Stats + Homework 2 Review

CS100 TAs
What’s on Homework 2?

- Confidence/Confidence intervals (mean, proportion, difference of each [all Z based])
- CLT, LOLN
- Some hypothesis testing (p-values)
- Statistical significance
- Sampling distributions, binomial and normal distributions, standard normal distribution
The Normal Distribution

- The Normal Distribution is a continuous distribution that’s also called a “Bell Curve”.
- What does the area under curve sum to?
The Normal Distribution

- Parameters:
  - $\mu$ (Mu, the mean)
  - $\sigma$ (Sigma, the standard deviation)
The Normal Distribution

- Standard deviation rule (empirical rule): This is also called the 68-95-99.7 rule.
- It represents the percentages that lie within one, two or three standard deviations from the mean.
Sampling distributions

- Sampling distributions are distributions of statistics taken after repeated sampling (of the same experiment).
- It’s often used for mean, proportions, etc.
- For example, in studio you created the sampling distribution of the mean number of heads for a certain number of fair coin flips.
Central Limit Theorem

- CLT states that if you have a sufficient number of samples (>30), the sampling distribution will be approximately normal.
Law of Large Numbers

- LOLN - as the number of trials in an experiment increases, the observed sample statistic approaches the expected value of the sample statistic
- If you flipped a coin 3 times and got all heads, is the coin fair? Maybe!
- If you flip a coin 100,000 times and you get 80,000 heads, is the coin fair?
- The LOLN states that we should get about ~50% heads with a fair coin
Confidence Intervals

- Confidence interval: given a confidence level, the interval in which that percentage of observed sample statistics should fall.
- If we find the range in which 95% of the measured statistics fall, we have a 95% confidence interval.
Confidence Intervals

- Example: 73% of Brown seniors surveyed admitted stealing food from a campus eatery. How accurate is this?
- The Brown Daily Herald surveyed 609 seniors. We’re going to assume this was a random sample.

http://www.browndailyherald.com/2016/05/26/senior-survey-2016/
Confidence Intervals

1. Get the variables:
   a. alpha = 1 - 0.95 = 0.05.
   b. Z score: look at table for alpha.
   c. Standard deviation ($\sigma$) How to calculate? $\sigma = \sqrt{p*(1-p)}$
   d. Sample size (n)

Wouldn’t it be nice if we could write a function to do this…? What parameters do you think you would need? What would your function return or print?
Confidence Intervals

- Step 1: Find the variables
  - For a 95% confidence interval, we give you the z-score: 1.96
  - \( p \) is the probability Brown students steal from dining halls (73%).
  - \( n = 609 \)
- Step 2: Calculate! 
  \[ \hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]
  \[ 1.96 \times \sqrt{0.73 \times 0.27 / 609} = 0.035 \]
- Upper bound: \( 0.73 + 0.035 = 0.765 \)
- Lower bound: \( 0.73 - 0.035 = 0.695 \)
- This means with 95% confidence we can say between 69.5%-76.5% Brown seniors have stolen food from the dining halls. Tsk tsk...
Margin of Error

- The margin of error expresses the random sampling error in a survey’s results
- Step 1: Get the variables
  - $p = 0.73$
  - $n = 609$
- $\sqrt{\frac{0.73 \times 0.27}{609}} \approx 0.018$
Helpful R Functions

- `dnorm(0)`: height of probability distribution at a point. The default mean is 0 and sd = 1. You can change this: `dnorm(x, mean=2.5, sd=0.1)`
- `pnorm(1)`: Given a number or list computes probability that a normally distributed number is less than that number. If you want to compute the probability that the number is above a given number, you can change it so `lower.tail=FALSE`. You can change the defaults similar to `dnorm`.
- `qnorm(0.25)`: Inverse of `pnorm`. You give it a probability, it returns the number whose cumulative distribution matches the probability. You can also change the default mean and sd: `qnorm(x, mean=3, sd=0.1)`. For example: `-1.0 * qnorm(0.05/2) = 1.96`
Other Helpful R Functions

- \( \text{sqrt}(4) = 2 \)
- \( \text{c(“Anna”, “Ben”, “Erin”) - creating a vector in R} \)
- \( \text{rbind(tas, Joon) - adding rows to a dataset} \)
- \( \text{newvalue <- drop[value]} \)
- \( \text{Sample - used to take a sample of a certain size from a vector} \)
- \( \text{select and filter (select to select columns of a dataset, filter to select rows based on a True/False statement)} \)
- \( \text{Append (adding values to a vector)} \)
- \( \text{Shuffled - used to shuffle a dataset randomly} \)
Hypothesis testing

- Testing a hypothesis (i.e., if we are flipping coins, and we get the proportion of heads to be 0.6, we might hypothesize that we have an unfair coin)
- We use the observed sample mean or proportion and compare it to the value we expect (in this case, 0.5)
Hypothesis testing

- Step 1: We find the z score (standardized to standard normal distribution)
- Step 2: Using the CLT (assuming sample is over 30), we know that the distribution is normal and we can find the p-value representing the probability that, given an expected value of 0.5 and an observed value of 0.6, we have a fair coin
- P-value - the probability, assuming that the coin is fair, that the observed sample was fair
Hypothesis testing

- What if we only have a small set of coin tosses to look at? So instead of 1000 coin tosses we have 10?
- We can use simulations like we did in Studio 6 but randomize the small dataset using shuffle.
Statistical significance

- At a certain confidence level, statistical significance occurs when confidence intervals do not overlap - you are 95% confident that there is a difference between the two statistics.
- If they do overlap, you don’t know that they aren’t statistically significant, you just can’t prove they are statistically significant.
- For hypothesis testing (ie. problem 3), statistical significance occurs when $p < 1$ - your confidence.