DSC 10, Spring 2018
Lecture 14

Statistics

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Review:
Distributions and Sampling

Probability Distribution

- Random quantity with various possible values
- “Probability distribution”:
  - All the possible values of the quantity
  - The probability of each of those values
- In some cases, the probability distribution can be worked out mathematically without ever generating (or simulating) the random quantity

Empirical Distribution

- Based on observations
- Observations can be from repetitions of an experiment
- “Empirical Distribution”
  - All observed values
  - The proportion of counts of each value

Law of Averages

If a chance experiment is repeated

- many times,
- independently,
- under the same conditions,

then the proportion of times that an event occurs
gets closer to the theoretical probability of the event.

Ex. As you roll a die repeatedly, the proportion of times you roll a 5 gets closer to $\frac{1}{6}$.

Large Random Samples

If the sample size is large,

then the empirical distribution of a uniform random sample

matches the distribution of the population,

with high probability.

(Demo)

At Least One Six

If you roll a die 4 times, what is the probability of getting at least one 6?

A. $\frac{5}{6}$
B. $1 - \frac{5}{6}$
C. $1 - (\frac{5}{6})^4$
D. $1 - (\frac{1}{6})^4$
E. None of the above.

What's the general formula, if you roll a die *n* times?

Statistics

Why sample?

Probability

Statistics

Sampling

Estimation

Statistical Inference:

Making conclusions based on data in random samples

Example:

fixed

Use the data to guess the value of an unknown number

depends on the random sample

Create an **estimate** of the unknown quantity

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)

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.

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

A: 291

B: 350

C: 470

D: Not enough information

E: Different guess

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 ~ 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 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.
- High variability makes it hard to estimate accurately.
- Good estimates typically have low variability.

Bias-variance trade-off

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

(Demo)