Homework #5

See Canvas for the HW #5 assignment due date. Complete all of the following problems. Ideally, the theoretical problems should be answered in a Markdown cell directly underneath the question. If you don't know LaTex/Markdown, you may submit separate handwritten solutions to the theoretical problems, but please see the class scanning policy in the syllabus. Please do not turn in messy work. Computational problems should be completed in this notebook (using the `R` kernel). Computational questions may require code, plots, analysis, interpretation, etc. Working in small groups is allowed, but it is important that you make an effort to master the material and hand in your own work.

A. Theoretical Problems

A.1 [6 points]

Unfortunately, the expected value of a random variable may not be defined. For example, if θ is an angle uniformly distributed in the interval $(-\pi/2, \pi/2)$, then $X = \alpha \tan(\theta)$ has the following distribution:

$$f(x) = \frac{\alpha}{\pi} \frac{1}{x^2 + \alpha^2},$$

where $\alpha > 0$ is a constant to ensure that $\int_{-\infty}^{\infty} f(x)\, dx = 1$. Show that $E(X)$ is undefined (HINT: compute $E(X)$ using a u-substitution and show that the integral does not converge).

We need compute the integral of $xf(x)$ and demonstrate nonconvergence. The expected value is:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

and

$$f(x) = \frac{\alpha}{\pi} \frac{1}{x^2 + \alpha^2}.$$

Computing the integral:

$$E(X) = \int_{-\infty}^{\infty} x \left(\frac{\alpha}{\pi} \frac{1}{x^2 + \alpha^2} \right) dx$$

Using u-sub:

$$u = x^2 + \alpha^2$$

$$du = 2x dx$$

The intergral is now expressed using u:

$$E(X) = \frac{\alpha}{\pi} \int_{-\infty}^{\infty} \frac{x}{x^2 + \alpha^2} dx = \frac{\alpha}{\pi} \int_{-\infty}^{\infty} \frac{1}{2} \frac{1}{u} du = \frac{\alpha}{2\pi} \int_{-\infty}^{\infty} \frac{1}{u} du$$

We see that this integral is improper, since limites are negative and positive infinity. Check if both the upper and lower limits of integration lead to a finite result. This gives convergence.

The integral of $\frac{1}{u}$ from negative to positive infinity does not converge:

$$\int_{-\infty}^{\infty} \frac{1}{u} du = \ln |u| \Big|_{-\infty}^{\infty} = \ln |\infty| - \ln | - \infty| = \infty - (-\infty) = \infty + \infty$$

This is an indeterminate form, and the integral does not yield a finite value and diverges. Therefore, the expected value is also undefined

A.2 Properties of expectation

Let X be a random variable such that

- $E[X] = 5$
- $sd[X] = 2.5$

and let Y be a random variable such that

- $E[Y] = 8$,
- $sd[Y] = 2$.

Also, suppose that X and Y are not independent. In fact, $\text{Corr}[X, Y] = 0.8$.

Compute the following.

A.2(a) [3 points] $E[4X + 3Y + 1]$

$$E[4X + 3Y + 1] = 4E[X] + 3E[Y] + E[1]$$

Given:

$$E[X] = 5$$

$$E[Y] = 8$$

Plugging in the values:

$$E[4X + 3Y + 1] = 4 \cdot 5 + 3 \cdot 8 + 1$$

$$E[4X + 3Y + 1] = 20 + 24 + 1 = 45$$

So,

$$E[4X + 3Y + 1] = 45$$

A.2(b) [6 points] $sd[4X + 2Y + 2]$

To compute the standard deviation:

$$sd[4X + 2Y + 2] = (\text{Var}[4X + 2Y + 2])^{0.5} = (\text{Var}[4X + 2Y])^{0.5}$$

To find $\text{Var}[aX + bY]$, use $a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab\text{Cov}[X, Y]$

$a = 4$ and $b = 2$

$$\text{Var}[X] = 2.5^2 = 6.25$$

$$\text{Var}[Y] = 2^2 = 4$$

$$\text{Cov}[X, Y] = \text{Corr}[X, Y] * sd[X] * sd[Y] = 4$$

plugging everything in,

$$(\text{Var}[4X + 2Y])^{0.5} = \sqrt{4^2 * 6.25 + 2^2 * 4 + 2 * 4 * 2 * 4} = 13.75$$

