PSTAT 10: Final Review/Practice

Here are some practice questions to help you prepare for the final exam. TAs will review these during the last section in Week 10, so be sure to bring any questions you may have!

You can find the answers here

Collaboration Opportunity: Feel free to add any questions or comments directly in the Google Doc. Helping each other out by contributing will make this review even more valuable for everyone. Let's support one another in preparing for the final!

Disclaimer: These practice questions are intended to guide your study, and I will draw a few questions with slight modifications from these for the final exam. While these are suggested areas of focus, they do not cover everything. Be sure to review all course material, including lectures, worksheets, homework, quizzes, and this practice set, for a comprehensive preparation.

Using the Answers: As you go through the practice questions, try to work through them on your own first. You may also use the sample scantron to bubble your answers in. Then, use the answers to check your work and understand any concepts you might have missed. If something isn't clear, bring it to the last discussion session or add a comment in the Google Doc for further clarification. You can also add your questions, comments directly in the google doc to help your peers.

---

1. What will be the output of the following R code?

```r
rep(c(1, 2, 3), times = 2, each = 2)
```

a. `1 1 2 2 3 3 1 1 2 2 3 3`

b. `1 2 3 1 2 3 1 2 3 1 2 3`

c. `1 1 2 2 3 3`

d. `1 2 3 1 2 3`

e. Error: invalid arguments in rep()

---

2. Which of the following statements about list in R is not true?

a. List allows for different data types to be included.

b. List allows for different data structures (even lists) to be included.

c. `list()` function is used to create a list in R.

d. An empty list is valid in R.

e. The output of the following R code is `c(1, 2, 3)`.

```
my_list <- list(x = c(1, 2, 3), y = "hello")
my_list$y
```

---

3. What will be the final value of `x` after running the following loop?

```
x <- 0
for (i in 1:5) {
  x <- x + 2
}
```

a. 5

b. 10

c. 0

d. 2

e. None of the above

---

4. Suppose that 70% of customers at a coffee shop order a latte. You randomly select 15 customers. Let X = total number of customers who order a latte. Which of the following statements about this situation in true in R?

a. X follows the Binomial distribution, $X \sim \text{Binom}(15, 0.3)$.

b. The expectation (mean) of X is $E(X) = 3.15$.

c. The probability that at most 10 customers order a latte is $P(X \leq 10) = \sum_{k=0}^{10} \binom{15}{k} (0.7)^k (0.3)^{15-k}$.

d. To calculate the probability $P(X = 5)$ in R, we can use `pbinom(5, size = 15, prob = 0.7)`.

e. To calculate the probability $P(X \geq 10)$ in R, we can use `pbinom(10, size = 15, prob = 0.7, lower.tail = FALSE)`.

---

5. Which of the following statements are incorrect about the normal distribution?

a. The normal density curve is bell-shaped.

b. The mean, median, and mode are all equal in a normal distribution.

c. There exists a normal distribution that is not symmetric about the mean.

d. For nearly normally distributed data, about 99.7% falls within 3 SD of the mean.

e. The total area that lies under the normal density curve is 1 or 100%.

---

6. Which of the following statements are incorrect about the key in SQL?

a. Each row in a table/relation has its own unique key.

b. A super key is a set of one or more attributes that uniquely identify a tuple in a relation.

c. Candidate keys should not have any redundant attributes.

d. The relationship between tables is expressed by primary keys and candidate keys.

e. The same primary key cannot be used for different relations

---

7. The price of a beverage selected from the Captain Fatty's brewery is normally distributed with a mean of $7.00 and a standard deviation of 1.50. A fellow walks into the brewery and asks the bartender for a drink. What is the price of a beverage that is at the 90th percentile?

a. `qnorm(0.9, mean = 7, sd = 1.5^2, lower.tail = T)`

b. `qnorm(0.1, mean = 7, sd = 1.5, lower.tail = F)`

c. `qnorm(0.1, mean = 7, sd = 1.5^2, lower.tail = T)`

d. `qnorm(0.1, mean = 7, sd = 1.5, lower.tail = T)`

e. `qnorm(0.1, mean = 7, sd = 1.5^2, lower.tail = F)`

---

8. Consider you are given the following vector

```
myVec <- 6:17
```

and the following loop:

```
for(i in 1:length(myVec)){
  print(myVec)
}
```

Which of the following are equivalent to the chunk above?

