PSTAT 10 - Final Exam Study Guide

Study Guide for Final Exam: Introduction to Data Science with R Programming and SQL

Vocabulary Terms and Definitions

R Programming

1. **Vector**
    - A sequence of data elements of the same basic type. Members in a vector are called components.
2. **Data Frame**
    - A table or two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column.
3. **Matrix**
    - A two-dimensional, homogeneous data structure in R, where all elements must be of the same type.
4. **List**
    - An ordered collection of objects (components), which can be of different types (numbers, strings, vectors, even lists).
5. **Function**
    - A block of code designed to perform a specific task, which can take inputs and return outputs.
6. **Factor**
    - A data structure used for fields that take only predefined, finite number of values (categorical data).

7. **Loop**
    - A control flow statement for specifying iteration, which allows code to be executed repeatedly.
8. **Conditional Statement**
    - A feature of a programming language that performs different computations or actions depending on whether a programmer-specified boolean condition evaluates to true or false.
9. **Data Manipulation**
    - The process of changing data to make it more organized and easier to read.
10. **Plotting**
    - The act of creating a visual representation of data, such as graphs and charts.

SQL (Structured Query Language)

1. **Database**
    - An organized collection of data, generally stored and accessed electronically from a computer system.
2. **Table**
    - A collection of related data held in a structured format within a database. It consists of rows and columns.
3. **Query**
    - A request for data or information from a database table or combination of tables.
4. **Join**
    - An operation that combines rows from two or more tables based on a related column between them.
5. **Primary Key**
    - A field in a table which uniquely identifies each row/record in that table.
6. **Foreign Key**
    - A field in one table that is a primary key in another table, used to link the two tables.

7. **Index**
    - A database object that improves the speed of data retrieval operations on a table at the cost of additional storage space and increased maintenance time.
8.

Relational Integrity

9.

Entity Integrity

10. **View**
    - A virtual table in SQL, which is the result of a stored query on the data.

Statistical Concepts

1. **Probability**
    - A measure of the likelihood that an event will occur.
2. **Random Variable**
    - A variable whose possible values are numerical outcomes of a random phenomenon.
3. **Distribution**
    - A mathematical description of the probabilities of occurrence of different possible outcomes.
4. **Binomial Distribution**
    - A discrete probability distribution of the number of successes in a sequence of n independent experiments.
5. **Normal Distribution**
    - A continuous probability distribution that is symmetrical around its mean, indicating that data near the mean are more frequent in occurrence.
6. **Uniform Distribution**
    - A type of probability distribution in which all outcomes are equally likely.

7.

Poison Distribution

Miscellaneous

1. **ETL (Extract, Transform, Load)**
    - The general procedure of copying data from one or more sources into a destination system that represents the data differently from the source(s).
2. **Big Data**
    - Large, complex data sets that traditional data processing software cannot adequately deal with.
3. **Machine Learning**
    - A method of data analysis that automates analytical model building, allowing computers to find hidden insights without being explicitly programmed.
4. **API (Application Programming Interface)**
    - A set of rules that allows different software entities to communicate with each other.
5. **Data Warehouse**
    - A system used for reporting and data analysis, and is considered a core component of business intelligence.

Practice Problems

R Programming

1. **Question**: How do you create a vector in R with the numbers 1 to 10?
    - **Solution**: `x <- c(1:10)`
2. **Question**: Write a function in R to calculate the square of a number.
    - **Solution**:

```r
square <- function(x) {
  return(x^2)
}
```

3. **Question**: How do you calculate the mean of a numeric vector `x` in R?
   - **Solution**: `mean(x)`
4. **Question**: How do you subset the first three elements of a vector `x` in R?
   - **Solution**: `x[1:3]`
5. **Question**: What is the output of `sum(c(1, 2, 3, NA), na.rm = TRUE)` in R?
   - **Solution**: `6`
6. **Question**: How do you create a data frame with columns `name` and `age` in R?
   - **Solution**:

```r
df <- data.frame(name = c("Alice", "Bob"), age = c(25, 30))
```

7. **Question**: Write a loop in R to print numbers from 1 to 5.
   - **Solution**:

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

8. **Question**: How do you read a CSV file named `data.csv` into R?
   - **Solution**: `df <- read.csv("data.csv")`
9. **Question**: How do you filter rows in a data frame `df` where the column `age` is greater than 25?
   - **Solution**: `df[df$age > 25, ]`
10. **Question**: How do you add a new column `height` to a data frame `df`?
    - **Solution**: `df$height <- c(160, 170)`

SQL

11. **Question**: How do you select all columns from a table `Employees` in SQL?
    - **Solution**: `SELECT * FROM Employees;`
12. **Question**: Write an SQL query to find the number of employees in the `Employees` table.

