

MS&E 125: Intro to Applied Statistics

Hypothesis Testing

Professor Udell

Management Science and Engineering
Stanford

April 24, 2023

Announcements

Outline

Hypothesis testing

Comparing two samples

Choosing a cutoff

Multiple hypotheses

Summary

Jury selection

Amendment VI of the United States Constitution states,

In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the State and district wherein the crime shall have been committed.

Swain vs. Alabama (1965)

- ▶ Robert Swain, a Black man, was convicted in Talladega County, Alabama, in 1962
- ▶ 26% of eligible jurors were Black
- ▶ jurors were selected from among 100 panelists
- ▶ only 8 of the 100 panelists were Black

Poll: was the jury rigged?

source: [https:](https://inferentialthinking.com/chapters/11/1/Assessing_a_Model.html)

[//inferentialthinking.com/chapters/11/1/Assessing_a_Model.html](https://inferentialthinking.com/chapters/11/1/Assessing_a_Model.html)

Hypothesis testing

how likely is this outcome, if the jury were selected at random?

- ▶ **null hypothesis:** the jury was selected at random from the eligible population
- ▶ **alternative hypothesis:** the jury was not selected at random from the eligible population
- ▶ **test statistic:** the number of Black jurors

Demo

approach:

- ▶ simulate the jury selection process many times
- ▶ visualize the **sampling distribution** of the test statistic using simulation
- ▶ compute the **p-value**: the proportion of simulations where the test statistic is at least as extreme as the observed value
- ▶ if the p-value is small (often, $< .05$), we **reject the null hypothesis**

`https://colab.research.google.com/github/
stanford-mse-125/demos/blob/main/testing.ipynb`

Statistics on the supreme court

Swain vs. Alabama (1965): “the overall percentage disparity has been small”

- ▶ how was the supreme court measuring the disparity?
- ▶ how would you suggest measuring it?

Outline

Hypothesis testing

Comparing two samples

Choosing a cutoff

Multiple hypotheses

Summary

Deflategate

the New England Patriots were accused of deflating footballs in the 2015 AFC Championship game

- ▶ NFL rules require footballs to be inflated in the range 12.5 to 13.5 psi
- ▶ each team is responsible for ensuring their footballs are properly inflated
- ▶ Colts intercepted a ball and measured its pressure < 12.5 psi
- ▶ the Patriots were accused of deflating the footballs to make them easier to grip

Comparing two samples

- ▶ **null hypothesis:** the two samples are drawn from the same population
- ▶ **alternative hypothesis:** the two samples are drawn from different populations
- ▶ **test statistic:** the difference between the two sample means

Demo

approach:

- ▶ simulate the process of assigning footballs to teams many times
- ▶ visualize the **sampling distribution** of the test statistic using simulation
- ▶ compute the **p-value**: the proportion of simulations where the test statistic is at least as extreme as the observed value
- ▶ if the p-value is small (often, $< .05$), we **reject the null hypothesis**

`https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/testing.ipynb`

Outline

Hypothesis testing

Comparing two samples

Choosing a cutoff

Multiple hypotheses

Summary

Choosing a cutoff

the p-value is the probability of observing a test statistic at least as extreme as the one observed, under the null hypothesis

- ▶ example: if the p-value is 0.05, then there is a 5% chance of observing a test statistic at least as extreme as the one observed, under the null hypothesis

in typical parlance, we say

- ▶ a p-value > 0.05 is **not statistically significant**
- ▶ a p-value < 0.05 is **statistically significant**
- ▶ a p-value < 0.01 is **highly statistically significant**

False positives vs false negatives

- ▶ **false positive:** we reject the null hypothesis when it is true
- ▶ **false negative:** we fail to reject the null hypothesis when it is false

example: cancer screening based on blood test

- ▶ cost of false positive: unnecessary treatment
- ▶ cost of false negative: cancer goes undetected

cost is different for different patients, so cutoff should also be different!