a.

```
for(a in myVec) paste(a)
```

b.

```
rec <- function(v) {
  if(length(v) == 1) {
    print(v[1])
  }
  else {
    print(v[length(v)])
    rec(v[1:(length(v) - 1)])
  }
}
```

c.

```
for(a in rev(myVec)) paste(a)
```

d.

```
rec <- function(v) {
  if(length(v) == 1) {
    print(v[1])
  }
  else {
    print(v[length(v)])
    rec(v[(length(v) - 1):1])
  }
}

rec(rev(myVec))
```

e.

```
for(a in rev(myVec)) print(a)
```

---

9. When querying a CUSTOMER table from an SQL database, I need to find the total number of transactions in the `UNITS_SOLD` column.
Which of the following is the correct query to use in R?

a. `dbGetQuery(myDB, "SELECT LENGTH(UNITS_SOLD) FROM CUSTOMER")`

b. `"SELECT COUNT(UNITS_SOLD) FROM CUSTOMER"`

c. `dbGetQuery(myDB, "SELECT COUNT(UNITS_SOLD) FROM CUSTOMER")`

d. `dbGetQuery(myDB, "SELECT TOTAL(UNITS_SOLD) FROM CUSTOMER")`

e. `"SELECT COUNT(UNITS_SOLD) FROM CUSTOMER"`

---

10. You have the following data frame in R. Which of the following options correctly extracts Bob's Score from the data frame?

```r
df <- data.frame(Name = c("Alice", "Bob", "Charlie"),
                 Age = c(25, 30, 22),
                 Score = c(90, 85, 88))
```

a. `df$Score[Bob]`

b. `df[2, "Score"]`

c. `df["Bob", "Score"]`

d. `df[2, 3]`

e. `df[[2]]`

---

11. What will be the output of the following R code?

```r
result <- c()
for (i in 1:5) {
  if (i %% 2 == 0) {
    result <- c(result, i * 2)
  } else {
    result <- c(result, i + 3)
  }
}

print(result)
```

a. 4 4 8 6 12

b. 4 8 6 12 8

c. 4 8 6 12 10

d. 4 4 6 6 8

e. 4 4 6 8 8

---

12. A candy factory produces sweets in two flavors: Chocolate and Strawberry. 70% of the candies are Chocolate-flavored. 30% of the candies are Strawberry-flavored. 10% of Chocolate-flavored candies have a misprint on the wrapper. 5% of Strawberry-flavored candies have a misprint on the wrapper. A candy is randomly selected from the factory's total production. What is the probability that it has a misprinted wrapper?

a. 0.065

b. 0.075

c. 0.080

d. 0.085

e. 0.07

---

13. A candy company produces bags of chocolates, and the weight of each bag follows a normal distribution with a mean of 500 grams and a standard deviation of 20 grams. The company wants to find the weight threshold that separates the top 5% of the heaviest bags from the rest. What R function should be used to calculate the minimum weight (95th percentile) for the top 5% of the heaviest bags?

a. `qnorm(0.95, mean = 500, sd = 20)`

b. `pnorm(0.95, mean = 500, sd = 20)`

c. `pnorm(500, mean = 20, sd = 500)`

d. `qnorm(0.05, mean = 500, sd = 20)`

e. `qnorm(0.95, mean = 500, sd = 20, lower.tail = FALSE)`

---

14. Which of the following best describes a foreign key in SQL?

a. A unique identifier for each row in a table.

b. A column that references the primary key of another table.

c. A command used to retrieve data from a database.

d. A constraint that ensures all values in a column are different.

e. A special type of index used for performance optimization.

---

15. In data analysis, terms can differ between data frames (R) and relational databases (SQL). Which of the following correctly matches the equivalent terms?

a. Column (Data Frame) - Table (Database)

b. Row (Data Frame) - Record (Database)

c. Data Frame (R) - Column (Database)

d. Index (Data Frame) - Foreign Key (Database)

e. Variable (Data Frame) - Schema (Database)

---

16. Find the probability that a student's final exam score, in a class with an average of 80 and a standard deviation of 5, falls between 70 and 90? Choose correct r code below.

