Homework 6 (due Tueday, March 20th)

By signing my name, I agree to abide by the Stern Code of Conduct _____

Question 1

Suppose $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) with the same distribution as X. Let $S_n = \sum_{i=1}^n X_i$, so $\bar{X}_n = \frac{1}{n} S_n$, this shorthand will be useful for parts (b) through (e).

a. What do we need to know or assume about $E[X]$ to conclude that $\bar{X}_n \to \mu$ as $n \to \infty$?

Solution: Since $\bar{X}_n \to E[X]$ as $n \to \infty$, we need to assume $\mu = E[X]$.

b. For this part only, suppose X can only be -1 or 1. What are the lowest and highest possible values of S_n?

Solution: If every $X_i = -1$ then $S_n = -n$ is the lowest possible value. Similarly, $S_n = n$ is the highest possible value of S_n.

c. Suppose instead now that a and b are the lowest and highest possible values of X, so $a \leq X \leq b$. What are the lowest and highest possible values for S_n? And for $\bar{X}_n$?

Solution: If all $X_i = a$ then $S_n = na$ is the smallest possible value, and similarly nb is the largest. For $\bar{X}_n$ we divide the answers above by n and conclude $a \leq \bar{X}_n \leq b$.

d. For this part only, suppose $n = 99$, $\bar{X}_{99} = 0$, we are going to add one new observation X_{100} to the sample, and we know $-1 \leq X_{100} \leq 1$. What are the lowest and highest possible values for $\bar{X}_{100}$?

Solution: If $X_{100} = -1$ then $\bar{X}_{100} = -1/100$ is the lowest, and similarly $1/100$ is the highest possible value.

e. Continuing part (c) above, suppose we get one more observation X_{n+1} to add to the sample, and compute a new sample average $\bar{X}_{n+1}$. We can write the new average in terms of the old average and the new observation this way: $\bar{X}_{n+1} = \frac{n}{n+1}\bar{X}_n + \frac{1}{n+1}X_{n+1}$. What is the largest amount that $\bar{X}_{n+1}$ can increase above $\frac{n}{n+1}\bar{X}_n$?.

Solution: Since b is the largest possible value for X_{n+1}, the largest increase happens if $X_{n+1} = b$. In that case,

$$\bar{X}_{n+1} = \frac{n}{n+1}\bar{X}_n + \frac{b}{n+1}$$

So $b/(n+1)$ is the largest possible increase.

Question 2

a. At a casino there is a game where the player bets some amount of money X and then spins a wheel. If the wheel says they win, they get X dollars. Otherwise they lose and pay X dollars to the casino. I have the perfect gambling strategy to guarantee that I win this game. On the first round I bet X_1. If I win, the payoff is X_1 dollars. If I lose, I go "double or nothing" and bet $X_2 = 2X_1$ on the second game. I repeat this strategy, doubling my bet every time until I win. If I win on the first round, my payoff will be X_1. If I win on the nth round, my payoff will be $2^n X_1$. Aside from the practical fact that I may run out of money before I win, this sequence of payoffs/losses doesn't satisfy the assumptions of the law of large numbers. Which assumption(s) does it violate?

Solution: The law of large numbers applies to sequences of independent and identically distributed random variables. The sequence of payoffs/losses in this game are not identically distributed, since they double in size each time.

b. Your roommate suggests rolling a standard 6-sided die to decide who does the cleaning chores each day. If the die comes up 3 or less, you win, and your roommate does the chores. If it's higher than 3, you do the chores. Let D represent the dice, so $D \in \{1, \ldots, 6\}$, and let $X = 1$ if $D > 3$ and 0 otherwise. Let D_i, X_i represent the result on day i, so $X_1, \ldots, X_n$ are i.i.d. with the same distribution as X (and similarly for the D's). What is $E[X]$? The proportion of times that you do the chores in n days is $\bar{X}_n$. What will that proportion be in the long run, as $n \to \infty$? Does this system seem fair to you?

Solution: $E[X] = P(D > 3) = 3/6 = 1/2$. In the long run $\bar{X}_n \to 1/2$ so I do the chores half of the time. This seems fair.

c. Next, your roommate suggests a small change in the rules: keep track of the average of the dice $\bar{D}_n = \sum_{i=1}^{n} D_i$, and then you only need to do the chores whenever this average is higher than 3. Based on this system, in the long run what proportion of the time will you be doing the chores?

Solution: Since $E[D] = 3.5$, we know $\bar{D}_n \to 3.5$ as $n \to \infty$. This means that eventually I will have to do the chores every day.

d. Continuing part (b), suppose you have had a streak of bad luck and had to do the chores every day for a week. Your roommate reassures you, "Don't worry, if you do the chores a bunch of times then it becomes more likely for me to do them a bunch of times, since it has to even out in the long run." Do you agree? Why or why not?

Solution: I don't agree. The dice rolls are *independent*, so it does not become any more or less likely for my roommate to do the chores based on previous days.

Question 3

This question is about the normal approximation to the Binomial distribution. You will need to use the `R` functions `dbinom`, `pbinom`, and `pnorm` to answer some parts of it.

