PSTAT 10 Midterm

Summer Session A 2024

For multiple choice questions, please circle the correct response unambiguously. For questions that ask you to write, please write legibly.

There are a total of 100 points to be had for the entire exam. Not all questions are worth the same number of points. Please review the point breakdown of entire exam before proceeding. It is given below. I highlighted the questions that are worth more than 4 points.

Point Breakdown

Part 1: R Programming and Data Manipulation

- **Question 1:** 4 points
- **Question 2:** 4 points
- **Question 3:** 4 points
- **Question 4:** 4 points

Part 2: Data Analysis Using the Iris Dataset

- **Question 5:** 4 points
- **Question 6:** 4 points
- **Question 7: 20 points**
- **Question 8:** 4 points
- **Question 9:** 4 points
- **Question 10: 10 points**
- **Question 11:** 4 points
- **Question 12: 6 points**
- **Question 13:** 4 points

Part 3: Probability and Simulation

- **Question 14: 12 points**
- **Question 15:** 4 points
- **Question 16:** 4 points
- **Question 17:** 4 points

Total: 100 points

Part 1: R Programming and Data Manipulation

Question 1 (4 points)

How would I access the values in the 1^{st} and 3^{rd} rows, along the 3^{rd} columnn of the below matrix? Assume the matrix is stored in a variable called `m`.

```
##      [,1] [,2] [,3] [,4]
## [1,]   10   48   22   64
## [2,]   40   83   82   99
## [3,]   67   28   55   72
## [4,]   88   19   93   63
```

A. `m[c(1,3),3]`

B. `m[1,3]`

C. `m[c(1,3,3)]`

D. `m[1,3,3]`

E. `m[3, c(1,3)]`

Question 2 (4 points)

What would the output of the below code be(what would be printed out in the right order)?

```
i <- 1
while(i <= 6) {
  if(i %% 3 == 0)  {
    print('a')
  } else if(i %% 2 == 0) {
    print('b')
  } else {
    print('c')
  }
  i <- i + 1
}
```

A.

```
[1] c
[1] b
[1] a
[1] b
[1] c
[1] a
```

B.

```
[1] c
[1] b
[1] a
[1] b
[1] c
[1] a
[1] b
```

C.

```
[1] a
[1] b
[1] c
[1] a
[1] b
[1] c
```

D.

```
[1] c
[1] b
[1] a
[1] b
[1] c
[1] a
```

E.

```
[1] c
[1] b
[1] a
[1] b
[1] c
[1] b
```

Question 3 (4 points)

Suppose the following tibble is created and saved in a variable called `scores`

```
## # A tibble: 4 x 3
##   name    grade1 grade2
##   <chr>    <dbl>  <dbl>
## 1 John        74     72
## 2 James       85     92
## 3 Michael     92     84
## 4 Howard      81     99
```

Which line of code would correctly filter for rows where `grade1` is higher than `grade2`?

A. `scores[grade1 > grade2,]`

B. `scores[scores$grade2 > scores$grade2]`

C. `scores[scores$grade2 > score$grade1,]`

D. `scores[scores$grade1 > scores$grade2,]`

E. `scores[grade2 > grade1]`

Question 4 (4 points)

What is the result of submitting the following R code?

```
x <- 1
repeat {
  print(x)
  x = x + 1
  if (x == 4) {
    break
  }
}
```

A. Error

B.
```
[1] 1
[1] 2
```

C.
```
[1] 4
```

D.
```
[1] 1
[1] 2
[1] 3
```

E.
```
[1] 1
[1] 2
[1] 3
[1] 4
```

Part 2: Data Analysis Using the Iris Dataset

Consider the following `Iris` dataset. I have converted it into a `tibble` here but I will use both the original dataframe and this tibble going forward.

```
library(tidyverse)
iris_tib <- as_tibble(iris)
print(iris_tib, n=5)
```

```
## # A tibble: 150 x 5
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##          <dbl>       <dbl>        <dbl>       <dbl> <fct>
## 1          5.1         3.5          1.4         0.2 setosa
## 2          4.9         3            1.4         0.2 setosa
## 3          4.7         3.2          1.3         0.2 setosa
## 4          4.6         3.1          1.5         0.2 setosa
## 5          5           3.6          1.4         0.2 setosa
## # i 145 more rows
```

Question 5 (4 points)

Which of the following is ***FALSE*** about `iris_tib`?

