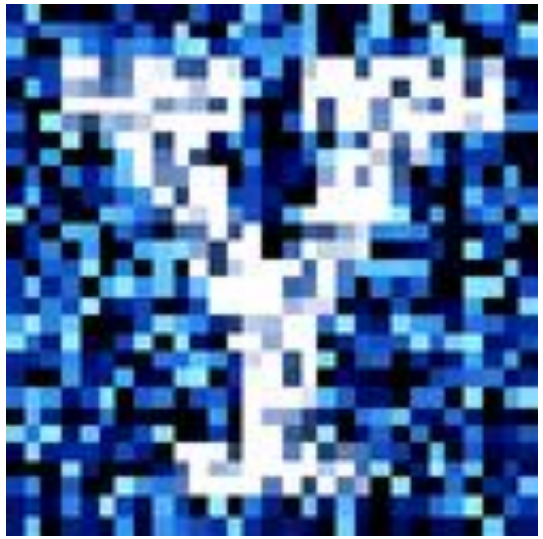
YData: Introduction to Data Science



Lecture 18: Decisions and Uncertainty

Overview

Hypothesis tests

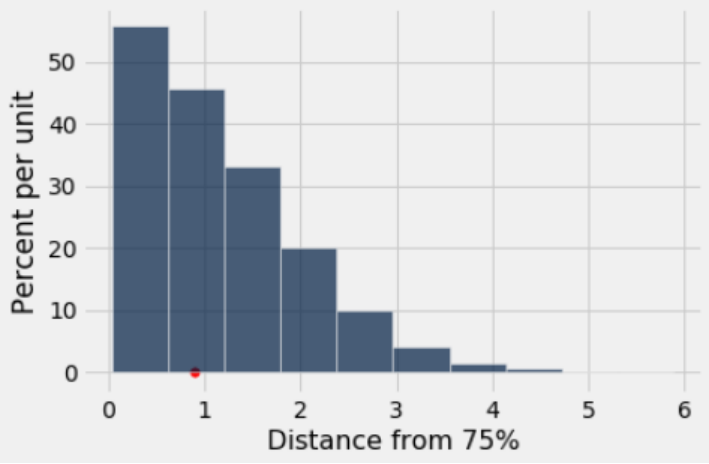
- Terminology
- Performing hypothesis tests
- p-values
- Error probabilities

Quick review: assessing models

The last two classes we assess models by:

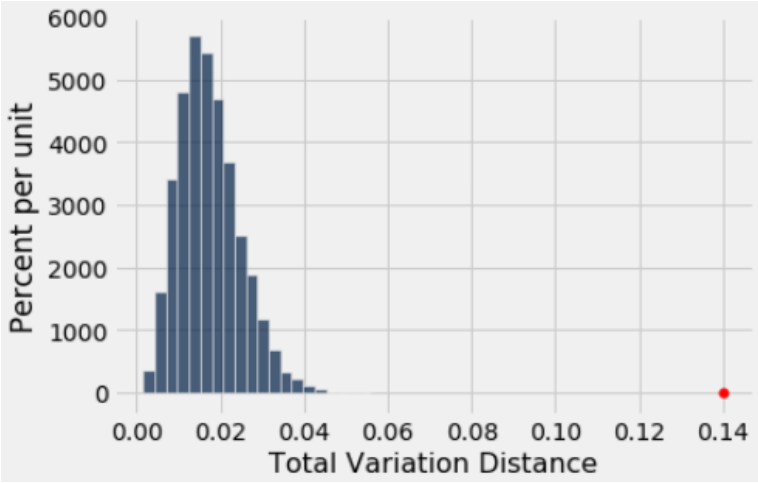
1. Choosing a statistic that will help you decide whether the model is consistent with observed data
2. Simulating the statistic under the assumptions of the model
3. Comparing the data to the model's predictions

Mendel's peas



statistic: absolute distance from 75%

Alameda county jury panels



statistic: TVD between distributions

Terminology

Testing hypotheses

In the examples we have seen we are choosing between two views

Mendel's peas

- 75% of peas have flowers that are purple vs.
- No, this is not the case

Almeda county jury selection

- Juries are selected randomly from a population vs.
- No, jury selection is biased