- **Solution**: `SELECT COUNT(*) FROM Employees;`

13. **Question**: How do you add a new column `salary` to a table `Employees` in SQL?
    - **Solution**: `ALTER TABLE Employees ADD COLUMN salary DECIMAL(10, 2);`
14. **Question**: Write an SQL query to select employees with a salary greater than 50000.
    - **Solution**: `SELECT * FROM Employees WHERE salary > 50000;`
15. **Question**: How do you update the salary of an employee with `id` 1 to 60000?
    - **Solution**: `UPDATE Employees SET salary = 60000 WHERE id = 1;`
16. **Question**: Write an SQL query to delete employees from the `Employees` table who are older than 65.
    - **Solution**: `DELETE FROM Employees WHERE age > 65;`
17. **Question**: How do you create a new table `Departments` with columns `id` and `name`?
    - **Solution**: `CREATE TABLE Departments (id INT PRIMARY KEY, name VARCHAR(50));`
18. **Question**: Write an SQL query to join `Employees` and `Departments` tables on the `department_id`.
    - **Solution**: `SELECT * FROM Employees JOIN Departments ON Employees.department_id = Departments.id;`
19. **Question**: How do you create a view `EmployeeView` that shows `name` and `salary` from `Employees`?
    - **Solution**: `CREATE VIEW EmployeeView AS SELECT name, salary FROM Employees;`
20. **Question**: Write an SQL query to find the maximum salary in the `Employees` table.
    - **Solution**: `SELECT MAX(salary) FROM Employees;`

Statistical Concepts

21. **Question**: What is the probability of getting a head in a single coin toss?
    - **Solution**: `0.5`
22. **Question**: How do you calculate the mean of the numbers 1, 2, 3, 4, and 5?
    - **Solution**: `(1 + 2 + 3 + 4 + 5) / 5 = 3`
23. **Question**: What is the binomial probability of getting exactly 2 heads in 3 coin tosses?
    - **Solution**:

```r
dbinom(2, size = 3, prob = 0.5)
```

24. **Question**: How do you calculate the standard deviation of the numbers 1, 2, 3, 4, and 5 in R?
    - **Solution**: `sd(c(1, 2, 3, 4, 5))`
25. **Question**: What is the z-score for a value of 70, with a mean of 50 and a standard deviation of 10?
    - **Solution**: `(70 - 50) / 10 = 2`
26. **Question**: How do you perform a t-test in R to compare the means of two vectors `x` and `y`?
    - **Solution**: `t.test(x, y)`
27. **Question**: What is the 95% confidence interval for a sample mean of 100 with a standard deviation of 15 and a sample size of 25?
    - **Solution**:

```r
mean <- 100
sd <- 15
n <- 25
error <- qnorm(0.975) * sd / sqrt(n)
c(mean - error, mean + error)
```

Multiple Choice Questions and Solutions

R Programming

1. **Question**: How do you create a vector in R with the numbers 1 to 10?
    - A) `x <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)`
    - B) `x <- c(1:10)`
    - C) `x <- 1; 10`
    - D) `x <- seq(1, 10)`
    - **Solution**: B) `x <- c(1:10)`
2. **Question**: Write a function in R to calculate the square of a number.
    - A) `square <- function(x) { x * x }`
    - B) `square <- function(x) { return(x^2) }`
    - C) `square <- function(x) { x**2 }`
    - D) `square <- function(x) { x ^ 2 }`

- **Solution**: B) `square <- function(x) { return(x^2) }`

3. **Question**: How do you calculate the mean of a numeric vector `x` in R?
   - A) `mean(x)`
   - B) `avg(x)`
   - C) `sum(x)/length(x)`
   - D) `average(x)`
   - **Solution**: A) `mean(x)`

4. **Question**: How do you subset the first three elements of a vector `x` in R?
   - A) `x[1, 2, 3]`
   - B) `x[c(1, 2, 3)]`
   - C) `x[1:3]`
   - D) `x[-4:length(x)]`
   - **Solution**: C) `x[1:3]`

5. **Question**: What is the output of `sum(c(1, 2, 3, NA), na.rm = TRUE)` in R?
   - A) `NA`
   - B) `6`
   - C) `7`
   - D) `NA, 6`
   - **Solution**: B) `6`

6. **Question**: How do you create a data frame with columns `name` and `age` in R?
   - A) `df <- data.frame(name = c("Alice", "Bob"), age = c(25, 30))`
   - B) `df <- data.frame(c("Alice", "Bob"), c(25, 30))`
   - C) `df <- data.frame(name = c("Alice", "Bob"), age = c("25", "30"))`
   - D) `df <- data.frame(names = c("Alice", "Bob"), ages = c(25, 30))`
   - **Solution**: A) `df <- data.frame(name = c("Alice", "Bob"), age = c(25, 30))`

7. **Question**: Write a loop in R to print numbers from 1 to 5.
   - A) `for (i in 1 to 5) { print(i) }`
   - B) `for (i in 1:5) { print(i) }`
   - C) `for (i in c(1, 2, 3, 4, 5)) { print(i) }`
   - D) `for (i = 1 to 5) { print(i) }`
   - **Solution**: B) `for (i in 1:5) { print(i) }`

