

Week 8: p-values

Given some observed data, it would be useful to know how surprising they are. Are these data consistent with what I'd expect by chance? If not, something more interesting might be going on.

Take Laplace's question of the birth rate of boy vs. girls in Paris, for example. He observed 251,527 boys in 493,472 births. Is this 0.51 frequency surprisingly different from what we'd expect by chance?

To be quantitative, we have to specify exactly what chance means. We formulate a hypothesis called the "null hypothesis", H_0 , so that we can calculate $P(D | H_0)$, the probability distribution over different data outcomes, given the null hypothesis.

In the Laplace example, the obvious null hypothesis is that boys and girls are equiprobable: $p = 0.5$. The possible data outcomes, in $n = 493472$ total births, are $c = 0 \dots 493472$ boys. The probability of any given count of boys c is a binomial distribution $P(c | p, n) = \binom{n}{c} p^c (1 - p)^{n-c}$.

As Laplace was aware, the probability of any specific outcome might be absurdly small, especially as n gets large. A specific outcome can be unlikely (in the sense that you wouldn't have bet on exactly that outcome beforehand), but unsurprising (in the sense that it's one of many outcomes that are consistent with the null hypothesis). If $p = 0.5$, the probability of getting exactly 50% boys ($c = 246736$) is tiny, 0.001. But the probability of getting a number within ± 1000 of 246736 is more than 99%.

Indeed, we may be able to find a range of data that are consistent with the null hypothesis, even if any particular one outcome is unlikely, and then ask if our observed data is outside that plausible range. This requires that the data can be arranged in some sort of linear order, so that it makes sense to talk about a range, and about data outside that range. That's true for our counts of boys c , and it's also true of a wide range of "statistics" that we might calculate to summarize a dataset (for example, the mean \bar{x} of a bunch of observations $x_1 \dots x_n$).

For example, what's the probability that we would observe c boys **or more** in Laplace's problem, if $p=0.5$? For this, we (and Laplace) use a **cumulative probability function (CDF)**, the probability of getting a result of x or less:

$\forall (P(X \leq x | \theta) = \sum_{-\infty}^x P(X = x | \theta))$ Our boy count c is discrete (and only defined on $c \geq 0$) so $P(C \geq 251527) = 1 - P(C \leq 251526 | p)$, which of course we can get in Python's SciPy stats.binom module:

```
import scipy.stats as stats

c = 251527
n = 493472
p = 0.5
1 - stats.binom.cdf(c-1,n,p)
```

which gives us $1.1e-16$. That's not what Laplace got!

The trouble with computers

This number is totally wrong. The math is right, but computers are annoying. The number is so close to zero, we get an artifact from a floating-point rounding error. When numbers get very small, we have to worry about pesky details. So let's take a detour for a second into how machines do arithmetic, and where it can go wrong if you're not paying enough attention.

On a machine, in floating point math, $1 + \epsilon = 1$ for some small threshold ϵ . In double-precision floating-point math (what Python uses internally), the machine ϵ is $1.1\text{e-}16$. This is the smallest relative unit of magnitude that two floating point numbers can differ by. The result of `stats.binom.cdf()` is so close to 1 that the machine can't keep track of the precision; it just left its return value at one epsilon less than 1, and $1 - (1 - \epsilon)$ gives us ϵ .

We have to make sure that we never try to represent $1 \pm x$ if we know x might be small; we need to use x instead. Here that means we want SciPy to tell us 1 - CDF instead of the CDF. That's got a name: the **survival function**, `.sf()` in SciPy. Let's try again:

```
c = 251527
n = 493472
p = 0.5
stats.binom.sf(c-1,n,p)
```

Now we get $1.2\text{e-}42$, which is right. There's a tiny probability that we'd observe 251,527 boys or more, if the boy-girl ratio is 50:50.

Definition of a p-value

A p-value is the probability that we would have gotten a result at least this extreme, if the null hypothesis were true.

We get the p-value from a cumulative probability function $P(X \leq x)$, so it has to make sense to calculate a CDF. *There has to be an order to the data, so that "more extreme" is meaningful.* Usually this means we're representing the data as a single number: either the data is itself a number (c , in the Laplace example), or a summary statistic like a mean.

