Problem Set 1

Matthew Krasnow

Due Friday, September 13, 2024 at 11:59 pm

Problem set policies. *Please provide concise, clear answers for each question while making sure to fully explain your reasoning. For problems asking for calculations, show your work in addition to clearly indicating the final answer. For problems involving* `R`*, be sure to include the code and output in your solution.*

Please submit the PDF of your knit solutions to Gradescope and be sure to assign which pages of your solution correspond to each problem. Make sure that the PDF is fully readable to the graders; e.g., make sure that lines don't run off the page margin.

We encourage you to discuss problems with other students (and, of course, with the teaching team), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution. Be aware that copying answers found online, whether human-generated or machine-generated, is a violation of the Honor Code.

Problem 1

Review the documents in the "Review" module on Canvas on RMarkdown, matrix algebra and probability.

Problem 2

Rename this file to include your first initial and last name – e.g., stat139_fall2024_hw1_jxenakis.Rmd. Then, in this document, edit the "author" portion of the header to include your name. Click *Knit.* This should produce a .pdf file located in the same folder as the .Rmd file, with a name like stat139_fall2024_hw1_jxenakis.pdf. Note that the file name for a PDF created from an Rmd document will be the same, except with a different file extension. The file name and title of the document seen in the header can be different.

If you did not get a PDF file, stop and try to diagnose the problem by looking at the error messages in the RMarkdown section of the Console. The error messages are not always helpful, so if you cannot solve the issue, ask the teaching staff for help or talk with another student.

Problem 3

Let's do some matrix algebra by hand and in R. Let:

$$\vec{c} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \quad \vec{x} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \vec{\mu} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

Calculate each of the following by hand and write your solution using LaTeX syntax. Then confirm your calculations in an R code chunk. You will need the `matrix()` function, the `solve()` function, and the `%*%` operator for matrix multiplication. Remember, if you are unfamiliar with an R function, type `?` in the console followed by the function name to access the manual pages.

(a) $\vec{c}^T(\vec{x} - \vec{\mu})$

(b) Σ^{-1}

(c) $\vec{c}^T \Sigma \vec{c}$

(d) $\vec{c}^T(\vec{x} - \vec{\mu})(\vec{c}^T \Sigma \vec{c})^{-1/2}$

$$

% Part (a) **(a)**

$$\mathbf{C}^T(\tilde{\mathbf{x}} - \tilde{\mu}) = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \cdot \left(\begin{bmatrix} 2 \\ 3 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= 1$$

% Part (b) **(b)**

$$\Sigma^{-1} \cdot \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$ad - bc \neq 0 \quad \Rightarrow \quad 8 - 1 = 7$$

$$\Sigma^{-1} = \frac{1}{7} \cdot \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{7} \cdot \begin{bmatrix} 4 & -1 \\ -1 & 2 \end{bmatrix}$$

% Part (c) **(c)**

$$\mathbf{C}^T \Sigma \mathbf{C} = \mathbf{C}^T (\Sigma \mathbf{C})$$

$$\Sigma \mathbf{C} = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

$$= \begin{bmatrix} 3/2 \\ 5/2 \end{bmatrix}$$

$$\mathbf{C}^T (\Sigma \mathbf{C}) = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \cdot \begin{bmatrix} 3/2 \\ 5/2 \end{bmatrix}$$

$$\mathbf{C}^T \Sigma \mathbf{C} = \frac{3}{4} + \frac{5}{4}$$

$$= \frac{8}{4} = 2$$

% Part (d) **(d)**

$$\mathbf{C}^T (\tilde{\mathbf{x}} - \tilde{\mu}) \left(\mathbf{C}^T \Sigma \mathbf{C} \right)^{-1/2} = 1 \times 2^{-1/2} = \frac{1}{\sqrt{2}}$$

