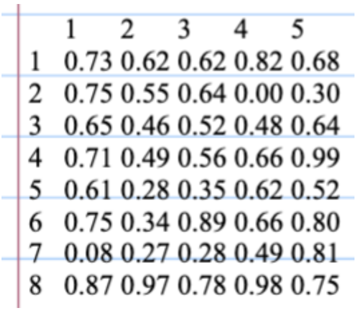
STAT 421 Homework 2

2023-10-03

Lecture 3 problems:

1. Consider the following data pertaining to a treatment factor, X, taking 8 levels denoted 1 through 8. There are 5 replications labeled 1 through 5. The response, Y, is a measure of accuracy (e.g. in weather prediction), and X denotes 8 different models used to make the predictions.

a. Make comparative boxplots that allow one to compare the performance/goodness of the 8 models, and discuss the results as thoroughly as you can (within half a page!). By R.



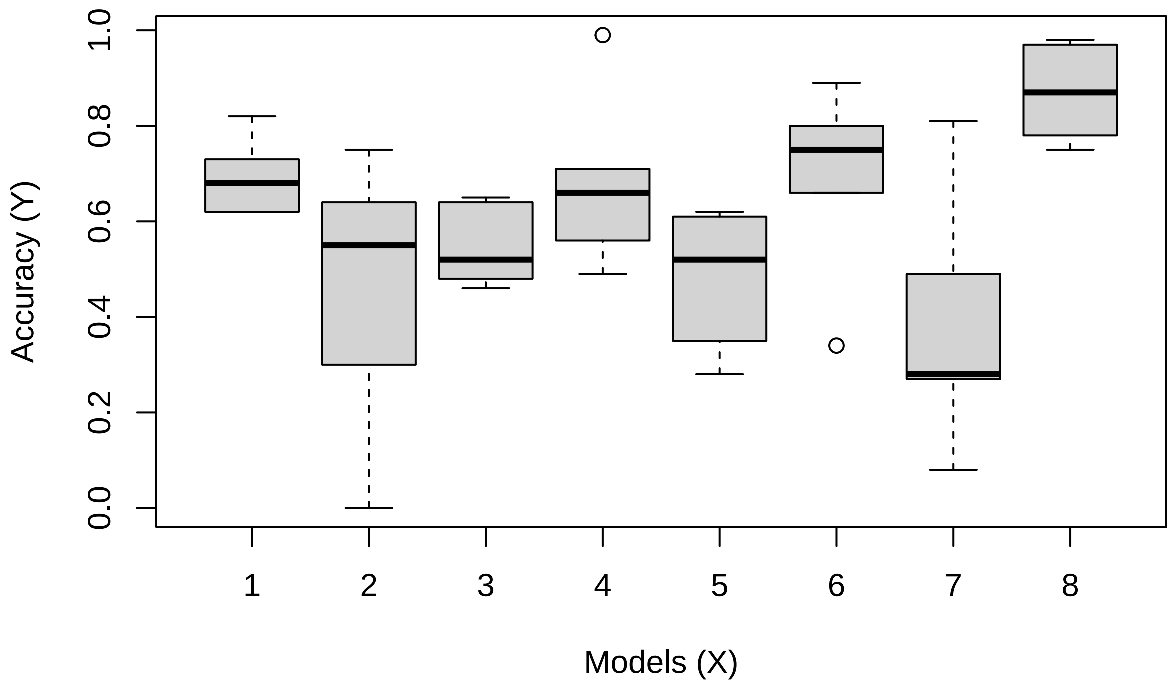
	1	2	3	4	5
1	0.73	0.62	0.62	0.82	0.68
2	0.75	0.55	0.64	0.00	0.30
3	0.65	0.46	0.52	0.48	0.64
4	0.71	0.49	0.56	0.66	0.99
5	0.61	0.28	0.35	0.62	0.52
6	0.75	0.34	0.89	0.66	0.80
7	0.08	0.27	0.28	0.49	0.81
8	0.87	0.97	0.78	0.98	0.75

Setup:
In order to analyze comparative boxplots of the data, we must first create the data and the boxplots themselves using R. We will start by creating a matrix of the data.

```r
# Create a vector of response data:
Y <- c(0.73, 0.62, 0.62, 0.82, 0.68, 0.75, 0.55, 0.64, 0.00, 0.30, 0.65, 0.46,
       0.52, 0.48, 0.64, 0.71, 0.49, 0.56, 0.66, 0.99, 0.61, 0.28, 0.35, 0.62,
       0.52, 0.75, 0.34, 0.89, 0.66, 0.80, 0.08, 0.27, 0.28, 0.49, 0.81, 0.87,
       0.97, 0.78, 0.98, 0.75)

# Turn this data into matrix form:
Y_matrix <- matrix(Y, nrow = 8, ncol = 5, byrow = TRUE)

# Transpose this data in order to put X factor levels as the columns:
Y_matrix_t <- t(Y_matrix)
```

Now that the data has been created in R, we will now construct comparative boxplots that allow us to compare the performance/goodness of the 8 models.

```r
# Create the comparative boxplots:
boxplot(Y_matrix_t, xlab="Models (X)", ylab="Accuracy (Y)")
```



Analysis:
The first thing that is important to analyze when looking at comparative boxplots is the spread of all of the boxplots. In this experiment, we can see that the boxplots of model 1, 3, 5, and 8 all have small spreads, meaning the 5 observations don't vary that much from each other, especially in comparison to the other models. Furthermore, the boxplots of model 2 and 7 have fairly large spreads, especially in comparison to the models mentioned previously. Lastly, the boxplots of model 4 and 6 have relatively small spreads, however, both of these boxplots have outliers present, which in turn makes the boxplots themselves have a larger spread. In general, the boxplots with narrower spread are preferred in terms of analyzing which one is "best".

Now that we have analyzed the spreads of the boxplots individually, we can use the placement of the boxplots to allow for comparisons between the models. For example, due to the overlap of many of the boxplots such as those for models 1, 3, 4, 5, and 6, it is difficult to decide which of these models is truly better. However, due to its position and low variability, the boxplot for model 8 provides evidence that it is better than most of, if not all of models. Due to the outlier in model 4, it is hard to decide if model 8 is truly better in terms of accuracy than model 4 is. However, based on the qualitative asessment from above, the leading candidates for the most accurate models are model 8, model 4, although model 8 would be the safest due to its high position and low variability.

b. We haven't yet studied the t-test, but I'm assuming that you do know about it, and how to perform it using the t-test() function in R. Perform a 2-sample, 2-sided t-test for whether models 1 and 5 have different accuracy. Turn in the R code, along with the p-value. Also at significance level $\alpha = 0.01$, is there a difference between the two models?

If we let μ_1 represent the mean accuracy score of model 1, and μ_5 represent the mean accuracy score of model 5, then in this problem we will use the t-test() function in R to test the hypotheses $H_0 : \mu_1 = \mu_5$

versus $H_1 : \mu_1 \neq \mu_5$ at the 0.01 level of significance. Since the boxplots of models 1 and 5 are relatively symmetric, the assumption that the accuracy scores come from a normal population is a valid one, hence we can continue with the t-test. However, since the spread of the two boxplots for model 1 and 5 are noticeably different, we will run the 2-sample t-test with the unequal variance assumption.

