



**190F**  
Fall 2018

# Foundations of Data Science

## Lecture 16

---

Decisions and Uncertainty

# **Announcements**

# **Review: Comparing Distributions**

# Summary of the Method

---

To assess whether a sample was drawn randomly from a known categorical distribution:

- Use TVD as the statistic because it measures the distance between categorical distributions
  - Sample at random from the population and compute the TVD from the random sample; repeat numerous times
  - Compare:
    - Empirical distribution of simulated TVDs
    - Actual TVD from the sample in the study
-

# **Decisions and Uncertainty**

# 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.
-

# Testing Hypotheses: Terminology

---

- 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 – simulating “under the null hypothesis”
  - **Alternative hypothesis**
    - A different view about the origin of the data
-



# Test Statistic

---

- The statistic that we choose to simulate, to decide between the two hypotheses

Questions before 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”. Try to avoid “both high and 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

---

To decide 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 – “rejects the null hypothesis”

Whether a value is consistent with a distribution:

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

# **Example: Exam Scores**

# The Problem

---

- A large class divided into 12 discussion sections
  - Teaching Assistants (TAs) lead the sections
  - After the midterm, students in Section 3 notice that the average score in their section is lower than in others
-

# The TA's Defense

---

## TA's position (Null Hypothesis):

- If we had picked the students in 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**

# Conventions About Inconsistency

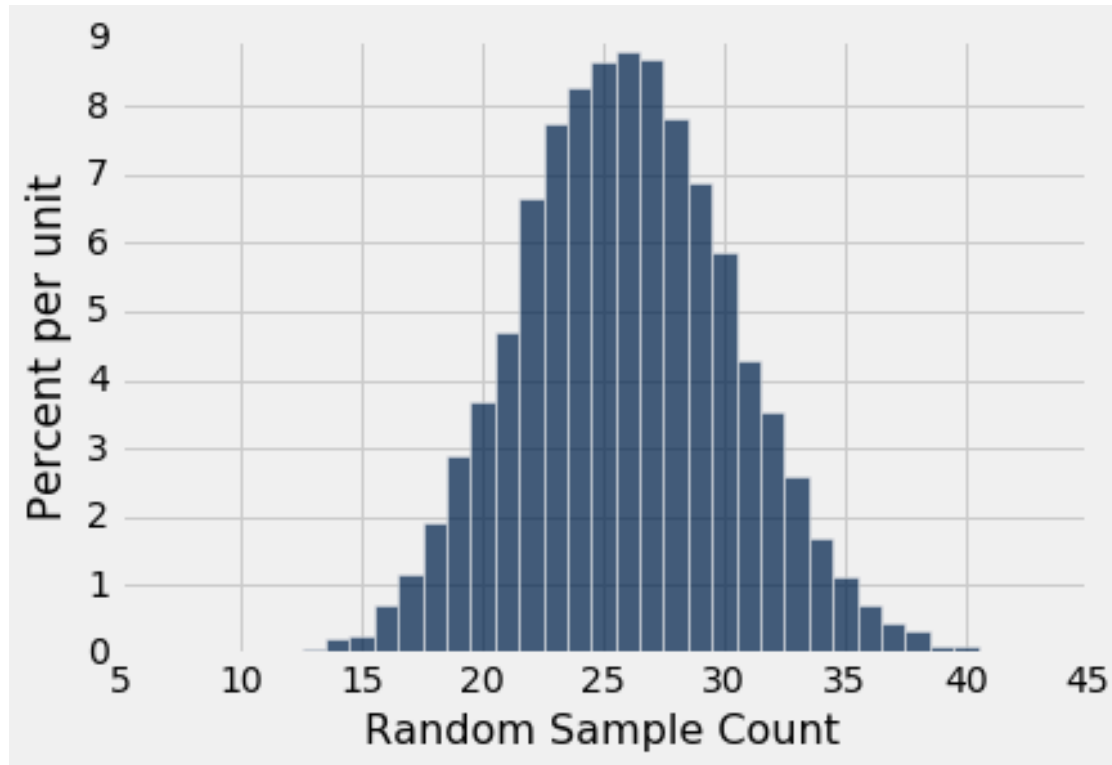
---

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



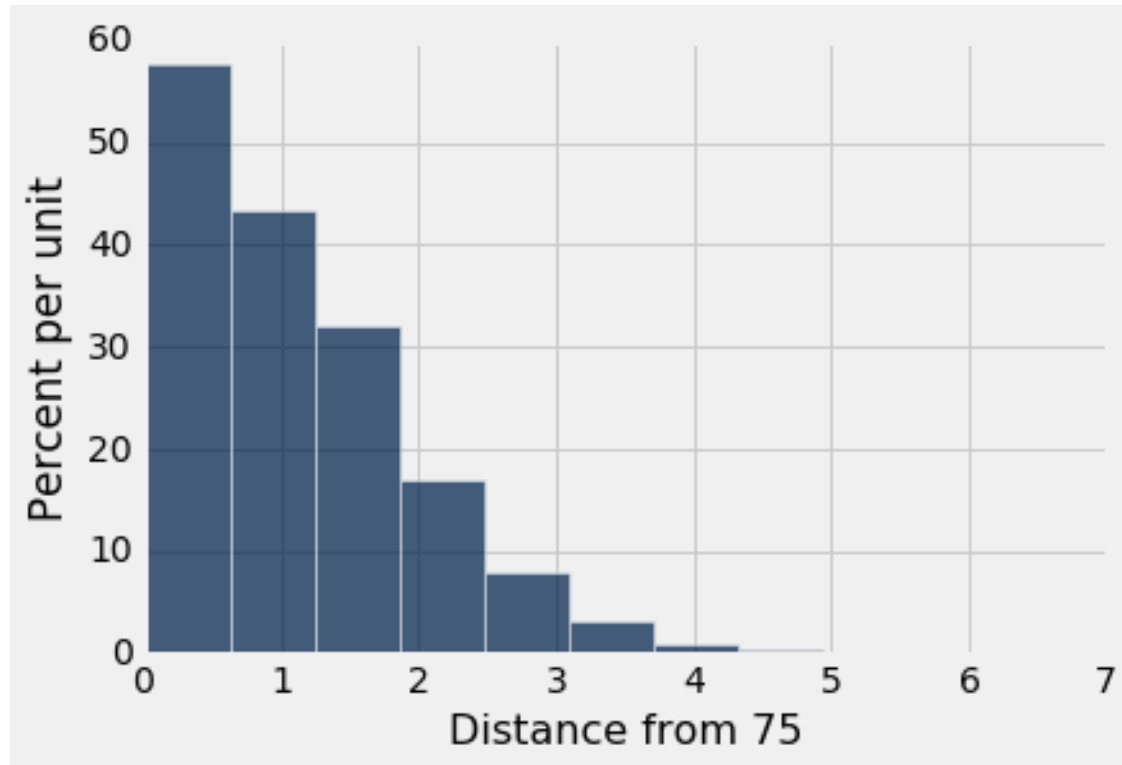
# Tail Areas

---



# Tail Areas

---



# 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”
-

# 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.
-

# **Error Probability of a Test**

# 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 about a 5% chance that your test will reject the null hypothesis.**
-

# **Origin of the Conventions**



# Sir Ronald Fisher, 1890-1962

---



*"We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions."*

*Ronald Fisher*

---

# 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 ...”

---