A.2(c) [4 points] $E[3X^2 + Y]$ (*hint: think about the shortcut formula for variance.*)

$$E[X^2] = \text{Var}(X) + (E[X])^2$$

We also know that $Var[X] = 6.25$

$$E[X^2] = \text{Var}(X) + (E[X])^2 = 6.25 + 5^2 = 31.25$$

$$E[3X^2 + Y] = 31.25 * 3 + 8 = 101.75$$

A.2(d) [4 points] $E[8XY]$

$$E[XY] = \text{Cov}[X, Y] + E[X] * E[Y]$$

Since Cov = 4,

$$E[XY] = 4 + 5 * 8 = 44$$

$$E[8XY] = 8 * 44 = 352$$

A.3 [4 points] Covariance and independence

In class, we introduced covariance and the correlation coefficient to measure how strongly two random variables are related to each other. We know that two dependent random variables can be uncorrelated, but it turns out that if X and Y are independent random variables, then they are uncorrelated:
$\mathrm{Corr}[X, Y] = \mathrm{Cov}[X, Y] = 0$.

Suppose that X and Y are independent random variables. Prove that
$E[XY] = E[X]E[Y]$.

The expected value of XY:

$$E[XY] = \sum_x \sum_y xy \cdot P(X = x, Y = y)$$

Since X and Y are independent:

$$E[XY] = \sum_x \sum_y xy \cdot P(X = x) \cdot P(Y = y)$$

Separate the summation into two parts:

$$E[XY] = \left(\sum_x x \cdot P(X = x) \right) \cdot \left(\sum_y y \cdot P(Y = y) \right)$$

The first part is the E[X]:

$$E[X] = \sum_x x \cdot P(X = x)$$

The second part is E[Y]:

$$E[Y] = \sum_y y \cdot P(Y = y)$$

So,

$$E[XY] = E[X] \cdot E[Y]$$

A.4 Properties of $\bar{X}$

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where the notation $\overset{iid}{\sim}$ means *independent and identically distributed*.

A.4(a) [4 points] Show (without integration), that $sd[\bar{X}] = \frac{\sigma}{\sqrt{n}}$

Use basic properties of variances and expectations.

Mean is defined as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Variance is defined as:

$$\mathrm{Var}[\bar{X}] = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)$$

Since X_i are independent and identically distributed, and the variance of a sum of independent random variables is sum of their variances,

$$\mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}[X_i]$$

Since X_i are independent and identically distributed, variances are the same:

$$\mathrm{Var}[\bar{X}] = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

The standard deviation of X bar is the square root of variance:

$$\mathrm{sd}[\bar{X}] = \sqrt{\mathrm{Var}[\bar{X}]} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

A.4(b) [5 points] Show that $\bar{X} = \mu + \frac{\sigma}{\sqrt{n}} Z$ where $Z \sim N(0,1)$.

To show that

$$\bar{X} = \mu + \frac{\sigma}{\sqrt{n}} Z$$

, where

$$Z \sim N(0,1)$$

, use properties of the normal distribution and the definitions of the variables:

Use the definition of the sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Where

$$X_i \sim N(\mu, \sigma^2)$$

and are independent and identically distributed.

Use properties of the normal distribution:

$$\bar{X} \sim N\left(\frac{1}{n}\sum_{i=1}^{n}\mu, \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2\right)$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The standardization of a normal distribution involves subtracting the mean and dividing by the standard deviation.

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\bar{X} = \mu + \frac{\sigma}{\sqrt{n}}Z$$

A.4.(c) [7 points] Use A.4(b) to find $Var[\bar{X}^2]$. Assume that $Cov[Z, Z^2] = 0$, and that $\mathrm{Var}[Z^2] = 2$ (you don't have to prove these facts).

$$\bar{X} = \mu + \frac{\sigma}{\sqrt{n}}Z$$

$$Var[\bar{X}] = \frac{\sigma^2}{n}$$

$$Var[\bar{X}^2] = Var\left(\left(\mu + \frac{\sigma}{\sqrt{n}}Z\right)^2\right)$$

$$= Var(\mu^2) + 2\mu\frac{\sigma}{\sqrt{n}}Var(Z) + Var\left(\left(\frac{\sigma}{\sqrt{n}}Z\right)^2\right)$$

