Stat 324 Homework #5

Student Name Here

*Submit your homework to Canvas by the due date and time. **Because this homework is due right before the exam, we will not be able to give extensions on it.**

*If an exercise asks you to use R, include a copy of the code and output. Please edit your code and output to be only the relevant portions.

*If a problem does not specify how to compute the answer, you many use any appropriate method. I may ask you to use R or use manually calculations on your exams, so practice accordingly.

*You must include an explanation and/or intermediate calculations for an exercise to be complete.

*Be sure to submit the HWK5 Autograde Quiz which will give you ~20 of your 40 accuracy points.

*50 points total: 40 points accuracy, and 10 points completion

Common Random Variables and Combining RV into Estimators

Exercise 1 Exit polling has been a controversial practice in recent elections, since early release of the resulting information appears to affect whether or not those who have not yet voted do so. Suppose that 90% of all registered Wisconsin voters favor banning the release of information from exit polls in presidential elections until after the polls in Wisconsin close. A random sample of 250 Wisconsin voters are selected (You can assume that the responses of those surveyed are independent). Let X be the count of people in the 250 who favor the ban.

a. Calculate the probability that exactly 230 people in the sample of 250 favor the ban, that is $P(X = 230)$.

```
choose(250, 230) * 0.9^(230) * (1 - 0.9)^20
```

```
## [1] 0.05122197
```

b. Calculate the **exact** probability that 230 or more people in the sample of 250 favor the ban, that is $P(X \geq 230)$. Hint: use a couple of R functions to help with this calculation.

```
1 - pbinom(229, 250, 0.9)
```

```
## [1] 0.1718898
```

c. What are the expected value (μ_X) and standard deviation (σ_X) of X?

```
250 * 0.9
```

```
## [1] 225
```

```r
sqrt(250 * 0.9 * 0.1)
```

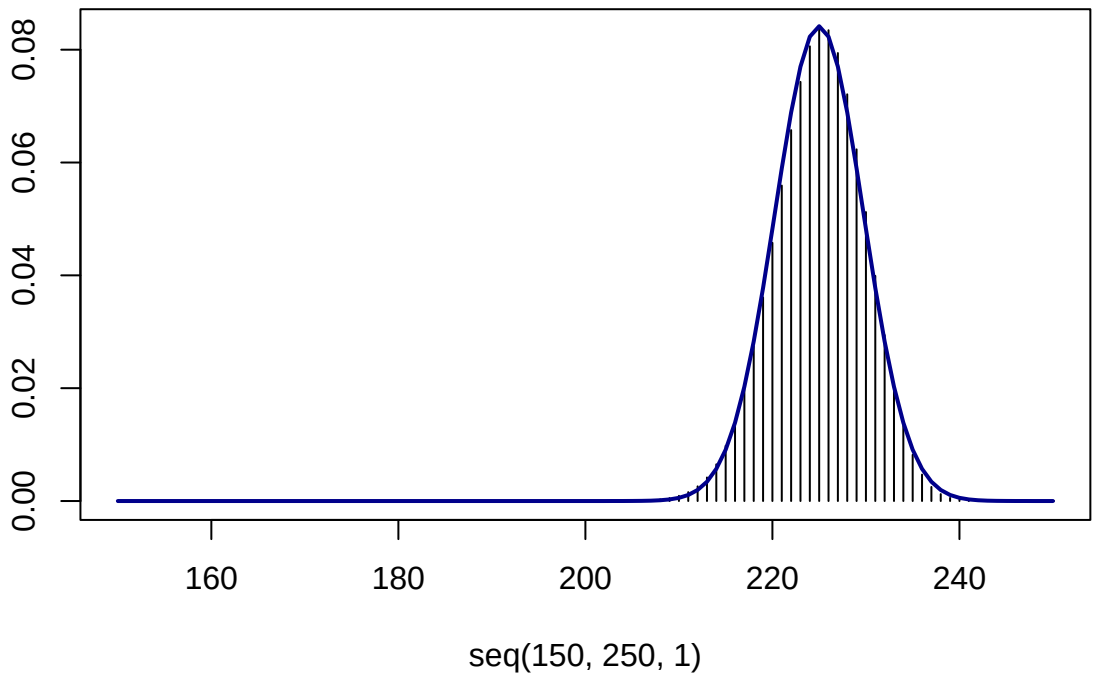
```
## [1] 4.743416
```

d. We can consider X the sum of 250 iid random draws from the population Y where P(Y=1)=0.90 and P(Y=0)=0.10. That is $X = Y_1 + Y_2 + .. + Y_{250}$. What do we think will be true about the shape of the distribution of X? What theorem are you using?

