Hypothesis Testing Basics

Release v10

Christopher Lee

October 03, 2015

Contents

Hypothesis testing. The words suggest two images: on the one hand, white-coated scientists sagely putting their own ideas to a serious test by doing the right experiment; on the other hand, bespectacled statisticians devising rigorous mathematical methods for scoring the results of such experiments. *Is the hypothesis accepted or rejected?* You'd expect that "hypothesis testing" methods would give a bulletproof answer to such questions.

Unfortunately, you would be wrong. Hypothesis testing is a nexus of confusion for people who try to use it, in large part because of the way it is taught and presented, even by professional statisticians. The gap between theory and practice is huge, and leads to fundamental errors in data interpretation not only in research, but in real-world dramas like murder trials. Underlying all this is basic disagreement among statistians themselves about the right way to do it. This is partly due to a long-standing "ideological" debate (frequentists vs. Bayesians), and partly due to the inherent difficulties of the problem: it is peppered with traps for the unwary, and stuck with the inconvenient truth that strictly speaking the standard methods don't give exactly what people want, namely the probability of a hidden model given a set of observations. The whole area should be emblazoned with a big red warning label shouting "DANGER! DANGER! DANGER!"

So what can we do about this? Hypothesis testing is such an important skill that we have to learn to do it well, or suffer the consequence of working hard, to produce wrong conclusions. This chapter will seek to expose the problems to the clear light of day, so that you can avoid the traps and know when you are on solid ground (vs. thin ice). Along the way you will learn the basic methods and see their uses and limitations.

1 The Basics

1.1 What is a Hypothesis Test?

A hypothesis test is an observable metric designed to predict a hidden state, namely whether the hypothesis is "true" or "false". The test itself is a discrete random variable usually with two possible states (typically called "positive" vs. "negative" or "accepted" vs. "rejected"). This random variable is itself derived from some set of observed data (i.e. the experimental data that are being used to test the hypothesis) according to a precisely defined rule. We consider it *observable* simply because once that rule is specified, there is zero uncertainty about what the value of the test variable will be given an input set of experimental data.

In this sense it is like a diagnostic test designed to test whether a patient has a disease or not: the result of the diagnostic test is observable, and predicts a hidden variable (whether the patient has disease). The only difference is that whereas the diagnostic test is a *physical* process (e.g. a strip of paper turns red vs. blue), a hypothesis test is a *calculation* that derives a new observable variable (the test result) from an input set of observable data (the experimental data).

Example

Say you wish to test whether a given coin is fair or biased. You define the following test: flip the coin ten times; if all ten flips give the same result, call the coin "biased", otherwise call it "fair". This fulfills our definition: the input data are the ten coin flip observations, and the rule specifies the value of the test variable with zero uncertainty.

Note that our definition of a test in no way guarantees that it will be accurate! In this case, the test clearly will not "detect" cases of *weak* bias (e.g. a coin that comes up heads 55% of the time).

1.2 Assessing the Accuracy of a Test

Let's consider a hidden variable H with two possible states, h^+ (e.g. "our hypothesis is true") vs. h^- (e.g. "our hypothesis is false"). Let's say our test variable T again has two states t^+ vs t^- that seek to predict the hidden state. Thus a perfect test would have $p(h^+|t^+) = 1, p(h^-|t^-) = 1$. We need precise terms for refering to all the possible combinations of cases for H vs. T:

Table 1: the four possible cases

	t^-	t^+
h^+	false negative aka "Type II error"	true positive
h^-	true negative	false positive aka "Type I error"

Clearly, a test can make two kinds of errors. Referring to our disease test, it might predict that some patients have the disease who actually do not. This is called a false positive. It might also predict that some patients do not have disease who actually do. This is called a false negative. A perfect test as defined above would produce zero false positives and zero false negatives. Thus we must report a test's accuracy in terms of both of these cases, i.e. $p(h^+, t^-), p(h^-, t^+)$.

Now that you are aware of the fact that conditional probability always has two possible directions, you can probably guess the first category of basic confusion that afflicts this field: one of these directions is relevant and meaningful ($p(H|T)$, because we know T and we want to know H), and the other is not ($p(T|H)$, because we do not know the value of H). So which direction shall we teach people to use? Unfortunately, what you will frequently encounter in talks, papers and even textbooks are the backwards forms.