$Var(\mu^2)$ is a constant, and its variance is zero

$Var(Z) = 1$ since $Z \sim N(0,1)$

$Var\left(\left(\frac{\sigma}{\sqrt{n}}Z\right)^2\right)$, we can simplify:

$$Var\left(\left(\frac{\sigma}{\sqrt{n}}Z\right)^2\right) = \left(\frac{\sigma}{\sqrt{n}}\right)^2 Var(Z^2)$$

$$Var(Z^2) = 2$$

$$\left(\frac{\sigma}{\sqrt{n}}\right)^2 Var(Z^2) = \left(\frac{\sigma}{\sqrt{n}}\right)^2 \cdot 2$$

Combining everything:

$$Var[\bar{X}^2] = 0 + 2\mu\frac{\sigma}{\sqrt{n}} + \left(\frac{\sigma}{\sqrt{n}}\right)^2 \cdot 2$$

$$Var[\bar{X}^2] = 2\mu\frac{\sigma}{\sqrt{n}} + \frac{2\sigma^2}{n}$$

A.5 Let's go bowling

Frank and Sue are bowling partners, but they have different skill levels. In particular,

- Frank's scores are normally distributed with a mean of $\mu_F = 110$ and a variance of $\sigma_F^2 = 75$
- Sue's scores are normally distributed with a mean of $\mu_S = 140$ and a variance of $\sigma_S^2 = 60$

Let X_F be the random variable that denotes Frank's score in a game, and X_S the random variable that denotes Sue's score. Answer the questions below.

A.5(a) [6 points] What is the distribution of $\bar{X}$, the average of the two bowling

scores? (Name the distribution and compute any parameters of the distribution; you may have an unknown covariance value in one of the parameters.)

The average $\bar{X}$:

$$\bar{X} = \frac{X_F + X_S}{2}$$

Determine the distribution of $\bar{X}$. Since both X_F and X_S are normally distributed, the sum of them is also normally distributed. The mean of the sum is the sum of the means, and the variance of the sum is the sum of the variances.

$\bar{X}$ is $\mu_{\bar{X}}$ and variance is $\sigma^2_{\bar{X}}$:

$$\mu_{\bar{X}} = \mu_F + \mu_S = 110 + 140 = 250$$

$$\sigma^2_{\bar{X}} = \sigma^2_F + \sigma^2_S + 2\text{Cov}(X_F, X_S)$$

Find covariance $\text{Cov}(X_F, X_S)$. If scores are independent, $\text{Cov}(X_F, X_S) = 0$, and $\sigma^2_{\bar{X}}$

$$\sigma^2_{\bar{X}} = \sigma^2_F + \sigma^2_S = 75 + 60 = 135$$

The distribution is a normal distribution with parameters:

$$\bar{X} \sim N(250, 135)$$

A.5(b) [3 points] Based on the information you have at this point, can we say that $Var[\bar{X}] = \frac{\sigma^2_F + \sigma^2_S}{4}$? Justify your answer.

We can't say $Var[\bar{X}] = \frac{\sigma^2_F + \sigma^2_S}{4}$

The equation assumes equal weighting of Frank's and Sue's scores in the average, which is not necessarily true, and it doesn't account for if there are dependencies.

A.5(c) [5 points] Assume that Frank and Ellen's scores are independent. What is the probability that the sample average score of the group will be between 120 and 140 points?

In [ ]:
```r
mu_F <- 110
sigma_F <- sqrt(75)
mu_S <- 140
sigma_S <- sqrt(60)
n <- 2

mu_X_bar <- (mu_F + mu_S) / n
sigma_X_bar <- sqrt((sigma_F^2 + sigma_S^2) / n)
```

```
p <- pnorm(140, mean = mu_X_bar, sd = sigma_X_bar) - pnorm(120, mean = mu_X_

p
```

0.694654195690399

The probability is 69.46%

A.5(d) [5 points] Assume that Frank and Ellen's scores are independent. What is the probability that both players bowl a score less than 135?

$$\begin{aligned} P(X_F \leq 135 \text{ and } X_S \leq 135) &= P(X_F \leq 135) \cdot P(X_S \leq 135) \\ &\approx 0.25879. \end{aligned}$$

