Empirical Distribution of a Statistic
Announcements

- Project is due 5 pm Tuesday Oct 4.
- Homework tonight!
- Midterm is on Friday Oct 14, two weeks away. No computers or calculators on the midterm.
- No alternate dates for the midterm.
Empirical Distribution of a Sample

If the sample size is large,

then the empirical distribution of a random sample

resembles the distribution of the population,

with high probability.
Roulette (Demo)
**Terminology**

- **Parameter**
  - A number associated with the population

- **Statistic**
  - A number calculated from the sample

- Sometimes, a statistic can be used as an *estimate* of a parameter.

(Demo)
Simulating a Statistic

Fix a sample size and choose your statistic.

1. Simulate the statistic once:
   a. Draw a random sample of the size you fixed.
   b. Calculate the statistic and keep a record of the value.

2. Repeat Step 1 numerous times (as many times as you have patience for; thousands are good).

3. You now have one value of the statistic for each repetition. Visualize the results.
How many enemy warplanes?
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

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?

**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 $N$.
- But it is also likely to underestimate $N$.

Another idea for an estimate:
Average of the serial numbers observed $\sim \frac{N}{2}$

New estimate: 2 times the average

(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 estimates typically have low bias.
Variability

- 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.
Bias-Variance Tradeoff

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