

Confidence Intervals and Hypothesis Testing.*

Greg Kochanski
<http://kochanski.org/gpk>

2005/02/28 23:22:25 UTC

1 What is a Hypothesis Test

If you have just two discrete hypotheses, then a hypothesis test is simply an application of Bayes' Theorem. You check to see whether your data can reasonably be explained by that boring old null hypothesis (conventionally called H_0), or not. To test H_0 , you need to compare it to one other model¹. One never tests a hypothesis in isolation; it is tested against an alternative (typically called H_1).

But, hypothesis testing is a biased, asymmetric operation, not just a comparison of which of the alternatives is more probable. Hypothesis testing is appropriate when one of the alternatives is special: simpler, more elegant, predicted by an eminent theorist or critical to some application or conclusion. You need some prior reason to choose the null hypothesis, because it will survive the test even if it's not the one most preferred by the data. In fact, in a hypothesis test, you will only reject H_0 if the alternative is about 100(!²) times more probable than H_0 , based on the data.

Because of this asymmetry, hypothesis testing is not really appropriate when there is no special choice that we can use for H_0 . In such a case, it's better to report a confidence interval, a list of reasonably probable hypotheses, or simply the most probable hypothesis.

Here's the recipe:

1. When you do a hypothesis test, you first get to choose H_0 . An sample hypothesis might be "The word 'verbify' is used as a verb, rather than a noun." To make H_0 a real hypothesis that can be disproven³, it needs to be clear and unambiguous. For instance, you need to state what "used as" means; perhaps it means that there is a some unarguable verb X, such that everywhere you can use X, you could replace it with "verbify" without changing grammaticality judgements.
2. Once H_0 is chosen, H_1 is normally not- H_0 . In this case, it might be "There is no un-arguable verb X such that X and 'verbify' always yield the same grammaticality judgements."
3. Finally, you get to choose your confidence level, C ⁴. The confidence level says how large $P(H_1|\text{Data})$ needs to be before you are willing to throw H_0 out. Recall that with two hypotheses, $P(H_0|\text{Data}) + P(H_1|\text{Data}) = 1$, so you are equivalently testing for $P(H_0) < 1 - C$.

*This work is licensed under the Creative Commons Attribution License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/1.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA. This work is available under <http://kochanski.org/gpk/teaching/04010xford>.

¹ One can test a hypothesis against many alternatives. Mathematically, it's very sensible. It will tell you which is the best alternative, and it can often be the right thing to do. However, sensible though it may be, it's not the same as conventional hypothesis testing, which is our topic for this lecture.

² That's an exclamation, not a factorial sign.

³ That is, you need to make it into a little theory in the sense of Karl Popper [Popper, 1959].

⁴ C is usually something like 95% or 99% or 99.9%. People sometimes express confidence levels using a phrase like "X is rejected at P=0.01," which is the same as $C = 0.99$ or the 99% confidence level. Since confidence levels near 50% are useless, there is little possible confusion.

4. Finally, you run Bayes' Theorem (BsT), compute $P(H_1|\text{Data})$, and compare it to your confidence level. If it exceeds your confidence level, you declare that " H_0 is rejected at the $C\%$ confidence level." Thus, if you want a 99% confidence in your rejection of H_0 , you are testing to see if $P(H_1|\text{Data})/P(H_0|\text{Data}) > 99$.

If you correctly reject H_0 , and H_0 was an interesting hypothesis, then you have achieved experimental nirvana. You have conducted an experiment which ruled out a plausible theory, so scientists will no longer have to waste their time and effort in believing or testing that theory.

Of course, sometimes H_0 is a boring hypothesis that no one really believes. That can be OK, if you need it to construct an argument or to help prove that your experiment is working properly. Just don't get too excited.

2 Hypothesis Tests with Many or Infinite Alternatives

When there are many alternatives, hypothesis testing is actually a different process, logically and mathematically, though people tend to talk about it in much the same way. There is an extra step added: selecting the one best alternative hypothesis to compare against.

