Statistics, Simulation And Hypothesis Testing

Logistics

- Midterm makeups due now
- Conditional probability worksheet due Thursday

Diagnostic Problem from Last Week

Will post an update online -- short answer is the slides were actually OK

Statistics

Why sample?

Probability

Why sample?

Probability

Statistics

Why sample?

Probability

Statistics

Sampling

Estimation

Statistical Inference:

Making conclusions based on data in random samples

Estimation

Statistical Inference:

Making conclusions based on data in random samples

Example:

Use the data to guess the value of an unknown number

Estimation

Statistical Inference:

Making conclusions based on data in random samples

Example:

Use the data to guess the value of an unknown number

Create an **estimate** of the unknown quantity

Estimation

Statistical Inference:

Making conclusions based on data in random samples

Example:

Use the data to guess the value of an unknown number

fixed

Create an **estimate** of the unknown quantity

Estimation

Statistical Inference:

Making conclusions based on data in random samples

Example:

Use the data to guess the value of an unknown number

fixed

depends on the random sample

Create an **estimate** of the unknown quantity

Terminology

Parameter

A number associated with the population

Terminology

Parameter

A number associated with the population

Statistic

A number calculated from the sample

Terminology

Parameter

A number associated with the population

Statistic

A number calculated from the sample

A statistic can be used as an **estimate** of a parameter

(Demo)

Estimation

How many enemy planes?

Assumptions

- Planes have serial numbers 1, 2, 3, …, N.
- We don’t know N.
- We would like to estimate N based on the serial numbers of the planes that we see.

Assumptions

- Planes have serial numbers 1, 2, 3, …, N.
- We don’t know N.
- We would like to estimate N based on the serial numbers of the planes that we see.

The main assumption

- The serial numbers of the planes that we see are a uniform random sample drawn with replacement from 1, 2, 3, …, N.

Discussion question

If you saw these serial numbers, what would be your estimate of N?

```
170   271   285   290   48
235    24    90   291   19
```

Discussion question

If you saw these serial numbers, what would be your estimate of N?

```
170   271   285   290   48
235    24    90   291   19
```

One idea: 291. Just go with the largest one.

The largest number observed

- Is it likely to be close to N?
  - How likely?
  - How close?

The largest number observed

- Is it likely to be close to N?
  - How likely?
  - How close?

Option 1. We could try to calculate the probabilities and draw a probability histogram.

The largest number observed

- Is it likely to be close to N?
  - How likely?
  - How close?

Option 1. We could try to calculate the probabilities and draw a probability histogram.

Option 2. We could simulate and draw an empirical histogram.

(Demo)

Verdict on the estimate

- The largest serial number observed is likely to be close to N.

Verdict on the estimate

- The largest serial number observed is likely to be close to N.
- But it is also likely to underestimate N.

Verdict on the estimate

- The largest serial number observed is likely to be close to N.
- But it is also likely to underestimate N.

Another idea for an estimate:

Average of the serial numbers observed ~ N/2

Verdict on the estimate

- The largest serial number observed is likely to be close to N.
- But it is also likely to underestimate N.

Another idea for an estimate:

Average of the serial numbers observed ~ N/2

New estimate: 2 times the average

(Demo)

Bias & Variance

Bias

- **Biased estimate:** On average across all possible samples, the estimate is either too high or too low.

Bias

- **Biased estimate:** On average across all possible samples, the estimate is either too high or too low.
- Bias creates a systematic error in one direction.

Bias

- **Biased estimate:** On average across all possible samples, the estimate is either too high or too low.
- Bias creates a systematic error in one direction.
- Good estimates typically have low bias.

Variability

- The degree to which the value of an estimate **varies** from one sample to another.

