# YData: An Introduction to Data Science

## Lecture 18: Decisions and Uncertainty

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

# Where are we?

- Topics this week: Hypothesis testing and causality
- Assignment 06 due on Thursday (start early!)
- Break day next Wednesday (no class or OH)
- Midterm next Friday
  - Practice midterm posted; another will follow
  - Review session on Monday (no new topics next week)
  - Exam available Friday during normal class time (no class)
- Assignment 07 posted Friday; due April 1 (no fooling...)

# Decisions and Uncertainty

# Reminder: Readings







**Decisions and Uncertainty**

We have seen several examples of assessing models that involve chance, by comparing observed data to the predictions made by the models. In all of our examples, there has been no doubt about whether the data were consistent with the model's predictions. The data were either very far away from the predictions, or very close to them.

But outcomes are not always so clear cut. How far is "far"? Exactly what does "close" mean? While these questions don't have universal answers, there are guidelines and conventions that you can follow. In this section we will describe some of them.

But first let us develop a general framework of decision making, into which all our examples will fit.

What we have developed while assessing models are some of the fundamental concepts of statistical tests of hypotheses. Using statistical tests as a way of making decisions is standard in many fields and has a standard terminology. Here is the sequence of the steps in most statistical tests, along with some terminology and examples. You will see that they are consistent with the sequence of steps we have used for assessing models.

**Step 1: The Hypotheses**

All statistical tests attempt to choose between two views of the world. Specifically, the choice is between two views about how the data were generated. These two views are

# Incomplete Information

- We are trying to choose between two views of the world, based on data in a sample.
- It is not always clear whether the data are consistent with one view or the other.
- Random samples can turn out quite extreme. It is unlikely, but possible.

# Terminology

# Testing Hypotheses

- A test chooses between two views of how data were generated
- The views are called **hypotheses**
- The test picks the hypothesis that is better supported by the observed data

# Null and Alternative

The method only works if we can simulate data under one of the hypotheses.

- **Null hypothesis**
  - A well defined chance model about how the data were generated
  - We can simulate data under the assumptions of this model – "under the null hypothesis"

- **Alternative hypothesis**
  - A different view about the origin of the data

# Test Statistic

- The statistic that we choose to compute, based on the simulated data, to decide between the two hypotheses

  Questions after choosing the statistic:

- What values of the statistic will make us lean towards the null hypothesis?

- What values will make us lean towards the alternative?
  - Preferably, the answer should be just "high" or just "low"

# Prediction Under the Null Hypothesis

- Simulate the test statistic under the null hypothesis; draw the histogram of the simulated values
- This displays the **empirical distribution of the statistic under the null hypothesis**
- It is a prediction about the statistic, made by the null hypothesis
  - It shows all the likely values of the statistic
  - Also how likely they are (if the null hypothesis is true)
- The probabilities are approximate, because we can't generate all the possible random samples

# Conclusion of the Test

Resolve choice between null and alternative hypotheses

- Compare the **observed test statistic** and its empirical distribution under the null hypothesis
- If the observed value is not consistent with the distribution, then the test favors the alternative – "reject the null hypothesis"

Evaluating consistency with respect to a distribution:

- A visualization may be sufficient
- If not, there are conventions about "consistency"

# Performing a Test

# The Problem

- Large Statistics class 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 other sections

# The GSI's Defense

**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.

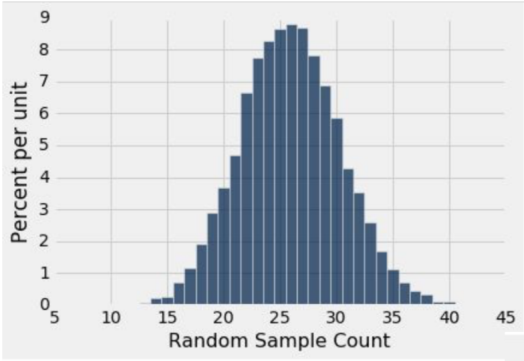
**Alternative:**

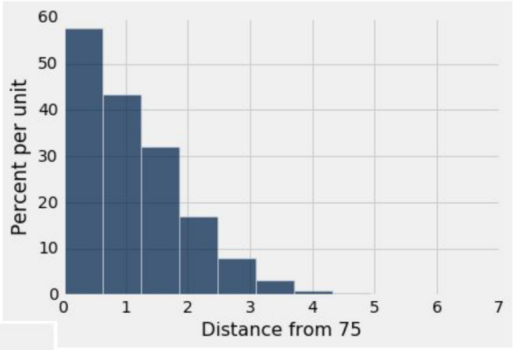
- No, the average score is too low. Randomness is not the only reason for the low scores.

