MS&E 226: “Small” Data

Lecture 14: Introduction to hypothesis testing (v2)

Ramesh Johari
ramesh.johari@stanford.edu

Hypotheses

Quantifying uncertainty

Recall the two key goals of inference:

- *Estimation*. What is our best guess for the process that generated the data?
- *Quantifying uncertainty*. What is our uncertainty about our guess?

Hypothesis testing provides another way to quantify our uncertainty.

Null and alternative hypotheses

In hypothesis testing, we quantify our uncertainty by asking whether it is likely that data came from a particular distribution.

We will focus on the following common type of hypothesis testing scenario:

- The data $\mathbf{Y}$ come from some distribution $f(Y|\theta)$, with parameter θ.
- There are two possibilities for θ: either $\theta = \theta_0$, or $\theta \neq \theta_0$.
- We call the case that $\theta = \theta_0$ the *null hypothesis*.[1]
- We call the case that $\theta \neq \theta_0$ the *alternative hypothesis*.[2]

[1] A hypothesis that is a single point is called *simple*.

[2] A hypothesis that is not a single point is called *composite*.

Examples

Some examples of null hypotheses:

- You observe n flips of a biased coin, and test whether it is likely that the bias of the coin is different from $1/2$.

Examples

Some examples of null hypotheses:

- You observe n flips of a biased coin, and test whether it is likely that the bias of the coin is different from $1/2$.
- You compute an ordinary least squares regression, and test whether it is likely that the true coefficient β_j is nonzero, given the data you have observed.

Examples

Some examples of null hypotheses:

- You observe n flips of a biased coin, and test whether it is likely that the bias of the coin is different from $1/2$.
- You compute an ordinary least squares regression, and test whether it is likely that the true coefficient β_j is nonzero, given the data you have observed.
- You compare two versions of a webpage, and test whether it is likely that the true conversion rates of the two pages are different, given the data you have observed. (Here you are testing whether the *difference* of the conversion rates is nonzero.)

Examples

Some examples of null hypotheses:

- You observe n flips of a biased coin, and test whether it is likely that the bias of the coin is different from $1/2$.
- You compute an ordinary least squares regression, and test whether it is likely that the true coefficient β_j is nonzero, given the data you have observed.
- You compare two versions of a webpage, and test whether it is likely that the true conversion rates of the two pages are different, given the data you have observed. (Here you are testing whether the *difference* of the conversion rates is nonzero.)

In each case, *you already know how to form an estimate of the desired quantity*; hypothesis tests gauge whether the estimate is meaningful.

The hypothesis testing recipe

The basic idea is:

- *If* the true parameter was θ_0...

The hypothesis testing recipe

The basic idea is:

- *If* the true parameter was θ_0...
- *then* $T(\mathbf{Y})$ should look like it came from $f(Y|\theta_0)$.

The hypothesis testing recipe

The basic idea is:

- *If* the true parameter was θ_0...
- *then* $T(\mathbf{Y})$ should look like it came from $f(Y|\theta_0)$.
- We compare the *observed* $T(\mathbf{Y})$ to the *sampling distribution under* θ_0.

The hypothesis testing recipe

The basic idea is:

- *If* the true parameter was θ_0...
- *then* $T(\mathbf{Y})$ should look like it came from $f(Y|\theta_0)$.
- We compare the *observed* $T(\mathbf{Y})$ to the *sampling distribution under* θ_0.
- If the observed $T(\mathbf{Y})$ is unlikely under the sampling distribution given θ_0, we *reject the null hypothesis that* $\theta = \theta_0$.

Note: Rejecting the null *does not mean* we accept the alternative!

Example: biased coin flipping

Suppose that we flip a coin 10 times. We observe 9 heads.

We estimate the bias as $\hat{q} = 0.8$. How likely are we to observe an estimate this extreme, *if* the coin really had bias $1/2$?

- In that case, the number of heads in ten flips is $\mathrm{Binomial}(10, 1/2)$.
- The chance of seeing at least 9 heads is ≈ 0.0107.

In other words, it is *very unlikely* that we would have seen so many heads if the true bias were $1/2$; seems reasonable to reject the null hypothesis.

Decision rules

In general, a hypothesis test is implemented using a *decision rule* given the test statistic. We focus on decision rules like the following::

"If $|T(\mathbf{Y})| \geq s$, then reject the null; otherwise accept the null."

In other words, the test statistics we consider will have the property that they are unlikely to have large magnitude under the null (e.g., $\hat{q}$ in the preceding example).

Errors

How do we evaluate a decision rule?

Note that hypothesis testing is just a version of a binary classification problem: *is the null true?*

Errors

How do we evaluate a decision rule?

Note that hypothesis testing is just a version of a binary classification problem: *is the null true?*

So there are two possible errors:

- *False positive*: In fact the null is true, but we mistakenly reject the null.
- *False negative*: In fact the null is false, but we mistakenly fail to reject the null.

Errors

How do we evaluate a decision rule?

Note that hypothesis testing is just a version of a binary classification problem: *is the null true?*

So there are two possible errors:

- *False positive*: In fact the null is true, but we mistakenly reject the null.
- *False negative*: In fact the null is false, but we mistakenly fail to reject the null.

The false positive probability $\mathbb{P}(\text{reject}|\theta_0)$ of a test is called its *size*.

For any specific alternative $\theta \neq \theta_0$, $\mathbb{P}(\text{reject}|\theta)$ is a called the *power* at θ.

“Good” hypothesis tests

So good hypothesis tests are those that:

- Have small false positive probability (small size)
- While providing small false negative probability (high power)