Assessing evidence: statistical hypothesis testing

Data Science Team

Stanford University, Department of Statistics
Summary of concepts so far

**Null hypothesis** status quo

**Test statistics** a random quantity that depends on the data and whose distribution we know under the null hypothesis

**p-value** probability of observing a value of the test statistics that is as contradictory of the null hypothesis or more than the observed value

We reject the null hypothesis when the p-value is small.
For a brief moment in 2010, Matt Motyl was on the brink of scientific glory: he had discovered that extremists quite literally see the world in black and white. The results were "plain as day," recalls Motyl, a psychology PhD student at the University of Virginia in Charlottesville. Data from a study of nearly 2,000 people seemed to show that political moderates saw shades of grey more accurately than did either left-wing or right-wing extremists. "The hypothesis was sexy," he says, "and the data provided clear support." The P value, a common index for the strength of evidence, was 0.01 — usually interpreted as 'very significant.' Publication in a high-impact journal seemed within Motyl's grasp.

But then reality intervened. Sensitive to controversies over reproducibility, Motyl and his adviser, Brian Nosek, decided to replicate the study. With extra data, the P value came out as 0.59 — not even close to the conventional level of significance, 0.05. The effect had disappeared, and with it, Motyl's dreams of youthful fame.

It turned out that the problem was not in the data or in Motyl's analyses. It lay in the surprisingly slippery nature of the P value, which is neither as reliable nor as objective as most scientists assume. "P values are not doing their job, because they can't," says Stephen Ziliak, an economist at Roosevelt University in Chicago, Illinois, and a frequent critic of the way statistics are used.

For many scientists, this is especially worrying in light of the reproducibility concerns. In 2005, epidemiologist John Ioannidis of Stanford University in California suggested that most published findings are false; since then, a string of high-profile replication problems has forced scientists to rethink how they evaluate results. At the same time, statisticians are looking for better ways of thinking about data, to help scientists to avoid missing important information or acting on false alarms. "Change your statistical philosophy and all of a sudden different things become important," says Steven Goodman, a physician and statistician at Stanford. "Then 'laws' handed down from God are no longer handed down from God. They're actually handed down to us by ourselves, through the methodology we adopt."

P values have always had critics. In their almost nine decades of existence, they have been likened to mosquitoes (annoying and impossible to swat away), the emperor's new clothes (fraught with obvious problems that everyone ignores) and the tool of a "sterile intellectual rake" who ravishes science but leaves it with no progeny. One researcher suggested rechristening the methodology "statistical hypothesis inference testing," presumably for the acronym it would yield. The irony is that when UK statistician Ronald Fisher introduced the P value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

- The chance of observing a result as unexpected as the recorded one if the null hypothesis is true.
- Not the probability that the null is true.
- Not the probability that the finding will not be replicated.
What is a *p*-value?

Suppose marital_status and sex really were independent and our marginal frequencies were good estimates.

- Let’s calculate the p-value for all the tables we have generated under independence,
What is a *p*-value?

We call the density with histograms as in the previous page *uniform* on the interval [0,1].

A random variable has a **uniform distribution** on the unit interval if

\[ P[X \in [a, b]] = |b - a| \]

for any \( 0 \leq a \leq b \leq 1 \).

- P-values are a random quantity, and they have **uniform distribution under the null**
- This is an example of the general result referred to as the probability integral transform (you don’t need to know this, just putting it here in case of interest)
What can we do with a *p*-value?

Suppose we decide that a *p*-value less than 0.05 is strong evidence against independence.

What do we know?

