CompSci 116: Hypotheses & Errors

Jeff Forbes

February 26, 2019

How do we assess a model?

- Simulate data according to the assumptions of the model
  - Learn what the model predicts.

- Compare the predictions to the data that were observed.

- If the data and the model's predictions are not consistent, that is evidence against the model.

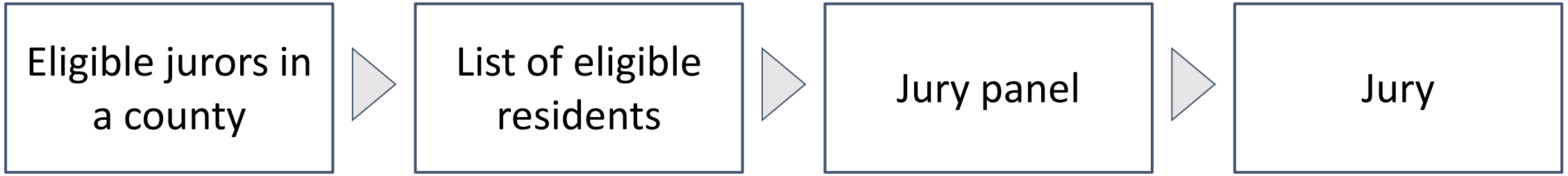
Robert Swain v. Alabama

1965 Supreme Court case about jury selection

- In Talladega, Alabama, 26% of residents were black
- In Swain's jury panel, 8 of 100 panelists were black
- All 8 were struck from the jury by the prosecution (using peremptory challenges)

Ruling: "The overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of [black men]."

Jury Panels



Section 197 of California's Code of Civil Procedure: All persons selected for jury service shall be selected at random, from a source or sources inclusive of a representative cross section of the population of the area served by the court.

Sixth Amendment to the US Constitution: … 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.

Sampling from a Distribution

- Sample at random from a categorical distribution

```
sample_proportions(sample_size, pop_distribution)
```

- Samples at random from the population
  - Returns an array containing the distribution of the categories in the sample

Terminology

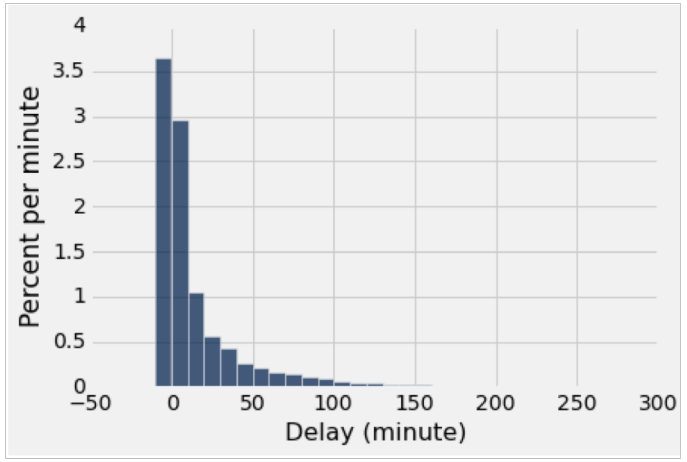
- **Population:** A collection of individuals
  - All United flights out of SFO in Summer 2015

- **Variable:** Something that varies in the population
  - airline *(categorical variable)*
  - amount of delay in departure *(quantitative variable)*

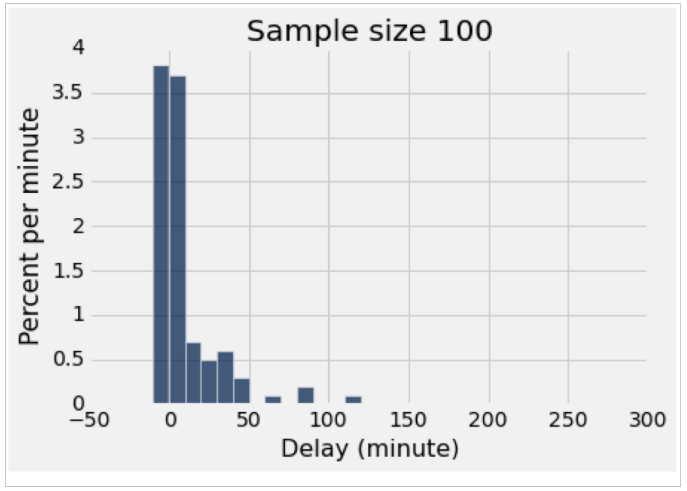
- **Sample:** A subset of the population