Why is this a problem? After all, can't we just convert $p(T|H)$ to $p(H|T)$ whenever we need to? No. You can't compute $p(H|T)$ from $p(T|H)$, because a crucial piece of information has been omitted, namely $p(H)$, which is often either not made available, or actually hard to estimate.

The Ugly Truth: Why False Positives Make Liars of Us All

I believe there is one underlying reason for this obfuscation: the problem of false positives. Usually we are interested in predicting "special events" (e.g. getting a specific disease), which are typically low probability events (i.e. a very small fraction of the total population gets that disease). The very small joint probability of true positives means that even a low joint probability of false positives will overwhelm it, rendering $p(h^+|t^+) << 1$. A researcher may work for years to develop a test that yields both $p(t^+|h^+) \approx 1$ and $p(t^-|h^-) \approx 1$. But due solely to the fact that $p(h^+)$ is tiny, his test's actual prediction accuracy is likely to be poor, i.e. $p(h^+|t^+) << 1$. Presenting his results in terms of $p(H|T)$ will expose that ugly truth. So it will be very tempting for that researcher to join all the other people who present their results strictly in terms of $p(T|H)$. It is surprising how this is the norm in some fields (sensitivity and specificity, anyone?). For example, even the term "false positive rate" actually is defined to mean the opposite of what it should mean (i.e. it is $p(t^+|h^-)$. the opposite of what matters for real-world prediction)! Caveat emptor.

The Problem of Jargon

It doesn't help that all of this is further obscured by a plethora of impressive-sounding jargon (see table below). For example, on the Wikipedia page for *Sensitivity and Specificity*, I count over a dozen separate terms for this problem that only has three degrees of freedom (e.g. if I give you $p(t^+, h^+), p(t^+, h^-), p(t^-, h^+)$ you can compute the only remaining entry $p(t^-, h^-)$ simply from normalization). People use and listen to this plethora of terms as if there was no need to question exactly what they mean. However, people's reasons for using these terms break down roughly as follows, in order of decreasing frequency:

- used to doing things a certain way, and no longer thinks about why.
- doesn't think about these things in terms of conditional probability or hidden vs. observable direction.
- could work out the conditional probability definitions for these terms, but why would you want to do that?
- understands what the problem is, but wants to use the same terms as everyone else.

Table 2: The Jargon

measurement target	"right direction" ($T \rightarrow H$)	backwards ($H \rightarrow T$)
false positives	$p(h^-\|t^+)$: False Discovery Rate (FDR)	$p(t^+\|h^-)$: False Positive Rate
false negatives	$p(h^+\|t^-)$	$p(t^-\|h^+)$: False Negative Rate
true positives	$p(h^+\|t^+)$: Positive Predictive Value (PPV), aka precision	$p(t^+\|h^+)$: sensitivity
true negatives	$p(h^-\|t^-)$: Negative Predictive Value (NPV)	$p(t^-\|h^-)$: specificity

Personally I find all the terminology hard to remember, and prefer to think and talk about these things just in terms of the conditional probabilities (e.g. $p(h^-|t^+)$), which make it instantly obvious whether the direction is from observed to hidden (or vice versa), without trying to remember a lot of jargon. On a positive note, it is refreshing that the False Discovery Rate is starting to be more widely used in the literature.

The Counter-argument: when should we ignore p(H)?

I do not wish to suggest that one conditional probability direction is inherently more virtuous than the other, but instead more relevant to real-world prediction. Calculating the converse probability $p(T|H)$ clearly has valid uses, as long as it is not represented as a measure of real-world prediction value. Let's consider when and why we use $p(T|H)$:

- $p(T|H)$ is a likelihood model; in other words we just assume a hidden model H and compute the likelihood of the observable under this model. So we can actually compute this!

- To state that another way, we can compute this *without* needing to know $p(H)$, which can be hard to estimate. Many statisticians who were skeptical about how to obtain such "priors" accurately, therefore prefered this pure likelihood approach.
- If we wish to characterize the accuracy of our test *independent* of $p(H)$, then this is the metric we would pick – precisely because it is independent of $p(H)$. For example, in data mining, researchers characterize their methods in terms of $p(T|H)$. If they presented the opposite metric $p(H|T)$ they would have to give tables of different numbers for different values of $p(H)$.

