Problem Set 2

Due October 9th, 2020

1. Normal distributions (3 pt)

   Suppose that X and Y are iid random variables distributed $\mathcal{N}(\mu,\ \sigma^2)$.

- Find $\mathbb{E}[5X-2Y+8]$
- Find $\mathbb{V}[2X+4Y]$
- Find $Cov[5X,\ 4Y]$

2. Jury selection (10 pt)

Just prior to jury selection for a murder trial, a poll finds that about 10% of the jury eligible population thinks the suspect should be found innocent. Let's **assume** that this is a known population parameter and the true percentage of people who think the main suspect is innocent prior to jury selection. Also assume that the twelve jurors are selected randomly and independently from the population.

- Let's suppose 100 jurors were called in for jury duty. How many ways are there to pick twelve of them?
- Now use simulation to draw a sample of 20 juries *from the population*, and find the probability that the jury had at least one member who believes in the suspect's innocence prior to jury selection.
- Now repeat your simulation, drawing 1000 juries. Is your answer the same as before? Why or why not?
- Calculate this probability analytically in R and verify your answer
- What is the probability that the **entire** jury thinks the suspect is innocent? Calculate this quantity either analytically or computationally

3. Playing cards in `R` (12 pt)

- Let's play poker. First, create a "deck" of cards with 4 suits, each numbered 2-14 (we will treat aces as high cards only for this problem) using the following code:

```
deck <- data.frame(rep(2:14,4),
                   c(rep(1,13), rep(2,13),
                     rep(3,13), rep(4,13)))
colnames(deck) <- c("number", "suit")
```

- Then deal yourself 100,000 hands of 5 cards each. Show that a flush (all cards of the same suit) is rarer than a straight (all cards in a row, regardless of suit).
- BONUS (5pt) This time we're playing (a slightly modified version of) blackjack. In blackjack, you get two cards to start the game and can continue drawing additional cards until you decide to stop or the sum of your cards exceeds 21 (which is called "busting"). Dealers have more restricted rules: they must draw if the sum of their cards is less than 17 and must stop drawing cards if the sum is 17 or greater. In real blackjack, aces can be either 1s or 10s, but here we'll treat them as all 1s. In addition, all face cards are worth 10 points, so we'll need a new deck:

```
deck <- data.frame(rep(c(1:10,10,10,10),4),
                   c(rep(1,13), rep(2,13),
                     rep(3,13), rep(4,13)))
colnames(deck) <- c("number", "suit")
```

Now, do 10,000 simulations of a dealer's blackjack hand. Using a while loop inside your simulation, continue drawing cards until the sum of the cards exceeds 17. Find out what percent of the time the sum of dealer's cards exceeds 21.

4. Polity score and income (15 pt)

Download the `demo.csv` dataset from the `GitHub` repository. The dataset contains the population of countries in the year 1990 and includes the following variables:

This data set is drawn from the Logic of Political Survival dataset by Bueno de Mesquita et. al.

Table 1: Contents of demo.csv

Variable	Content
Country	Country name
Polity2	The country's polity score (like Freedom House scores, a measure of how democratic it is)
GDP	Real GDP per capita
Regime	A variable coded from the polity2 that codes whether a country is an autocracy (1), an anocracy, which is between autocracy and democracy (2), or a democracy (3)
Wealth	A variable coded from gdp that codes whether a country is poor (1), middle income (2), or rich (3)

- Produce an appropriately named and labeled scatterplot of `gdp` on `polity2`. Do you think there is a relationship between the two variables?
- Let X be a random variable for `regime` and Y be a random variable for `gdp`. Calculate the conditional expectations $\mathbb{E}(Y|X)$ for the three regime types. (Conditional expectation is just the random variable analogue of

conditional probability, the expectation of one thing given another). What do you notice? Does this confirm or contradict your answer to the previous part?

- Now calculate the conditional variances $\mathbb{V}(Y|X)$ for the three regime types. What do you notice? Briefly explain what might be going on, using whatever knowledge about the world you may have.
- Produce two histograms side by side. The first histogram should be the conditional distribution of `gdp` for autocracies and the second should be the conditional distribution of `gdp` for democracies. On each histogram, add a vertical line indicating the conditional means. Change the x-axes so that they are on the same scale for both plots in order to compare them. Do the conditional distributions look approximately normal?
- Remembering that the dataset represents the population of countries, let Z denote the wealth status of a country (poor, middle income, or rich). Give the joint probability (PMF) of X and Z by filling in Table 2 below with probabilities (2 decimal places is fine).

Table 2: Joint Probability of X and Z

Wealth	Regime Type		
	Autocracy	Anocracy	Democracy
Poor	?	?	?
Middle Income	?	?	?
Rich	?	?	?

- What is the joint probability that a country is middle income and an anocracy?
- What is the marginal probability that a country is an autocracy?
- What is the conditional probability that a country is democratic given that it is rich?
- Find the correlation and the covariance of `polity2` and `gdp` in R. What do they tell you about the two variables? Which measure do you prefer and why?

5. Central limit theorem (15 pt)

In the middle of an engrossing discussion about the normal distribution, your friend laments "my statistics professor claims that under repeated sampling, the distribution of sample means will resemble a normal distribution. But I have population data that doesn't look at all normal. If I repeatedly sample from the population, take the mean of those samples and plot the distribution

of those means, the distribution of sample means is just going to look like the distribution of my original data, not like a normal distribution." You decide to use your friend's data and see if she is right.

- Download the data set `population.csv` and load it into R. Provide graphical evidence that the population is not distributed normally.
- Next, simulate "repeated sampling" from the population. First, write a for loop which, in each iteration, draws a sample of size 5 from the population, calculates the sample mean of the draw, and stores the sample mean in one slot of a vector. Use 1000 iterations so that at the end of your loop you have a vector of length 1000 containing the sample means of 1000 samples drawn from the population.
  - Plot the sampling distribution of the sample mean.
  - Overlay a normal distribution centered at the mean of your sampling distribution with standard deviation equal to the standard deviation of your sampling distribution.
  - Comment on what your plot suggests about your friend's claim.
- Repeat the simulation from above, this time drawing samples of size 50. Again, compare the sampling distribution to a normal distribution and briefly comment on whether on how your result differs from before.
- What do these simulations have to do with the Central Limit Theorem?