8. **Question**: How do you read a CSV file named `data.csv` into R?
   - A) `df <- read.csv(file = "data.csv")`
   - B) `df <- read(file = "data.csv")`
   - C) `df <- read.csv("data.csv")`

- D) `df <- read.file("data.csv")`
- **Solution**: C) `df <- read.csv("data.csv")`

9. **Question**: How do you filter rows in a data frame `df` where the column `age` is greater than 25?
   - A) `df[df[,"age"] > 25,]`
   - B) `df[df$age > 25,]`
   - C) `df[age > 25,]`
   - D) `subset(df, age > 25)`
   - **Solution**: B) `df[df$age > 25,]`

10. **Question**: How do you add a new column `height` to a data frame `df`?
    - A) `df$height <- c(160, 170)`
    - B) `df[,"height"] <- c(160, 170)`
    - C) `df <- cbind(df, height = c(160, 170))`
    - D) `df <- rbind(df, height = c(160, 170))`
    - **Solution**: A) `df$height <- c(160, 170)`

SQL

11. **Question**: How do you select all columns from a table `Employees` in SQL?
    - A) `SELECT ALL FROM Employees;`
    - B) `SELECT * FROM Employees;`
    - C) `SELECT COLUMNS FROM Employees;`
    - D) `SELECT [ALL] FROM Employees;`
    - **Solution**: B) `SELECT * FROM Employees;`

12. **Question**: Write an SQL query to find the number of employees in the `Employees` table.
    - A) `SELECT COUNT ALL FROM Employees;`
    - B) `SELECT COUNT COLUMNS FROM Employees;`
    - C) `SELECT COUNT(*) FROM Employees;`
    - D) `SELECT COUNT FROM Employees;`
    - **Solution**: C) `SELECT COUNT(*) FROM Employees;`

13. **Question**: How do you add a new column `salary` to a table `Employees` in SQL?
    - A) `ALTER TABLE Employees ADD COLUMN salary DECIMAL(10, 2);`
    - B) `ADD COLUMN salary DECIMAL(10, 2) TO Employees;`
    - C) `ALTER Employees ADD COLUMN salary DECIMAL(10, 2);`
    - D) `ADD salary DECIMAL(10, 2) TO Employees;`
    - **Solution**: A) `ALTER TABLE Employees ADD COLUMN salary DECIMAL(10, 2);`

14. **Question**: Write an SQL query to select employees with a salary greater than 50000.

- A) `SELECT * FROM Employees WHERE salary > 50000;`
- B) `SELECT * FROM Employees WHERE salary >= 50000;`
- C) `SELECT * FROM Employees WHERE salary > 50k;`
- D) `SELECT * FROM Employees WHERE salary => 50000;`
- **Solution**: A) `SELECT * FROM Employees WHERE salary > 50000;`

15. **Question**: How do you update the salary of an employee with `id` 1 to 60000?
    - A) `UPDATE Employees SET salary = 60000 WHERE id IS 1;`
    - B) `UPDATE Employees SET salary = 60000 WHERE id = 1;`
    - C) `UPDATE Employees SET salary = 60000 WHERE employee_id = 1;`
    - D) `SET salary = 60000 WHERE id = 1 IN Employees;`
    - **Solution**: B) `UPDATE Employees SET salary = 60000 WHERE id = 1;`
16. **Question**: Write an SQL query to delete employees from the `Employees` table who are older than 65.
    - A) `DELETE FROM Employees WHERE age > 65;`
    - B) `REMOVE FROM Employees WHERE age > 65;`
    - C) `DELETE Employees WHERE age > 65;`
    - D) `DELETE FROM Employees WHERE age IS GREATER THAN 65;`
    - **Solution**: A) `DELETE FROM Employees WHERE age > 65;`
17. **Question**: How do you create a new table `Departments` with columns `id` and `name`?
    - A) `CREATE TABLE Departments (id INT, name VARCHAR(50));`
    - B) `CREATE Departments (id INT, name VARCHAR(50));`
    - C) `CREATE TABLE Departments (id INTEGER, name VARCHAR(50));`
    - D) `CREATE TABLE Departments (id INT PRIMARY KEY, name VARCHAR(50));`
    - **Solution**: D) `CREATE TABLE Departments (id INT PRIMARY KEY, name VARCHAR(50));`
18. **Question**: Write an SQL query to join `Employees` and `Departments` tables on the `department_id`.
    - A) `SELECT * FROM Employees INNER JOIN Departments ON Employees.department_id = Departments.id;`
    - B) `SELECT * FROM Employees JOIN Departments ON Employees.department_id = Departments.id;`
    - C) `SELECT * FROM Employees LEFT JOIN Departments ON Employees.department_id = Departments.id;`
    - D) `SELECT * FROM Employees RIGHT JOIN Departments ON Employees.department_id = Departments.id;`
    - **Solution**: B) `SELECT * FROM Employees JOIN Departments ON Employees.department_id = Departments.id;`
19. **Question**: How do you create a view `EmployeeView` that shows `name` and `salary` from `Employees`?

