Statistics 324 Homework #4

Student's Name Here

*Submit your homework to Canvas by the due date and time. Email your instructor if you have extenuating circumstances and need to request an extension.

*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 manual 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 HWK4 Autograde Quiz which will give you ~20 of your 40 accuracy points.

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

Exercise 1 For each of the following questions, say whether the random variable is reasonably approximated by a binomial random variable or not, and explain your answer. Comment on the reasonableness of each of things that must be true for a variable to be a binomial random variable (ex: identify n : the number of Bernoulli trials, π the probability of success, etc).

a. A fair die is rolled until a 1 appears, and X denotes the number of rolls.

There is not a fixed number of trials, so the random variable is not reasonably approximated by a binomial random variable.

b. Twenty of the different Badger basketball players each attempt 1 free throw and X is the total number of successful attempts.

Even though there is a fixed number of trials, the probability of success π is not constant since each player has a different free throw percentage. Therefoer, the random variable is not reasonably approximated by a binomial random variable.

c. A die is rolled 40 times. Let X be the face that lands up.

Since the outcome isn't binary, the random variable cannot be approximated as a binomial random variable.

d. In a bag of 10 batteries, I know 2 are old. Let X be the number of old batteries I choose when taking a sample of 4 (without replacement) to put into my calculator.

Since the probability changes with each battery you take out of the bag (without replacement), the trials are not independent of one another, and therefore the RV cannot be approximated as a binomial random variable.

e. It is reported that 20% of Madison homeowners have installed a home security system. Let X be the number of homes without home security systems installed in a random sample of 100 houses in the Madison city limits.

Since there is a fixed number of trials, each trial is independent, the probability stays the same, and there is a binary outcome, the RV can be approximated as a binomial random variable.

Exercise 2: A chemical supply company ships a certain solvent in 10-gallon drums. Let X represent the number of drums ordered by a randomly chosen customer. Assume X has the following probability mass function (pmf). The mean and variance of X is : $\mu_X = 2.2$ and $\sigma_X^2 = 1.76 = 1.32665^2$:

x	P(X=x)
1	0.4
2	0.3
3	0.1
4	0.1
5	0.1

a. Calculate $P(X \leq 2)$, and describe what it means in the context of the problem.

$P(X \leq 2) = 0.4 + 0.3 = \boxed{0.7}$

This means that the probability that when choosing a random customer, the number of drums that customer will have ordered has a 70% chance of being less than or equal to 2.

b. Let Y be the number of gallons ordered, so $Y = 10X$. Complete the probability mass function of Y.

y	P(Y=y)
10	0.4
20	0.3
30	0.1
40	0.1
50	0.1

c. Calculate the mean number of gallons ordered μ_Y.

$\mu_Y = \boxed{22}$

d. Calculate the standard deviation of the number of gallons ordered, σ_Y.

$\sigma_Y = \boxed{13.2665}$

Exercise 3: The bonding strength A of a drop of plastic glue from a particular manufacturer is thought to be well approximated by a normal distribution with mean 98 lbs and standard deviation 7.5 lbs. $A \sim N(98, 7.5^2)$. Compute the following values using a normal model assumption.

a. What proportion of drops of plastic glue will have a bonding strength between 95 and 104 lbs according to this model?

```r
prob_between_95_and_104 <- pnorm(104, mean = 98, sd = 7.5) - pnorm(95, mean = 98, sd = 7.5)
prob_between_95_and_104
```

```
## [1] 0.4435663
```

b. A single drop of that glue had a bonding strength that is 0.5 standard deviations above the mean. What proportion of glue drops have a bonding strength that is higher ?

```r
mean <- 98
sd <- 7.5
above_mean <- mean + 0.5 * sd

# Proportion of glue drops with bonding strength higher than above_mean
proportion_higher <- 1 - pnorm(above_mean, mean = mean, sd = sd)
proportion_higher
```

```
## [1] 0.3085375
```

c. What bonding strength did a drop of glue have that is at the 90th percentile?

```r
# Given data
mean <- 98
sd <- 7.5

# Find bonding strength at the 90th percentile
strength_at_90th_percentile <- qnorm(0.90, mean = mean, sd = sd)
strength_at_90th_percentile
```

```
## [1] 107.6116
```