$$

```r
c <- matrix(c(0.5, 0.5), nrow = 2, ncol = 1)
x <- matrix(c(2, 3), nrow = 2, ncol = 1)
mu <- matrix(c(1, 2), nrow = 2, ncol = 1)
Sigma <- matrix(c(2, 1, 1, 4), nrow = 2, ncol = 2)

cat("c:\n")
```

c:

```r
print(c)
```

```
##      [,1]
## [1,]  0.5
## [2,]  0.5
```

```r
cat("\nx:\n")
```

```
## 
## x:
```

```r
print(x)
```

```
##      [,1]
## [1,]    2
## [2,]    3
```

```r
cat("\nmu:\n")
```

```
## 
## mu:
```

```r
print(mu)
```

```
##      [,1]
## [1,]    1
## [2,]    2
```

```r
cat("\nSigma:\n")
```

```
## 
## Sigma:
```

```r
print(Sigma)
```

```
##      [,1] [,2]
## [1,]    2    1
## [2,]    1    4
```

```r
# A

x_minus_mu <- x - mu

# Transpose
c_transpose <- t(c)

result <- c_transpose %*% x_minus_mu

cat("\nResult of c^T * (x - mu):\n")
```

```
## 
## Result of c^T * (x - mu):
```

```r
print(result);
```

```
##      [,1]
## [1,]    1
```

```r
# B

Sigma_inverse <- solve(Sigma)