- A) `CREATE VIEW EmployeeView AS SELECT name, salary FROM Employees;`
- B) `CREATE VIEW EmployeeView AS SELECT * FROM Employees;`
- C) `CREATE VIEW EmployeeView AS SELECT name, salary FROM Employees WHERE salary > 50000;`
- D) `CREATE VIEW EmployeeView AS SELECT name, salary FROM Employees WHERE department_id = 1;`
- **Solution**: A) `CREATE VIEW EmployeeView AS SELECT name, salary FROM Employees;`

20. **Question**: Write an SQL query to find the maximum salary in the `Employees` table.
    - A) `SELECT MAX(salary) FROM Employees;`
    - B) `SELECT salary FROM Employees ORDER BY salary DESC LIMIT 1;`
    - C) `SELECT MAX(salary) FROM Employees WHERE department_id = 1;`
    - D) `SELECT MAX(salary) FROM Employees GROUP BY department_id;`
    - **Solution**: A) `SELECT MAX(salary) FROM Employees;`

Statistical Concepts

21. **Question**: What is the probability of getting a head in a single coin toss?
    - A) `0.25`
    - B) `0.5`
    - C) `0.75`
    - D) `1.0`
    - **Solution**: B) `0.5`

22. **Question**: What is the binomial probability of getting exactly 2 heads in 3 coin tosses?
    - A) `dbinom(2, size = 3, prob = 0.5)`
    - B) `dbinom(2, size = 3, prob = 0.25)`
    - C) `pbinom(2, size = 3, prob = 0.5)`
    - D) `pbinom(2, size = 3, prob = 0.25)`
    - **Solution**: A) `dbinom(2, size = 3, prob = 0.5)`

23. **Question**: How do you calculate the standard deviation of the numbers 1, 2, 3, 4, and 5 in R?
    - A) `var(c(1, 2, 3, 4, 5))`
    - B) `sqrt(var(c(1, 2, 3, 4, 5)))`
    - C) `sd(c(1, 2, 3, 4, 5))`
    - D) `mean(c(1, 2, 3, 4, 5))`
    - **Solution**: C) `sd(c(1, 2, 3, 4, 5))`

24. **Question**: What is the z-score for a value of 70, with a mean of 50 and a standard deviation of 10?

- A) `2`
- B) `1.5`
- C) `1`
- D) `0.5`
- **Solution**: A) `2`

25. **Question**: How do you perform a t-test in R to compare the means of two vectors `x` and `y`?

    - A) `t.test(x, y)`
    - B) `t.test(x ~ y)`
    - C) `test.t(x, y)`
    - D) `t_test(x, y)`
    - **Solution**: A) `t.test(x, y)`

26. **Question**: What is the 95% confidence interval for a sample mean of 100 with a standard deviation of 15 and a sample size of 25?

    - A) `[95, 105]`
    - B) `[90, 110]`
    - C) `[85, 115]`
    - D) `[80, 120]`
    - **Solution**: B) `[90, 110]`

Multiple Choice Practice Final Exam Questions and Solutions

Modified Questions from Midterm 2 Version A and B

12. **Question**: Which of the following is NOT true about foreign keys in a relational database?

    - A) A foreign key contains attributes that point to another relation's primary key.
    - B) The primary key of one relation could be the foreign key of another relation.
    - C) A foreign key establishes a link between two relations.
    - D) A foreign key always consists of more than one attribute.
    - **Solution**: D) A foreign key always consists of more than one attribute.

13. **Question**: Which of the following is NOT part of the relational model?

    - A) Structural
    - B) Security
    - C) Manipulative
    - D) Integrity
    - **Solution**: B) Security

14. **Question**: Which one of the following provides the ability to query information from the database and to insert tuples into, delete tuples from, and modify tuples in the database?

- A) DML (Data Manipulation Language)
- B) DDL (Data Definition Language)
- C) Query
- D) Relational Schema
- **Solution**: A) DML (Data Manipulation Language)

15. **Question**: Assuming you are working in the database `PSTAT10db`, which of the following queries does not contain an error? The relation `INVOICES` contains an attribute `QUANTITY`.

- A) `dbGetQuery(PSTAT10db, 'SELECT AVG (QUANTITY) FROM INVOICES')`
- B) `dbGetQuery(PSTAT10db, 'SELECT AVG QUANTITY FROM INVOICES')`
- C) `dbGetQuery(PSTAT10db, 'SELECT AVERAGE (QUANTITY) FROM INVOICES')`
- D) `dbGetQuery(PSTAT10db, 'SELECT MEAN (QUANTITY) FROM INVOICES')`
- **Solution**: A) `dbGetQuery(PSTAT10db, 'SELECT AVG (QUANTITY) FROM INVOICES')`

16. **Question**: Which of the following is NOT true for independent events?