In [ ]:
```
mu_F <- 110
sigma_F <- sqrt(75)
mu_E <- 140
sigma_E <- sqrt(60)

p_F <- pnorm(135, mean = mu_F, sd = sigma_F)
p_E <- pnorm(135, mean = mu_E, sd = sigma_E)


p_total <- p_F * p_E

p_total
```

0.258797851452886

B. Computational Problems

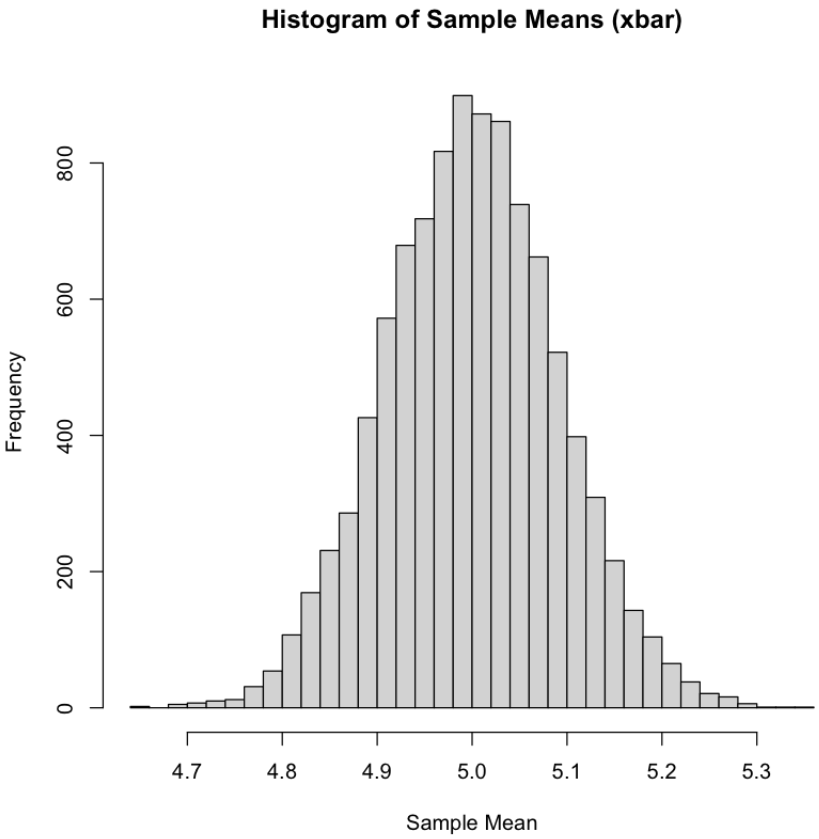
B.1

Let $X_1, X_2, \ldots, X_{500} \overset{iid}{\sim} N(\mu = 5, \sigma^2 = 4)$.

B.1(a) [7 points] Construct a matrix `x` with $m = 10,000$ columns, where each column is a different sample (i.e., set of realizations) from $X_1, X_2, \ldots, X_{500}$. Then, use `colMeans()` to find the mean of each sample/column. Store these means in `xbar`. Based on lecture, what is the distribution of `xbar`? Is a histogram of `xbar` consistent with that?

In [ ]:
```
mu <- 5
sigma <- sqrt(4)
n_samples <- 10000
n_variables <- 500
```

```r
x <- matrix(rnorm(n_samples * n_variables, mean = mu, sd = sigma), ncol = n_

xbar <- colMeans(x)

hist(xbar, breaks = 30, main = "Histogram of Sample Means (xbar)", xlab = "S
```



The distribution of xbar is normal. This is consistent with the Central Limit Theorem

B.1(b) [5 points] Numerically verify the results from A.4(a)-(c). For example, for A.4(a), show that the sample standard deviation of `xbar` is *close* to μ (say within ≈ 0.1 of the true value). How could you make `xbar` closer?