Example: prostate cancer screening

PSA (prostate-specific antigen) is a protein produced by the prostate gland

- ▶ most men w/o prostate cancer have $\text{PSA} < 4 \text{ ng/ml}$
- ▶ men with PSA between 4 and 10 ng/ml have a 25% chance of having prostate cancer
- ▶ men with $\text{PSA} > 10 \text{ ng/ml}$ have a 50% chance of having prostate cancer

Poll: who has a higher cost for a false positive? for a false negative? what cutoff would you use for follow-up testing in

- ▶ 80yo patient
- ▶ 40yo patient

Outline

Hypothesis testing

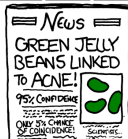
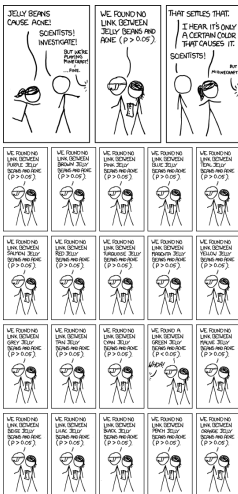
Comparing two samples

Choosing a cutoff

Multiple hypotheses

Summary

Multiple hypothesis testing



Multiple hypothesis testing

imagine the experiment:

- ▶ scientists divide students into **test** and **control** population

Multiple hypothesis testing

imagine the experiment:

- ▶ scientists divide students into **test** and **control** population
- ▶ scientists give the test group jellybeans

Multiple hypothesis testing

imagine the experiment:

- ▶ scientists divide students into **test** and **control** population
- ▶ scientists give the test group jellybeans
- ▶ (what do they give the control group?)

Multiple hypothesis testing

imagine the experiment:

- ▶ scientists divide students into **test** and **control** population
- ▶ scientists give the test group jellybeans
- ▶ (what do they give the control group?)
- ▶ scientists compare the students' acne levels

Multiple hypothesis testing

imagine the experiment:

- ▶ scientists divide students into **test** and **control** population
 - ▶ scientists give the test group jellybeans
 - ▶ (what do they give the control group?)
 - ▶ scientists compare the students' acne levels
-
- ▶ null hypothesis: jellybeans have no effect on acne

Multiple hypothesis testing

imagine the experiment:

- ▶ scientists divide students into **test** and **control** population
 - ▶ scientists give the test group jellybeans
 - ▶ (what do they give the control group?)
 - ▶ scientists compare the students' acne levels
-
- ▶ null hypothesis: jellybeans have no effect on acne
 - ▶ alternative hypothesis: jellybeans have an effect on acne

Multiple hypothesis testing

imagine the experiment:

- ▶ scientists divide students into **test** and **control** population
 - ▶ scientists give the test group jellybeans
 - ▶ (what do they give the control group?)
 - ▶ scientists compare the students' acne levels
-
- ▶ null hypothesis: jellybeans have no effect on acne
 - ▶ alternative hypothesis: jellybeans have an effect on acne
 - ▶ test statistic: proportion of patients with acne in test vs control group

Demo

`https://colab.research.google.com/github/
stanford-mse-125/demos/blob/main/testing.ipynb`

What to do?

if you read a scientific finding, consider

- ▶ how many hypotheses do you think they tested to find this result?
- ▶ how many similar hypotheses did other research groups test?

What to do?

if you read a scientific finding, consider

- ▶ how many hypotheses do you think they tested to find this result?
- ▶ how many similar hypotheses did other research groups test?

in your own work, consider methods to control the **false discovery rate** (FDR)

- ▶ **Bonferonni correction:** divide the cutoff significance level by the number of hypotheses tested
- ▶ ... many more!

Outline

Hypothesis testing

Comparing two samples

Choosing a cutoff

Multiple hypotheses

Summary

Demo

`https://colab.research.google.com/github/
stanford-mse-125/demos/blob/main/bootstrap_
cheatsheet.ipynb`