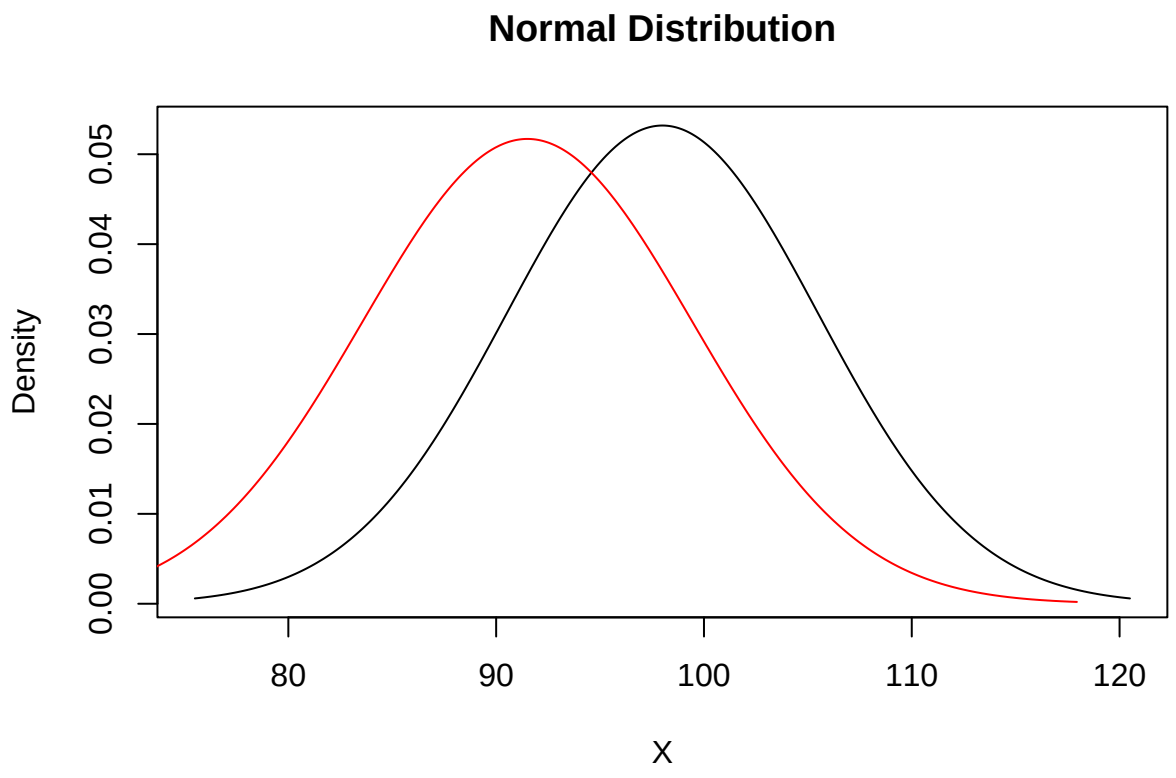
d. What is the IQR of bonding strength for drops of glue from this manufacturer?

```r
# Given data
mean <- 98
sd <- 7.5

# Calculate Q1 and Q3 using mean +/- 0.675 * SD
Q1 <- mean - 0.675 * sd
Q3 <- mean + 0.675 * sd

# Calculate IQR
IQR <- Q3 - Q1
IQR
```

```
## [1] 10.125
```

e. Drops of a similar plastic glue from another manufacturer (manufacturer B) is claimed to have bonding strength well approximated by a normal distribution with mean 43 kg and standard deviation of 3.5 kg $B_{KG} \sim N(43, 3.5^2)$. Transform the bonding strength of manufacturer B into lbs using the conversion: 1 kg $\approx$ 2.20462 lbs. You can use the transformation: $B_{LBS} = 2.20462 * B_{KG}$. Compare the shape, center, and spread of the two glues' bonding strength.

```r
# Parameters of the normal distribution
mean <- 98  # Replace with your mean value
sd <- 7.5    # Replace with your standard deviation value

# Generate x values from mean - 3*sd to mean + 3*sd (for a range of 6 standard deviations)
x <- seq(mean - 3*sd, mean + 3*sd, length=1000)

# Calculate the corresponding y values (PDF)
y <- dnorm(x, mean = mean, sd = sd)

# Parameters of the normal distribution
mean1 <- 43 * 2.20462  # Replace with your mean value
sd1 <-  3.5 * 2.20462    # Replace with your standard deviation value

# Generate x values from mean - 3*sd to mean + 3*sd (for a range of 6 standard deviations)
x1 <- seq(mean1 - 3*sd1, mean1 + 3*sd1, length=1000)

# Calculate the corresponding y values (PDF)
y1 <- dnorm(x, mean = mean1, sd = sd1)

# Plot the normal distribution
plot(x, y, type="l", xlab="X", ylab="Density", main="Normal Distribution")
lines(x1, y1, col="red")  # You can also use points() to plot points instead of lines
```



The center of the distribution of the similar plastic glue is lesser than the original distribution. Additionally, the spread of the new distribution is slightly bigger, signifying a bigger standard deviation. The shapes are

similar since both distributions are normal.