In [ ]:
```r
mu <- 5
sigma <- 2
n_samples <- 10000
n_variables <- 500

x <- matrix(rnorm(n_samples * n_variables, mean = mu, sd = sigma), ncol = n_

xbar <- colMeans(x)
sample_std <- sd(xbar)
expected_std <- sigma / sqrt(n_variables)

cat("Sample Standard Deviation:", sample_std, "\n")
cat("Expected Standard Deviation:", expected_std, "\n")
```

```
Sample Standard Deviation: 0.09046211
Expected Standard Deviation: 0.08944272
```

In [ ]:
```r
x <- matrix(rnorm(n_samples * n_variables, mean = mu, sd = sigma), ncol = n_

xbar <- colMeans(x)

std_dev <- sd(xbar)

Z <- (xbar - mu) / (sigma / sqrt(n_variables))

expected_Z <- rnorm(n_samples, mean = 0, sd = 1)

cat("Standard Deviation:", std_dev, "\n")
cat("Expected Standard Deviation:", sigma / sqrt(n_variables), "\n")
```

```
Standard Deviation: 0.08948883
Expected Standard Deviation: 0.08944272
```

In [ ]:
```r
sample_var <- var(xbar)

Z <- (xbar - mu) / (sigma / sqrt(n_variables))

sample_var_Z <- var(Z)

expected_var_Z <- 1  # Since Z ~ N(0, 1)

expected_var_sigma_Z_sq <- (sigma^2 / n_variables) * 2

expected_var_xbar_sq <- 0 + 2 * mu * (sigma / sqrt(n_variables)) + expected_

# Print the results
cat("Sample Variance of xbar:", sample_var, "\n")
cat("Expected Variance:", (sigma^2 / n_variables), "\n")
cat("Sample Variance of Z:", sample_var_Z, "\n")
cat("Expected Variance of Z :", expected_var_Z, "\n")
cat("Expected Variance of xbar^2:", expected_var_xbar_sq, "\n")
```

```
Sample Variance of xbar: 0.00800825
Expected Variance: 0.008
Sample Variance of Z: 1.001031
Expected Variance of Z : 1
Expected Variance of xbar^2: 0.9104272
```

B. 2

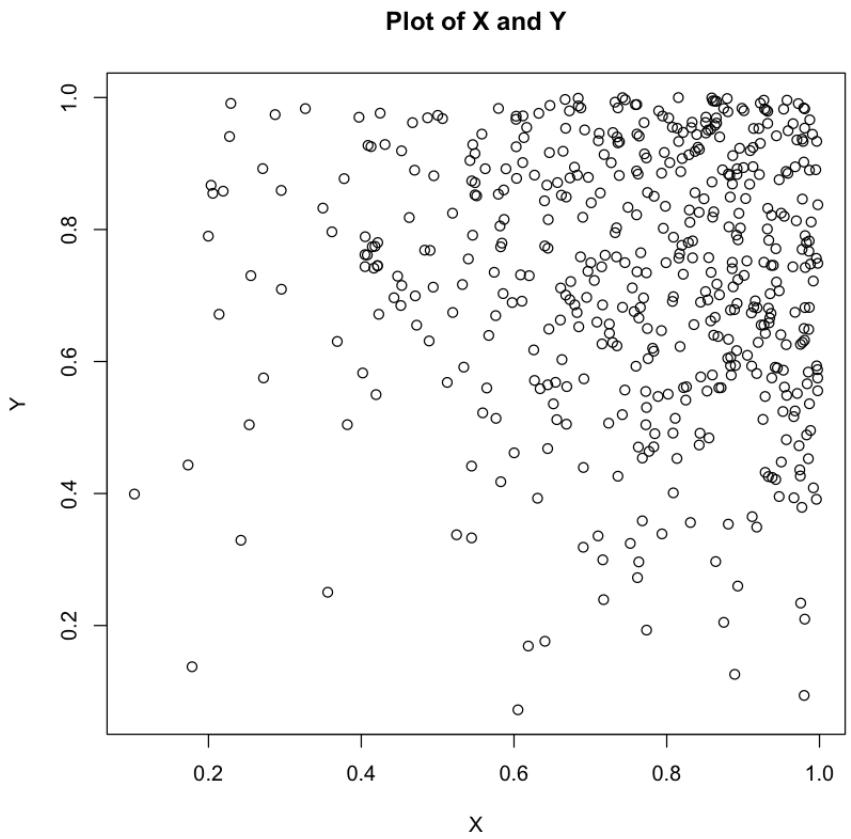
B.2(a) [4 points] Generate three random vectors of length $n = 500$ from:

1. $X \sim \text{Beta}(3, 1)$
2. $Y \sim \text{Beta}(3, 1)$
3. $Z \sim \text{Beta}(2, 3)$

Present numerical (using `cor()`) and visual evidence that X and Y are (effectively) uncorrelated.