We will now perform the t-test in R and show its output.

```r
# Extract the model 1 accuracy scores:
acc_mod_1 <- Y_matrix[1,]

# Extract the model 5 accuracy scores:
acc_mod_5 <- Y_matrix[5,]

# Run the 2-sample 2-sided t-test in R:
t.test(x = acc_mod_5, y = acc_mod_1, alternative = "two.sided",
       var.equal = FALSE, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  acc_mod_5 and acc_mod_1
## t = -2.7771, df = 6.1917, p-value = 0.03108
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -0.50536217  0.06936217
## sample estimates:
## mean of x mean of y
##     0.476     0.694
```

As seen from the above 2-sample t-test with the unequal variance assumption, the p-value obtained was 0.03108. Hence at the 0.01 significance level, we fail to reject the null hypothesis. Thus we have found no significant evidence to suggest that the difference in average accuracy between the two models is different than zero. More concretely, we have found no significant evidence that the two models are different.

c. Now, suppose I reveal to you that in the first replicate, models 1 and 5 are run under absolutely identical conditions. Similarly, models 1 and 5 are run under identical conditions as we perform the 2nd replicate, (i.e. it's like making observations of y on ONE coupon). Etc. Perform the t-test for this new design. Again, report the p-value, and state your conclusion pertaining to whether models 1 and 5 have different accuracy. Hint/question: what is the block factor?

As mentioned in the problem description, if we know that in the first and second replicate, models 1 and 5 are run under absolutely identical conditions, this is theoretically the same as making observations of y on one coupon as we did in the tip testing example. Hence, if we assume that these identical conditions of model 1 and 5 carry out throughout all replicates, we can block on replication and run a paired t-test instead of a 2-sample t-test.

If we let μ_d represent the mean difference between the accuracy score of model 1, and the accuracy score of model 5, then in this problem we will use the t-test() function in R to test the hypotheses $H_0 : \mu_d = 0$ versus $H_1 : \mu_d \neq 0$ at the 0.01 level of significance.

We will now perform the t-test in R and show its output.

```r
# Run the paired t-test in R:
t.test(x = acc_mod_5, y = acc_mod_1, alternative = "two.sided",
       paired = TRUE, conf.level = 0.99)
```

```
##
##  Paired t-test
##
## data:  acc_mod_5 and acc_mod_1
## t = -5.548, df = 4, p-value = 0.005164
## alternative hypothesis: true mean difference is not equal to 0
## 99 percent confidence interval:
##  -0.39891222 -0.03708778
## sample estimates:
## mean difference
##          -0.218
```

As seen from the above paired t-test, the p-value obtained was 0.005164. Hence at the 0.01 significance level, we reject the null hypothesis that the average difference between the accuracy scores of the two models is zero. Hence we have found significant evidence to suggest that the there is a difference between the two models, in terms of their mean accuracy score. As mentioned in the textbook, blocking is a noise reduction technique, hence leading to higher power, which in this case allowed us to detect the difference.

2. Consider the problem of testing the following pair of hypotheses: H_0 : x is related to y and H_1 : x is not related to y. One possible parametric approach to testing those hypotheses look like: $H_0 : \mu_x = \mu_y$ versus $H_1 : \mu_x > \mu_y$, where μ_x and μ_y are the μ parameters of two normal distributions for x and y, respectively. Suppose the σ parameter of the two distributions are known to be $\sigma_x = 1$ and $\sigma_y = 1$. Also, suppose this is our observed data:

```r
set.seed(12)
nsample <- 10
x_obs <- rnorm(nsample, 0, 1)
y_obs <- rnorm(nsample, 0, 1)
```

a. Technically, since we know the true values of σ_x and σ_y, we should perform a z-test, but here use the R function t.test() to perform the appropriate test of the H_0/H_1, above. Report/write the p-value.

Since the above data comes from a normal distribution the use of a t-test is justified (even though we should use a z-test as explained above). Furthermore, since we know that the variances of x and y are identical we will use the equal variance version of the 2-sample t-test. We will perform this using the t.test() function in R below.

```r
# Run the 2-sample 1-sided t-test in R:
t.test(x = x_obs, y = y_obs, alternative = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x_obs and y_obs
## t = -0.69194, df = 18, p-value = 0.7511
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.9536862        Inf
## sample estimates:
##  mean of x  mean of y
## -0.4672139 -0.1952043
```

As found from the above 2-sample 1-sided t-test, the p-value is 0.7511. This p-value is extremely high and wouldn't be evidence to reject the null at any reasonable significance level.

b. I can tell you that, under H_0, the statistic $\delta = \bar{X} - \bar{Y}$ has an approximately normal distribution with parameters $\mu = 0$, $\sigma = \sqrt{\frac{\sigma_x}{n} + \frac{\sigma_y}{n}}$, where n is the sample size. Use the R function pnorm() to find the p-value. Report the number you get. Hint: the p-value is the area under the aforementioned distribution to a certain side of the observed δ.

Under H_0, the statistic $\delta = \bar{X} - \bar{Y}$ has an approximately normal distribution with parameters $\mu = 0$, $\sigma = \sqrt{\frac{\sigma_x}{n} + \frac{\sigma_y}{n}}$, where n is the sample size. From the above data, we know that $\sigma_x = \sigma_y = 1$, and $n = 10$. Hence we can see that $\delta \sim Norm\left(\mu = 0, \sigma = \sqrt{\frac{1}{5}}\right)$. Below we will use R to calculate our observed δ value, and use the pnorm() function to find the p-value of our observed δ value. Since the order of subtraction was x minus y, we will be looking at the upper tail of this normal distribution as it points to extreme values in the direction of the alternative.

```r
# Find the observed difference in means:
delta_obs <- mean(x_obs) - mean(y_obs)

# Find the p-value using pnorm():
norm_p_val <- pnorm(delta_obs, mean = 0, sd = sqrt(1/5), lower.tail = FALSE)
```

As found from the distribution of δ, the p-value is 0.7284832. This p-value is extremely high and wouldn't be evidence to reject the null at any reasonable significance level.