- A) P(A|B) = P(A)
- B) P(B|A) = P(B)
- C) P(A or B) = P(A) + P(B)
- D) P(A and B) = P(A)P(B)
- **Solution**: C) P(A or B) = P(A) + P(B)

17. **Question**: Which of the following R statements will find the probability of having exactly two successes given a binomial random variable bin(6, 0.3)?

- A) `dbinom(2, size=6, prob=0.3)`
- B) `dbinom(6, size=2, prob=0.3)`
- C) `pbinom(2, size=6, prob=0.3)`
- D) `pbinom(6, size=2, prob=0.3)`
- **Solution**: A) `dbinom(2, size=6, prob=0.3)`

18. **Question**: Suppose that the area under the normal curve (X N(50, 4)) to the left of some unknown x-value is 0.95. Which code will return the value of x?

- A) `1 - qnorm(0.95, mean=50, sd=4)`
- B) `qnorm(0.95, mean=50, sd=4)`
- C) `pnorm(0.95, mean=50, sd=4)`
- D) `1 - pnorm(0.95, mean=50, sd=4)`
- **Solution**: B) `qnorm(0.95, mean=50, sd=4)`

19. **Question**: Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Which of the following R statements will find the probability of having four or fewer correct answers if a student attempts to answer every question at random?

- A) `pbinom(4, size=12, prob=0.2)`
- B) `dbinom(4, size=12, prob=0.2)`
- C) `pbinom(5, size=12, prob=4/12)`
- D) `dbinom(5, size=12, prob=0.2)`
- **Solution**: A) `pbinom(4, size=12, prob=0.2)`

20. **Question**: Which code will generate 2000 random variates of a standard normal distribution?

- A) `rvariates <- rnorm(0 - 2000, mean = 1, sd = 0)`
- B) `rvariates <- rnorm(n = 2000, mean = 1, sd = 1)`
- C) `variates <- rnorm(0, 2000, mean = 0, sd = 1)`
- D) `rvariates <- rnorm(n = 2000, mean = 0, sd = 1)`
- **Solution**: D) `rvariates <- rnorm(n = 2000, mean = 0, sd = 1)`

21. **Question**: What is the result of submitting the following R code? `R x <- 1 repeat { print(x) x = x + 1 if (x == 4) { break } }`

- A) Error
- B) `[1] 4`
- C) `[1] 1 [1] 2 [1] 3`
- D) `[1] 1 [1] 2 [1] 3 [1] 4`
- **Solution**: C) `[1] 1 [1] 2 [1] 3`

22. **Question**: Which of the following is NOT true regarding a binomial experiment?

- A) All the trials must be independent.
- B) The number of trials is fixed.
- C) Trials can be a success or a failure only.
- D) The probability of a success varies for each trial.
- **Solution**: D) The probability of a success varies for each trial.

23. **Question**: Given the relation `employee`: `R employee_id name salary 1001 Annie 6000 1009 Ross 4500 1018 Zeith 7000` Which employee_id will be displayed for the given query? `SQL SELECT * FROM employee WHERE employee_id > 1009`

- A) 1001, 1009, 1018
- B) 1009, 1018

- C) 1001
- D) 1018
- **Solution**: D) 1018

24. **Question**: The number of visits on a web page is known to follow a Poisson distribution with mean 15 visits per hour. Which of the answers will NOT give the probability of getting 10 or fewer visits in an hour?

- A) `ppois(10, lambda = 15)`
- B) `1 - ppois(10, lambda = 15, lower.tail = FALSE)`
- C) `sum(ppois(0:10, lambda = 15))`
- D) `sum(dpois(0:10, lambda = 15))`
- **Solution**: C) `sum(ppois(0:10, lambda = 15))`

25. **Question**: An FBI survey shows that about 80% of all property crimes go unsolved. Suppose that in your town 3 such crimes are committed, and they are each deemed independent of each other. This is best represented by the:

- A) Uniform distribution
- B) Normal distribution
- C) Exponential distribution
- D) Binomial distribution
- **Solution**: D) Binomial distribution

26. **Question**: Which of the following R statements will find the probability of having exactly two successes given a binomial random variable bin(6, 0.3)?

- A) `dbinom(2, size=6, prob=0.3)`
- B) `dbinom(6, size=2, prob=0.3)`
- C) `pbinom(2, size=6, prob=0.3)`
- D) `pbinom(6, size=2, prob=0.3)`
- **Solution**: A) `dbinom(2, size=6, prob=0.3)`

27. **Question**: Suppose that the area under the normal curve (X N(50, 4)) to the left of some unknown x-value is 0.95. Which code will return the value of x?