- Throughout this problem X refers to a Bin(n, p) random variable and Z is normal: $N(\mu, \sigma^2)$.
- Setting `size <-` n and `prob <-` p, the function `dbinom(x, size, prob)` computes $P(X = x)$, and the function `pbinom(x, size, prob)` computes $P(X \leq x)$.
- For Z, using `mean <-` μ and `sd <-` σ in `pnorm(z, mean, sd)` computes $P(Z \leq z)$.
- Finally, note that for a discrete random variable like the Binomial, `pbinom(b) - pbinom(a)` computes $P(a < X \leq b)$. The strict inequality doesn't matter for continuous random variables like the normal.

a. Let $n = 20$, $p = 0.4$. Compute $P(X \leq 10)$. Let $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ and compute $P(Z \leq 10)$.

Solution:

```
c(pbinom(10, 20, .4),
pnorm(10, 20*.4, sqrt(20*.4*.6)))
```

```
## [1] 0.8724788 0.8193448
```

b. You may notice these numbers are a little different. Compute $P(Z \leq 10.5)$ (this is called a *continuity correction* because the normal distribution is continuous while the Binomial is not). Is this a more accurate approximation?

Solution:

```
pnorm(10.5, 20*.4, sqrt(20*.4*.6))
```

```
## [1] 0.8730835
```

Yes, this is more accurate.

c. Use `dbinom` to compute $P(X = 10)$. Use `pnorm` to compute $P(9.5 \leq Z \leq 10.5)$. (Hint: you may need to use the function twice and do subtraction).

Solution:

```
c(dbinom(10, 20, .4),
pnorm(10.5, 20*.4, sqrt(20*.4*.6)) - pnorm(9.5, 20*.4, sqrt(20*.4*.6)))
```

```
## [1] 0.1171416 0.1198649
```

d. Using `pbinom` again, compute $P(4 < X \leq 7)$ and $P(4.5 \leq Z \leq 7.5)$. Note: there is a plot on the last page illustrating graphically the numbers you are computing here (you don't need to print this plot).

Solution:

```
c(pbinom(7, 20, .4) - pbinom(4, 20, .4),
pnorm(7.5, 20*.4, sqrt(20*.4*.6)) - pnorm(4.5, 20*.4, sqrt(20*.4*.6)))
```

```
## [1] 0.364941 0.354664
```

Question 4

According to a Marketplace/Edison survey in April of 2017, about 23.4% of survey responders agreed with the statement "the economic system in the U.S. is fair to all Americans." In this question we'll use a Bernoulli probability model to analyze this number. Suppose that there were 1,000 survey respondents and 234 agreed with the above quotation. Define a Bernoulli random variable which is 1 if a person agrees and 0 otherwise. Assume the survey was done with independent sampling (with replacement), so these Bernoulli random variables are independent. Then the number of people in the sample of 1,000 who agree is a Binomial random variable.

- We have X_i i.i.d Ber(p) for $i = 1, \ldots, 1000$.
- Let $S_n = \sum_{i=1}^{n} X_i$, so S_n is Bin(n, p).

a. Using the fact that $n\bar{X}_n = S_n$, how could you use the Binomial distribution to calculate $P(a \leq \bar{X}_n \leq b)$? How would you use `pbinom` with the given values of a, b, n, and p?

Solution: We know $a \leq \bar{X}_n \leq b$ is the same as $na \leq S_n \leq nb$ so we could compute the probability that a Binomial random variable is between na and nb. This is:

```
pbinom(n*b, size = n, prob = p) - pbinom(n*a, size = n, prob = p)
```

b. Instead of the Binomial distribution, how would we use the central limit theorem to calculate the same probabilities? Hint: your answer should use `pnorm` and involve $\sqrt{n}$ (and a, b, and p).

Solution: The CLT says $\bar{X}_n$ is approximately normal with mean p and standard deviation $\sqrt{p(1-p)/n}$. In `R` we could do:

```
pnorm(b, mean = p, sd = sqrt(p*(1-p)/n)) - pnorm(a, mean = p, sd = sqrt(p*(1-p)/n))
```

c. Now let $n = 1000$, $p = 0.234$, $b = 0.250, a = 0.239$ and compute the desired probability with both methods.

```
n <- 1000
p <- 0.234
a <- 0.239
b <- 0.250
c(pbinom(n*b, size = n, prob = p) - pbinom(n*a, size = n, prob = p),
pnorm(b, mean = p, sd = sqrt(p*(1-p)/n)) - pnorm(a, mean = p, sd = sqrt(p*(1-p)/n)))
```

```
## [1] 0.2291025 0.2383744
```

d. And last, repeat (c) but with $p = 0.3$.

```
p <- 0.3
c(pbinom(n*b, size = n, prob = p) - pbinom(n*a, size = n, prob = p),
pnorm(b, mean = p, sd = sqrt(p*(1-p)/n)) - pnorm(a, mean = p, sd = sqrt(p*(1-p)/n)))
```

```
## [1] 0.0002496295 0.0002671506
```

```r
df <- data.frame(x = 0:14, y = dbinom(0:14, size = 20, prob = .4))
ggplot(df, aes(x, y)) +
  stat_function(fun = dnorm, args = list(mean = 20*.4, sd = sqrt(20*.4*.6))) +
  geom_bar(stat = "identity", alpha = .7) + ylab("Probability") +
  geom_bar(stat = "identity", fill = "red", alpha = .8, data = subset(df, x > 4 & x <= 7)) +
  stat_function(fun = dnorm, args = list(mean = 20*.4, sd = sqrt(20*.4*.6)),
                xlim = c(4.5, 7.5), geom = "area", alpha = .6, fill = "blue") +
  theme_minimal() + ggtitle("Normal approximation (blue area) to Binomial (red area)")
```