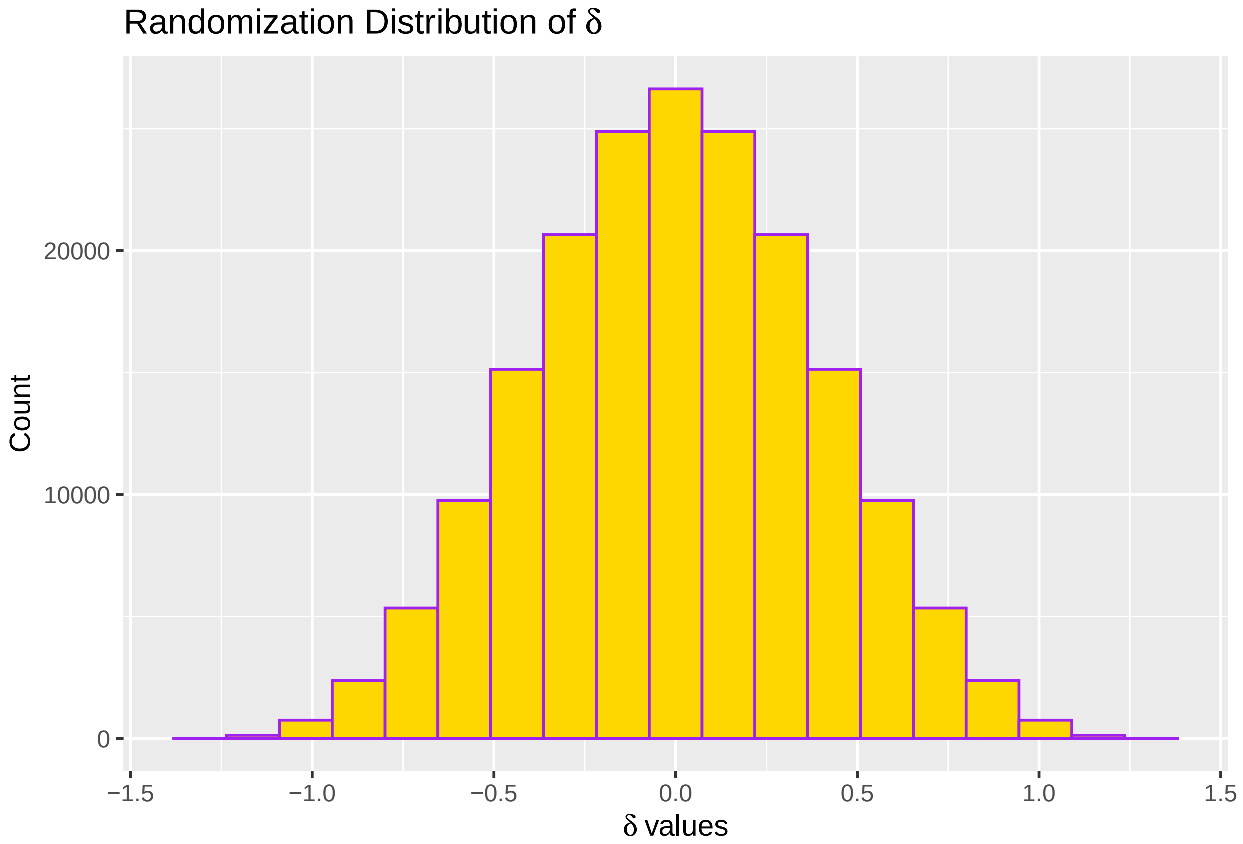
c. Now, the randomization test. Let me, first, introduce you to some basic commands that you can then use to perform the randomization test. Study the following code:

```r
N <- 6
index <- 1:N # Think of these as the case number for 6 observations.
M <- combn(index, N/2) # You can check the help pages, but you can also see
M # that it returns all 6 choose 3 combinations.
choose(N, N/2) # BTW, 6 choose 3 is 20 which is ncol(M).
M[,1] # These are the 1st 3 selected cases, and
index[-M[,1]] # these are the remaining 3 cases, not selected.
dat <- c(10, 8, 2, 9, 1, 5) # So, if these are our data (y), on 2 tips (A and B), then
dat[M[,1]] # these 3 cases correspond to tip A,
dat[index[-M[,1]]] # while these 3 cases correspond to tip B.
```

Using the above tricks, write code to generate the randomization distribution of δ. Hint: the question is asking for the histogram of the $\binom{20}{10}$ δ values. Also, dat = c(x_obs, y_obs).

```r
# Set up the randomization test:
N <- 20
index <- 1:N
comb <- combn(index, N/2)
num_combs <- choose(N, N/2)
data <- c(x_obs, y_obs)
delta_values <- rep(0, num_combs)

# Perform the randomization test:
```

```
for (x in 1:num_combs) {
  new_x = data[comb[,x]]
  new_y = data[index[-comb[,x]]]
  delta_values[x] = mean(new_x) - mean(new_y)
}

# Create and display the randomization distribution of delta (number of bins
# calculated using Sturges rule):
rand_dist <- ggplot() +
  geom_histogram(mapping = aes(x = delta_values), bins = 19, color = "purple",
                 fill = "gold") +
  labs(title = expression("Randomization Distribution of" ~ delta),
       y = "Count",
       x = expression(delta ~ "values"))

rand_dist
```



d. Now, find/report the p-value. Hint: this is number of δ values exceeding δ_{obs}, divided by the total number of δs.

Based on the randomization test coded in part c, and our observed δ value computed in part b, we will use R to compute the p-value of the randomization test. This is calculated as the number of δ values exceeding δ_{obs}, divided by the total number of δs.

```r
# Compute the p-value of the randomization test:
rand_p_val <- length(delta_values[delta_values > delta_obs]) / length(delta_values)
```

As found from the randomization test, the p-value is 0.7497727. This p-value is extremely high and wouldn't be evidence to reject the null at any reasonable significance level.

Lecture 4 problems:

1. Now that I've mentioned the importance of the subscript on $E[\,]$, let me show you one reason for dropping it! It can be shown that $X \sim f_X$, and $Y = x^2$, Then

   $$f_Y(t) = \begin{cases} \frac{1}{2\sqrt{t}}\left(f_X(\sqrt{t}) + f_X(-\sqrt{t})\right), & t > 0 \\ 0, & \text{else} \end{cases} \tag{1}$$

   You can (but don't have to) confirm that $\int\limits_{-\infty}^{\infty} f_Y(t)dt = 1$. Let's play with the following 3 expectations:

   $$E_X\left[X^2\right], \quad E_{X^2}[X], \quad E_Y[Y]$$

   The middle one can only mean $E_{X^2}[X] = E_Y[\pm\sqrt{Y}] = \pm\int\limits_0^{\infty} f_Y(t)dt$. We can work it out, but it's not too interesting. Instead, substitute (1) into the definition of $E_Y[Y]$, to show that $E_Y[Y] = E_X\left[X^2\right]$.