The views are called hypotheses

Null and Alternative hypotheses

Null hypothesis

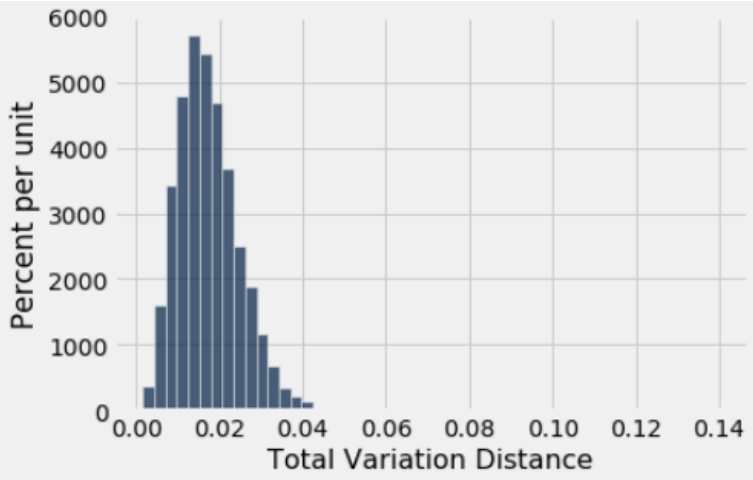
- A hypothesis that gives a well-defined chance model about how the data were generated
- We can simulate data under the assumptions of this model to get a "null distribution" of statistics

Alternative hypothesis

- A different view about the origin of the data

A test statistic is the statistic we choose to simulate in order decide between the two hypotheses

Alameda county jury "null distribution"

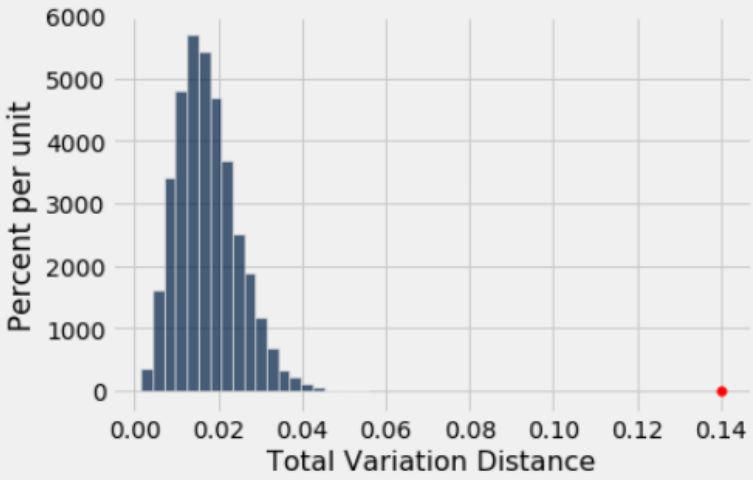


Testing the null hypothesis

To resolve choice between null and alternative hypotheses:

- We compare the observed test statistic to the statistic values in the null distribution
- If the observed statistic is not consistent with the null distribution, then we can reject the null hypothesis
  - And we accept the alternative hypothesis

Alameda county jury "null distribution"



Performing a test

Example problem

A large Statistics class at Berkeley was divided into 12 discussion sections

Graduate Student Instructors (GSIs) lead the sections

After the midterm, students in Section 3 notice that the average score in their section is lower than in others

The GSI's defense

Section 3 GSI's position (Null Hypothesis):

- If we had picked my section at random from the whole class, we could have got an average like this one

Alternative:

- No, the average score is too low. Randomness is not the only reason for the low scores

Let's explore this in Jupyter!

