statistics, sampling, simulation and R

as a data scientist you might be asked

- (at an insurance company) what's the largest claim we might expect this year?

- (in advertising) I just a/b tested these 2 creative campaigns, is my new one better?

- (sports data scientist) historical 43% 3pt shooter has 36% 3pt % in first month of the season, is something wrong?

what is a discrete random variable?

e.g.: Rolling a die

discrete random variable is one which may take on only a countable number of distinct values such as 0,1,2,3,4,...,infinity
We use discrete random variables in machine learning to classify an image.

You can sum up the probabilities

what is a continuous random variable?

e.g.: normal (Gaussian) distribution, uniform distribution

... where the data can take infinitely many values. For example, a random variable measuring the time taken for something to be done is continuous since there are an infinite number of possible times that can be taken.

You cannot sum the probabilities, but you can use integral.

what is a normal random variable?

The number of cars in a parking lot, the average daily rainfall in inches, the number of defective tires in a production line, and the weight in kilograms of an African elephant cub are all examples of quantitative variables.

central limit theorem

basketball shooter has z 3pt % in a month of the season
(50 attempts)

then z will be approximately Normal (Gaussian)

**THE SAMPLE MEAN APPROACHES
A NORMAL DISTRIBUTION AS
THE SAMPLE SIZE GETS LARGE**

hypothesis testing in R

Evals last semester:

y1=c(2,2,4,5,5,4,4,3,5)

This semester:

y2=c(2,4,4,5,3,2,4,3,4,2)

t-distribution!

coffee or tea?

testing p=.5 using simulation

- sample many binom(N,p) random variables, how many are as extreme as our class split on coffee or tea?

- could also use a proportions test

bootstrap

- related to simulation
- create “copies” of your data set by sampling with replacement
- calculate statistic (or model) on each copy and look at the variation
- allows you to calculate confidence intervals, do hypothesis testing, or create better ML models

bootstrap example

Evals last semester:
y1=c(2,2,4,5,5,4,4,3,5)

This semester:
y2=c(2,4,4,5,3,2,4,3,4,2)

- calculate 90% CI for y1
- test difference of means

in class exercise

- download loss_premium.csv from canvas
- pretend you are a data scientist at State Farm and you have last years loss/premium results. The CEO wants to know how the company is doing, and the standard metric is loss ratio (loss/premium). 70% is good, over 100% is very bad. Use bootstrap to give the CEO a likely range of loss ratios.