To show that $\int\limits_{-\infty}^{\infty} f_Y(t)dt = 1$, we will make the following substitutions: $u = \sqrt{t} \implies du = \frac{1}{2\sqrt{t}}dt$, $v = -u \implies du = -dv$. Thus we can see that

$$\begin{aligned}
\int\limits_{-\infty}^{\infty} f_Y(t)dt &= \int\limits_0^{\infty} f_Y(t)dt \\
&= \int\limits_0^{\infty} \frac{1}{2\sqrt{t}}\left(f_X(\sqrt{t}) + f_X(-\sqrt{t})\right)dt \\
&= \int\limits_0^{\infty} f_X(u)du - \int\limits_0^{-\infty} f_X(v)dv \\
&= \int\limits_0^{\infty} f_X(u)du + \int\limits_{-\infty}^{0} f_X(v)dv \\
&= \int\limits_{-\infty}^{\infty} f_X(x)dx \\
&= 1
\end{aligned}$$

Now we will show that, $E_Y[Y] = E_X\left[X^2\right]$.

$$\begin{aligned}
E_Y[Y] &= \int\limits_{-\infty}^{\infty} t f_Y(t)dt \\
&= \int\limits_{0}^{\infty} \frac{t}{2\sqrt{t}}\left(f_X(\sqrt{t}) + f_X(-\sqrt{t})\right)dt \\
&= \int\limits_{0}^{\infty} \frac{t}{2\sqrt{t}} f_X(\sqrt{t})dt + \int\limits_{0}^{\infty} \frac{t}{2\sqrt{t}} f_X(-\sqrt{t})dt
\end{aligned}$$

We will now make the following substitutions: $u = \sqrt{t} \implies du = \frac{1}{2\sqrt{t}}dt$ and $v = -\sqrt{t} \implies dv = \frac{-1}{2\sqrt{t}}dt$. Furthermore, we can do some algebra and see that $t = u^2$ and $t = v^2$, respectively. Going back to the derivation we see that

$$\begin{aligned}
E_Y[Y] &= \int\limits_{0}^{\infty} u^2 f_X(u)du - \int\limits_{0}^{-\infty} v^2 f_X(v)dv \\
&= \int\limits_{0}^{\infty} u^2 f_X(u)du + \int\limits_{-\infty}^{0} v^2 f_X(v)dv \\
&= \int\limits_{-\infty}^{\infty} x^2 f_X(x)dx \\
&= E_X[X^2]
\end{aligned}$$

2. Suppose $Y \sim N(\mu, \sigma)$, i.e. $f_Y(t) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{\frac{-1}{2}\left(\frac{t-\mu_Y}{\sigma_Y}\right)^2}$, $-\infty < t < \infty$. Use only the definition of $E_Y[g(Y)]$ given above to find $\mu_y \equiv E_Y[Y]$ and $E_Y\left[Y^2\right]$. Hint: use $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-1}{2}x^2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{\frac{-1}{2}x^2}dx = 1$, $\int\limits_{-\infty}^{\infty} xe^{\frac{-1}{2}x^2}dx = 0$.

a. $\mu_y \equiv E_Y[Y]$.

$$\begin{aligned}
E_Y[Y] &= \int\limits_{-\infty}^{\infty} t f_Y(t)dt \\
&= \int\limits_{-\infty}^{\infty} \frac{t}{\sqrt{2\pi\sigma_Y^2}} e^{\frac{-1}{2}\left(\frac{t-\mu_Y}{\sigma_Y}\right)^2}
\end{aligned}$$

If we let $u = \frac{t-\mu_Y}{\sigma_Y} \implies du = \frac{1}{\sigma_Y}dt$. Also, we can see that $t = \sigma_Y u + \mu_Y$. Going back to the main proof we can see that

$$\begin{aligned}
E_Y[Y] &= \int\limits_{-\infty}^{\infty} \frac{\sigma_Y u + \mu_y}{\sqrt{2\pi}} e^{\frac{-1}{2}u^2}du \\
&= \frac{\sigma_Y}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} ue^{\frac{-1}{2}u^2}du + \frac{\mu_Y}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} e^{\frac{-1}{2}u^2}du
\end{aligned}$$

As given in the hint, $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-1}{2}x^2}dx = 1$ and $\int\limits_{-\infty}^{\infty} xe^{\frac{-1}{2}x^2}dx = 0$. Hence we can see that

$$\begin{aligned} E_Y[Y] &= \frac{\sigma_Y}{\sqrt{2\pi}} \cdot 0 + \mu_Y \cdot 1 \\ &= \mu_Y \end{aligned}$$

b. $E_Y\left[Y^2\right]$.

$$\begin{aligned} E_Y\left[Y^2\right] &= \int\limits_{-\infty}^{\infty} t^2 f_Y(t)dt \\ &= \int\limits_{-\infty}^{\infty} \frac{t^2}{\sqrt{2\pi\sigma_Y^2}} e^{\frac{-1}{2}\left(\frac{t-\mu_Y}{\sigma_Y}\right)^2} \end{aligned}$$

If we let $u = \frac{t-\mu_Y}{\sigma_Y} \implies du = \frac{1}{\sigma_Y}dt$. Also, we can see that $t = \sigma_Y u + \mu_Y \implies t^2 = (\sigma_Y u + \mu_Y)^2 = \sigma_Y^2 u^2 + \mu_Y^2 + 2\sigma_Y\mu_Y u$. Going back to the main proof we can see that

$$\begin{aligned} E_Y[Y] &= \int\limits_{-\infty}^{\infty} \frac{\sigma_Y^2 u^2 + \mu_Y^2 + 2\sigma_Y\mu_Y u}{\sqrt{2\pi}} e^{\frac{-1}{2}u^2} du \\ &= \frac{\sigma_Y^2}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} u^2 e^{\frac{-1}{2}u^2} du + \frac{\mu_Y^2}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} e^{\frac{-1}{2}^2} du + \frac{2\sigma_Y\mu_Y}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} ue^{\frac{-1}{2}^2} du \end{aligned}$$

As given in the hint, $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-1}{2}x^2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{\frac{-1}{2}x^2}dx = 1$, and $\int\limits_{-\infty}^{\infty} xe^{\frac{-1}{2}x^2}dx = 0$. Hence we can see that

$$\begin{aligned} E_Y[Y] &= \sigma_Y^2 \cdot 1 + \mu_Y^2 \cdot 1 + \frac{2\sigma_Y\mu_y}{\sqrt{2\pi}} \cdot 0 \\ &= \sigma_Y^2 + \mu_Y^2 \end{aligned}$$

3. Starting from the integral definition of V and Cov, show the following. In your proof **clearly** show the subscripts on V, Cov, and the pdf; don't skip any steps.

a. $V[X_1 + X_2] = V[X_1] + V[X_2] + 2Cov[X_1, X_2]$.