a. `dnorm(90, 80, 5) - dnorm(70, 80, 5)`

b. `2 * pnorm(2)`

c. `pnorm(2) - pnorm(-2)`

d. `pnorm(90) - pnorm(70)`

e. `qnorm(90, 80, 5) - qnorm(70, 80, 5)`

---

17. Which of the following functions have different output from the loop below?

```r
for (i in 1:5) {
  print(i)
}
```

a.

```r
i <- 1
while (i <= 5) {
  print(i)
  i <- i + 1
}
```

b.

```r
i <- 1
repeat {
  print(i)
  i <- i + 1
  if (i > 5) {
    break
  }
}
```

c.

```r
i <- 0
repeat {
  i <- i + 1
  print(i)
  if (i >= 5) {
    break
  }
}
```

d.

```r
i <- 0
repeat {
  i <- i + 1
  print(i)
  if (i >= 5) {
    break}

}
```

e.

```r
i <- 0
repeat {
  i <- i + 1
  if (i >= 5) {
    break
  }
  print(i)
}
```

---

18. Some of these structures are homogeneous, while others are heterogeneous. Which of the following answer choices correctly categorizes these data structures into homogeneous and heterogeneous??

a. Homogeneous: Matrix, Dataframe Heterogeneous: List, Vector, Array

b. Homogeneous: Vector, Matrix, Array Heterogeneous: List, Dataframe

c. Homogeneous: List, Matrix, Dataframe Heterogeneous: Vector, Array

d. Homogeneous: Vector, Dataframe, List Heterogeneous: Matrix, Array

e. Homogeneous: Vector, Matrix, List Heterogeneous: Dataframe, Array

---

19. We want to:
Retrieve only the student records with scores greater than 60.
Group the result by the class column and calculate the average score for each class.
Keep only those groups whose average score is greater than 80.
Sort them in descending order by the average score (from highest to lowest).
Display only the first 5 matching groups.
Choose the correct SQL query below:

```sql
SELECT class, AVG(score) AS avg_score
FROM students
WHERE score > 60
GROUP BY class
HAVING AVG(score) > 80
ORDER BY avg_score DESC
LIMIT 5;
```

a. Uses WHERE AVG(score) $>$ 80 instead of HAVING AVG(score) $>$ 80. Filtering on a grouped aggregate must be done with HAVING, not WHERE. Hence, A is incorrect.

b. Correct. Uses HAVING for aggregation filtering.

c. Groups the data by score instead of class, and orders by class DESC rather than the average score. This fails the requirement to sort by the average score in descending order and to group by class. Hence, C is incorrect.

d. Selects id, name, AVG(score) AS avg_score. We only need each class and its average score. Including id and name without proper aggregation would typically cause an error or produce incorrect grouping. Hence, D is incorrect.

e. Applies WHERE score > 80, which filters out all students scoring 60–80. The requirement is specifically to select students with scores greater than 60 but not necessarily over 80. Hence, E is incorrect.

---

20. Given a diagnostic test with a true positive rate of 95% and a specificity of 90%, what is the probability that a randomly selected individual who tests positive actually has the disease?

```
P_A <- 0.02
P_B_given_A <- 0.95
P_B_given_Ac <- 0.10
P_B <- (P_B_given_A * P_A) + (P_B_given_Ac * (1 - P_A))
P_A_given_B <- (P_B_given_A * P_A) / P_B
P_A_given_B
```