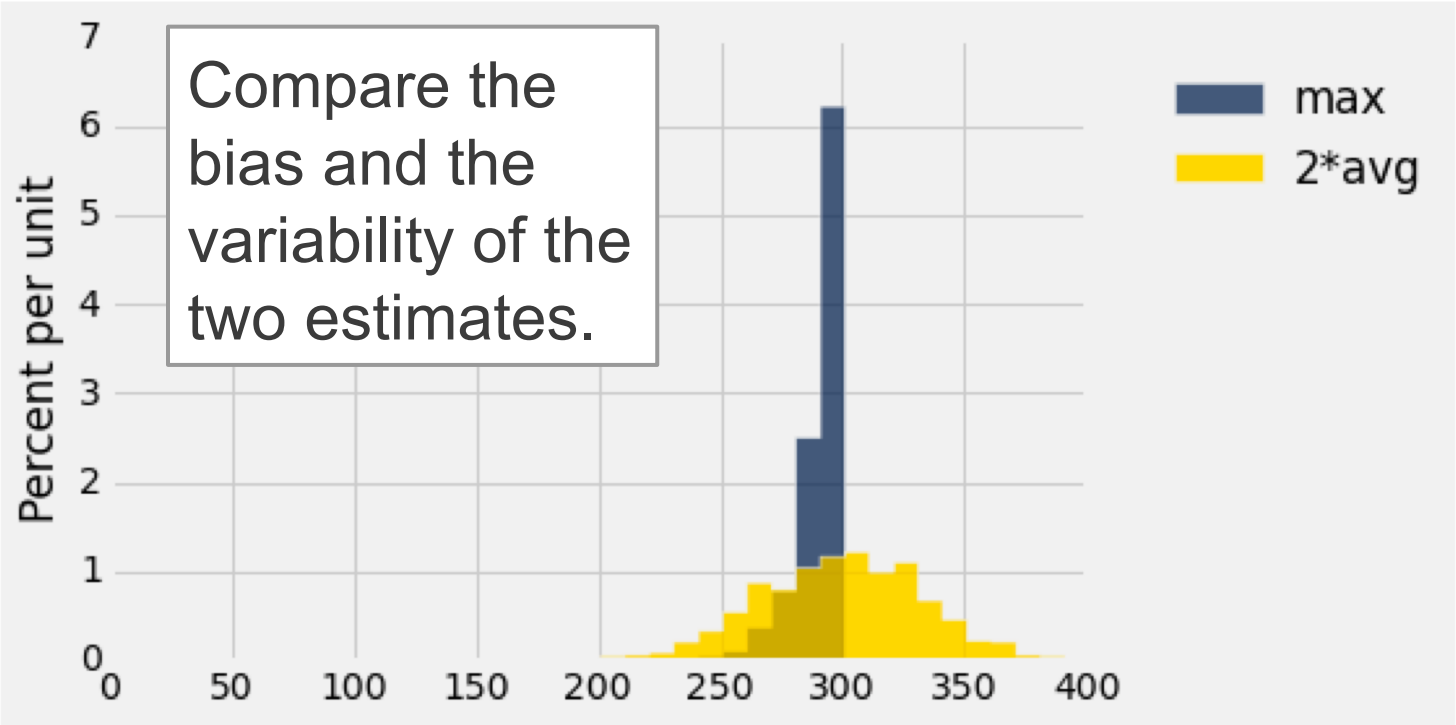
Variability

- The degree to which the value of an estimate **varies** from one sample to another.
- High variability makes it hard to estimate accurately.

Variability

- The degree to which the value of an estimate **varies** from one sample to another.
- High variability makes it hard to estimate accurately.
- Good estimates typically have low variability.

Discussion question



Bias-variance trade-off

- The **max** has low variability, but it is biased.
- **2*average** has little bias, but it is highly variable.
- Life is tough.

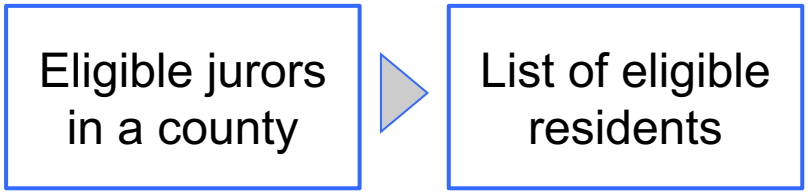
(Demo)