Why? As one adds more and more alternative hypotheses, the probability of H_0 will tend to decrease, simply because the total probability of all the hypotheses is one. Thus, anytime when we add another possible alternative, any probability in that alternative has to come from somewhere, and some of it typically comes from H_0 . In an extreme case, where there are an infinity of alternatives, it's quite possible for the total probability of all the alternatives to go to 100%, and $P(H_0) \rightarrow 0$. If we did our hypothesis test like the simple one-alternative case, comparing H_0 to not- H_0 , we'd then reject any hypothesis⁵.

Let's look at a concrete example. Suppose we have a table that is listed in a catalogue as 80 cm wide; we'll take that as H_0 . To test H_0 , we will measure it with a meter-stick and find it 80.1 cm wide. Let's say that we know our measurement is only accurate to 2 mm. Now, intuitively, we know that the measurement is close to H_0 and isn't accurate enough to reject H_0 . So, H_0 is still a plausible model for the true table. But, what is the probability that H_0 is exactly correct? Zero! No factory will be able to make a table *exactly* 80 cm wide. Machines have errors, and no one at the factory cares whether it's 80.000 cm or 80.001 cm wide. Indeed, the table will shrink and expand as the temperature and humidity change, so even if it were 80.0000 cm when it left the factory, there's no reason to expect it to be exactly the same size today. Therefore, the probability that the table is *exactly* 80 cm wide ($P(H_0)$) must be zero, and the total of all the alternative hypotheses must be 100%. If we compared the probabilities of H_0 to all of not- H_0 , then H_0 would lose, no matter what we chose for H_0 .

We get this counter-intuitive result if we compare a single point (H_0) with something that is an entire region (not- H_0). Instead, to make the comparison fair, we need to pick the single best alternative hypothesis and compare H_0 to it.

The model you compare it against should be one that has the same form as H_0 and is the best fit to your data in the family of models that includes H_0 . We will call this H_* . Now, because H_* is defined to be the best fit, $P(H_0|\text{Data}) < P(H_*|\text{Data})$, so H_* is always the MAP⁶ or Maximum Likelihood choice; However, hypothesis testing is a conservative procedure, which won't throw out H_0 unless $P(H_0|\text{Data})$ proves itself to be untenably small.

A hypothesis test has three steps:

1. Choose H_0 . For H_0 , people often assume that the experimental conditions have no effect. An example might be "People are equally good at understanding sentences if they are spoken slowly or rapidly." To make H_0 a real hypothesis that can be disproven⁷, it needs to be clear and unambiguous. For instance, you need to state what "equally good" means; we'll assume it means equal scores on some standardised comprehension test.

⁵ If there isn't really anything special about H_0 , such an attitude may actually be entirely reasonable.

⁶ Maximum À-posteriori Probability. In other words, one good way to pick the best alternative hypothesis is to take the one judged most probable by Bayes' Theorem. The output of BsT is the à-posteriori probability, which translates into "probability estimate after you've seen the data."

⁷ That is, you need to make it into a little theory in the sense of Karl Popper [Popper, 1959].

Even better, if there is a well-established published result, you might choose that as H_0 .

2. Once H_0 is chosen, H_* normally follows⁸. H_* is set by your data, and in this case, it might be “People score five points higher on a comprehension test if the test sentences are read slowly.” (Assuming that your data show a five point shift, of course.)
3. Choose the confidence level, C .
4. Then, you run BsTwith H_0 and H_* as the two alternatives, compute $P(H_*|\text{Data})$, and compare it to your confidence level. If it exceeds your confidence level, you declare that “ H_0 is rejected at the $C \cdot 100\%$ confidence level.”

To go back to our example, no one really believes that comprehension is completely independent of the speech rate. So, H_0 is somewhat of a fiction. However it may be convenient to assume H_0 to simplify your analysis or your presentation. If so, you may be asking whether H_0 is “close enough” to true so that you can get away with assuming that it is zero. Contrariwise, you may be expecting H_0 to fail and looking for confirmation that comprehension changes in the expected direction and by a plausible amount. Rejection of H_0 might provide evidence that your experiment is working well.