A. Calling `class(iris_tib)` would show that it is both a tibble and a dataframe.

B. Its dimensions are 150 rows and 5 columns.

C. Calling `is.matrix(iris_tib)` would return TRUE.

D. Each row represents an observation.

E. It has four numeric variables.

Question 6 (4 points)

You want to understand how `Sepal.Length` is distributed. What code could you use for an appropriate visualization?

A. `boxplot(iris$Sepal.Length, main = "Boxplot of Sepal Length")`

B. `barplot(iris$Sepal.Length, main = "Barplot of Sepal Length")`

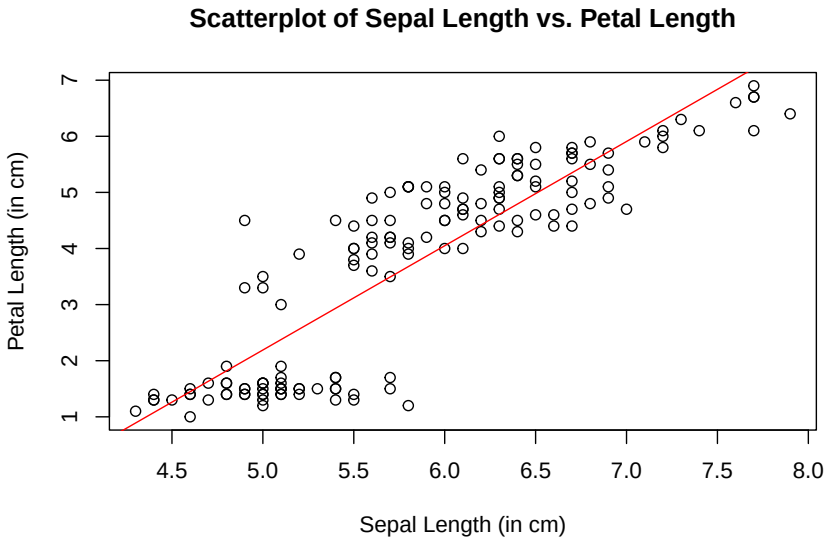
C. `plot(iris$Sepal.Length, main = "Plot of Sepal Length")`

D. `histogram(iris$Sepal.Length, main = "Histogram of Sepal Length")`

E. A and B

Question 7 (Total for this question: 20 points, 4 points per blank.)

You suspect that `Sepal.Length` and `Petal.Length` are somehow related and decide to create a scatterplot. Fill in the blanks below to produce this scatterplot. You may use either version of the `iris` dataset: `iris` or `iris_tib`.



Note that the size of the blank does not necessarily correspond to the length of the correct content.

```
____1____(____2_____, _____3____,
     main = "Scatterplot of Sepal Length vs. Petal Length",
     xlab = "Sepal Length (in cm)",
     ylab = "Petal Length (in cm)")
model <- ___4___(____5_____ ~ _____6____)
_____7____(model, col = "red")
```

1: _____

2: _____

3: _____

4: _____

5: _____

Question 8 (4 points)

You find that the correlation between Sepal Length and Petal Length is 0.872. Considering this value and the scatterplot, how would you describe the relationship between these two variables?

A. There is a weak positive linear relationship between Sepal Length and Petal Length.

B. There is a strong positive linear relationship between Sepal Length and Petal Length.

C. There is a strong negative linear relationship between Sepal Length and Petal Length.

D. There is no relationship between Sepal Length and Petal Length.

E. There is a moderate negative linear relationship between Sepal Length and Petal Length.

Question 9 (4 points)

What base R code could you run to obtain the correlation coefficient of `Sepal Length` and `Petal Length`?

