Homework 1 (33 points)

INF511

Purnabhishek Sripathi Sateesh Nuthalapati Mounika Maddi

You can complete this assignment in teams of **one to three students**. However, **each student** will submit a copy of their team's completed assignment, via BbLearn. The **names of all team members** who participated on the assignment must be included in the `author` section of the `YAML`. Only one assignment per team will be selected and graded, with all team members receiving the same score, which will be recorded in BbLearn. Please ensure that all team members indicate the same names on their submitted assignments. Also note that any assignment that was completed in a team of four or more individuals will receive zero points.

You must submit this assignment as a `.qmd` file rendered as a `.pdf`. Submit both the `.qmd` and the `.pdf` to Bblearn. This will help in grading, especially to figure out if you had a bug in the code.

NOTE: All homeworks will be scaled to 100 points so that each homework is equally weighted in your grade.

1 Some R Basics

1.1 `faraway` package (1 point)

Install the `faraway` package on your laptop (*do not show the code for this*). Then, in the code chunk below, load the `faraway` package for use.

```
# Load the `faraway` package
library(faraway)
```

1.2 Working with `data.frame` objects

Here is an example data frame:

```
toy_df<- data.frame(
    name=c("Fred", "Ethyl", "Ricky", "Lucy","Babalu"),
    program=c("surfing", "surfing", "singing","singing", "dancing"),
    gpa=c(3.2, 3.4, 3.4, 3.3, 4.0),
    sat=c(1200, -999, 1300, 1250, 1600))
summary(toy_df)
```

```
    name              program                gpa            sat
Length:5           Length:5           Min.   :3.20   Min.   :-999.0
Class :character   Class :character   1st Qu.:3.30   1st Qu.:1200.0
Mode  :character   Mode  :character   Median :3.40   Median :1250.0
                                      Mean   :3.46   Mean   : 870.2
                                      3rd Qu.:3.40   3rd Qu.:1300.0
                                      Max.   :4.00   Max.   :1600.0
```

```
toy_df$gpa
```

```
[1] 3.2 3.4 3.4 3.3 4.0
```

```
mean(toy_df$gpa)
```

```
[1] 3.46
```

```
toy_df$program
```

```
[1] "surfing" "surfing" "singing" "singing" "dancing"
```

```
toy_df2 <- toy_df
class(toy_df2$program)
```

```
[1] "character"
```

```
toy_df2$program <- as.factor(toy_df2$program)
class(toy_df2$program)
```

```
[1] "factor"
```

```
levels(toy_df2$program)
```

```
[1] "dancing" "singing" "surfing"
```

1.2.1 Summarize (1 point)

Run the `summary()` function on the `toy_df` object.

1.2.2 Subsetting (2 points)

Use the `$` notation to subset the `gpa` column of the `toy_df` object, and use the `mean()` function to calculate the average of the column.

1.2.3 Levels (4 points)

You will need to use the `$` notation again to subset the `program` column of the `toy_df` object. Convert the `program` column, which is currently a `character` string, to a `factor` variable. Then, use the `levels()` function to print the levels of this new factor object.

1.3 Vectorized functions

Use the following vector to complete the sub-tasks below:

```
my_vec = seq(from=1,to=5,by=0.8)
my_vec
```

```
[1] 1.0 1.8 2.6 3.4 4.2 5.0
```

```
print(paste("length of my_vec is:",length(my_vec)))
```

```
[1] "length of my_vec is: 6"
```

```
log(my_vec)
```

```
[1] 0.0000000 0.5877867 0.9555114 1.2237754 1.4350845 1.6094379
```

1.3.1 Calculate the length of `my_vec`. (1 point)

Use a built-in function within R to output the length of `my_vec`.

1.3.2 Calculate the natural log of each element of `my_vec`. (1 point)

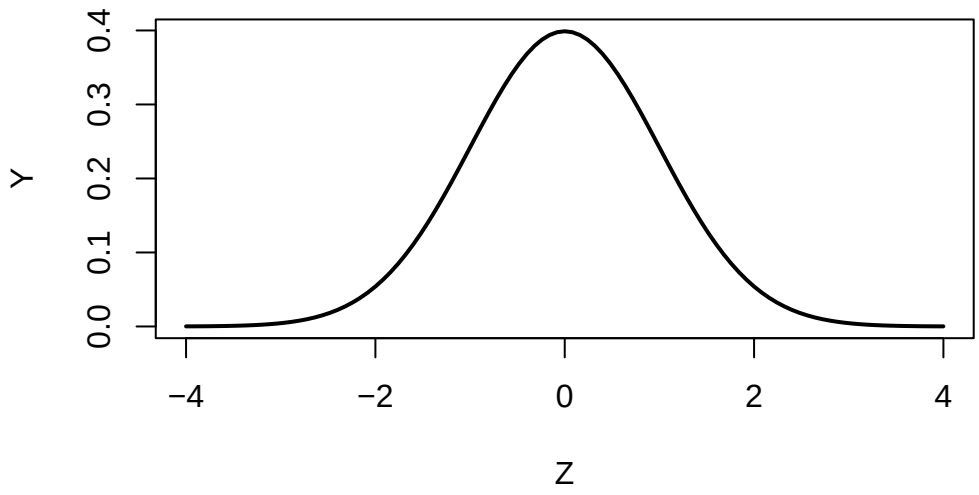
Use a built-in function within R to output the natural log of each element of `my_vec`. This should be a single function on a single line of code (i.e., a vectorized function).

