DATA 505 (Statistics Using R): Exam 1 Practice

Part 1: Closed Computer

Name: _____

Instructions:

This is a practice exam designed to help you prepare for the actual Exam 1. The format, question types, and topics covered are similar to what you will encounter on the real exam. Use this to identify areas where you need additional study.

The actual exam (like this practice) will have two parts:

- *Part 1 (30 points, approximately 45 minutes):*
  - *Closed computer and closed notes*
  - *Handwritten answers*
- *Part 2 (20 points, approximately 30-45 minutes):*
  - *Open computer. Any online resource that is "not alive" may be used (i.e. you may freely use online resources, but you may not communicate with any other person).*
  - *Electronic submission on Moodle: you will fill in a Quarto script with R code and written answers*

Once you turn in Part 1, you may open your computer and download the blank Quarto script for Part 2. You may not return to Part 1 after turning it in and beginning the open-computer portion.

PRACTICE EXAM - NOT FOR CREDIT

Part 1A (Closed computer): Multiple Choice

1

Which of the following creates a vector with the values 1, 3, 5, 7, 9 in R? *(Circle all that apply)*

A. `c(1, 3, 5, 7, 9)`

B. `seq(1, 9, by = 2)`

C. `1:9[c(1,3,5,7,9)]`

D. `rep(c(1,3,5,7,9), times = 1)`

2

Match each R expression with its corresponding `class()` output. Write the letter of the correct output next to each number.

Possible outputs:

A. `"numeric"`

B. `"factor"`

C. `"data.frame"`

D. `"character"`

E. `"logical"`

Expressions:

1. `class(3.14)` _____**A**
2. `class(factor(c("A", "B", "C")))` _____**B**
3. `class(c("hello", "world"))` _____**D**
4. `class(5 > 3)` _____**E**
5. `class(data.frame(x = 1:3, y = 4:6))` _____**C**

3

You have a vector called `ages` with values `c(25, 30, 35, 40, 45)`. Which of the following would extract the elements greater than 32?

A. `ages[ages > 32]`

B. `subset(ages, ages > 32)`

C. `ages[c(3, 4, 5)]`

D. **Both A and B are correct**

E. All of the above are correct

4

In hypothesis testing, what does it mean to "reject the null hypothesis"?

A. We have proven the null hypothesis is false

B. **We have strong evidence against the null hypothesis**

C. The alternative hypothesis is definitely true

D. We made a Type II error

E. The p-value is greater than the significance level

5

Which of the following best describes the interpretation of a 95% confidence interval for a population mean?

A. 95% of the data values fall within this interval

B. There is a 95% chance the sample mean falls in this interval

C. **We are 95% confident the population mean falls within this interval**

D. 95% of all possible confidence intervals will contain our sample mean

E. The population mean has a 95% probability of being in this interval

6

What is the primary purpose of using factors in R?

A. To perform faster mathematical computations

B. **To represent categorical variables with specific levels**

C. To store large amounts of text data efficiently

D. To create multi-dimensional arrays

E. To handle missing values automatically

7

Which measure of spread is MOST affected by outliers?

A. Standard deviation

B. Interquartile range (IQR)

C. Median absolute deviation

D. **Range**

E. Both A and D are equally affected

8

Consider the following analysis examining whether flower species affects petal length, i.e. testing

$$H_0 : \mu_1 = \mu_2$$

against

$$H_A : \mu_1 \neq \mu_2,$$

where μ_1 and μ_2 are the mean sepal length for the setosa and versicolor species, respectively.

```r
# Load iris data and create subset
data("iris")
petal_data <- iris[iris$Species %in% c("setosa", "versicolor"), ]

# Perform t-test comparing petal lengths
t.test(Petal.Length ~ Species, data = petal_data)
```

```
    Welch Two Sample t-test

data:  Petal.Length by Species
t = -39.493, df = 62.14, p-value < 2.2e-16
alternative hypothesis: true difference in means between group setosa and group versicolor is
95 percent confidence interval:
 -2.939618 -2.656382
sample estimates:
    mean in group setosa mean in group versicolor
                   1.462                    4.260
```

Based on this output, which conclusion is correct at the $\alpha = 0.05$ significance level?