Exercise 4: A serving of a specific type of yogurt has a sugar content that is well approximated by a Normally distributed random variable X with mean 13 g and variance: $1.3^2 g^2$. We can consider each serving as an independent and identical draw from X.

a. In what percent of servings will the sugar content be above 13.3 g?

```
# Given data
mean <- 13
sd <- 1.3
threshold <- 13.3

# Calculate Z-score
z <- (threshold - mean) / sd

# Calculate probability that X > 13.3
probability_above_13.3 <- 1 - pnorm(z)
percentage_above_13.3 <- probability_above_13.3 * 100
percentage_above_13.3
```

```
## [1] 40.8747
```

b. What is the probability that a randomly chosen serving will have a sugar content between 13.877 and 12.123? What do we call the difference: 13.877-12.123=1.754?

```
# Given data
mean <- 13
sd <- 1.3
upper_value <- 13.877
lower_value <- 12.123

# Calculate Z-scores
z_upper <- (upper_value - mean) / sd
z_lower <- (lower_value - mean) / sd

# Calculate probability that 12.123 < X < 13.877
probability_between <- pnorm(z_upper) - pnorm(z_lower)
probability_between
```

```
## [1] 0.5000798
```

The difference is the IQR.

c. Calculate the probability that in 6 servings, only 1 has a sugar content below 13 g.

```
# Given data
mean <- 13
sd <- 1.3
```

```r
n <- 6
k <- 1

# Calculate p (probability of a single serving below 13 g)
p <- pnorm(13, mean = mean, sd = sd)

# Calculate binomial probability
binomial_probability <- choose(n, k) * p^k * (1 - p)^(n - k)
binomial_probability
```

```
## [1] 0.09375
```

d. Describe the sampling distribution for the mean sugar content of 6 servings $\bar{X}$. (Give shape, mean, and standard deviation or variance, if possible)

```r
1.3/sqrt(6) # population sd over sqrt(trials) = variance
```

```
## [1] 0.5307228
```

the sampling distribution of the mean sugar content of 6 servings $\bar{x}$ is approximately normal with a mean of 13 g and a variance of approximately 0.530 g.

e. What is the interquartile range of the sampling distribution for the sample mean $\bar{X}$ when n=6? Is that value larger or smaller than the IQR implied in part b? Why does the relative sizes of the IQRs make sense?

```r
mu = 13
sd = 1.3
q1 = mean - (sd * 0.675)
q3 = mean + (sd * 0.675)
q3 - q1
```

```
## [1] 1.755
```

The IQR for the sampling distribution of the sample mean $\bar{x}$ when $n = 6$ is approximately 1.755 g. This is slightly larger than the IQR calculated in part B. This makes sense because the probability calculated in part B is slightly larger than 0.5, which means that the IQR should be a little bigger for the probability to be exactly 0.5.

f. What is the probability that the mean sugar content in 6 servings is more than 13.3 g?

```r
# Given data
mean_sampling <- 13
sd_sampling <- 0.530
threshold <- 13.3

# Calculate Z-score
z <- (threshold - mean_sampling) / sd_sampling

# Calculate probability that mean sugar content is more than 13.3 g
probability_more_than_13.3 <- 1 - pnorm(z)
probability_more_than_13.3
```

```
## [1] 0.2856841
```

g. Is it more or less likely that the mean sugar content is above 13.3 g in 10 servings or 6 servings (as computed in f)? Can you explain it without actually computing the new probability?

Since the standard deviation of the sampling distribution decreases as the sample size increases, the probability of observing a mean sugar content above 13.3 g decreases as the sample size increases. Therefore, it is less likely that the mean sugar content is above 13.3 g in 10 servings rather than 6 servings.

h. Suppose each large yogurt container of this type contains 10 servings and consider the total sugar content in each container as a sum of 10 iid random draws from $X \sim N(13, 1.3^2)$. If you were to eat a whole large container of yogurt, **above what total sugar content** would you consume with 95% probability? Show and briefly explain your calculations.

```r
# Given data
mean_sum <- 10 * 13
sd_sum <- sqrt(10) * 1.3
percentile <- 0.95

# Find z-score for the 95th percentile
z <- qnorm(percentile)

# Calculate value above which you would consume with 95% probability
value <- z * sd_sum + mean_sum
value
```

```
## [1] 136.7619
```

Exercise 5: You will be comparing the sampling distributions for two different estimators of σ, the population standard deviation.