a. 2%

b. 10%

c. 16.2%

d. 50%

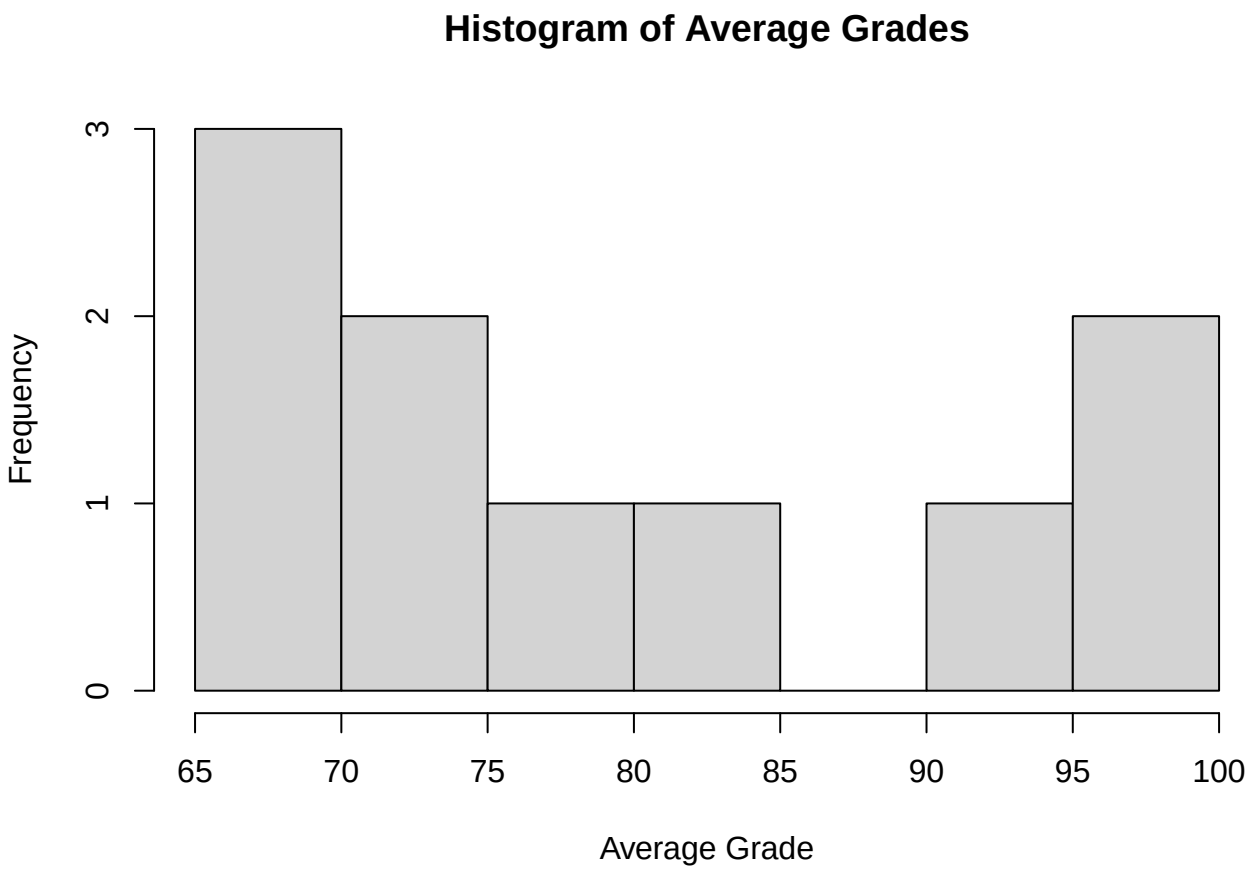
e. 95%

---

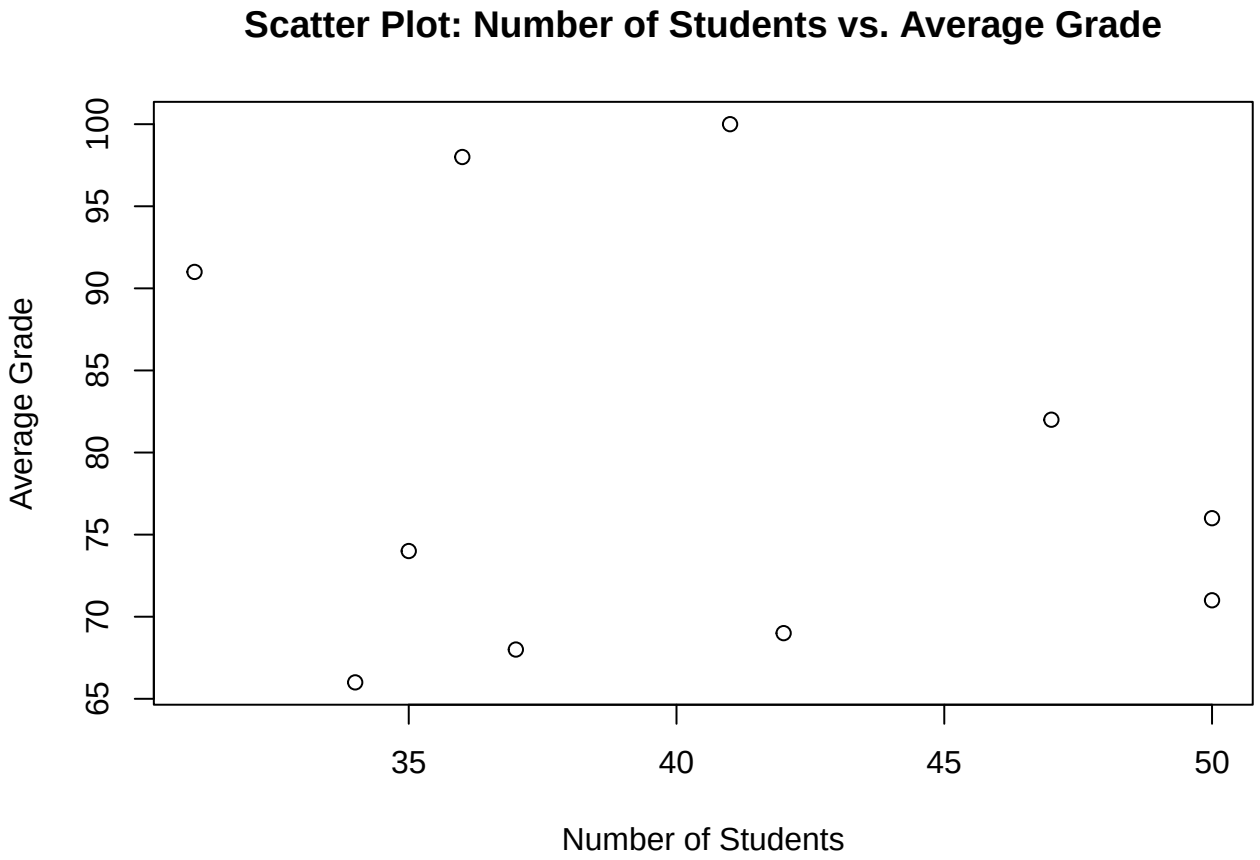
21. Which plot is least suitable for visualizing the relationship between class size and average score?

```
set.seed(123)       # For reproducibility (optional)
avg_grade <- c(76, 98, 71, 74, 91, 66, 68, 100, 69, 82)
num_students <- c(50, 36, 50, 35, 31, 34, 37, 41, 42, 47)
```