- A) `1 - qnorm(0.95, mean=50, sd=4)`
- B) `qnorm(0.95, mean=50, sd=4)`
- C) `pnorm(0.95, mean=50, sd=4)`
- D) `1 - pnorm(0.95, mean=50, sd=4)`
- **Solution**: B) `qnorm(0.95, mean=50, sd=4)`

28. **Question**: Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Which of the following R statements will find the probability of having four or fewer correct answers if a student attempts to answer every question at random?

- A) `pbinom(4, size=12, prob=0.2)`
- B) `dbinom(4, size=12, prob=0.2)`
- C) `pbinom(5, size=12, prob=4/12)`
- D) `dbinom(5, size=12, prob=0.2)`
- **Solution**: A) `pbinom(4, size=12, prob=0.2)`

29. **Question**: Which code will generate 2000 random variates of a standard normal distribution?

- A) `rvariates <- rnorm(0 - 2000, mean = 1, sd = 0)`
- B) `rvariates <- rnorm(n = 2000, mean = 1, sd = 1)`
- C) `variates <- rnorm(0, 2000, mean = 0, sd = 1)`
- D) `rvariates <- rnorm(n = 2000, mean = 0, sd = 1)`
- **Solution**: D) `rvariates <- rnorm(n = 2000, mean = 0, sd = 1)`

30. **Question**: What is the result of submitting the following R code? `R x <- 1 repeat { print(x) x = x + 1 if (x == 4) { break } }`

- A) Error
- B) `[1] 4`
- C) `[1] 1 [1] 2 [1] 3`
- D) `[1] 1 [1] 2 [1] 3 [1] 4`
- **Solution**: C) `[1] 1 [1] 2 [1] 3`

31. **Question**: Which of the following is NOT true regarding a binomial experiment?

- A) All the trials must be independent.
- B) The number of trials is fixed.
- C) Trials can be a success or a failure only.
- D) The probability of a success varies for each trial.
- **Solution**: D) The probability of a success varies for each trial.

32. **Question**: Given the relation `employee`: `R employee_id name salary 1001 Annie 6000 1009 Ross 4500 1018 Zeith 7000` Which employee_id will be displayed for the given query? `SQL SELECT * FROM employee WHERE employee_id > 1009`

- A) 1001, 1009, 1018
- B) 1009, 1018
- C) 1001
- D) 1018
- **Solution**: D) 1018

33. **Question**: The number of visits on a web page is known to follow a Poisson distribution with mean 15 visits per hour. Which of the answers will NOT give the probability of getting 10 or fewer visits in an hour?

- A) `ppois(10, lambda = 15)`
- B) `1 - ppois(10, lambda = 15, lower.tail = FALSE)`
- C) `sum(ppois(0:10, lambda = 15))`
- D) `sum(dpois(0:10, lambda = 15))`
- **Solution**: C) `sum(ppois(0:10, lambda = 15))`

34. **Question**: An FBI survey shows that about 80% of all property crimes go unsolved. Suppose that in your town 3 such crimes are committed, and they are each deemed independent of each other. This is best represented by the:

- A) Uniform distribution
- B) Normal distribution
- C) Exponential distribution
- D) Binomial distribution
- **Solution**: D) Binomial distribution

35. **Question**: Which of the following R statements will find the probability of having exactly two successes given a binomial random variable bin(6, 0.3)?

- A) `dbinom(2, size=6, prob=0.3)`
- B) `dbinom(6, size=2, prob=0.3)`
- C) `pbinom(2, size=6, prob=0.3)`
- D) `pbinom(6, size=2, prob=0.3)`
- **Solution**: A) `dbinom(2, size=6, prob=0.3)`

36. **Question**: Suppose that the area under the normal curve (X N(50, 4)) to the left of some unknown x-value is 0.95. Which code will return the value of x?

- A) `1 - qnorm(0.95, mean=50, sd=4)`
- B) `qnorm(0.95, mean=50, sd=4)`
- C) `pnorm(0.95, mean=50, sd=4)`
- D) `1 - pnorm(0.95, mean=50, sd=4)`
- **Solution**: B) `qnorm(0.95, mean=50, sd=4)`

37. **Question**: Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Which of the following R statements will find the probability of having four or fewer correct answers if a student attempts to answer every question at random?

- A) `pbinom(4, size=12, prob=0.2)`
- B) `dbinom(4, size=12, prob=0.2)`

- C) `pbinom(5, size=12, prob=4/12)`
- D) `dbinom(5, size=12, prob=0.2)`
- **Solution**: A) `pbinom(4, size=12, prob=0.2)`

38. **Question**: Which code will generate 2000 random variates of a standard normal distribution?