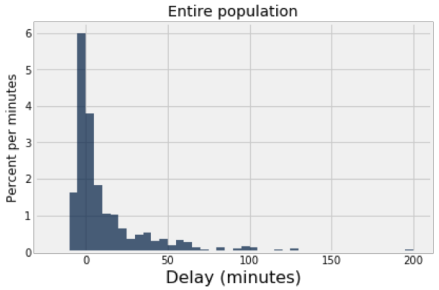
Two distributions

distribution of the population



empirical distribution of a sample

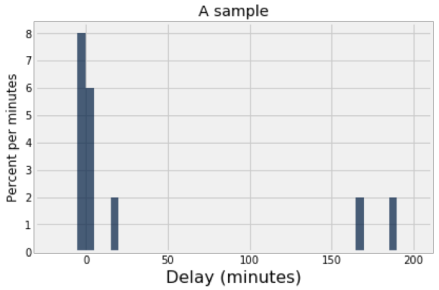


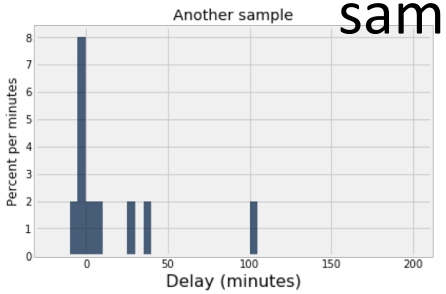


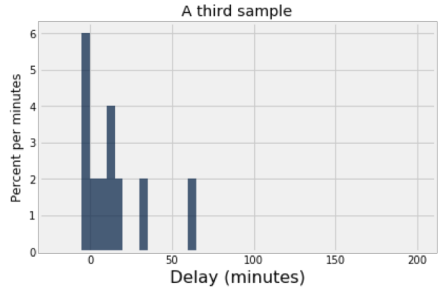
randomly sample

randomly sample

randomly sample



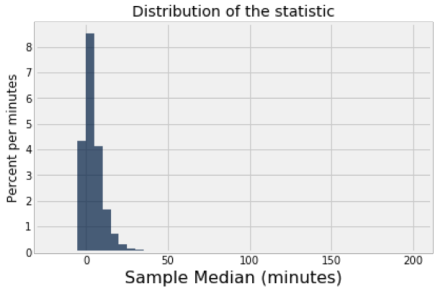




median = 2.0

median = -0.5

median = 9.5



Gary's Game

- Flip a fair (?) coin 10 times
  - **`if`** number of heads >= 5, we win
  - **`else`** Gary wins

- Play the game once
  - There's two heads
  - Was the game rigged?

Testing a Hypothesis

Step 1: Select Two Hypotheses

- A test chooses between two views of how data were generated: *Null hypothesis* proposes that data were generated at random; *Alternative hypothesis* proposes some effect other than chance

Step 2: Choose a Test Statistic

- A value that can be computed from the data

Step 3: Compute What The Null Hypothesis Predicts

- Compute the distribution of the test statistic: what the test statistic might be if the null hypothesis were true.

Step 4: Compare the Prediction to the Observed Data

Step 1: Smoking and Babies

Researchers are interested in whether there is an association between smoking mothers and the health of their babies. For each birth, they record the baby's birth weight and whether the mother smokes or not.

- **Birth weights aren't affected by maternal smoking.** The birthweight distribution for babies of smokers is same as that of babies of non-smokers.

Null

- They are affected.

Alternative

Example: Smoking and Babies

Researchers are interested in whether there is an association between smoking mothers and the health of their babies. For each birth, they record the baby's birth weight and whether the mother smokes or not.

- **Birth weights aren't affected by maternal smoking.** The birthweight distribution for babies of smokers is same as that of babies of non-smokers.
- **They are lower.** Birthweight of babies of smokers are lower than birthweights of babies of non-smokers.