In [ ]:
```r
n <- 500

x <- rbeta(n, 3, 1)
y <- rbeta(n, 3, 1)
z <- rbeta(n, 2, 3)

correlation_XY <- cor(x, y)

plot(x, y, main = "Plot of X and Y", xlab = "X", ylab = "Y")

correlation_XY
```

-0.0382580724958073

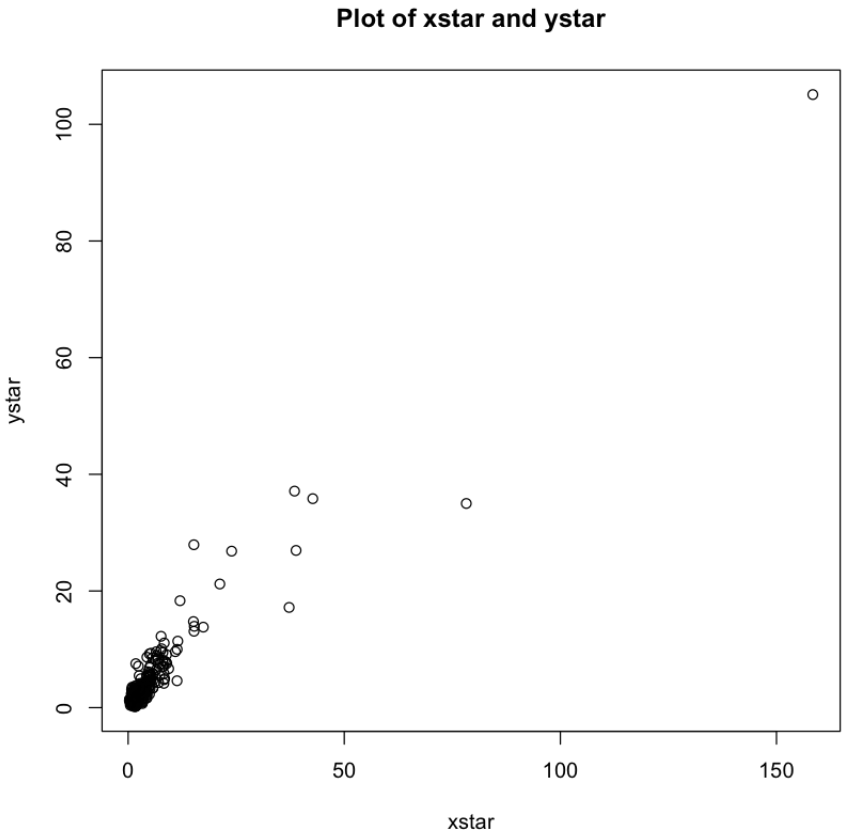


B.2(b) [4 points] Create vectors `xstar` and `ystar` by taking `x` and `y` and dividing by `z`. Are `xstar` and `ystar` correlated? Support your claim with numerical and graphical evidence.

In [ ]:
```r
xstar <- x / z
ystar <- y / z

correlation_xstar_ystar <- cor(xstar, ystar)

plot(xstar, ystar, main = "Plot of xstar and ystar", xlab = "xstar", ylab =
```

```
correlation_xstar_ystar
```