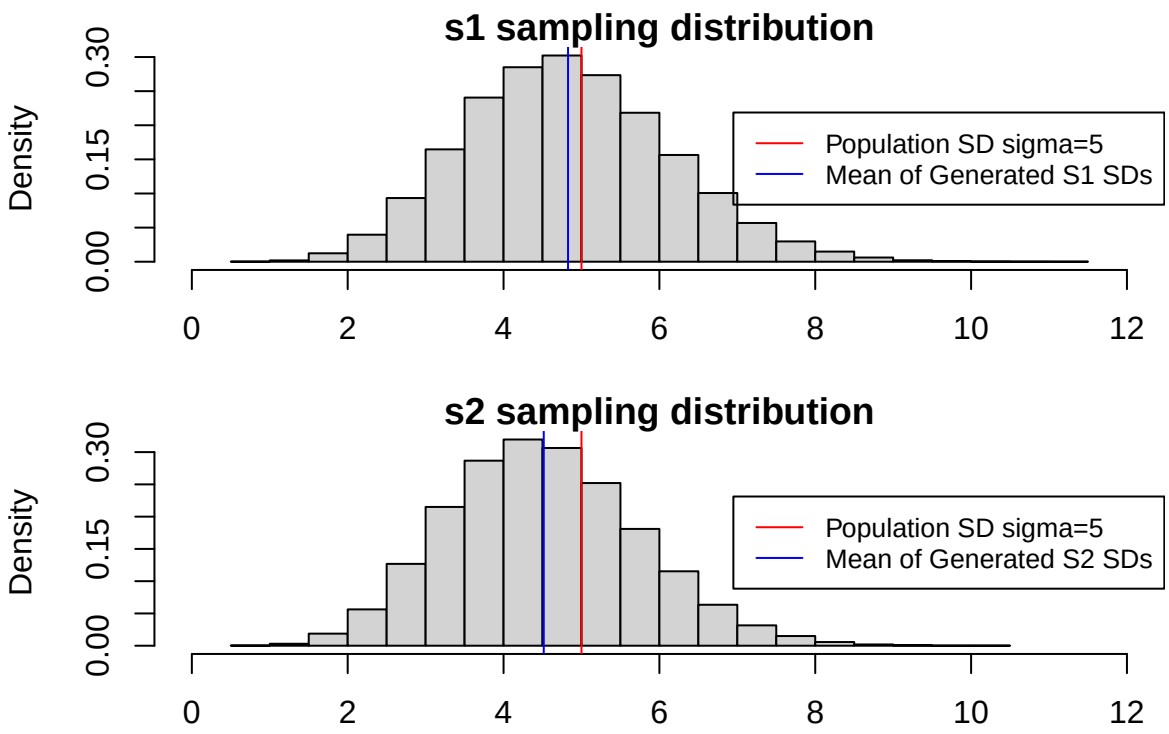
When trying to estimate the standard deviation of a population (σ) from a sample we could use:

$$s_1 = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}} \text{ or } s_2 = \sqrt{\frac{\sum(X-\bar{X})^2}{n}}$$

The graphs below give the sampling distributions produced by these estimators when drawing a sample of size 8 from a normal population with mean $\mu_x = 3$ and standard deviation $\sigma_X = 5$.

a. What do you notice about the mean of the standard deviations produced using the s_1 estimator compared to the s_2 estimator compared to the true population standard deviation? Why do we prefer to use the s_1 formulation when we have a sample of data and are interested in estimating the population standard deviation? (You should use the resulting histograms to help you answer the question and use the word "bias".)

The mean of the standard deviations in s1 is more centered around the true standard deviation compared to s2. We prefer to use the s1 formulation when we have a sample of data and are interested in estimating the population standard deviation because it introduces bias if you use n rather than n-1, since it tends to underestimate the standard deviation.


s1 sampling distribution
Density
0.00
0.15
0.30
0
2
4
6
8
10
12
Population SD sigma=5
Mean of Generated S1 SDs

s2 sampling distribution
Density
0.00
0.15
0.30
0
2
4
6
8
10
12
Population SD sigma=5
Mean of Generated S2 SDs