andrew B. Gelman, John S. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, first edition, 1995. ISBN 0-412-03991-5.
- Larry Gonick and Woolcott Smith. *The Cartoon Guide to Statistics*. HarperCollins, New York, 1994. ISBN 0062731025.
- Bernard W. Lindgren. *Statistical Theory*. MacMillan, third edition, 1976. ISBN 0-02-370830-1.
- Sheldon Ross. *A First Course in Probability*. Macmillan, New York, London, 1984. ISBN 0-02-403910-1.