- A) `rvariates <- rnorm(0 - 2000, mean = 1, sd = 0)`
- B) `rvariates <- rnorm(n = 2000, mean = 1, sd = 1)`
- C) `variates <- rnorm(0, 2000, mean = 0, sd = 1)`
- D) `rvariates <- rnorm(n = 2000, mean = 0, sd = 1)`
- **Solution**: D) `rvariates <- rnorm(n = 2000, mean = 0, sd = 1)`

39. **Question**: What is the result of submitting the following R code? `R x <- 1 repeat { print(x) x = x + 1 if (x == 4) { break } }`

- A) Error
- B) `[1] 4`
- C) `[1] 1 [1] 2 [1] 3`
- D) `[1] 1 [1] 2 [1] 3 [1] 4`
- **Solution**: C) `[1] 1 [1] 2 [1] 3`

40. **Question**: Which of the following is NOT true regarding a binomial experiment?

- A) All the trials must be independent.
- B) The number of trials is fixed.
- C) Trials can be a success or a failure only.
- D) The probability of a success varies for each trial.
- **Solution**: D) The probability of a success varies for each trial.

41. **Question**: Given the relation `employee`: `R employee_id name salary 1001 Annie 6000 1009 Ross 4500 1018 Zeith 7000` Which employee_id will be displayed for the given query? `SQL SELECT * FROM employee WHERE employee_id > 1009`

- A) 1001, 1009, 1018
- B) 1009, 1018
- C) 1001
- D) 1018
- **Solution**: D) 1018

42. **Question**: The number of visits on a web page is known to follow a Poisson distribution with mean 15 visits per hour. Which of the answers will NOT give the probability of getting 10 or fewer visits in an hour?

- A) `ppois(10, lambda = 15)`
- B) `1 - ppois(10, lambda = 15, lower.tail = FALSE)`
- C) `sum(ppois(0:10, lambda = 15))`
- D) `sum(dpois(0:10, lambda = 15))`
- **Solution**: C) `sum(ppois(0:10, lambda = 15))`

43. **Question**: An FBI survey shows that about 80% of all property crimes go unsolved. Suppose that in your town 3 such crimes are committed, and they are each deemed independent of each other. This is best represented by the:

- A) Uniform distribution
- B) Normal distribution
- C) Exponential distribution
- D) Binomial distribution
- **Solution**: D) Binomial distribution

44. **Question**: Which of the following R statements will find the probability of having exactly two successes given a binomial random variable bin(6, 0.3)?

- A) `dbinom(2, size=6, prob=0.3)`
- B) `dbinom(6, size=2, prob=0.3)`
- C) `pbinom(2, size=6, prob=0.3)`
- D) `pbinom(6, size=2, prob=0.3)`
- **Solution**: A) `dbinom(2, size=6, prob=0.3)`

45. **Question**: Suppose that the area under the normal curve (X N(50, 4)) to the left of some unknown x-value is 0.95. Which code will return the value of x?

- A) `1 - qnorm(0.95, mean=50, sd=4)`
- B) `qnorm(0.95, mean=50, sd=4)`
- C) `pnorm(0.95, mean=50, sd=4)`
- D) `1 - pnorm(0.95, mean=50, sd=4)`
- **Solution**: B) `qnorm(0.95, mean=50, sd=4)`

46. **Question**: Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Which of the following R statements will find the probability of having four or fewer correct answers if a student attempts to answer every question at random?

- A) `pbinom(4, size=12, prob=0.2)`
- B) `dbinom(4, size=12, prob=0.2)`
- C) `pbinom(5, size=12, prob=4/12)`
- D) `dbinom(5, size=12, prob=0.2)`
- **Solution**: A) `pbinom(4, size=12, prob=0.2)`

47. **Question**: Which code will generate 2000 random variates of a standard normal distribution?

- A) `rvariates <- rnorm(0 - 2000, mean = 1, sd = 0)`
- B) `rvariates <- rnorm(n = 2000, mean = 1, sd = 1)`
- C) `variates <- rnorm(0, 2000, mean = 0, sd = 1)`
- D) `rvariates <- rnorm(n = 2000, mean = 0, sd = 1)`
- **Solution**: D) `rvariates <- rnorm(n = 2000, mean = 0, sd = 1)`

48. **Question**: What is the result of submitting the following R code? `R x <- 1 repeat { print(x) x = x + 1 if (x == 4) { break } }`

- A) Error
- B) `[1] 4`
- C) `[1] 1 [1] 2 [1] 3`
- D) `[1] 1 [1] 2 [1] 3 [1] 4`
- **Solution**: C) `[1] 1 [1] 2 [1] 3`

49. **Question**: Which of the following is NOT true regarding a binomial experiment?

- A) All the trials must be independent.
- B) The number of trials is fixed.
- C) Trials can be a success or a failure only.
- D) The probability of a success varies for each trial.
- **Solution**: D) The probability of a success varies for each trial.

50. **Question**: Given the relation `employee`: `R employee_id name salary 1001 Annie 6000 1009 Ross 4500 1018 Zeith 7000` Which employee_id will be displayed for the given query? `SQL SELECT * FROM employee WHERE employee_id > 1009`

- A) 1001, 1009, 1018
- B) 1009, 1018
- C) 1001
- D) 1018
- **Solution**: D) 1018

Data Analysis

31. **Question**: What is Exploratory Data Analysis (EDA)?