$$\begin{aligned}
V[X_1 + X_2] &= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \left(t_1 + t_2 - E_{X_1,X_2}[X_1 + X_2]\right)^2 f_{X_1,X_2}(t_1,t_2)dt_1dt_2 \\
&= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \left(t_1 + t_2 - E_{X_1}[X_1] - E_{X_2}[X_2]\right)^2 f_{X_1,X_2}(t_1,t_2)dt_1dt_2 \\
&= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \left((t_1 - E_{X_1}[X_1]) + (t_2 - E_{X_2}[X_2])\right)^2 f_{X_1,X_2}(t_1,t_2)dt_1dt_2 \\
&= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \Big(\left(t_1 - E_{X_1}[X_1]\right)^2 + \left(t_2 - E_{X_2}[X_2]\right)^2 + 2\left(t_1 + -E_{X_1}[X_1]\right)\left(t_2 - E_{X_2}[X_2]\right)\Big) f_{X_1,X_2}(t_1,t_2)dt_1dt_2 \\
&= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \left(t_1 - E_{X_1}[X_1]\right)^2 f_{X_1,X_2}(t_1,t_2)dt_1dt_2 + \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \left(t_2 - E_{X_2}[X_2]\right)^2 f_{X_1,X_2}(t_1,t_2)dt_1dt_2 \\
&\quad + 2\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \left(t_1 + -E_{X_1}[X_1]\right)\left(t_2 - E_{X_2}[X_2]\right) f_{X_1,X_2}(t_1,t_2)dt_1dt_2 \\
&= \int\limits_{-\infty}^{\infty} \left(t_1 - E_{X_1}[X_1]\right)^2 \left(\int\limits_{-\infty}^{\infty} f_{X_1,X_2}(t_1,t_2)dt_2\right) dt_1 + \int\limits_{-\infty}^{\infty} \left(t_2 - E_{X_2}[X_2]\right)^2 \left(\int\limits_{-\infty}^{\infty} f_{X_1,X_2}(t_1,t_2)dt_1\right) dt_2 \\
&\quad + 2\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \left(t_1 + -E_{X_1}[X_1]\right)\left(t_2 - E_{X_2}[X_2]\right) f_{X_1,X_2}(t_1,t_2)dt_1dt_2 \\
&= \int\limits_{-\infty}^{\infty} \left(t_1 - E_{X_1}[X_1]\right)^2 f_{X_1}(t_1)dt_1 + \int\limits_{-\infty}^{\infty} \left(t_2 - E_{X_2}[X_2]\right)^2 f_{X_2}(t_2)dt_2 \\
&\quad + 2\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \left(t_1 + -E_{X_1}[X_1]\right)\left(t_2 - E_{X_2}[X_2]\right) f_{X_1,X_2}(t_1,t_2)dt_1dt_2 \quad \text{(Using marginalization)} \\
&= V[X_1] + V[X_2] + 2Cov[X_1, X_2]
\end{aligned}$$

b. $Cov[X_1, X_2 + X_3] = Cov[X_1, X_2] + Cov[X_1, X_3]$

$$\begin{aligned}
Cov[X_1, X_2 + X_3] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(t_1 - E_{X_1}[X_1])(t_2 + t_3 - E_{X_2,X_3}[X_2 + X_3])f_{X_1,X_2,X_3}(t_1, t_2, t_3)dt_1dt_2dt_3 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(t_1 - E_{X_1}[X_1])(t_2 + t_3 - E_{X_2}[X_2] - E_{X_3}[X_3])f_{X_1,X_2,X_3}(t_1, t_2, t_3)dt_1dt_2dt_3 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(t_1 - E_{X_1}[X_1])((t_2 - E_{X_2}[X_2]) + (t_3 - E_{X_3}[X_3]))f_{X_1,X_2,X_3}(t_1, t_2, t_3)dt_1dt_2dt_3 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(t_1 - E_{X_1}[X_1])(t_2 - E_{X_2}[X_2])f_{X_1,X_2,X_3}(t_1, t_2, t_3)dt_1dt_2dt_3 \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(t_1 - E_{X_1}[X_1])(t_3 - E_{X_3}[X_3])f_{X_1,X_2,X_3}(t_1, t_2, t_3)dt_1dt_2dt_3 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(t_1 - E_{X_1}[X_1])(t_2 - E_{X_2}[X_2])\left(\int_{-\infty}^{\infty} f_{X_1,X_2,X_3}(t_1, t_2, t_3)dt_3\right)dt_1dt_2 \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(t_1 - E_{X_1}[X_1])(t_3 - E_{X_3}[X_3])\left(\int_{-\infty}^{\infty} f_{X_1,X_2,X_3}(t_1, t_2, t_3)dt_2\right)dt_1dt_3 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(t_1 - E_{X_1}[X_1])(t_2 - E_{X_2}[X_2])f_{X_1,X_2}(t_1, t_2)dt_1dt_2 \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(t_1 - E_{X_1}[X_1])(t_3 - E_{X_3}[X_3])f_{X_1,X_3}(t_1, t_3)dt_1dt_3 \quad \text{(Using marginalization)} \\
&= Cov[X_1, X_2] + Cov[X_1, X_3]
\end{aligned}$$

4. Let Y_i be iid with mean μ_Y and variance σ_Y^2. Show $Z_i \equiv \frac{Y_i - \mu_Y}{\sigma_Y}$ satisfies the following properties. Use the properties $E[\,]$ and $V[\,]$ listed in the lecture. You may also drop the subscripts. Note that the assumption of normality is NOT made here. Show that $E[Z_i] = 0$, $V[Z_i] = 1$, $E\left[\sum_i (Z_i - \bar{Z})^2\right] = n-1$. Hint for c: use $E[S_Z^2] = \sigma_Z^2$, otherwise, it's too hard.

a. $E[Z_i] = 0$.

$$\begin{aligned}
E[Z_i] &= E\left[\frac{Y_i - \mu_Y}{\sigma_Y}\right] \\
&= \frac{1}{\sigma_Y}E[Y_i - \mu_Y] \\
&= \frac{1}{\sigma_Y}E[Y_i] - \frac{1}{\sigma_Y}E[\mu_Y] \\
&= \frac{1}{\sigma_Y}E[Y] - \frac{1}{\sigma_Y}E[\mu_Y] \quad \text{(Since the } Y_i\text{'s are iid)} \\
&= \frac{1}{\sigma_Y}\mu_Y - \frac{1}{\sigma_Y}\mu_Y \\
&= 0
\end{aligned}$$

b. $V[Z_i] = 1$.

$$\begin{aligned} V[Z_i] &= V\left[\frac{Y_i - \mu_Y}{\sigma_Y}\right] \\ &= \frac{1}{\sigma_Y^2} V[Y_i - \mu_Y] \\ &= \frac{1}{\sigma_Y^2}\left(V[Y_i] + V[\mu_y] - 2Cov[Y_i, \mu_Y]\right) \\ &= \frac{1}{\sigma_Y^2}\left(V[Y_i] + 0 - 2Cov[Y_i, \mu_Y]\right) \\ &= \frac{1}{\sigma_Y^2}(V[Y_i] + 0 - 2 \cdot 0) \\ &= \frac{V[Y_i]}{\sigma_Y^2} \\ &= \frac{V[Y]}{\sigma_Y^2} \quad \text{(Since the } Y_i\text{'s are iid)} \\ &= \frac{\sigma_Y^2}{\sigma_Y^2} \\ &= 1 \end{aligned}$$

c. $E\left[\sum_i (Z_i - \bar{Z})^2\right] = n - 1$.