To state the matter simply: there is nothing wrong with quoting $p(T|H)$ values as long as you *also* present $p(H|T)$ values as the explicit measure of real-world prediction performance. However, *only* presenting $p(T|H)$ values implies that they are a valid measure of real-world prediction value, which they are not.

Take-home Lessons

- in your own work, always think in terms of explicit conditional probabilities.
- assess your real-world prediction results in the "right direction", i.e. $T \rightarrow H$.
- present your final results in terms of $p(H|T)$ unless there is a very good reason for the opposite direction.
- when you hear someone else presenting their results in terms of $p(T|H)$, watch out. Try to compute $p(H|T)$ from their data, to see what their data really mean. Ask yourself *why* they presented their assessment results "backwards". Oftentimes what you'll find is that the results are actually not as good as they were made to sound.

1.3 The Next Level: Scoring Functions

A test is defined by a specific rule applied to input data. We now generalize this by considering how to implement such a rule. We adopt the general model of a *scoring function* that maps every possible value of the input variable(s) to a scalar value called the *score*. We then map this to a two-valued test variable simply by choosing a *threshold score* s_t that divides the possible scores into two subsets (those above vs. below the threshold).

This generalization suggests that we should be interested in the total performance of a given scoring function over the whole range of possible threshold values. We can do this by simply plotting the two possible error rates (false positives vs. false negatives) as a function of the threshold value. A clever way to do this is to plot the two functions as the X and Y coordinates of a graph, so the resulting curve completely characterizes the performance of the scoring function, in terms of its ability to cleanly distinguish positives vs. negatives.

The ROC Curve

The standard way of doing this is called the *ROC curve*. It plots $X = p(t^+|h^-)$ vs. $Y = p(t^+|h^+)$ as parametric functions of the score threshold s_t. A perfect ROC curve would jump to $Y = 1$ immediately at $X \geq 0$. A completely random predictor in which T, H are independent would simply yield $p(t^+|h^-) = p(t^+|h^+) = p(t^+)$ across all possible values of s_t, in other words a diagonal line $Y = X$. Real prediction methods generally fall somewhere between these two extremes. Thus, one common measure of the overall performance of a scoring function is simply the integrated *area under its ROC curve*, which goes from 0.5 for a purely random predictor, to 1.0 for a perfect predictor.

Again it should be emphasized that these metrics are defined to be independent of $p(H)$, and thus do not measure real-world prediction performance. For that, you would need to use the opposite conditional probabilities, e.g. $p(h^+|t^+)$.

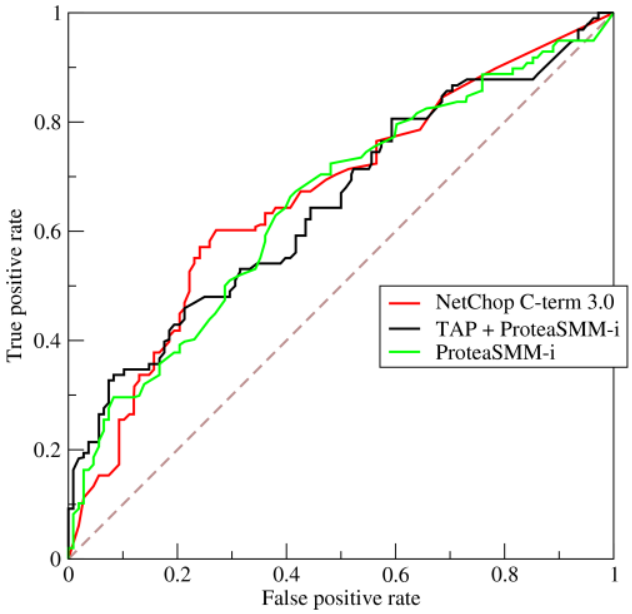

Figure 1: An example ROC curve, care of Wikipedia

2 Classical Hypothesis Testing with Extreme Value Tests

For the moment, let's assume that we want to construct a test based strictly on likelihoods, i.e. $p(obs|H)$. What scoring function should we use?