- Because of the uniform distribution, 5% of the time, even when marital_status and sex are independent, we will say that there is strong evidence they are not independent.
- This type of error is called a **Type I error**: rejecting the null when the null is true.
- Thresholding the *p*-value at 5% controls the **Type I error rate** at 5%.
- In other words, suppose we test *many* hypotheses by thresholding the *p*-value at 5%, and in **every case**, the null was actually true, then in about 5 percent of those cases, we will **incorrectly reject the null**.
Recap of p-values and hypothesis testing

So far this week, we have discussed doing inference in the following way: 1. State a null hypothesis $H_0$ 2. Define a test statistic and compute it on the data 3. Quantify evidence about the hypothesis by comparing to a reference distribution (the distribution of the test statistic given that the hypothesis is true). We often do this by computing a p-value

- We talked about **Type I errors**: rejecting the null hypothesis when the null hypothesis is true, and said that if we reject the null when the p-value $p < \alpha$, then we will have a type I error rate of $\alpha$, which we call **level** of the test.
- But what happens when the null **isn’t true**?
The alternative hypothesis $H_1$

- Often, this is just the complement (or negation) of the null. For example, when we were testing for independence of $X$ and $Y$, the alternative was just that $X$ and $Y$ were dependent.
- Sometimes the alternative is more specific. Ex. $H_0 : \mu = 0$ and $H_1 : \mu = 2$.
- We don’t want to reject the null when the null is true... but ideally we would like to reject the null when the alternative is true as often as possible.
- Failing to reject the null when the alternative is true is called a **Type II error**: false negative.
- The **power** of a test is, instead, the probability of rejecting the null hypothesis when it is false.

\[
\text{Power} = 1 - \mathbb{P}(\text{type 2 error})
\]
Putting all the terms together

We need to decide between $H_0$ and $H_1$ and there are two ways in which we can go wrong

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ is true</th>
<th>$H_1$ is true</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>Type I error</td>
<td>-</td>
</tr>
<tr>
<td>Accept $H_0$</td>
<td>-</td>
<td>Type II error</td>
</tr>
</tbody>
</table>

- A statistical test is a rule depending on the data that makes us decide for $H_0$ or $H_1$
- The probability of type I error is called **level** of the test
- 1-probability of type II error is called **power** of the test
Suppose $x$ is normally distribution with mean $\mu$ and variance 1,

$$x \sim \mathcal{N}(\mu, 1)$$

We will visualize the power of the one-sided test

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu > 0$$

at fixed level $\alpha = 0.05$
Let’s construct a test

It makes sense to use $X$ as a test statistics

Which values of $X$ “contradict” the null hypothesis?

- The larger $X$ is, the more the null seems contradicted
- $\mathbb{P}(X > x_{\text{obs}}|\mu = 0)$ is the p-value.
- Decision rule: reject $H_0$ when p-value is smaller than 0.05.
- Equivalent decision rule: reject $H_0$ when $x_{\text{obs}} > 1.65$
High power regime
\( \mu = 3, \text{ power } = 0.91 \)

Power is 0.91 and \( P(\text{type 2 error}) = P(\text{false negative}) = 0.09 \)
Power is 0.44 and $\mathbb{P}(\text{type 2 error}) = \mathbb{P}(\text{false negative}) = 0.56$
Low power regime
$\mu = 0.25$, power = 0.08

Power is 0.08 and $\mathbb{P}(\text{type 2 error}) = \mathbb{P}(\text{false negative}) = 0.92$
Power for one sided test of level 0.05

```r
mu = seq(-4, 4, length=51)
power = 1 - pnorm(1.65, mu, 1)
plot(mu, power, type='l', lwd=2, col='red', ylim=c(0,1))
abline(h=0.05)
```
Power for two sided test of level 0.05

```r
power = pnorm(-1.96, mu, 1) + 1 - pnorm(1.96, mu, 1)
plot(mu, power, type='l', lwd=2, col='red', ylim=c(0,1))
abline(h=0.05)
```
This makes sense – we have trouble distinguishing the null from the alternative when they correspond to very similar values of $\mu$. When the values they correspond to are very different, it is easier to distinguish the null and the alternative, because the data we see tend to look very different from what we would expect under one or the other hypothesis.