3 Rules for using Hypothesis Tests.

The hard part about hypothesis testing, which is often violated to some degree, is that you can only use your data once. You can only test one hypothesis on one set of data.

The above rule has some exceptions; for instance, it is acceptable⁹ to use data for more than one hypothesis test for tutorial purposes, or to help the reader understand the experiment or the data. These are OK so long as they are not an essential part of the main argument.

However, it is not acceptable practice to use the same data for several tests, then deduce something from the combined result¹⁰. One must not test two hypotheses, H_0 and h_0 with the same data set, prove them to be false, then use those two facts to prove X as if they were independently established.

Why? The procedure is not correct because the rejection of H_0 is a random event, as is the rejection of h_0 , but they are not independent random events. They are correlated because they are derived from the same set of data. For instance, if there is a wild point among the data, it will affect both hypotheses tests. Consequently, the probability of X will not be what you think it is: you may think you are 90% certain of X , but instead, there might actually be only 75% certain¹¹.

The proper way to deduce X is to combine everything into one big hypothesis test. If normal statistical methods can't handle it, you can always do something like a Bootstrap re-sampling test or a Monte Carlo simulation.

It can sometimes be reasonably safe to recycle your data for more than one hypothesis test, if the two tests are looking at the data set in a very different manner, and if the two results are used in such a way that any correlations between them won't matter. However, this should not be done lightly or without thought about all the ways it can go wrong. Again, this is not strictly correct use of data; for one thing, you have no control over what someone who reads your papers may do with your data.

Another tempting technique that should be avoided is using a single data set both to search for an interesting H_0 and then to test the result, using statistical tests based on the same data. At worst (if you hide it) that action is incompetence, pathological science, or fraud. At best (if you make your methods clear), it robs your statistical tests of their power; your results become mere suggestions rather than conclusions, ripe for proof or disproof by someone else.

⁸ You may have a little freedom in choosing H_* , such as defining it by taking the median of the data instead of the mean.

⁹ But not strictly correct.

¹⁰ You can, actually, but you will be forced into some rather ambitious Monte-Carlo simulations to calculate the confidence levels of your conclusions.

¹¹ Of course, depending on the correlation between H_0 and h_0 and the details of the deduction, you might get lucky and end up with only a 5% chance of X being false (95% confidence), but don't count on such luck.

Why? The definition of a confidence level allows a certain chance that the hypothesis test will get the answer wrong. At 95% confidence, the test will wrongly reject H_0 once in 20 data sets. Similarly, if you try 20 different hypothesis tests on one data set, you can reasonably expect that one of the tests will falsely reject its null hypothesis.

This is indeed a classic technique for lying with statistics [Huff, 1954]. For instance, if you wanted to say that “Four out of five dentists recommend” your toothpaste, a simple (but subtly dishonest) strategy is to start with five dentists. Then, if you get fewer than four recommendations, simply pick another group of five, then another, then another. Eventually, you’ll get lucky and find five dentists who recommend your product; only then do you publish.

Such behaviour is very tempting when you have a data set that is actually junk, but when one very much wants to get *something* out of it. So, you make up a hypothesis and test it. Darn! That’s not significant. Make up another and test. Rats! That’s not significant either. You begin to worry if you’ve just wasted the last few months and to wonder if you’ll have any results for that conference you were planning to go to. Then, after a few more hypotheses and tests, you get it! A statistically significant result. You write, publish, and relax. By dint of hard work, you managed to extract a good conclusion from the data.

Or did you? No. Actually, you just kept taking a 5% chance of winning (i.e. a false rejection of H_0) until you got lucky. With enough imagination and persistence, anyone can get results that seem statistically significant from a random number generator.

3.1 Exploratory Data Analysis.

But, what to do? Is it possible to explore the data before a final analysis? If you need exploratory work on your data, and you can’t get more data, you have two choices.