The Central Limit Theorem states that the sum (or average) of a large number of i.i.d. random variables, each with finite mean and variance, will be approximately normally distributed, regardless of the original distribution of the random variables. The distribution will, therefore, be approximately normal.

e. To look at how well a normal curve approximates the distribution of X with n=250, $\pi = 0.90$ run the following code with the mean and sd values computed in (c) substituted for "MEAN_VALUE" and "SD_VALUE", respectively. You'll also need to change eval=TRUE for the r chunk to run.

```r
plot(x=seq(150,250,1), y=dbinom(150:250, 250, prob=0.90), type='h', ylab="")
curve(dnorm(x, mean=225, sd=4.74), col="darkblue", lwd=2, add=TRUE, yaxt="n")
```



f. Calculate the approximate probability that at least 230 people in a sample of 250 favor the ban, that is $P(X \geq 230)$, assuming a Normal Distribution for X centered at the mean and sd found in c. Compare the value to that found in b. and explain why they are not exactly equal.

```r
1 - pnorm(230, 225, 4.74)
```

```
## [1] 0.1457464
```

The reason that the two values are not exactly equal is that an actual binomial distribution is discrete unlike a normal distribution, which is smooth and continuous. Since a normal approximation basically smooths out the binomial distribution, some discrepancies arise if the sample size is still small, which 250 is.

g. Consider the sample proportion. Calculate the approximate probability that at least $\hat{p} = \frac{230}{250}$ favor the ban, that is $P(\hat{p} \geq 0.92)$, assuming a Normal Distribution of $\hat{p}$ centered at the appropriate mean and standard deviation. Compare the value to that found in f. and explain the relationship between the values.

```r
sample_proportion <- 230 / 250
sd_sample_proportion <- sqrt(0.9 * (1 - 0.9) / 250)
1 - pnorm(sample_proportion, mean = 0.9, sd = sd_sample_proportion)
```

```
## [1] 0.1459203
```

This value measures the same probability as in problem f, just a different way of going around it. While f measures the probability of the actually value, this calculation measures the probability of the probability itself. > h. Give at least one suggestion to the data collection team to make the independence of offered opinion a better assumption.

Stop participants from discussing the survey with others until the survey is complete.

Exercise 2 Let X denote the number of flaws in a 1 in. length of copper wire. The probability mass function of X is given in the table below. It has mean: $\mu = 0.66$ and variance $\sigma^2 = 0.5244$.

Number of Flaws in length of wire (X)	Probability
x=0	0.48
x=1	0.39
x=2	0.12
x=3	0.01

```r
vals=c(0,1,2,3)
probs=c(0.48, 0.39, 0.12, 0.01)
(EV_pop=sum(vals*probs))
```

```
## [1] 0.66
```

```r
(Var_pop=sum(probs*(vals-EV_pop)^2))
```

```
## [1] 0.5244
```

a. Is the distribution of X left skewed, symmetric, or right skewed? How do you know?

The distribution appears to be left skewed, since $P(x)$ is biggest at $x = 0$ and steadily decreases.

b. In what percent of 1 in. length of copper wire will 1 or more flaws be observed?