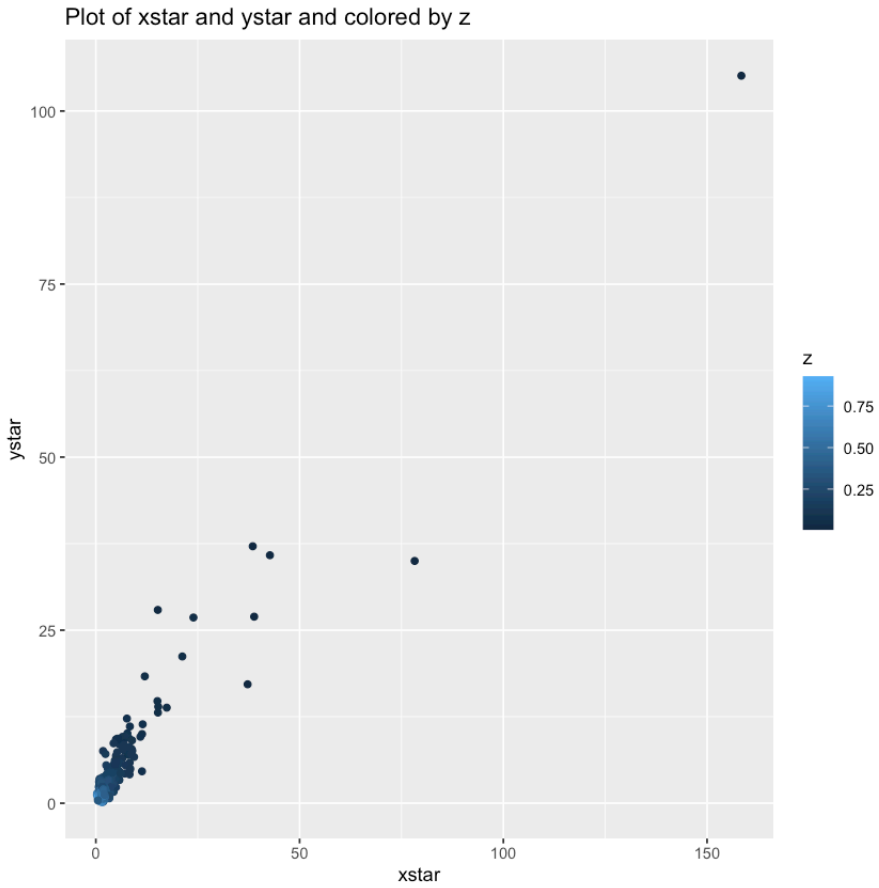
0.951658191510581



The vectors are strongly correlated.

B.2(c) [4 points] Store these vectors in a data frame called `df`. Load the `ggplot2` library, and use the `ggplot() + geom_point()` function to create a scatter plot of `xstar` and `ystar`, and color each point relative to the value of `z`.

In [ ]:
```
library(ggplot2)

df <- data.frame(xstar, ystar, z)

scatter_plot <- ggplot(df, aes(x = xstar, y = ystar, color = z)) +
  geom_point() +
  labs(title = "Plot of xstar and ystar and colored by z",
       x = "xstar",
       y = "ystar")

print(scatter_plot)
```



(d) [4 points] Interpret and analyze the results from the previous part. What practical lessons can you learn from this example?

The scatter plot with color-coding provides a visual and quantitative means to understand how multiple variables interact and influence each other.

The distribution of the data, colored by z, shows that the lighter the color (higher z value), the closer the xstar, ystar values are to origin.

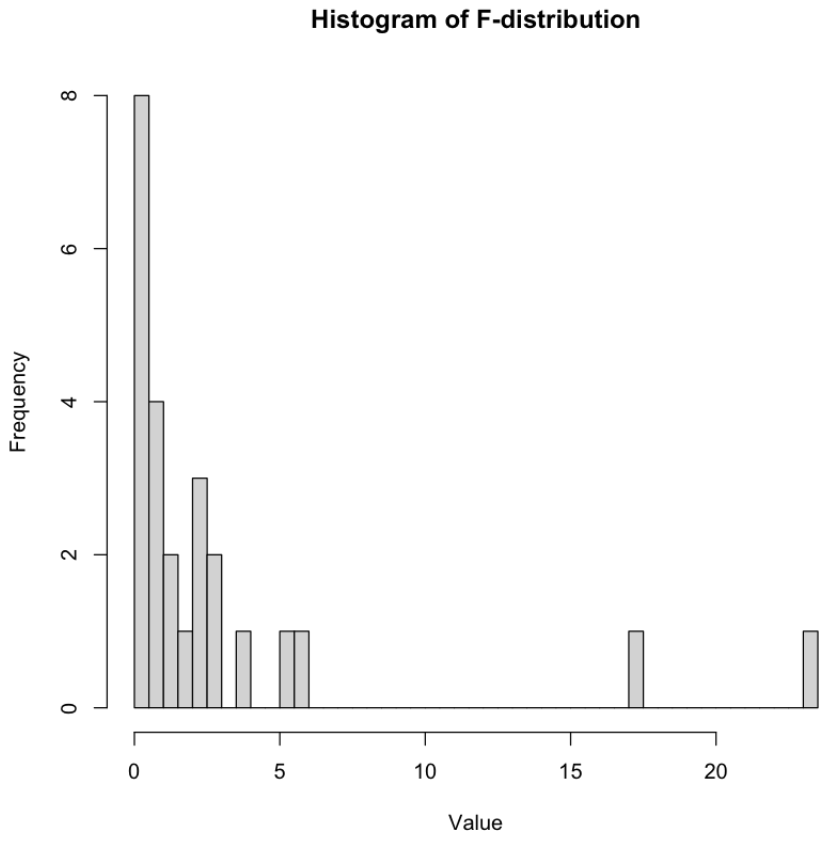
Practical lessons from this include the need to consider the context of coloring, dependencies, and variations in multivariate data analysis