cat("\nSigma^(-1):\n");

##
## Sigma^(-1):
print(Sigma_inverse);

##              [,1]         [,2]
## [1,]  0.5714286 -0.1428571
## [2,] -0.1428571  0.2857143
# C

c_transpose <- t(c);

step2_result <- c_transpose %*% Sigma;

final_result <- step2_result %*% c;

cat("\nResult of c^T * Sigma * c:\n");

##
## Result of c^T * Sigma * c:
print(final_result);

##      [,1]
## [1,]    2
# D
c_transpose <- t(c)
x_minus_mu <- x - mu
part_a_result <- c_transpose %*% x_minus_mu;

part_c_result <- c_transpose %*% Sigma %*% c

inverse_sqrt <- as.numeric(part_c_result)^(-1/2)

final_result <- part_a_result * inverse_sqrt

cat("\nResult of c^T(x-mu) * (c^T * Sigma * c)^(-1/2):\n")

##
## Result of c^T(x-mu) * (c^T * Sigma * c)^(-1/2):
print(final_result)
```
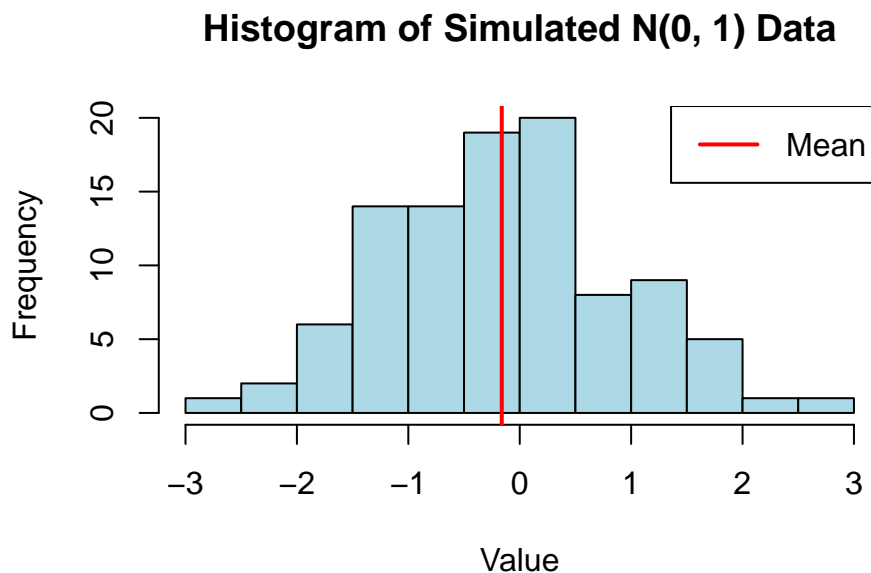
```
##              [,1]
## [1,] 0.7071068
```
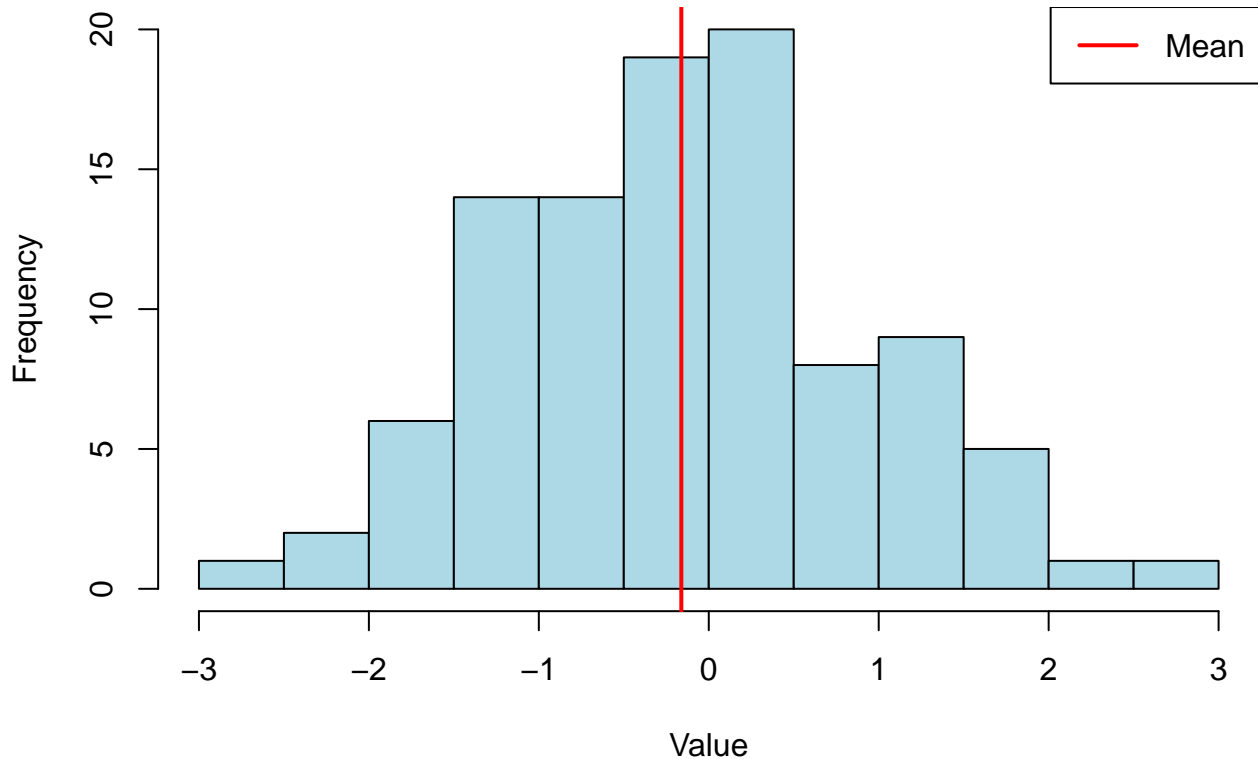
Problem 4