Since $S_Z^2 = \frac{\sum_i (Z_i - \bar{Z})^2}{n-1}$ it follows that $\sum_i (Z_i - \bar{Z})^2 = (n-1)S_Z^2$. Hence we can see that

$$\begin{aligned} E\left[\sum_i (Z_i - \bar{Z})^2\right] &= E\left[(n-1)\frac{\sum_i (Z_i - \bar{Z})^2}{n-1}\right] \\ &= E[(n-1)S_Z^2] \\ &= (n-1)E[S_Z^2] \\ &= (n-1)\sigma_Z^2 \\ &= (n-1) \cdot 1 \quad \text{(Since } V[Z_i] = 1\text{)} \\ &= n - 1 \end{aligned}$$

Lecture 5 problems:

1. Here is yet another way of finding $E[SS]$. Show the steps that say “show,” and clearly specify whether you are using the assumption of identically distributed, or the assumption of independence, separately.

All parts that need to be shown are included below the main proof.

$$\begin{aligned}
E\left[\sum_i (y_i - \bar{y})^2\right] &= \sum_i E\left[(y_i - \bar{y})^2\right] \\
&= \sum_i \left(E\left[y_i^2\right] - 2E[y_i\bar{y} + E\left[\bar{y}^2\right]\right) \\
&= \sum_i \left(E\left[y_i^2\right] - \frac{2}{n}\left(E\left[y_i^2\right] + (n-1)\mu_y^2\right) + E\left[\bar{y}^2\right]\right) \\
&= \sum_i \left(\left(1 - \frac{2}{n}\right) E\left[y_i^2\right] - \frac{2(n-1)}{n}\mu_y^2 + E\left[\bar{y}^2\right]\right) \\
&= \sum_i \left(\left(1 - \frac{2}{n}\right)\left(\sigma_y^2 + \mu_y^2\right) - \frac{2(n-1)}{n}\mu_y^2 + \left(\frac{\sigma_y^2}{n} + \mu_y^2\right)\right) \\
&= n\left[\sigma_y^2\left(1 - \frac{2}{n} + \frac{1}{n}\right) + \mu_y^2\left(1 - \frac{2}{n} - \frac{2(n-1)}{n} + 1\right)\right] \\
&= n\left(1 - \frac{1}{n}\right)\sigma_y^2 \\
&= (n-1)\sigma_y^2
\end{aligned}$$

Show 1:

$$\begin{aligned}
\sum_i E\left[(y_i - \bar{y})^2\right] &= \sum_i E\left[y_i^2 - 2y_i\bar{y} + \bar{y}^2\right] \\
&= \sum_i \left(E\left[y_i^2\right] - 2E[y_i\bar{y}] + E\left[\bar{y}^2\right]\right)
\end{aligned}$$

Show 2:

$$\begin{aligned}
E[y_i\bar{y}] &= E\left[y_i \frac{1}{n}\sum_{j=1}^{n} y_j\right] \\
&= \frac{1}{n}\sum_{j=1}^{n} E[y_i y_j] \\
&= \frac{1}{n}\left(E[y_i y_i] + \sum_{j=1, j\neq i}^{n} E[y_i y_j]\right) \\
&= \frac{1}{n}\left(E[y_i^2] + (n-1)E[y_i y_j]\right) \\
&= \frac{1}{n}\left(E[y_i^2] + (n-1)E[y_i]E[y_j]\right) \quad \text{(Assuming indepencence of } y_i \text{ and } y_j\text{)} \\
&= \frac{1}{n}\left(E[y_i^2] + (n-1)E[y]E[y]\right) \quad \text{(Assuming the } y_i\text{'s and } y_j\text{'s are identically distributed)} \\
&= \frac{1}{n}\left(E[y_i^2] + (n-1)\mu_y^2\right)
\end{aligned}$$

Show 3:

$$\begin{aligned}
E\left[y_i^2\right] &= V[y_i] + E[y_i]^2 \\
&= V[y] + E[y]^2 \quad \text{(Assuming the } y_i\text{'s are identically distributed)} \\
&= \sigma_y^2 + \mu_y^2
\end{aligned}$$

Show 4:

$$\begin{aligned} E\left[\bar{y}^2\right] &= V[\bar{y}] + E[\bar{y}]^2 \\ &= \frac{\sigma_y^2}{n} + \mu_y^2 \quad \text{(Proved in Lecture 4, assuming iid)} \end{aligned}$$

2. In one of the proofs above, we started with $\sum_i (y_i - \mu_y)^2$, "injected" $\bar{y}$, and decomposed in order to get $\sum_i (y_i - \bar{y})^2$. Now, look what would happen if we started with $\sum_i (y_i - \bar{y})^2$, injected μ_y, …

$$\begin{aligned} \frac{1}{\sigma_y^2}\sum_i (y_i - \bar{y})^2 &= \frac{1}{\sigma_y^2}\sum_i (y_i - \bar{y} + \mu_y - \mu_y)^2 \\ &= \frac{1}{\sigma_y^2}\sum_i (y_i - \mu_y)^2 + \frac{1}{\sigma_y^2}\sum_i (\bar{y} - \mu_y)^2 + \frac{2}{\sigma_y^2}\sum_i (y_i - \mu_y)(\bar{y} - \mu_y) \\ &= \frac{1}{\sigma_y^2}\sum_i (y_i - \mu_y)^2 + \left(\frac{\bar{y} - \mu_y}{\sigma_y/\sqrt{n}}\right)^2 + \frac{2}{\sigma_y^2}(\bar{y} - \mu_y)\sum_i (y_i - \mu_y) \\ &= \chi_n^2 + \chi_1^2 + 0 \\ \therefore \frac{1}{\sigma_y^2}\sum_i (y_i - \bar{y})^2 &\sim \chi_{n+1}^2 \end{aligned}$$

The conclusion is clearly wrong. Point out where is/are the fallacies.