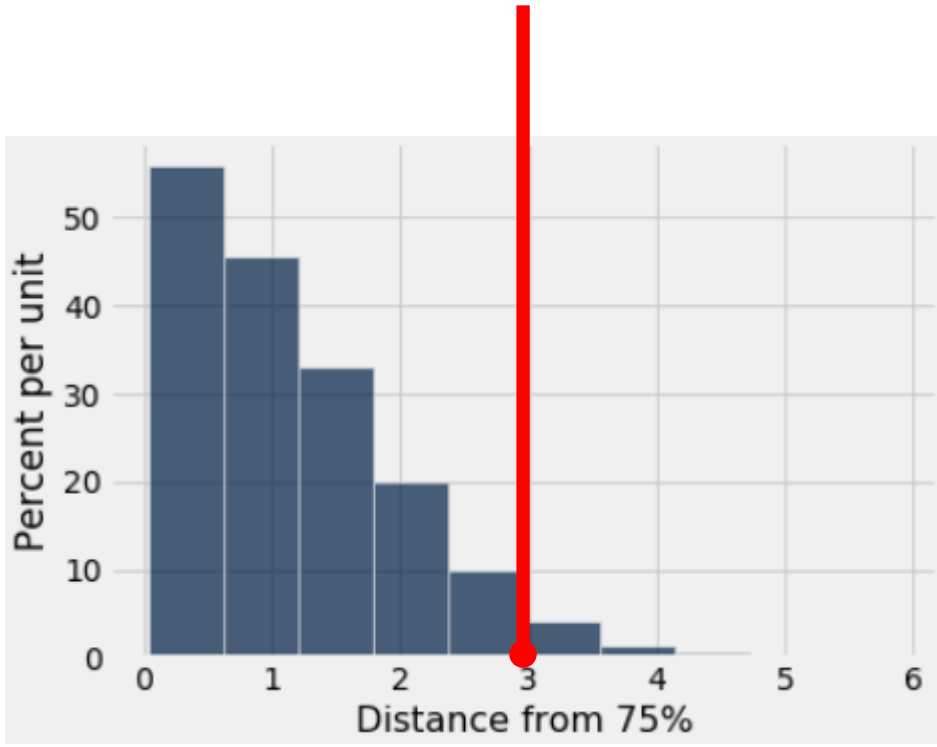
P-values

Definition of the p-value

"Inconsistent": The test statistic is in the tail of the empirical distribution under the null hypothesis

The p-value is the probability, that we get a statistic as or more extreme than the observed statistic from the null distribution

- P(Null_Stat ≥ obs_stat | H_0)



Let's explore this in Jupyter!

The error probability

Can the conclusion be wrong?

Yes!

	Null is true

The error probability

If we reject the null hypothesis when the p-value is less than a strict cutoff value (called α), then this cutoff value give the probability we will make a type I error.

e.g., if we reject the null hypothesis when the p-values is less than 0.05, then when the null hypothesis is true, there is a 5% probability that your test will falsely reject it

- i.e., if you ran 100 hypothesis tests, then approximately 5% would be type I errors

Statistical significance

Historically, if a p-value was less than a particular threshold (α) than the results were called statistically significant

A commonly used threshold is that the p-value is $\alpha = 0.05$

- i.e., the p-value should be less than 0.05 to reject the null hypothesis

Use of the $\alpha = 0.05$ is often attributed to Ronald Fisher

- "It is convenient to take this point [5%] as a limit in judging whether a deviation is to be considered significant or not."

  - Statistical Methods for Research Workers

The replication crisis

Recently there has been a "replication crisis" in several scientific fields where many results can not be replicated

One reason some statisticians have given for the replication crisis is due to the use of a strict cutoff threshold on p-values

- i.e., researchers might be running many experiments and the publishing any result that appears "statistically significant"

American Statistical Association

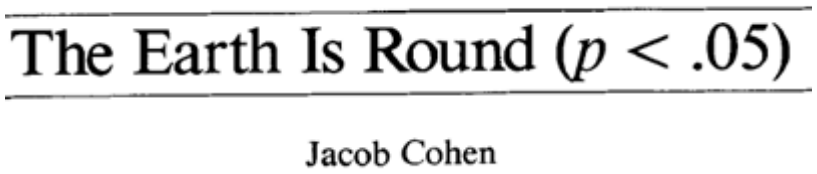
The American Statistical Association has recommended that researchers try to avoid using the term "statistically significant" and instead just report the p-value.


THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose


The Earth Is Round ($p < .05$)

Jacob Cohen