- The best choice is to randomly split your data into an exploratory set and a testing set. You can do whatever you want to the exploratory set. When you have explored and settled on a hypothesis, then you do one single test of that hypothesis using the testing set. And, sometimes you will find that the hypothesis you found by exploring the data is rejected on the test set. If so, you should write a paper that announces a null result, especially if the data is data is good, plentiful and clean. If everyone refused to write papers unless they had a positive result (and that’s nearly the case), science would be cluttered by thousands of spurious positive results.
- Another procedure that can work, but is slightly more vulnerable to self-delusion is as follows: Decide, in advance, how many exploratory tests you will allow yourself. Set your confidence level correspondingly tighter and explore away. For instance, if you want to allow yourself 10 exploratory tests, you should use a 99.5% confidence level instead of a 95% level. This is known as the Bonferroni correction [Bonferroni, 1936] (See the discussion in Bland [2000] and the digression on the Bonferroni Correction.) Then, when

Bonferroni Correction: Note that the Bonferroni correction assumes that the different tests are independent: i.e. if you know that one test was significant it wouldn’t help you to predict whether or not other tests were significant. If the tests are strongly correlated, then the Bonferroni correction can be too strong (if the test results would be positively correlated) or too weak (if the test results were negatively correlated.) Because the correlations between tests are generally not well known, the Bonferroni correction is sensible only when the number of tests (and therefore the size of the correction) is fairly small. Ten tests might be a reasonable upper limit. See Bland [2000] (most conveniently <http://www-users.york.ac.uk/~mb55/intro/bonf.htm>) for a discussion.

you publish, you should either:

- mention all the non-significant tests you did (good), or
- report your significant result, but at the lower (*e.g.* 95%) confidence level (acceptable). The lower confidence level accounts for the fact that (assuming H_0 were really true and the data were really non-significant), you gave the test 10 chances to make an error, instead of one.

Finally, if you get to the end of the 10 tests you allowed yourself, you're just out of luck. This procedure can get you into trouble if you are particularly insightful and can find the patterns in the data that allow you to pick a hypothesis that matches the data. Eyeballing the data may count as a hypothesis test, depending on what you see and what you know.

These points are discussed in some length in *How to Lie with Statistics* [Huff, 1954] and *Cargo Cult Science* [Feynman, 1985]. Feynman, especially, makes the point that you should publish null results just as you would publish a positive result; suppressing the failures of a theory is just a gentler form of fabricating successes.

4 What is a confidence interval?

Confidence intervals are equivalent to encapsulating the results of many hypothesis tests, and often they are just as easy to compute as a single hypothesis test. They explicitly show the region where you're likely to find the true answer.

A Wide Confidence Interval?: The experimenter comes running into the theorist's office, excitedly carrying a plot of the results of his latest experiment. "Ah!" says the theorist, "that's exactly where you should get that peak." He launches into a detailed and convincing explanation of the data. The experimenter looks puzzled: "Wait a minute—" he looks at the chart, turns it over and says "Oh, sorry." "Ah!" says the theorist, "You'd expect to get a dip in exactly that position. Here's why...".

Confidence intervals make sense any time your hypothesis is a real-valued measurement. For example:

- **Q:** What fraction of the people could find the subject of the sentence?

A: Between 40% and 45% could find it¹².

NOTE: Here, we are talking about the fraction of a large population that can accomplish the task. The fraction is extrapolated from a small survey, and the uncertainty comes about because if you have a small group, one or two extra people who are good at grammar can push the average up by a noticeable amount.

- **Q:** What is the mean f_0 of the group during a subordinate clause?

A: I'm 90% sure it was between 100 and 107 Hz.

NOTE: Here, we are talking about a sampling situation, where we are (as above) extrapolating from a small set of measured subordinate clauses to the hypothetical set of all subordinate clauses that the speaker will produce.

- **Q:** What was the mean f_0 of his voice during this subordinate clause?

A: I'm 90% sure it was between 100 and 107 Hz.

NOTE: Here, we are talking about a particular subordinate clause, and the uncertainty perhaps arises from measurement problems. Should you count the f_0 of a certain voiced fricative where the acoustic signal is not very periodic or not? If you count it, perhaps you get an average of 100 Hz, if not, perhaps you get 107 Hz.

- **Q:** What is the length of a sentence in the book?

A: 90% of the sentences were 3 to 18 words long.