Comparing Samples to Distributions

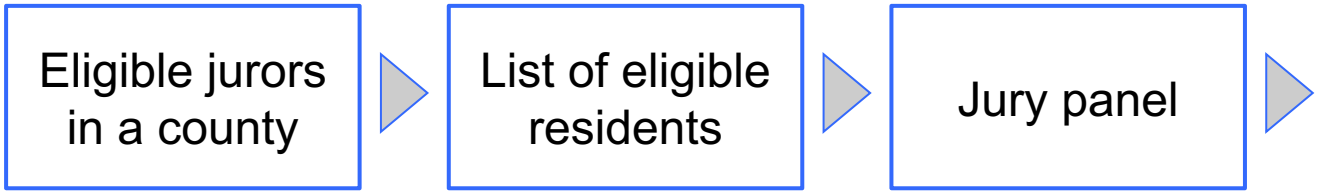
Jury Panels

Eligible jurors in a county

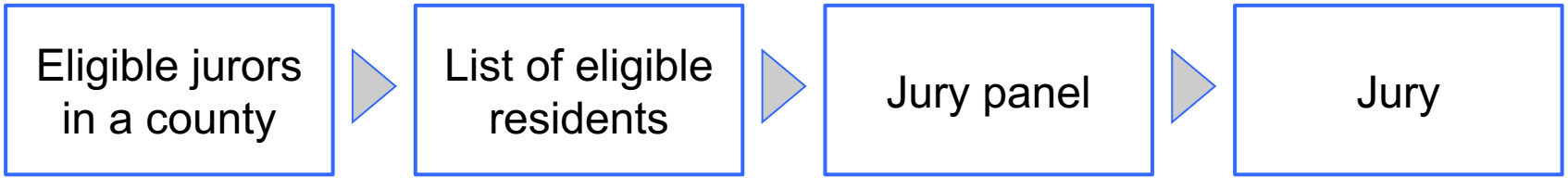
Jury Panels



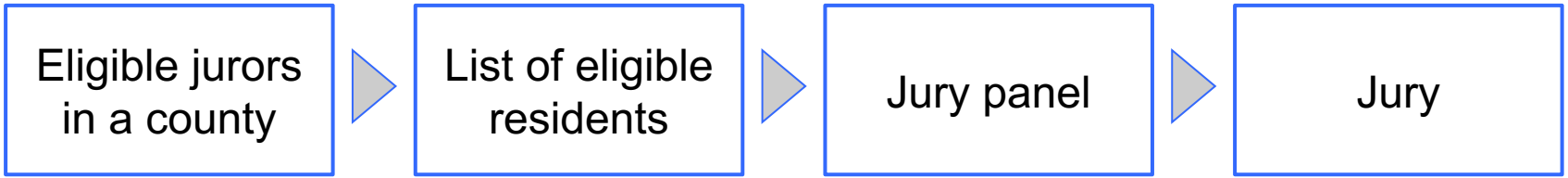
Jury Panels


Eligible jurors in a county
List of eligible residents
Jury panel

Jury Panels


Eligible jurors in a county
List of eligible residents
Jury panel
Jury

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.

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)

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]."

Is the actual jury panel likely?

(Demo)

Hypothesis Testing

Testing a Hypothesis

Step 1: The Hypotheses

- A test chooses between two views of how data were generated

Testing a Hypothesis

Step 1: The Hypotheses

- A test chooses between two views of how data were generated
- *Null hypothesis* proposes that data were generated at random

Testing a Hypothesis

Step 1: The 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

Testing a Hypothesis

Step 1: The 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: The Test Statistic

- A value that can be computed for the data and for samples

Testing a Hypothesis

Step 1: The 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: The Test Statistic

- A value that can be computed for the data and for samples

Step 3: The Sampling Distribution of the Test Statistic

- What the test statistic might be if the null hypothesis were true

Testing a Hypothesis

Step 1: The 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: The Test Statistic

- A value that can be computed for the data and for samples

Step 3: The Sampling Distribution of the Test Statistic

- What the test statistic might be if the null hypothesis were true
- Approximate the sampling distribution by an empirical distribution

Jury Hypothesis Test

Step 1: The Hypotheses

- *Null hypothesis:* The panel of jurors was selected at random from the eligible juror population. Any differences between the ethnicity distributions of panel and population are due to randomness.
- *Alternative hypothesis*: The panel was not selected at random.

Step 2: The Test Statistic

- TVD between a panel and the population distribution

Step 3: The Sampling Distribution of the Test Statistic

- Simulate drawing panels from population. See if the actual panel TVD is likely.