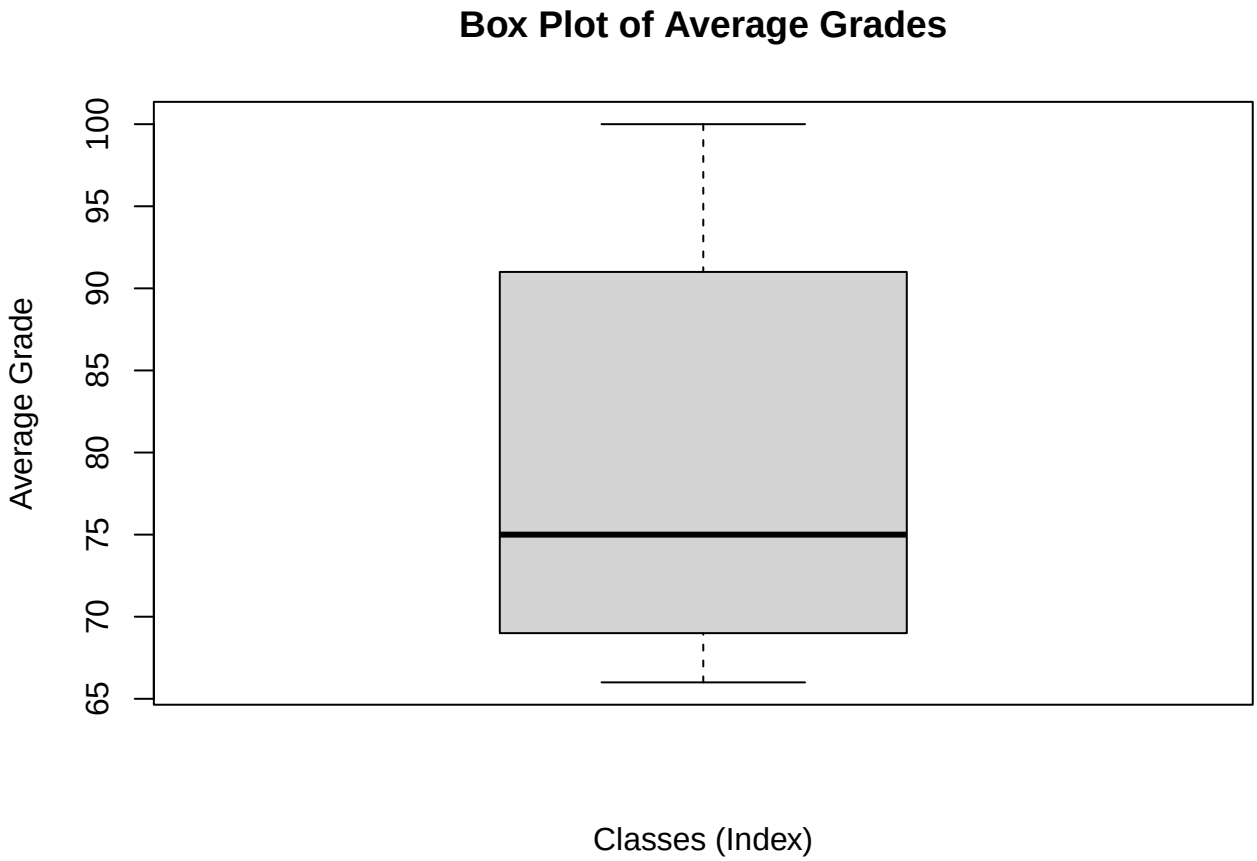
a. Histogram



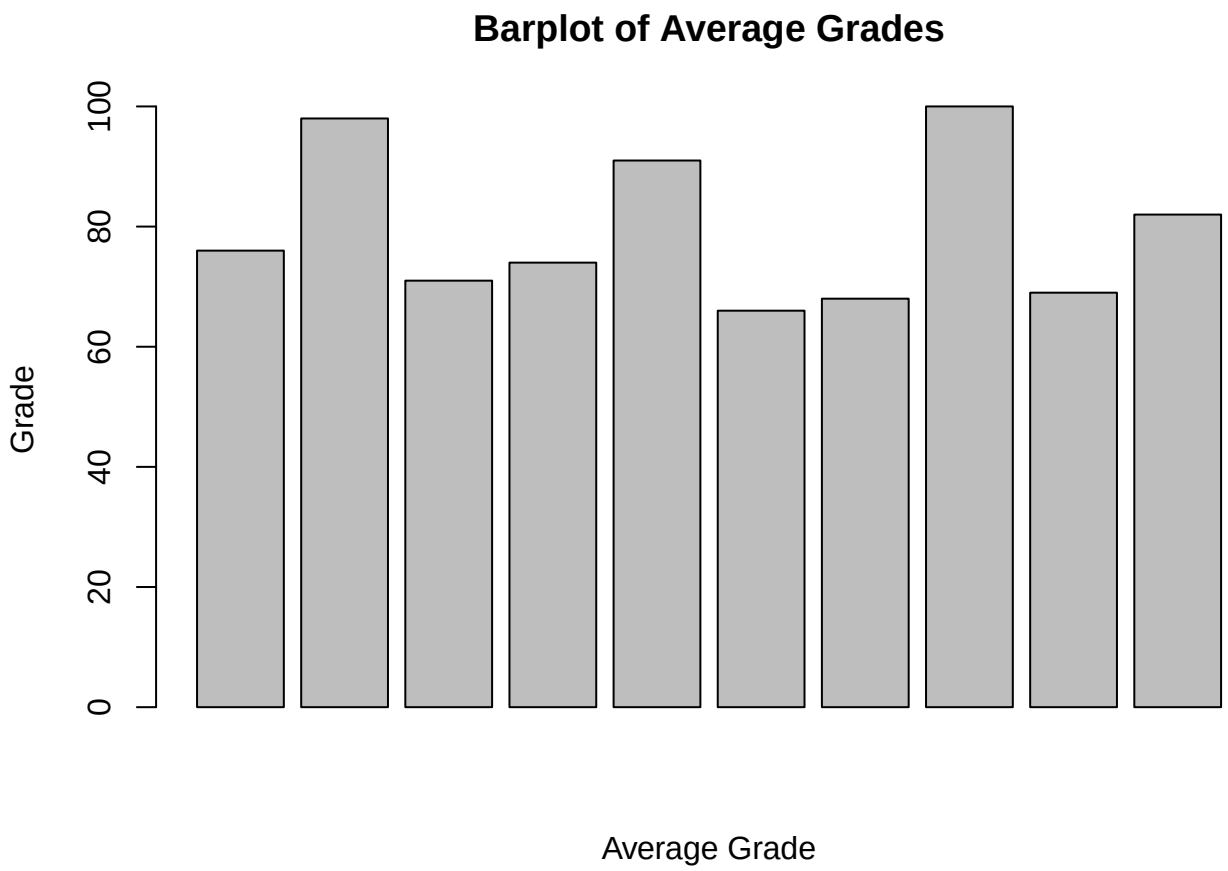
b. Scatter plot


Scatter Plot: Number of Students vs. Average Grade
Average Grade
65
70
75
80
85
90
95
100
35
40
45
50
Number of Students