    - A) A technique to explore unknown data.
    - B) A technique to summarize the main characteristics of the data.

- C) A technique to create predictive models.
- D) A technique to clean the data.
- **Solution**: B) A technique to summarize the main characteristics of the data.

32. **Question**: How do you create a histogram of a numeric vector `x` in R?

    - A) `barplot(x)`
    - B) `plot(x)`
    - C) `hist(x)`
    - D) `pie(x)`
    - **Solution**: C) `hist(x)`

33. **Question**: Write R code to create a scatter plot of `x` vs `y`.

    - A) `plot(y ~ x)`
    - B) `plot(x, y)`
    - C) `scatter(x, y)`
    - D) `plot(x ~ y)`
    - **Solution**: B) `plot(x, y)`

34. **Question**: Write R code to split a data frame `df` into a training set (70%) and a test set (30%).

    - A)

    ```r
    set.seed(123)
    sample <- sample.int(n = nrow(df), size = floor(.70 * nrow(df)), replace = F)
    train <- df[sample, ]
    test <- df[-sample, ]
    ```

    - B)

    ```r
    set.seed(123)
    sample <- sample.int(n = nrow(df), size = floor(.70 * nrow(df)), replace = T)
    train <- df[sample, ]
    test <- df[-sample, ]
    ```

    - C)

    ```r
    set.seed(123)
    sample <- sample.int(n = nrow(df), size = floor(.70 * nrow(df)), replace = F)
    train <- df[-sample, ]
    test <- df[sample, ]
    ```

    - D)

```
set.seed(123)
sample <- sample.int(n = nrow(df), size = floor(.70 * nrow(df)), replace = T)
train <- df[-sample, ]
test <- df[sample, ]
```

- **Solution**: A)

```
set.seed(123)
sample <- sample.int(n = nrow(df), size = floor(.70 * nrow(df)), replace = F)
train <- df[sample, ]
test <- df[-sample, ]
```

35. **Question**: How do you normalize a numeric vector `x` in R?
    - A) `(x - min(x)) / (max(x) - min(x))`
    - B) `(x - mean(x)) / sd(x)`
    - C) `(x - min(x)) / sd(x)`
    - D) `(x - mean(x)) / (max(x) - min(x))`
    - **Solution**: A) `(x - min(x)) / (max(x) - min(x))`

Miscellaneous

41. **Question**: What does ETL stand for in data processing?
    - A) Extract, Transform, Load
    - B) Extract, Transfer, Load
    - C) Extract, Transform, Link
    - D) Extract, Transfer, Link
    - **Solution**: A) Extract, Transform, Load
42. **Question**: Which of the following is NOT a type of join in SQL?
    - A) INNER JOIN
    - B) LEFT JOIN
    - C) RIGHT JOIN
    - D) TOP JOIN
    - **Solution**: D) TOP JOIN
43. **Question**: How do you find the median of a numeric vector `x` in R?
    - A) `median(x)`
    - B) `mean(x)`
    - C) `mid(x)`
    - D) `mode(x)`
    - **Solution**: A) `median(x)`

44. **Question**: Which of the following is used to handle missing values in R?
    - A) `is.na()`
    - B) `na.omit()`
    - C) `complete.cases()`
    - D) All of the above
    - **Solution**: D) All of the above

45. **Question**: What is the function to compute the variance of a numeric vector `x` in R?
    - A) `var(x)`
    - B) `variance(x)`
    - C) `sd(x)`
    - D) `sqrt(var(x))`
    - **Solution**: A) `var(x)`

46. **Question**: Which SQL statement is used to remove a table?
    - A) `REMOVE TABLE table_name;`
    - B) `DROP TABLE table_name;`
    - C) `DELETE TABLE table_name;`
    - D) `TRUNCATE TABLE table_name;`
    - **Solution**: B) `DROP TABLE table_name;`

47. **Question**: What package in R is used for data manipulation with data frames?
    - A) `ggplot2`
    - B) `dplyr`
    - C) `tidyr`
    - D) `shiny`
    - **Solution**: B) `dplyr`

48. **Question**: Which function in R is used to create a boxplot?
    - A) `boxplot()`
    - B) `plot()`
    - C) `hist()`
    - D) `barplot()`
    - **Solution**: A) `boxplot()`

49. **Question**: How do you find the number of rows in a data frame `df` in R?
    - A) `length(df)`
    - B) `nrow(df)`
    - C) `rows(df)`
    - D) `dim(df)[1]`
    - **Solution**: B) `nrow(df)`

50. **Question**: What is a foreign key in SQL?

- A) A field in a table that uniquely identifies each row/record in that table.
- B) A field in a table that is the primary key in another table.
- C) A field in a table that stores foreign data.
- D) A field in a table that refers to a foreign record.
- **Solution**: B) A field in a table that is the primary key in another table.