Step 2: Choosing a Test Statistic

Test statistic: The statistic that you have chosen to calculate, to help you decide between the two hypotheses

Goal: If the null hypothesis is false, then you expect that measuring the test statistic will allow you to reject the null

Choosing a Test

If the alternative hypothesis is true, will test statistic be *larger* than prediction, *smaller*, or *could be either way*?

For the baby birth weights e
would be reasonable test st
(Choose all that are OK.)

1. The average birth weight of all the babies.
2. The proportion of babies whose mother smoked.
3. The average birth weight of babies of smokers, minus the average birth weight of babies of non-smokers.
4. The absolute value of the previous difference.

Absolute Values & Alternatives

- Choose a test statistic where alternative hypothesis suggests which direction statistic will go.
- Alternative: Smoking causes poor health.
  - Test statistic: Average birth weight for smokers, minus average for non-smokers.
- Alternative: Smoking has some relation to health.
  - Test statistic: Absolute value of that difference.

Conclusion of a Test

Resolve choice between null and alternative hypotheses

- Compare observed test statistic to its empirical distribution under the null hypothesis
- If the observed value is **consistent** with the distribution, then the test *does not* reject the null hypothesis

Whether a value is consistent with a distribution:

- A visualization may be sufficient
- Convention: The observed significance level (P-value)

Quantifying Conclusions

Step 0: Go find some data. These are the *observations.*

Step 1: Two descriptions of the world:

- Null: Data come from a well-defined random process
- Alternative: Something else is going on

We evaluate how unusual the data would be under the null

Step 2: Choose a test statistic to summarize the data.

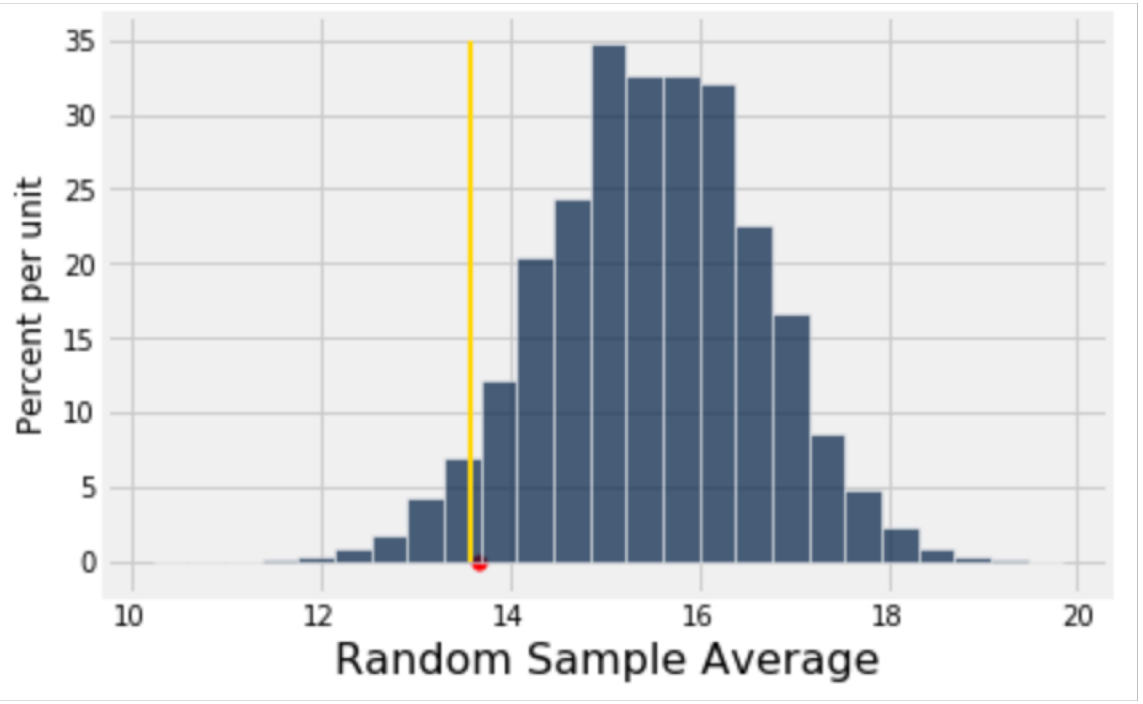
Step 3: Compute the following probability (p-value)

P(the test statistic would be equal to or more extreme than the observed test statistic under the null hypothesis)