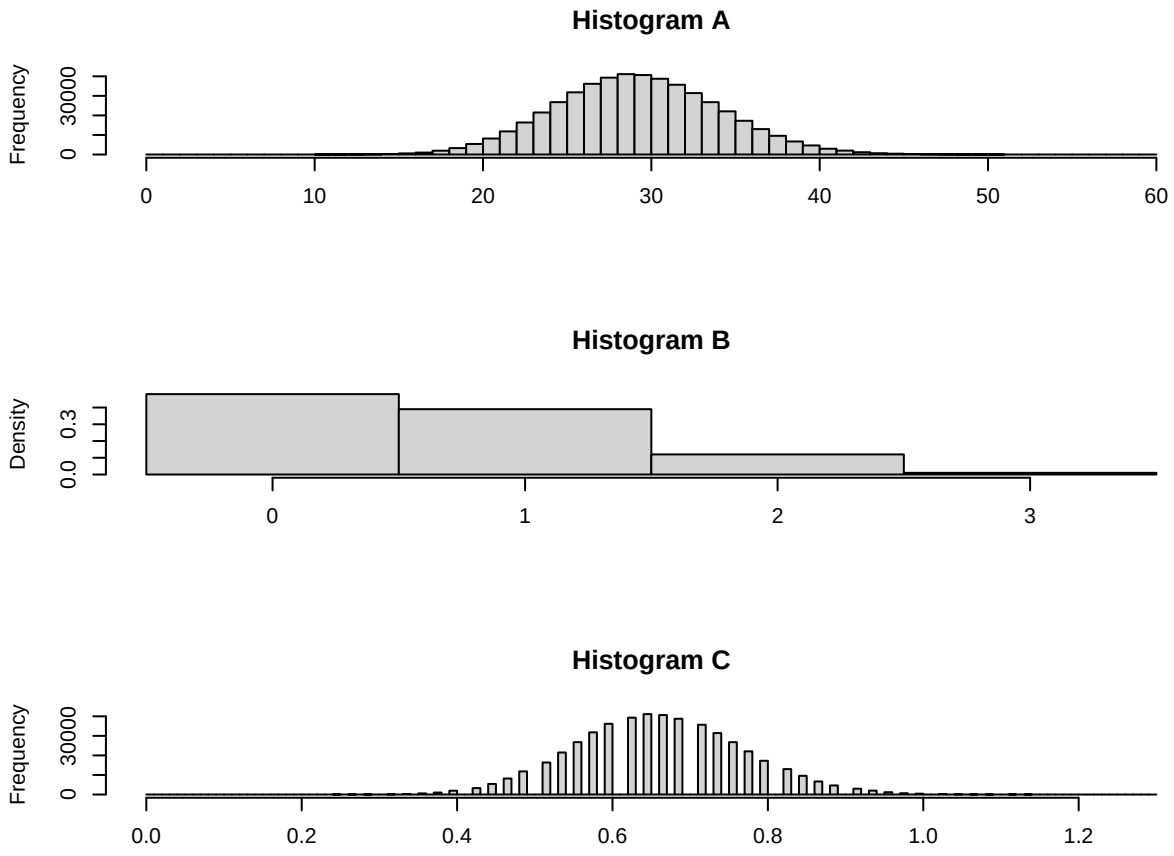
$0.39 + 0.12 + 0.01 = \boxed{52}$

A random sample of 45 1 in. lengths of the copper wire are selected for review. Since this is a SRS from a very large population, we can consider the number of flaws on multiple draws from the population $X_1, X_2, ..X_n$ iid to X.

c. The simulation below selects 45 lengths of copper wire from the population with replacement and computes the sample mean and sample sum. It then repeats this process manytimes, stores the sample mean and sample sum values in vectors and then creates histograms of those vectors of values. Identify which histogram diplays (1) the population X values, (2) the simulated sampling distribution of the sample mean $\bar{X}$, (3) the simulated sampling distribution of the sample sum S. Briefly explain how you know.

```
items=c(rep(0, 48), rep(1, 39), rep(2, 12), rep(3, 1))
manytimes=500000
samp_mean=rep(9, manytimes)
samp_sum=rep(9, manytimes)
for (i in 1:manytimes){
  samp=sample(items, size=45, replace=TRUE)
  samp_mean[i]=mean(samp)
  samp_sum[i]=sum(samp)
}

par(mfrow=c(3,1))
hist(samp_sum, breaks=seq(0, 60, 1), main="Histogram A", xlab="")
hist(items, breaks=seq(-0.5, 3.5, 1), freq=FALSE, main="Histogram B", xlab="")
hist(samp_mean, breaks=seq(0,1.3, 0.01), main="Histogram C", xlab="")
```




```
par(mfrow=c(1,1))
```

Histogram B displays the population X values, Histogram C displays the sample means, and Histogram A displays the simulated sampling distribution of the sample sum S.

I know Histogram B displays the population values since the x-axis only includes the values 0-3, which are the values in `items`.

Histogram C displays the sample means because the x-axis is smaller than Histogram A, and it makes more sense for them to be the means rather than the sum. Histogram A shows the sums by process of elimination.

> d. Describe the sampling distribution (shape, mean, and standard deviation) of the sample mean number of flaws in 45 1 in. length of copper wire $\bar{X} = \frac{X_1+X_2+...+X_{45}}{45}$ according to theory. Make sure to name any theorems you are using. (You can compute the mean and sd of one of the vectors constructed above to make sure your theoretical values are close to what you get in the simulation.)

The mean of the sampling distribution (μ_x) is equal to the population mean (μ). In this case, $\mu = \mu_x = 0.66$. The variance is equal to the population sd over the sample size. So, $\sqrt{0.5244}/45$

According to the Central Limit Theorem, the sampling distribution of the sample mean becomes approximately normal.

e. According to your theoretical distribution in (d), what is the probability that the mean number of flaws in the 45 1 inch lengths of wire reviewed will be 1 or more flaws?

```
z = (1 - 0.66) / 0.1079506
1 - pnorm(z)
```

```
## [1] 0.0008175021
```

f. Explain why the value you found in e. was so much smaller than the value found in b.

The normal approximation allows for fractional numbers, so the probability is much more

g. Consider the total number of flaws in the 45 1 in. lengths of copper wire. Describe the sampling distribution (shape, mean, and standard deviation) of $Sum = X_1 + X_2 + ... + X_{45}$ according to theory. Make sure to name any theorems you are using. (You can compute the mean and sd of one of the vectors constructed above to make sure your theoretical values are close to what you get in the simulation.)

Shape: The shape of the sampling distribution of the sum of i.i.d. random variables tends to be approximately normal (due to the CLT) if the original distribution of flaws is not heavily skewed and the sample size is reasonably large.

Mean: simply size of sum times mean.

```
45 * 0.66
```

```
## [1] 29.7
```

Variance:

```
sqrt(45 * sqrt(0.5244))
```

```
## [1] 5.708499
```

h. Find an upper bound b such that the **total number of flaws in 45 1 in. lengths of copper wire** will be less than b with probability 0.95.

```
1.645 * sqrt(45 * 0.5244) + 45 * 0.66
```

```
## [1] 37.69104
```