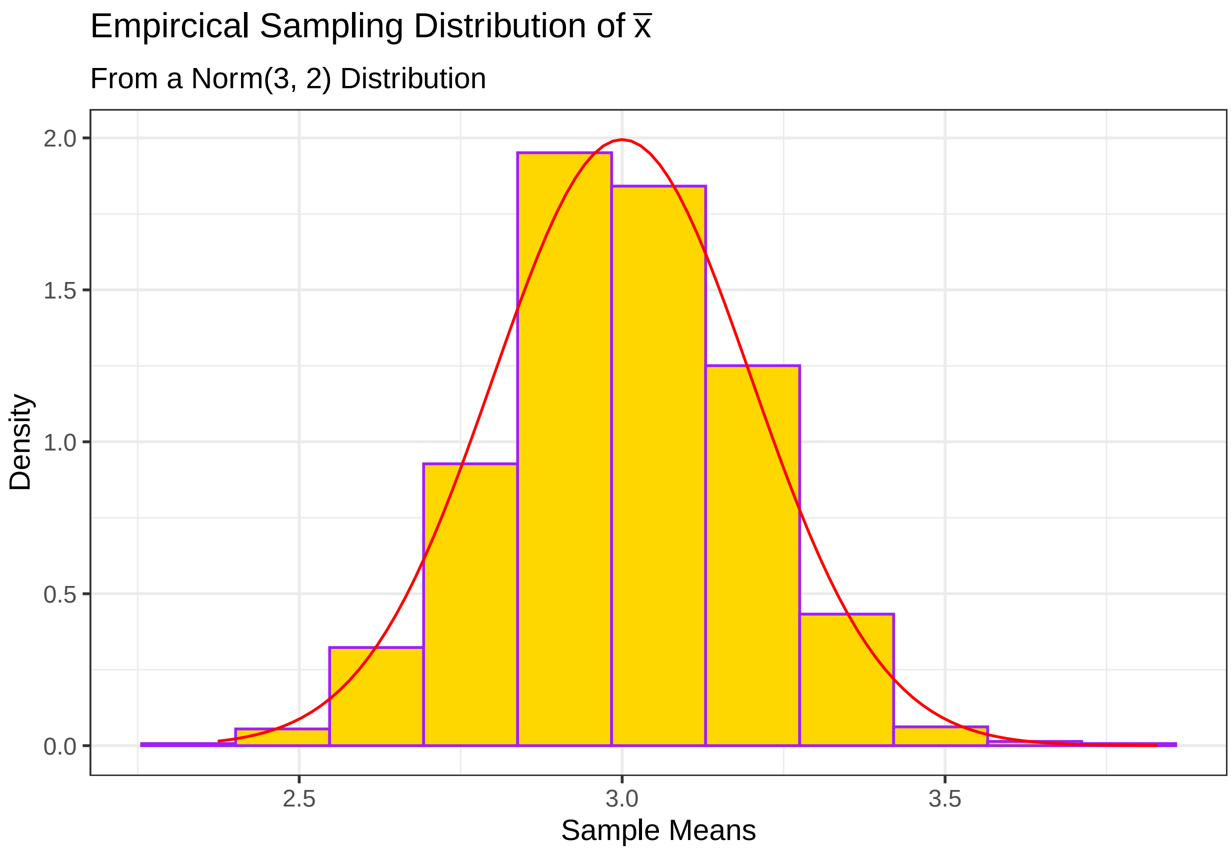
The one and only fallacy in this proof is the statement that $\sum_i (y_i - \mu_y) = 0$. This is a misuse of the fact that $\sum_i (y_i - \bar{y}) = 0$. This is a fallacy, since there are cases in which $\sum_i (y_i - \mu_y) \neq 0$. This inequality is simply due to the fact that μ_y is the population mean, and isn't controlled by samples, unlike $\bar{y}$.

3. In this hw, we will simulate some of the results mentioned above, in R.

a. First, the central limit theorem: Write code to draw 1000 samples of size n=100 from a normal distribution with $\mu = 3$, $\sigma = 2$, make a relative frequency (on density scale) histogram of the 1000 sample means, and superimpose on the histogram the pdf of a normal distribution with mean μ and standard deviation $\sigma/\sqrt{100}$. The histogram is called the *empirical sampling distribution* of sample means. The two should agree. Note: density scale histograms have a total area of 1 under them.

```r
set.seed(123)

# Number of simulations:
B <- 1000

samp_sim <- lapply(1:B, FUN = function(i) {
  # Make sample:
  sample_1 <- rnorm(100, 3, 2)

  # Create data frame:
  data.frame(ybar = mean(sample_1))
})

# Turn the data into 1 big data frame:
samp_sim_df <- do.call(rbind, samp_sim)

# Histogram with density superimposed (number of bins calculated using Sturges
# rule):
```

```r
emp_dist_1 <- ggplot(data = samp_sim_df) +
  geom_histogram(mapping = aes(x = ybar, y = ..density..), bins = 11,
                 color = "purple", fill = "gold") +
  stat_function(fun = dnorm, args = list(mean = 3, sd = 1/5), color = "red") +
  labs(title = expression("Empircical Sampling Distribution of" ~ bar(x)),
       subtitle = "From a Norm(3, 2) Distribution",
       x = "Sample Means",
       y = "Density") +
  theme_bw()

emp_dist_1
```

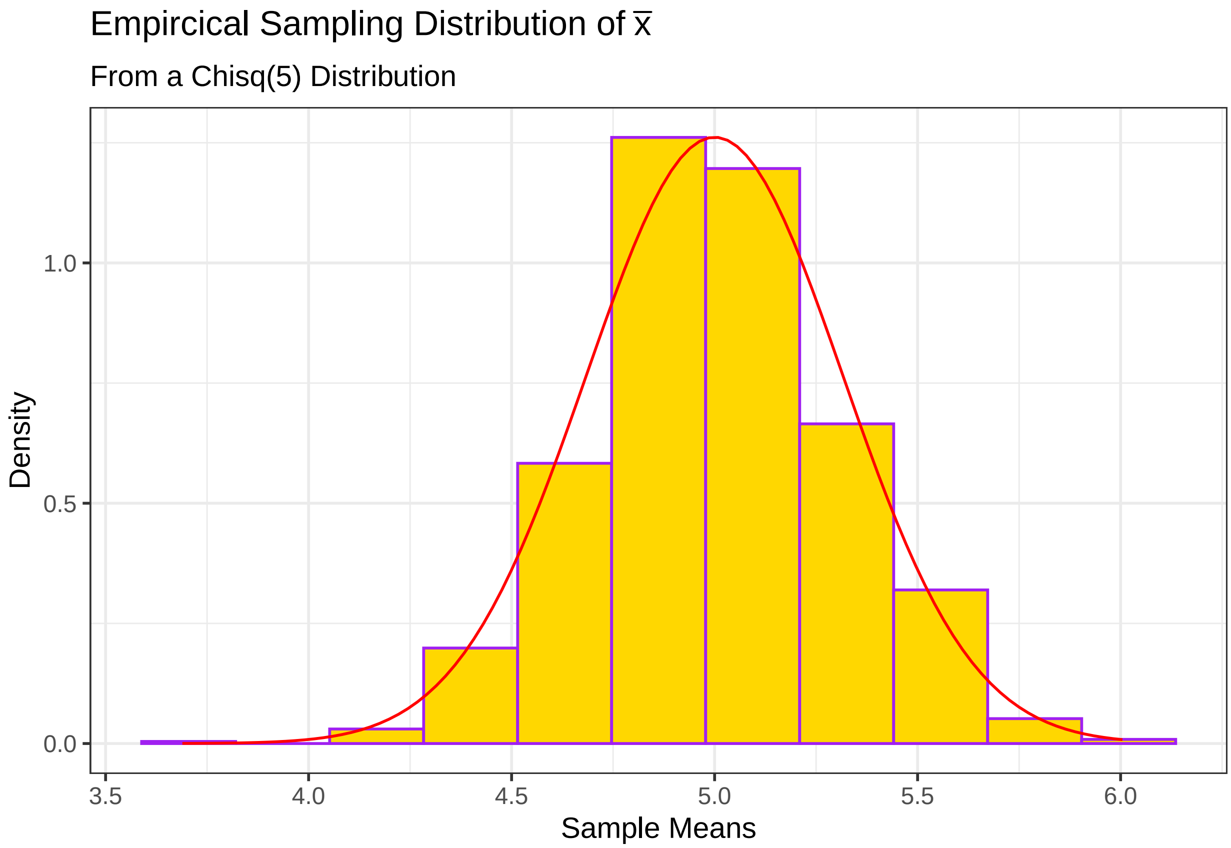


b. Repeat part a, but taking sample from a chi-squared distribution with 5 degrees of freedom. Recall that for a chi-squared distribution with df degrees of freedom, the expected value and variance are df and 2*df, respectively. Even though the population (i.e., the chi-squared distribution) is not normal at all, the empirical sampling distribution of the sample means is still normal with mean μ and standard deviation $\sigma/\sqrt{n}$.