a) In an R code chunk, simulate a vector of length 100 from a $N(0,1)$ distribution and then plot the distribution of your simulated data as a histogram. You will need the `rnorm()` and `hist()` functions.

```r
set.seed(139)

simulated_data <- rnorm(n = 100, mean = 0, sd = 1)

hist(simulated_data,
     main = "Histogram of Simulated N(0, 1) Data",
     xlab = "Value",
     ylab = "Frequency",
     col = "lightblue",
     border = "black")

abline(v = mean(simulated_data), col = "red", lwd = 2)

legend("topright", legend = "Mean", col = "red", lwd = 2)
```

Histogram of Simulated N(0, 1) Data

b) Now repeat your R code chunk, and make a simple change in the chunk header so that only the output (i.e, the histogram) is output, not the code itself.

Histogram of Simulated N(0, 1) Data

Problem 5

a) What are two useful facts about moment generating functions (MGF's)? Use Markdown syntax to write your answer as an unordered list. Make some text in bold, and some in italics.

They can be used to characterize distributions. Often, if we are trying to prove the equality of two random variables, it is easier to prove the equality of their MGFs. 2. They can be used to generate the moments of the distribution. Note this is not always the best way to generate the moments! 3. The MGF of the sum of independnet RVs is the product of their MGFs (we'll use this one below)

b) A Markdown logo was included in this assignment as a .pdf file. Attach it here using Markdown syntax (resize it to make it visually appealing if necessary).

c) Restate the two useful facts about moment generating functions here as an **ordered** list using LaTeX syntax. Make some text in bold and some in italics.

1. The moment generating function **uniquely determines** the distribution. In other words, if two random variables have the same moment generating function, they must have the *same probability distribution.*

2. If $X_1, X_2, \ldots, X_n$ are **independent** random variables, then the moment generating function of their sum is the *product* of their individual moment generating functions:

$$M_{X_1+X_2+\ldots+X_n}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdot \ldots \cdot M_{X_n}(t)$$

d) The LaTeX logo was included in this assignment as a .png file. Attach it here using LaTeX syntax (resize it to make it visually appealing if necessary).
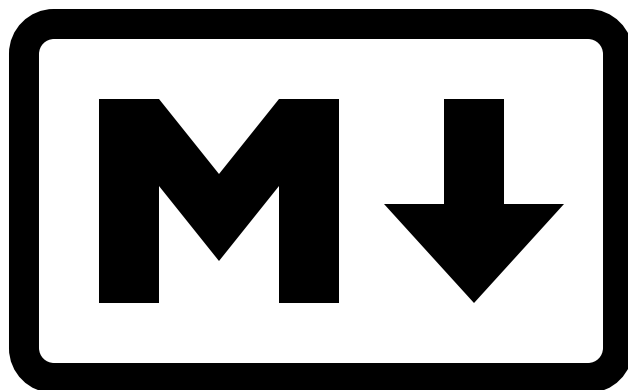
Figure 1: Markdown Logo

The remainder of the problems are a perhaps challenging review of some STAT 110 concepts that will remain important in STAT 139. But this is not meant to be a challenging probability homework, it is meant to be an introduction to RMarkdown and LaTeX. For this reason, I will post the answers to questions 6-9 in a few days and at that point you can copy them to practice your RMarkdown and LaTeX skills. Feel free to challenge yourself by trying them beforehand though.

Problem 6

Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. The sample mean is defined as

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

and the sample variance as

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

a) Why is there an $n-1$ and not an n in the denominator of the S^2?

It makes the estimator unbiased. The reason for it is that we don't know the mean of the distribution, and so we plug in an estimate. We are therefore wasting a degree of freedom by estimating that parameter, and the $n-1$ accounts for that. In the future, we will waste more degrees of freedom estimating more parameters, and that correction factor will likewise be further adjusted.

b) What is the sampling distribution of $\bar{X}$?

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

c) What is the expected value of S^2?

$$\mathbb{E}[S^2] = \sigma^2$$

d) What is the sampling distribution of S^2?

It follows a scaled chi-square distribution:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

e) What is the expected value of a χ^2 random variable with n degrees of freedom?

$$\mathbb{E}[\chi^2_n] = n$$

Problem 7

The MGF for a random variable $X \sim \chi_n^2$ random variable is defined as:

$$M_X(t) = (1 - 2t)^{-\frac{n}{2}}, \quad t < \frac{1}{2}$$

How would you have derived this if I hadn't given it to you? (No need to derive it, just set it up).