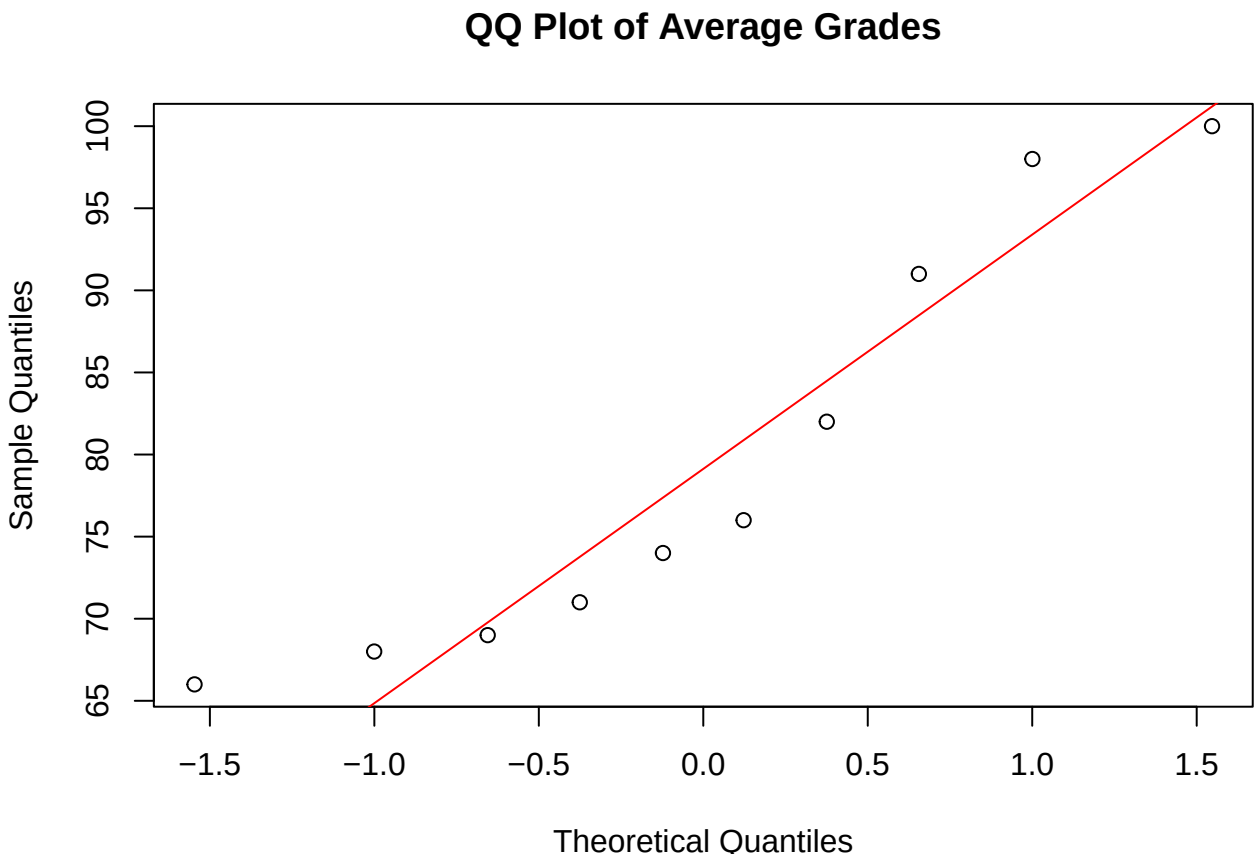
c. Box plot


Box Plot of Average Grades
Average Grade
65
70
75
80
85
90
95
100
Classes (Index)

d. Bar plot


Barplot of Average Grades
Grade
0
20
40
60
80
100
Average Grade

e. QQ plot



---

22. Consider the following R object `my_string`. Which of the following answer choices will produce the output "PSTAT 10"?

```
my_string <- "I love PSTAT 10"
```

a. `nchar(my_string)`

b. `subset(my_string, 7, 14)`

c. `substr(my_string, 8, 15)`

d. `subset(my_string, 8, 15)`

e. `substr(my_string, 7, 14)`

---

23. Suppose $X \sim \text{Unif}(-1, 6)$. Then $\mathbb{P}(X > 1)$ is given by

a. `punif(1, min = -1, max = 6, lower.tail = T)`

b. `1 - punif(1, min = -1, max = 6, lower.tail = F)`

c. `qunif(1, min = -1, max = 6, lower.tail = F)`

d. `1 - punif(1, min = -1, max = 6, lower.tail = T)`

e. `1 - qunif(1, min = -1, max = 6, lower.tail = F)`

---

24. Consider the following scenario:

In Worksheet 14, Exercise 2, you identified that `SupportRepId` in the `Customer` table is a foreign key to the primary key of the `Employee` table, which is `EmployeeId`.

