DSFA

Spring 2020

Lecture 13

Estimation

Project 1 due 3/6 at 5:59pm

Distribution

- A distribution is a description of the likelihood of *events or outcomes*

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

- Empirical distribution:
  - Experimental: made from observations
  - Proportion of each event in sample

Large Random Samples

If the sample size is large,

then the empirical distribution of a simple random sample

resembles the population distribution,

with high probability.

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

Parameter

A number associated with the population

Statistic

A number associated with the sample

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, …, 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, …, 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)

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 * average(samples)

(Demo)

Bias

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

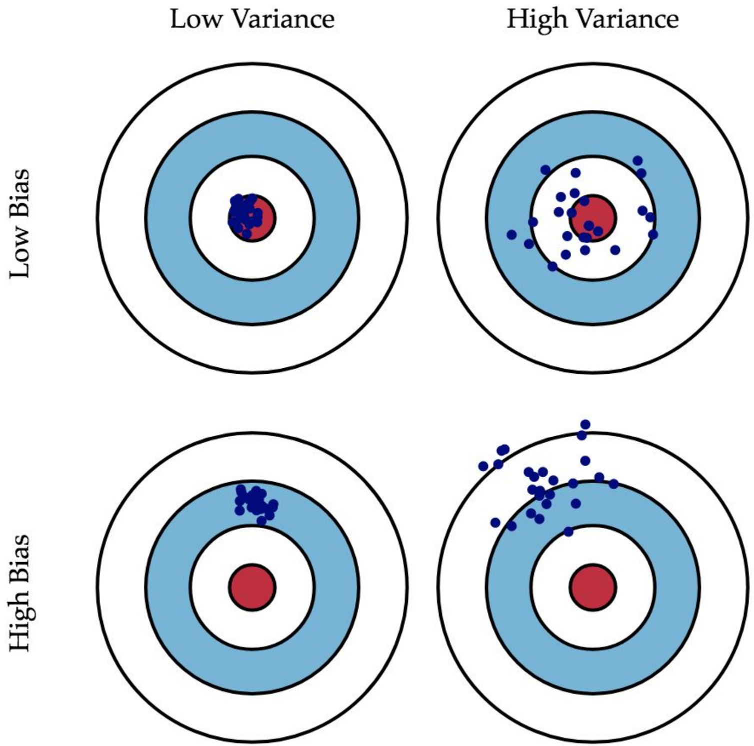
- Bias creates a systematic error in one direction

- 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



http://scott.fortmann-roe.com/docs/BiasVariance.html

Bias-Variance Tradeoff

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