For example, it wouldn't make sense to talk about the p-value of the result of rolling a die n times. The observed data are six values $c_1 \dots c_6$, and it's not obvious how to order them. We could calculate the p-value of observing c_6 sixes **or more** out of n rolls, though. Similarly, it wouldn't make sense to talk about the p-value of a specific poker hand, but you could talk about the p-value of drawing a pair or better, because the value of a poker hand is orderable.


A p-value is a false positive rate

We will get to this in more detail when we cover classifiers, but there is something called a false positive rate. If you have a group of items belonging to one class ("0") or another ("1"), and then you apply some strategy to guess the class of each item, you can assess the performance of your strategy by calculating the false positive rate: the fraction of false positives out of all negatives: $\frac{\text{FP}}{\text{FP} + \text{TN}}$. If we consider our test statistic x to be the threshold for defining positives, i.e. everything that scores at least x is called positive, then the p-value and the false positive rate are the same thing: for data samples generated by the null hypothesis (negatives), what fraction of the time do they nonetheless score x or greater?

This idea leads to a simple way of calculating p-values called *order statistics*. Generate N synthetic negative datasets, calculate the score (test statistic) for each of them, and count the fraction of times that you get x or more; that's the p-value for score x .

Any experimentalist is familiar with this idea. *Do negative controls*. Simulate negative datasets and count how frequently a negative dataset gets a score of your threshold x or more.

p-values are uniformly distributed on (0,1)



If the data were actually generated by the null hypothesis, and you did repeated experiments, calculating a p-value for each observed data sample, you would see that the p-value is uniformly distributed. By construction – simply because it's a cumulative distribution! 5% of the time, if the null hypothesis is true, we'll get a p-value of < 0.05 ; 50% of the time, we'll get a p-value < 0.5 .

Understanding this uniform distribution of p-values is important. Sometimes people say that a result with a p-value of 0.7 is “less significant” than a result with a p-value of 0.3, but in repeated samples from the null hypothesis, you expect to obtain the full range of possible p-values from 0..1 in a uniform distribution. Seeing a p-value of 0.7 is literally *equally* probable as seeing a p-value of 0.3, or 0.999, or 0.001, under the null hypothesis. Indeed, seeing a uniform distribution under the null hypothesis is a good check that you're calculating p-values correctly.

Null hypothesis significance testing

P-values were introduced in the 1920's by the biologist and statistician Ronald Fisher. He intended them to be used as a tool for detecting unusual results:

“Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.”

There are three important things in this passage. First, it introduced $P < 0.05$ as a standard of scientific evidence. Second, Fisher recognized that this was a “low standard”. Third, by saying “rarely fails”, Fisher meant it to be used in the context of *repeated* experiments, not a single experiment: a true effect should *reproducibly* and *repeatedly* be distinguishable from chance.

Many fields of science promptly forgot about the second two points and adopted $P < 0.05$ as a hard standard of scientific evidence. A result is said to be “statistically significant” if it achieves $P < 0.05$. Sometimes, contrary to both logic and what Fisher intended, a single result with $P < 0.05$ is publishable in some fields.

Nowadays there's a backlash. Some people want to change the 0.05 threshold to 0.005, which rather misses the point. Some people want to ban P-values altogether.

P-values are useful, if you're using them the way Fisher intended. It *is* useful to know when the observed data aren't matching well to an expected null hypothesis, alerting you to the possibility that something else may be going on. But 5% is a low standard – even if the null hypothesis is true, 5% of the time you're going to get results with $P < 0.05$. You need to see your unusual result reproduce consistently before you're going to believe in it.

- When you say that you've rejected the null hypothesis H_0 , and therefore your hypothesis H_1 is true. A tiny p-value doesn't necessarily mean the data support some other hypothesis of yours, just because the data don't agree with the null hypothesis. Nothing about a p-value calculation tests any other hypothesis, other than the null hypothesis.
- When you equate “statistical significance” with effect size. A miniscule difference can become statistically significant, given large sample sizes. The p-value is a function of both the sample size and the effect size. In a sufficiently large dataset, it is easy to get small p-values, because real data always depart from simple null hypotheses. This is often the case in large, complex biological datasets.
- When you do multiple tests but you don't correct for it. Remember that the p-value is the probability that your test statistic would be at least this extreme if the null hypothesis is true. If you chose $\alpha = 0.05$ (the “standard” significance threshold), you're going to get values that small 5% of the time, even if the null hypothesis is true: that is, **you are setting your expected false positive rate to 5%**. Suppose there's nothing going on and your samples are generated by the null hypothesis. If you test one sample, you have a 5% chance of erroneously rejecting the null. But if you test a million samples, 50,000 of them will be “statistically significant”.