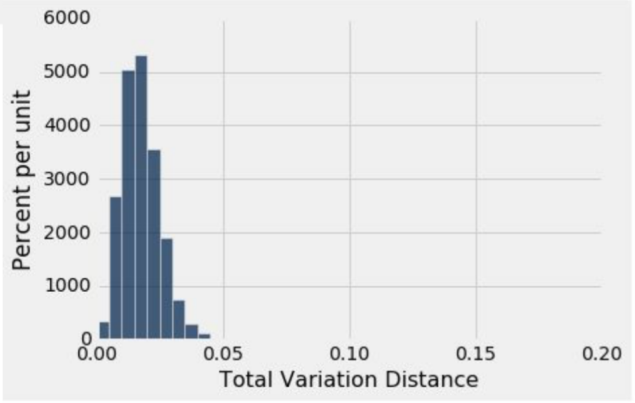
(DEMO)

# Statistical Significance

# Tail Areas


Percent per unit
Random Sample Count
0
1
2
3
4
5
6
7
8
9
5
10
15
20
25
30
35
40
45


Percent per unit
Distance from 75
0
10
20
30
40
50
60
0
1
2
3
4
5
6
7


Percent per unit
Total Variation Distance
0
1000
2000
3000
4000
5000
6000
0.00
0.05
0.10
0.15
0.20

# Conventions About Inconsistency

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

- **"In the tail," first convention**:
  - The area in the tail is less than 5%
  - The result is "statistically significant"

- **"In the tail," second convention**:
  - The area in the tail is less than 1%
  - The result is "highly statistically significant"

(DEMO)

# Definition of the P-value

Formal name: **observed significance level**

**The P-value is the chance, under the null hypothesis, that the test statistic is equal to the value that was observed in the data or is even further in the direction of the alternative.**

# An Error Probability

# Can the Conclusion be Wrong?

**Yes.**

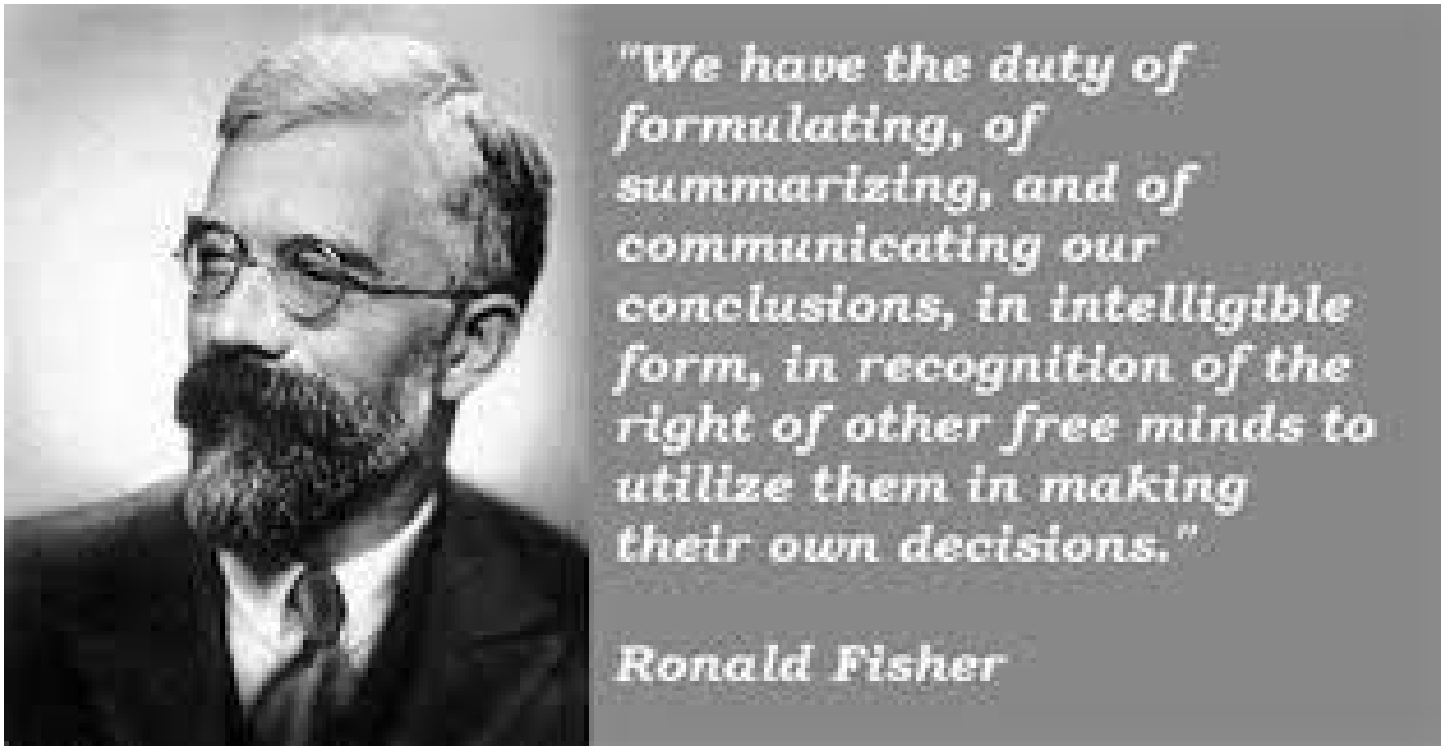
| | Null is true | Alternative is true |
|---|---|---|
| **Test rejects the null** | ✗ | ✓ |
| **Test doesn't reject the null** | ✓ | ✗ |

# An Error Probability

- The cutoff for the P-value is an error probability.
- If:
  - your **cutoff is 5%**
  - and the **null hypothesis happens to be true**
- then there is a **5% chance** that **your test will reject the null hypothesis**.

# Origin of the Conventions

# Sir Ronald Fisher, 1890-1962



# Sir Ronald Fisher, 1925

"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

# Sir Ronald Fisher, 1926

"If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the author prefers to set a low standard of significance at the 5 percent point ..."