A. `correlation(iris$Sepal.Length, iris$Petal.Length)`

B. `relationship(iris$Sepal.Length, iris$Petal.Length)`

C. `cor(iris$Sepal.Length, iris$Petal.Length)`

D. `linear_model(iris$Sepal.Length ~ iris$Petal.Length) %>% summary()`

E. `cor_coef(iris$Sepal.Length, iris$Petal.Length)`

Question 10 (10 points)

You decide to measure the total combined area of the Sepal and the Petal for each flower.

First, write a function called `total_area` that takes 4 arguments: `Sepal.Length`, `Sepal.Width`, `Petal.Length`, and `Petal.Width`. It should compute the area of the Sepal and the area of the Petal and return the sum of the two. Assume that the Sepal and Petal are rectangular.

```
# write your code below




# end
```

A simple test for this function is

```
total_area(1, 2, 2, 2)
```

```
[1] 6
```

Question 11 (4 points)

Now you want to create a new column in your tibble which contains the total area using the function you wrote above. Use tidyverse functions and the pipe to create this new column. Which tidyverse verb goes into the blank?

```
iris <- iris %>%
  ____1____(total_area = total_area(Sepal.Length,
                                    Sepal.Width,
                                    Petal.Length,
                                    Petal.Width))
```

A. filter

B. select

C. arrange

D. mutate

E. summarize

Question 12 (6 points)

You suspect that the relationship between `Sepal.Length` and `Sepal.Width` is different for each Iris species. Write code below to create a subset dataframe, based on the `Species` column which should equal to `setosa`. Save this subset dataframe as a new object setosa_df. **Use base R (not the tidyverse) to subset.**

```
# write base R code below to subset the iris dataframe
#
setosa_df <-
```

Question 13 (4 points)

You could have done the same operation using `tidyverse` functions. What `dplyr` function is suitable to subset a data frame, retaining all rows that satisfy your conditions?

A. filter

B. select

C. mutate

D. summary

E. summarize

Part 3: Probability and Simulation

Question 14 (12 points)

Estimating the Probability of Successfully Picking a Lock

Background: In the tabletop role-playing game Dungeons & Dragons (D&D), players often find themselves in situations where they need to perform tasks that require a bit of luck and skill. One common task is picking a lock. To determine whether a character successfully picks a lock, players roll dice and add bonuses based on their character's abilities.

Objective: Imagine you are a character trying to pick a lock. You have a high level of skill in this task, represented by a +5 bonus.

Task: ***Use simulations to estimate the probability of successfully picking a lock that requires a total score of 15 or higher.*** This involves rolling a 20-sided die (also known as "d20") and adding a +5 bonus.

Steps to Simulate:

- Roll a 20-sided die (d20) to represent your initial attempt.
- Add a bonus of 5 to the result of the d20 roll to account for your skill.
- Repeat the above steps 10,000 times.
- Calculate the proportion of times the total score is 15 or higher (successfully picking the lock).

```
# write your code below




# end
```

Question 15 (4 points)

What would happen if you increased the numer of replications from 10,000 to 1,000,000?

A. The estimated probability would become exactly 1.

B. The estimated probability would become exactly 0.

C. The estimated probability would become more accurate.

D. The estimated probability would remain the same, but the computation would be faster.

E. The estimated probability would vary more significantly with each run.

Question 16 (4 points)

The result of rolling a die or a combination of dice are random variables. Which of the following best describes a random variable in the context of probability theory?

A. A fixed numerical outcome of an experiment.

B. A variable that can take on any value within a given range.

C. A numerical outcome of a random experiment that can vary from trial to trial.

D. The average outcome of an experiment over many trials.

E. A deterministic function of an experiment's parameters.

Question 17 (4 Points)

We want to find the missing values A and B for the below PMF for a discrete random variable X, which represent P(X=10) and P(X=20) respectively. We are also given information that P(X=10) = 2 * P(X=20). Find the values for A and B.

x	0	5	10	15	20
P(X=x)	0.14	0.05	A	0.21	B

A. `A=0.20,B=0.40`

B. `A=0.40,B=0.20`

C. `A=0.50,B=0.50`

D. `A=0.40,B=0.40`

E. `A=0.20,B=0.20`