Most importantly, using a p-value to test whether your favorite hypothesis H_1 is supported by the data is fundamentally illogical. A p-value test never even considers H_1 ; it only considers the null hypothesis H_0 . “Your model is unlikely; therefore my model is right!” is just not the way logic works.

Multiple testing correction

Suppose you do test $n =$ one million things. What do you need your p-value to be (per test), to decide that any positive result you get in N tests is statistically significant?

Well, you expect np false positives. The probability of obtaining one or more false positives is (by Poisson) $1 - e^{-np}$. This is still a p-value, but with a different meaning, conditioned on the fact that we did n tests: now we’re asking, what is the probability that we get result at least this extreme (at least one positive prediction), given the null hypothesis, when we do n independent experiments? For small x , $1 - e^{-x} \simeq x$, so the multiple-test-corrected p-value is approximately np . That is, multiply your per-test p-value by the number of tests you did to get a “corrected p-value”. Like many simple ideas, this simple idea has a fancy name: it’s called a **Bonferroni correction**. It’s considered to be a very conservative correction.

The false discovery rate (FDR)

One reason that the Bonferroni correction is conservative is the following. Suppose you run a genome-wide screen and you make 80,000 predictions. Do you really need all of them to be “statistically significant” on their own? That is, do you really need to know that the probability of even one false positive in that search is < 0.05 or whatever? More reasonably, you might say you’d like to know that 99% of your 80,000 results are true positives, and 1% or less of them are false positives.

Suppose you tested a million samples to get your 80,000 positives, at a per-test p-value threshold of < 0.05 . By the definition of the p-value you expected up to 50,000 false positives, because in the worst case, all million samples are in fact from the null hypothesis, and at a significance threshold $\alpha = 0.05$, you expect 5% of them to be called as (false) positives. So if you trust your numbers, at least 30,000 of your 80,000 predictions (80000 positives - 50000 false positives) are expected to be true positives. You could say that the expected fraction of false positives in your 80,000 positives is $50000/80000 = 62.5\%$.

This is called a **false discovery rate** calculation – specifically, it is called the Benjamini-Hochberg FDR.

The false discovery rate (FDR) is the proportion of your called “positives” that are expected to be false positives, given your p-value threshold, the number of samples you tested, and the number of positives that were “statistically significant”.

Getting to the FDR from the p-value is straightforward. Suppose we’ve ranked all n tests by their p-value, and we set a cutoff threshold at the r th best test – i.e. we take the top r tests and call them statistically significant. Let the p-value of the r th test be p_r ; then we expect up to np_r false positives with this p-value or better (because the p-value is literally the false positive rate). We made r predictions, and we expect up to np_r to be false positives... the FDR is the fraction of the r predictions that we expect to be false, $< \frac{np_r}{r}$. People typically choose FDR thresholds of 0.05 or so.

What Bayes says about p-values

A good way to see the issues with using p-values for hypothesis testing is to look at a Bayesian posterior probability calculation. Suppose we’re testing our favorite hypothesis H_1 against a null hypothesis H_0 , and we’ve collected some data D . What’s the probability that H_1 is true? That’s its posterior:

$$P(H_1 | D) = \frac{P(D | H_1)P(H_1)}{P(D | H_1)P(H_1) + P(D | H_0)P(H_0)}$$

To put numbers into this, we need to be able to calculate the likelihoods $P(D \mid H_1)$ and $P(D \mid H_0)$, and we need to know how the priors $P(H_0)$ and $P(H_1)$ – how likely H_0 and H_1 were before the data arrived.

What does p-value testing give us? It gives us $P(s(D) \geq x \mid H_0)$: the cumulative probability that some statistic of the data $s(D)$ has a value at least as extreme as x , under the null hypothesis.

We don't know anything about how likely the data are under our hypothesis H_1 . We don't know how likely H_0 or H_1 were in the first place. And we don't even know $P(D \mid H_0)$, really, because all we know is a related cumulative probability function of H_0 and the data.

Therefore it is utterly impossible (in general) to calculate a Bayesian posterior probability, given a p-value – which means, a p-value simply cannot tell you how much evidence your data give in support of your hypothesis H_1 .

(This was the fancy way of saying that just because the data are unlikely given H_0 does not logically mean that H_1 must be true.)