B.3 The Central Limit Theorem revisited

The code `rf(n, 2, 3)` can be used to generate n random numbers from an "F-distribution" with parameters `2` and `3` (don't worry to much about what the F-distribution is or models at this point; it's just a probability distribution!).

B.3(a) [5 points] Construct a matrix `x` with $m = 10,000$ columns, where each column is a different sample of size $n = 25$ from the F-distribution with parameters `2` and `3`. Construct a histogram of one of the columns. What can you say about the shape of this distribution?

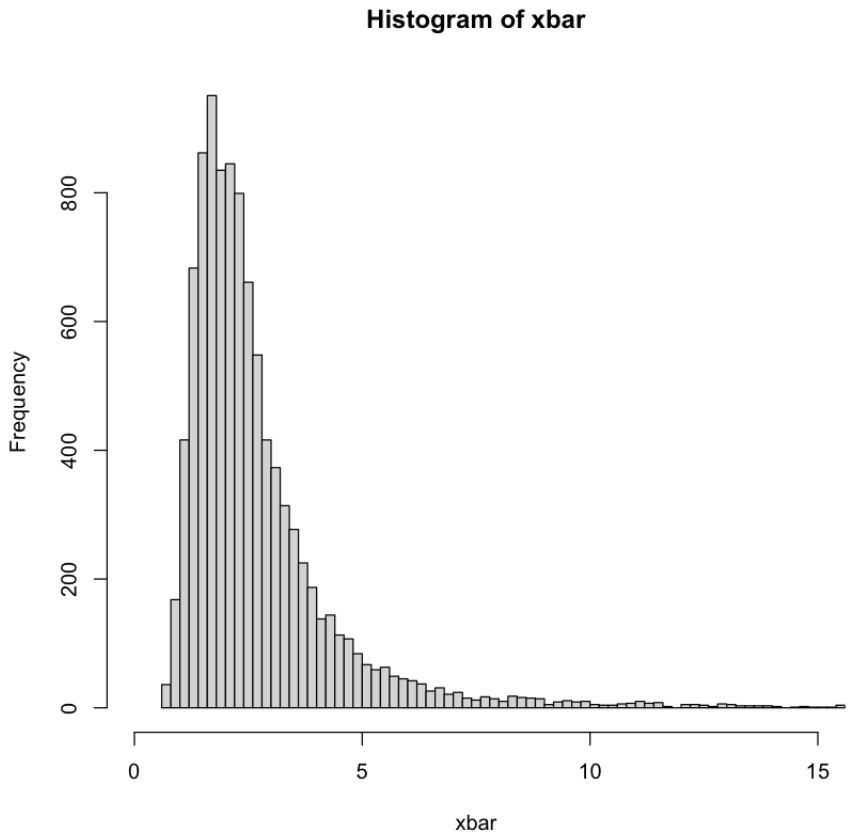
In [ ]:
```
m <- 10000
n <- 25

x <- matrix(rf(m * n, 2, 3), ncol = m)

column <- 2000

hist(x[, column], main = "Histogram of F-distribution", xlab = "Value", ylab
```



The shape of the F-distribution is right-skewed. It approaches a normal distribution as the degrees of freedom increase

B.3(b) [5 points] Use `colMeans()` to find the mean of each sample (each column of `x`). Store these means in `xbar`. Construct a histogram of `xbar`. Does the histogram look normal (or approximately normal)? If yes, why? If no, why not?

In [ ]:
```
xbar <- colMeans(x)

hist(xbar, main = "Histogram of xbar", xlim = c(0,15), breaks = 30000)
```



The histogram looks normal, but it is right skewed. This is because of the central limit theorem. Furthermore, the right skew is most likely due to the low degree of freedom.