

**DSFA**  
Spring 2021

# Lecture 16

---

Estimation

# Announcements

---

- Project 1, part 1 deadline now Sunday 3/21, midnight
  - Project 1, part 2 deadline Friday 3/26, 5:59PM
  - Prelim 1 grading in progress
-

**Does sample look like  
population?**

(Demo)

# Large Random Samples

---

If the sample size is large,

then the **empirical distribution** of a **uniform random** sample

resembles the **population distribution**,

**with high probability.**

---

# Distribution

---

- A **distribution** is a description of the likelihood of *events*
- **Empirical** distribution:
  - Experimental: made from observations
  - Proportion of each event in sample

vs.

- **Probability** distribution:
    - Theoretical: made from mathematics
    - Probability of each event
-

# Law of Large Numbers

---

If an experiment is repeated many times, independently and under the same conditions, then the proportion of times that an event occurs gets closer to the theoretical probability of the event

Sometimes called *Law of Averages*

---

# Terminology

---

## Statistic

A number associated with the sample

## Parameter

A number associated with the population

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

---

# How many enemy planes?

---





# Estimating enemy planes

---

- Population: planes with serial numbers  $1, 2, 3, \dots, N$ .
- Parameter:  $N$ , which we don't know
- Sample: planes spotted by our troops
- Statistic: ???

**Assumption:** The serial numbers of the planes that are spotted are a uniform random sample drawn with replacement from  $1, 2, 3, \dots, N$ .

---

# 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

**One idea:** 291. Just go with the maximum.

(Demo)

---

# Is max a good estimator?

---

Is it likely to be close to  $N$ ?

- How likely?
- How close?

**Option 1.** Calculate the probabilities and draw a *probability histogram*.

**Option 2.** Simulate and draw an *empirical histogram*.

(Demo)

---

When poll is active, respond at [pollev.com/dsfa](https://pollev.com/dsfa)

Text **DSFA** to **22333** once to join

# What would happen if we changed the sample size to 10 instead of 30?

The histogram would spread out, with more area to the left of 300.

The histogram would spread out, with more area to the right of 300.

The histogram would compress, with more area concentrated around 300.

Not predictable

None of the above



# Verdict on max

---

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

# New idea

---

- Maybe the average of the sample resembles the average of the population
- Average of population is about  $N/2$

**New statistic:**  $2 * \text{average}(\text{samples})$

(Demo)

---

# 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 estimators have low bias
-

# Variance

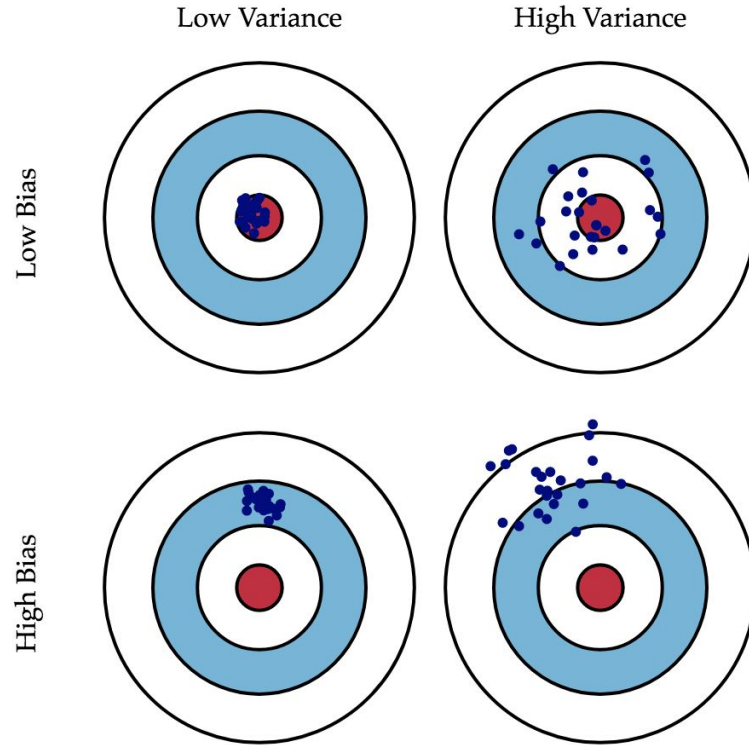
---

- Value of an estimate **varies** from one sample to another
  - High variability makes it hard to estimate accurately
  - Good estimators have low variance
-



# Bias vs Variance

---



# Bias-Variance Tradeoff

---

- **max** has low variability, but is biased
  - **2\*average** has little bias, but is highly variable
  - Life is tough!
-