Using the information above, $\mu_Y = 5$, $\sigma_Y = \sqrt{10}$. Thus it follows that $\bar{Y} \sim Norm(5, \sqrt{1/10})$.

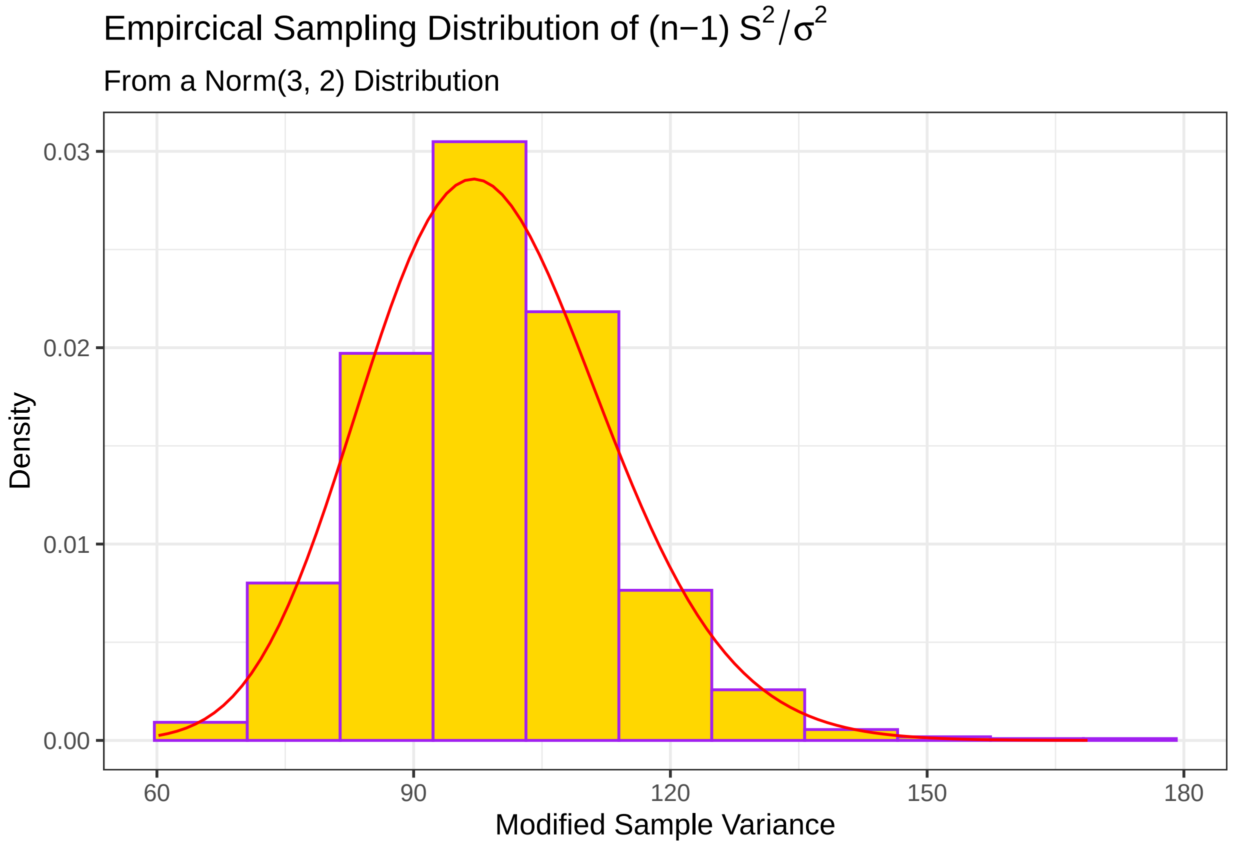
```r
set.seed(123)

# Number of simulations:
B <- 1000
```

```r
samp_sim_2 <- lapply(1:B, FUN = function(i) {
  # Make sample:
  sample_2 <- rchisq(100, 5)

  # Create data frame:
  data.frame(ybar = mean(sample_2))
})

# Turn the data into 1 big data frame:
samp_sim_df_2 <- do.call(rbind, samp_sim_2)

# Histogram with density superimposed (number of bins calculated using Sturges
# rule):
emp_dist_2 <- ggplot(data = samp_sim_df_2) +
  geom_histogram(mapping = aes(x = ybar, y = ..density..), bins = 11,
                 color = "purple", fill = "gold") +
  stat_function(fun = dnorm, args = list(mean = 5, sd = sqrt(1/10)), color = "red") +
  labs(title = expression("Empircical Sampling Distribution of" ~ bar(x)),
       subtitle = "From a Chisq(5) Distribution",
       x = "Sample Means",
       y = "Density") +
  theme_bw()

emp_dist_2
```



c. Now, we showed that $(n-1)s^2/\sigma^2$ has a chi-squared distribution with df = n-1. So, let's simulate that result. Specifically, write code to draw 1000 samples of size n = 100 from a normal distribution with $\mu = 3$, $\sigma = 2$, make a density scale histogram of the 1000 sample variances times $(n-1)/\sigma^2$. Superimpose on the histogram the pdf of a chi-squared distribution with df = n-1. The histogram (i.e., the empirical sampling distribution of $(n-1)s^2/\sigma^2$) should agree with the pdf.

```r
set.seed(123)

# Number of simulations:
B <- 1000

samp_sim_3 <- lapply(1:B, FUN = function(i) {
  # Make sample:
  sample_3 <- rnorm(100, 3, 2)

  # Create data frame:
  data.frame(var_mod = (99/4) * var(sample_3))
})

# Turn the data into 1 big data frame:
samp_sim_df_3 <- do.call(rbind, samp_sim_3)

# Histogram with density superimposed (number of bins calculated using Sturges
# rule):
emp_dist_3 <- ggplot(data = samp_sim_df_3) +
  geom_histogram(mapping = aes(x = var_mod, y = ..density..), bins = 11,
                 color = "purple", fill = "gold") +
  stat_function(fun = dchisq, args = list(df = 99), color = "red") +
  labs(title = expression("Empircical Sampling Distribution of (n-1)" ~ S^2/sigma^2),
       subtitle = "From a Norm(3, 2) Distribution",
       x = "Modified Sample Variance",
       y = "Density") +
  theme_bw()

emp_dist_3
```



Empircical Sampling Distribution of (n−1) S^2/σ^2

From a Norm(3, 2) Distribution