Definition of *P*-value

The P-value is the chance,

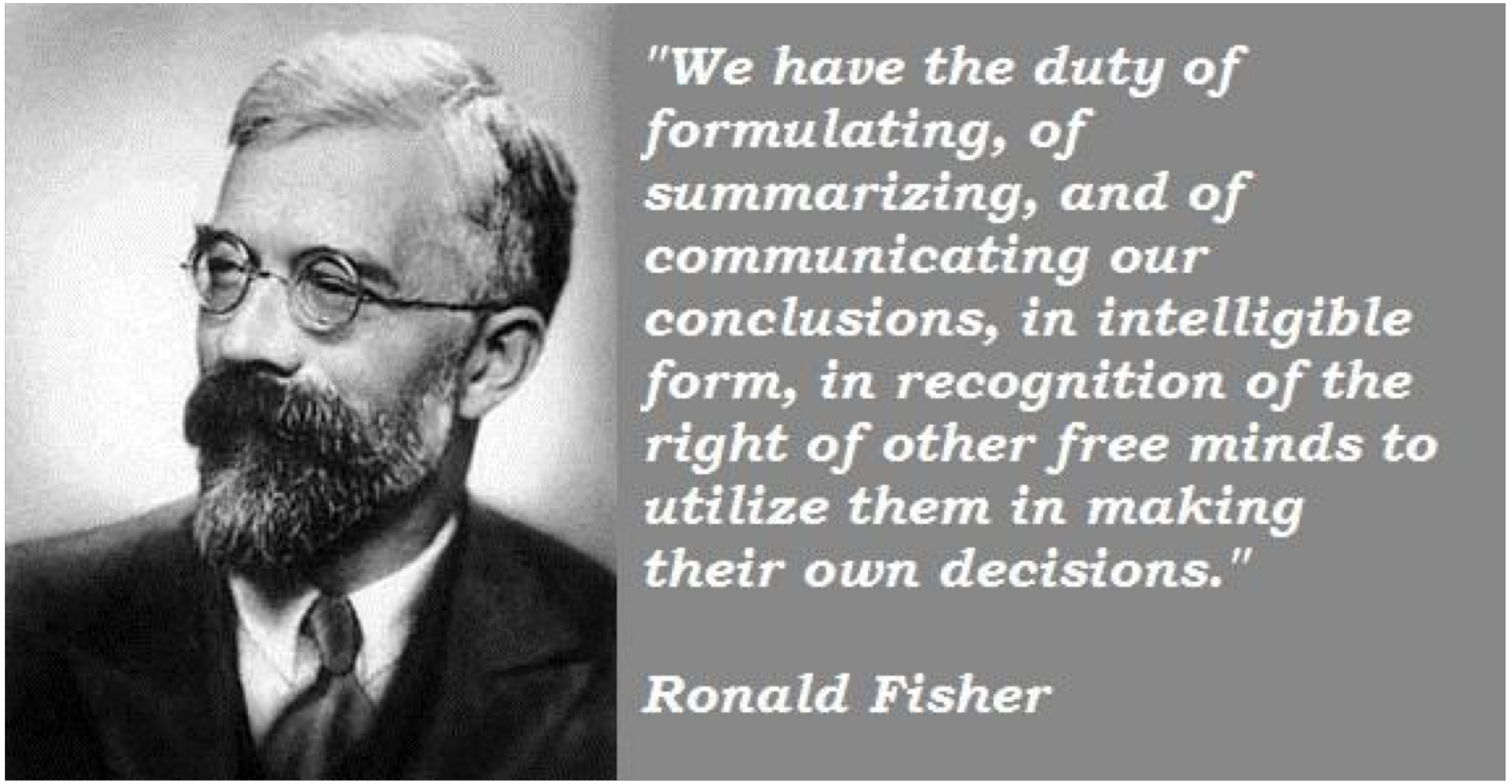
- under the null hypothesis,
- that the test statistic
- is equal to the value that was observed in the data or is even further in the direction of the alternative.



Conventions of Consistency

- **"Inconsistent":** The test statistic is in the tail of the null distribution.
- **"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."

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 …”

Sir Ronald Fisher, 1935

“No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon.”

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
  - (but you don't know that)
- then there is about a 5% chance that your test will reject the null hypothesis anyway.

(Demo)