- Let's begin with an obvious candidate: why not use $p(obs|H)$ itself? On first glance it seems to have the right property: as Bayes' Law shows, a bigger value of $p(obs|H = h)$ tends to increase our belief in $H = h$, whereas a smaller value tends to reduce our belief. The problem is how to formalize this. And here we run immediately into a basic problem: there is no "absolute scale" on which to assess the observed value of $p(obs|H = h)$. Instead it is only meaningful to interpret $p(obs|H = h)$ in terms of the range of values we would *expect* to see if $H = h$ were true (i.e. under that specific model h). Thus the observed value of $p(obs|H = h)$ is not meaningful in an absolute sense, but only *relative* to the specific model h, in two ways:
  - different likelihood models h will yield different "expected likelihoods". For example, a model that splits the likelihood equally between just two possible observable states yields an expected likelihood of 1/2, whereas one that splits equally between 100 possible states yields an expected likelihood of 1/100.
  - the size of the *obs* dataset also matters critically. If *obs* consists of a single scalar variable observed once, its likelihood will be much larger than if *obs* were a vector of many variables. Remember, *p(obs)* must sum to 1 over all possible combinations of values of the variables contained in *obs*, and the number of possible combinations goes up exponentially with the number of variables.

  To summarize: observing $p(obs|H = h) = 10^{-10}$ might be "good" in one case (i.e. matches what we would expect if $H = h$ were true), and "bad" in another case (not at all what we would expect if $H = h$ were true).

For these reasons, statisticians sought a score function whose value would always mean the same thing regardless of what the model was, or the size of the *obs* dataset.

2.1 Extreme Value Tests

It turns out that a small change in computation can resolve this problem. Say our *obs* are mapped to a scalar variable O; thus our likelihood model $p(obs|H = h)$ can be transformed to a distribution $p(O|H = h)$. If we see a value of O

that falls right in the middle of the predicted distribution, that seems consistent with $H = h$, but if we get a value far out in the low probability "tail", that seems inconsistent. We can quantify this in terms of the *probability of getting a result at least this extreme given the model*. For example say we observe a value o that is much greater than expected. We can define the "p-value" for a continuous variable O

$$p_> = p(O \geq o|H = h) = \int_o^\infty p(O|H = h)dO$$

Why is this useful? Because it solves both of the above problems:

- the probability distribution of $p_>$ is always the same regardless of what the original likelihood model h was. Specifically, it is a uniform density $p(p_>) = 1$ for all possible values $0 \leq p_> \leq 1$.
- the probability distribution of $p_>$ is always the same regardless of what the *obs* are, for exactly the same reason.

So how exactly do we use this as a hypothesis test? This flows straight from the definition. We just define $p_>$ as our scoring function, and choose a threshold value α. This yields a test T: if $p_> \leq \alpha$, then t^- ("reject h"); otherwise t^+ ("accept h"). From the definition we know immediately that

$$p(t^-|H = h) = p(p_> \leq \alpha|H = h) = \alpha$$

In other words, for *obs* that were actually emitted from model h, the test will erroneously reject h only α fraction of the time. By choosing α to be small, we can make this false rejection rate as low as we want. α is usually refered to as the *significance level* of the test.

Caveats

- You should first note that this conditional probability is backwards to what we want to know for assessing real-world prediction performance, i.e. $p(H = h|T)$. We will come back to this vexing question in the next reading.
- Note that this is not truly a general hypothesis test, but only an extreme value test. That is, rather than checking that the whole distribution of the *obs* matches that predicted by the model, it simply checks whether the value of O goes *outside* the range predicted by the model. So if the *obs* differ from the model without going outside its range, the extreme value test will not detect that.
- Therefore, we must give some thought to deciding what measurable value to perform an extreme value test on. Usually this means choosing a measurable value that is predicted to differ strongly between two models (ideally, the two likelihood models would have minimal overlap), and to define an extreme value test based on one of those models.

For example, for the problem of phenotype sequencing of a pool of n independent mutant strains, we have a *non-target gene* model vs. a *target gene* model. The latter is predicted to have a much larger observed count of mutations per gene K. We can therefore construct an extreme value test based on the non-target gene model, $p(K \geq k|\lambda)$. If we define h^- to mean the non-target model, and t^+ to mean $p(K \geq k|\lambda) \leq \alpha$, then we are assured that $p(t^+|h^-) = \alpha$. We typically want the expected number of such "false positives" to be less than one *over all non-target genes in the genome*, so we choose $\alpha < 1/G$, where G is the total number of genes in the genome. Note that this is one case where we have a pretty good idea about $p(h^-)$, namely that nearly all genes in the genome are non-target genes.