NOTE: Here, we are not extrapolating at all; we are just describing the document. In this case, the sample and the population are the same: both are the set of words in the book, so there are no statistical errors from sampling. Further, there are no measurement errors.

¹² To be complete and pedantic, you would say "I am 95% certain that, based on my survey, between 40% and 45% of the adult population of Oxfordshire could find the subject of sentence number twelve."

As you can see, a confidence interval is usually described by its endpoints¹³ and a indication of how confident you are that the result is between those endpoints.

5 How are confidence intervals related to Hypothesis Tests?

They are very much equivalent. If the confidence interval doesn't include H_0 , then a hypothesis test will reject H_0 , and vice versa. In fact, if you made up a vast number of different null hypotheses, one after the other, then tested them, the accepted null hypotheses would mark out the confidence interval.

Confidence intervals put the focus on the measured number; hypothesis testing puts the focus on the theory that you are testing. Which is best depends on what you are trying to say. Generally, though, I prefer confidence intervals because they help answer the question of how big the effect is, which is the question an alert reader will ask immediately after you convince him or her that the effect exists at all.

You can work from either end:

- From the confidence interval end, you can pick a confidence level, C , and then compute a region¹⁴ where your measurement would land $C\%$ of the time, if the null hypothesis were actually true.
- Or, you can simply repeat your hypothesis test again and again, using a new null hypothesis each time (using the same data). If a particular null hypothesis is allowed, put down a mark. The confidence interval is the smallest region that contains all your marks and leaves out all points you didn't mark.

In our table example above successive null hypotheses that the table was 79.5 cm long, 79.6, 79.7, 79.8, ..., 80.2, 80.3, Then, test each of these null hypotheses and find out that (for example), the hypotheses for very short tables (length < 79.7 cm) and very long tables (length > 80.5 cm) were rejected at the 95% level. You could then say that your 95%-confidence interval for the table length was 79.7 cm–80.5 cm.

6 Recommended Reading

Read the section on confidence intervals in *Cartoon Guide* [Gonick and Smith, 1994].

References

Martin Bland. *An Introduction to Medical Statistics*. Oxford University Press, Oxford, UK, third edition, 2000. URL <http://www-users.york.ac.uk/~mb55/intro/introcon.htm#sig>.

C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936. Bonferroni adjustment for multiple statistical tests using the same data.

Richard Feynman. *Surely You're Joking Mr. Feynman!*, chapter Cargo Cult Science, pages 308–317. Bantam, 1985. URL http://www.physics.brocku.ca/etc/cargo_cult_science.html.

¹³ Endpoints are usually sufficient for one-dimensional confidence intervals. Confidence regions also make sense in two or more dimensions. For instance, a two-dimensional confidence region is usually the inside of a closed curve. A two-dimensional confidence interval is something you might want to know if you're just had your property surveyed and you want to know where one corner is. The surveyor could scratch an ellipse on the ground and say "I'm 90% certain that the corner is inside *that*." A three-dimensional confidence interval is typically the inside of a blob.

¹⁴ There are more than one possibility for the confidence interval, actually. The only strict requirement for a $C\%$ -confidence interval is that it contain $C\%$ of the probability. Conventionally, this is done symmetrically, so that you have $(100 - C)/2\%$ chance of the confidence interval being too low, and the same chance of it being too high. Sometimes, people use a "single-sided confidence interval", where all the chance of error is at one end. A single-sided confidence interval is something you might use when determining the height of a door frame; it's OK for it to be too tall, but it's not OK for the door frame to be too short. It would be phrased as "99% of humans are less than 2 meters tall." Other confidence intervals are mathematically sensible, but are never used, as they cause too much confusion.

Larry Gonick and Woollcott Smith. *The Cartoon Guide to Statistics*. HarperCollins, New York, 1994. ISBN 0062731025.

Darrell Huff. *How to Lie with Statistics*. W. W. Norton & Co., New York, 1954. ISBN 0-393-09426-X (republished in 1984).

Karl Popper. *The Logic of Scientific Discovery*. Routledge, London, New York, 1959.