2 Probability distributions

2.1 Standard normal (5 points)

Plot the probability distribution function (as a curve) that describes the "standard normal," which is the normal distribution with mean zero and standard deviation equal to one. In other words, plot $P(z|\mu = 0, \sigma = 1)$ for a range of continuous random variable z, where $z \sim N(\mu = 0, \sigma^2 = 1)$. Make sure that z ranges from -4 to 4. Label the axes appropriately.

```
curve(dnorm, -4, 4, lwd=2, axes = TRUE, xlab = "Z", ylab = "Y")
```



2.2 CDF (2 points)

Use R to calculate $P(z \leq 1.645|\mu = 0, \sigma = 1)$ (i.e., from the standard normal).

```
pnorm(1.645, mean=0, sd=1, lower.tail=TRUE)
```

```
[1] 0.9500151
```

2.3 Inverse CDF (2 points)

Use the `qnorm()` to calculate the value of z that delineates that 95% of the standard normal probability distribution falls below this value of z. This demonstrates the inverse CDF. You should see a relationship with the answer of the above question.

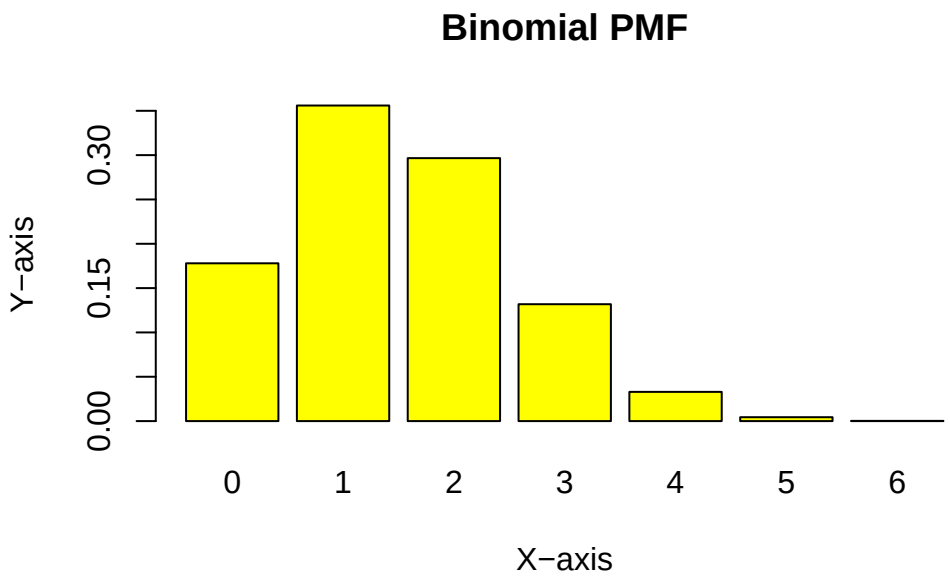
```
z <- qnorm(.95, mean=0,sd=1)
print(paste("Qnorm is:", z))
```

```
[1] "Qnorm is: 1.64485362695147"
```

2.4 Binomial distribution (5 points)

Use R to plot the binomial probability mass function with $n = 6$ (i.e., `size`=6) and $p = .25$. Because the binomial is a discrete probability distribution, your plot should be formatted similarly to the Poisson example from class. Be sure to label axes appropriately.

```
binomial_pmf <- dbinom(0:6, size=6, prob=.25)
barplot(binomial_pmf,xlab="X-axis",ylab="Y-axis",main = "Binomial PMF", names.arg=0:6,col="yellow")
```



2.5 CDF of Binomial (2 points)

Use R to compute $P(Y \geq 2)$ when $Y \sim binomial(n = 6, p = 0.25)$. Be careful with the sign of the inequality.

```
print(pbinom(1, size=6, prob=0.25))
```

```
[1] 0.5339355
```

3 Algebraic expressions

You will need to have read Dr. Barber's Appendix A to complete these tasks. Consider Y_1 and Y_2, which are *independent* random variables with means (i.e., *expectations*) equal to μ_1 and μ_2, respectively, and variances σ_1^2 and σ_2^2, respectively. Answer the following:

3.1 What is the mean of the linear expression? (2 points)

What is the mean of $2Y_1 + 5 + 8Y_2$? Use a LaTex-style equation to express your answer.

Here 'E' represent the mean.

$$2E(Y_1) + 5 + 8E(Y_2) = 2\mu_1 + 8\mu_2 + 5$$

3.2 What is the variance? (2 points)

What is the variance of $2Y_1 + 5 + 8Y_2$? Use a LaTex-style equation to express your answer.

The variance of $2Y_1 + 5 + 8Y_2$ .Here 'Var' represents the variance.

$$Var(2Y_1 + 5 + 8Y_2) = 4Var(Y_1) + 64Var(Y_2) + 2 * 2 * 8 * Cov(Y_1, Y_2)$$

3.3 What is the distribution? (2 points)

If Y_1 and Y_2 are both normally distributed, what is the distribution of the linear combination $2Y_1+5+8Y_2$? Moreover, what are the parameters that describe this distribution?

The Linear Distribution is also normally distributed and the Parameters are :

Mean:

$$2\mu_1 + 8\mu_2 + 5$$

3.4

Variance:

$$4\sigma_1 + 64\sigma_2$$

3.5 What is the covariance? (1 point)

What is the covariance between Y_1 and Y_2?

The covariance of Y_1 and Y_2 is 0. Where both Y1 and Y2 are independent random variables.