Now, in a single query, retrieve a list of customers who are from **California**, **USA**, including **their first name**, along with the **first name of their assigned sales representative**. Your output should look like this:

CustomerFirst	EmployeeFirst
Frank	Margaret
Tim	Jane
Dan	Margaret

a.

```
dbGetQuery(chinook_db,
           "Select Customer.FirstName AS CustomerFirst,
           Employee.FirstName AS EmployeeFirst
           FROM Employee
           INNER JOIN customer ON Employee.employeeId = Customer.SupportRepId
           WHERE CUSTOMER.COUNTRY = 'USA' and customer.STate = 'CA'")
```

b.

```
dbGetQuery(chinook_db,
           "select Customer.CustomerFirst AS Customer.FirstName,
           Employee.EmployeeFirst AS Employee.FirstName
           From Employee
           INNER JOIN Customer ON employee.EmployeeId = Customer.SupportRepId
           WHERE Customer.Country = 'USA' and Customer.State = 'CA'")
```

c.

```
dbGetQuery(chinook_db,
           "SELECT Customer.FirstName AS CustomerFirst,
           Employee.FirstName AS EmployeeFirst
           FROM Employee
           INNER JOIN customer ON Employee.EmployeeId = Customer.SupportRepId
           where Customer.country = 'Usa' and Customer.state = 'CA'")
```

d.

```
dbGetQuery(chinook_db,
           "Select c.FirstName AS CustomerFirst,
           e.FirstName AS EmployeeFirst
           FROM Employee e
           inner join customer c e.EmployeeId = c.SupportRepId
           WHERE c.COUNTRY = 'USA' and c.state = 'CA'")
```

e.

```
dbGetQuery(chinook_db,
           "SELECT c.CustomerFirst AS c.FirstName,
           e.EmployeeFirst AS e.FirstName
           FROM Employee e
           INNER JOIN Customer c ON e.EmployeeId = c.SupportRepId
           WHERE c.country = 'USA' AND c.STATE = 'CA'")
```

---

25. Which query returns the first **5** records for `CustomerId`, `FirstName`, `LastName` and `Address` from the Customer table?

```
a. dbGetQuery(chinook_db, "select CustomerId, FirstName, LastName, Address
            from customer
            max 5")
b. dbGetQuery(chinook_db, "select CustomerId, FirstName, LastName, Address
            from customer
            limit 5")
c. dbGetQuery(chinook_db, "select CustomerId, FirstName, LastName, Address
            from customer
            greatest 5")
d. dbGetQuery(chinook_db, "select CustomerId, FirstName, LastName, Address
            from customer
            total 5")
e. dbGetQuery(chinook_db, "select CustomerId, FirstName, LastName, Address
            from customer
            head 5")
```

---

26. Which definition best describes a "factor" (as in data type) in R?

```
a. A data structure used for storing categorical variables.

b. A function that calculates the factorial of a number

c. type of matrix used for performing linear algebra operations.

d. A numeric vector that contains real numbers only

e. A special data structure used exclusively for storing dates
```

---

27. I flip a fair coin 10 times. I am interested in calculating the probability of getting 4 tails. Which code would return the correct answer in R?

```
a. prob_4_heads <- pbinom(4, size = 10, prob = 0.5)

b. prob_4_heads <- pnorm(4, size = 10, prob = 0.5)

c. prob_4_heads <- dbinom(4, size = 10, prob = 0.5)

d. prob_4_heads <- dnorm(4, size = 10, prob = 0.5)

e. prob_4_heads <- dunif(4, size = 10, prob = 0.5)
```

---

28. The probability distribution of a discrete random variable X is as shown in the table:

x	1	2	3	4	5
$\mathbb{P}(X = x)$	0.0	a	0.25	0.20	0.12

What is the value of a?

a. 0.38

b. 0.36

c. 0.41

d. 0.43

e. 0.50

---

29. Identify the the **"Event"** in the following activity:

we draw a card from a standard deck and look for a diamond.

a. Drawing a red card

b. Drawing a diamond

c. Drawing a card from the standard deck

d. Drawing anything but a diamond

e. Drawing a heart

---

30. A machine in a factory takes a uniformly distributed amount of time between 4 and 10 hours to complete a task. A task is chosen at random, and the time it takes to complete it is recorded. What is the probability that the machine will take exactly 7 hours to complete the task?

a. 0

b. 0.25

c. 0.1

d. 1

e. 0.5