Take the following expectation:

$$\mathbb{E}(e^{tX})$$

(provided the expectation exists for t in some neighborhood around zero).

Problem 8

Now we will derive the distribution of S^2 together. Let

$$A = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \sim \chi_n^2,$$

$$B = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

and

$$C = n \frac{(\bar{X} - \mu)^2}{\sigma^2}$$

a) What are the distributions of A and C?

Note that there was a typo in the original homework. I asked for the distribution of B, but I meant C.

$$A = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2$$

$$C = \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2 \sim \chi_1^2$$

b) Can you show that $A = B + C$?

$$A = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X} + \bar{X} - \mu)^2$$

10

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} \left((X_i - \bar{X}) + (\bar{X} - \mu) \right)^2$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[(X_i - \bar{X})^2 + 2(X_i - \bar{X})(\bar{X} - \mu) + (\bar{X} - \mu)^2 \right]$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 + \frac{1}{\sigma^2} \sum_{i=1}^{n} 2(X_i - \bar{X})(\bar{X} - \mu) + \frac{1}{\sigma^2} \sum_{i=1}^{n} (\bar{X} - \mu)^2$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 + \frac{n}{\sigma^2} (\bar{X} - \mu)^2$$

$$= B + C$$

c) Assuming that B and C are independent, can you use moment generating functions (see problem 5) to show that the moment generating function of B is that of χ^2_{n-1} random variable?

Specifically, note that the MGF of a random variable $X \sim \chi^2_k$ is:

$$M_X(t) = (1 - 2t)^{-\frac{k}{2}}$$

Further, the MGF of the sum of two independent RV's is the product of the MGFs. That is:

$$M_A(t) = M_B(t) M_C(t)$$

$$\Rightarrow (1 - 2t)^{-\frac{n}{2}} = M_B(t)(1 - 2t)^{-\frac{1}{2}}$$

$$\Rightarrow M_B(t) = (1 - 2t)^{-\frac{n-1}{2}}$$

Which we recognize as the MGF of a χ^2_{n-1} distribution. That is, $B \sim \chi^2_{n-1}$.

Problem 9

a) What is the distribution (just the name of the distribution) of the vector $(\bar{X}, X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$?

Multivariate Normal, since any linear combination of its components can be written as a linear combination of $X_1, X_2, \ldots, X_n$.

b) What is the expected value of each component?

The expected value of the first component is μ. For every other component:

$$E(X_j - \bar{X}) = \mu - \mu = 0 \quad \text{(by linearity of expectation)}$$

c) If we can show that $\mathrm{Cov}(X_j, X_j - \bar{X}) = 0$ for any element j of the vector, does that prove $\bar{X}$ is independent of every element j? Why or why not?

Yes, that's a property of the MVN distribution. I actually should have asked you to prove it, it's not hard: First, note that

$$\mathrm{Cov}(X_j, X_j - \bar{X}) = \mathrm{Cov}(X_j, X_j) - \mathrm{Cov}(X_j, \bar{X})$$

$$\mathrm{Cov}(X_j, \bar{X}) = \mathrm{Cov}\left(X_j, \frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n\right)$$

$$= \mathrm{Cov}\left(X_j, \frac{1}{n}X_j\right)$$

$$= \frac{1}{n}\mathrm{Var}(X_j)$$

$$= \frac{\sigma^2}{n}$$

$$\mathrm{Cov}(X_j, \bar{X}) = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Therefore, $\mathrm{Cov}(X_j, X_j - \bar{X}) = 0$.

d) What other distributions would that be true for?

None that I can think of.

e) Now why is S^2 independent of $\bar{X}$?

We have shown that $\bar{X}$ is uncorrelated and therefore independent (that's a property of Normality!) of every element of $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$. Since S^2 is just a function of $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$, $\bar{X}$ is also independent of S^2.