A. We fail to reject H_0. There is no difference in mean petal length between species.

B. **We reject H_0. There is significant evidence of a difference in mean petal length between species.**

C. We cannot make a conclusion because the sample sizes are too small.

D. The test is invalid because the data are not normally distributed.

E. We have proven that the species have identical mean petal lengths.

Part 1B (Closed computer): Case Study

For the following questions, write your answers in complete sentences. Show all R code where requested.

We will analyze data from the built-in `mtcars` dataset to investigate whether the proportion of high-performance cars (defined as having 8 cylinders) differs between automatic and manual transmission vehicles. Here is the data structure:

```
# Load and examine the mtcars data
data("mtcars")
# Create categorical variables
mtcars$transmission <- factor(mtcars$am,
                              levels = c(1, 0),
                              labels = c("Manual", "Automatic"))
mtcars$high_performance <- factor(mtcars$cyl == 8,
                                  levels = c(TRUE, FALSE),
                                  labels = c("High Performance", "Standard"))
str(mtcars[, c("transmission", "high_performance", "cyl")])
```

```
'data.frame':   32 obs. of  3 variables:
 $ transmission    : Factor w/ 2 levels "Manual","Automatic": 1 1 1 2 2 2 2 2 2 2 ...
 $ high_performance: Factor w/ 2 levels "High Performance",..: 2 2 2 2 1 2 1 2 2 2 ...
 $ cyl             : num  6 6 4 6 8 6 8 4 4 6 ...
```

1 (3 points)

Write R code that would create a frequency table showing the counts of high-performance versus standard cars in the dataset.

```
table(mtcars$high_performance)
```

```
High Performance         Standard
              14               18
```

2 (3 points)

Write R code that would allow you to determine the proportion of high-performance cars specifically among manual transmission vehicles.

```r
# example 1
mean(mtcars$high_performance[mtcars$transmission == "Manual"] == "High Performance")
```

```
[1] 0.1538462
```

```r
# example 2
with(mtcars, prop.table(table(transmission, high_performance), margin = 1))
```

```
            high_performance
transmission High Performance  Standard
   Manual           0.1538462 0.8461538
   Automatic        0.6315789 0.3684211
```

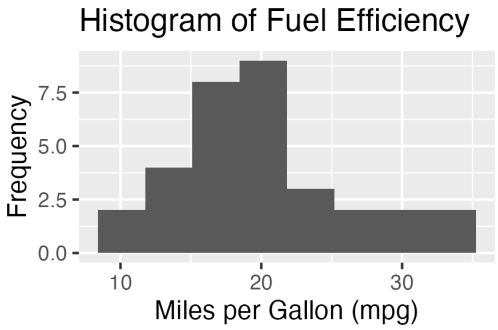
```r
# example 3
manual_cars <- mtcars[mtcars$transmission == "Manual",]
prop.table(table(manual_cars$high_performance))
```

```
High Performance         Standard
       0.1538462        0.8461538
```

3 (4 points)

Based on the histogram and summary statistics shown below, describe the distribution of fuel efficiency (mpg) in this dataset. What do you notice about the shape, center, and spread of the distribution?

Graph:



Descriptive statistics:

Statistic	Value
Mean	20.1
Median	19.2
Standard Deviation	6.0
Q1 (25th percentile)	15.4
Q3 (75th percentile)	22.8
Minimum	10.4
Maximum	33.9

The distribution is roughly symmetric. The median, 19.2 MPG, is representative of the center of the distribution. The middle 50% of cars have gas mileage between 15.4 and 22.8 MPG, giving a sense of spread.

4 (3 points)

If we hypothesize that manual transmission cars are more likely to be high-performance than automatic transmission cars, write the null and alternative hypotheses using proper statistical notation. Define your parameters clearly.

Let p_M and p_A be the proportion of high-performance cars among manual and automatic transmission cars, respectively. We wish to test

$$H_0 : p_M = p_A$$

against

$$H_A : p_M > p_A$$

5 (4 points)

Based on the proportion test results shown below, draw a conclusion about your hypotheses from the previous question. Interpret your results in the context of car performance and transmission type. Use $\alpha = 0.05$.

```
# Create table for prop.test
perf_table <- table(mtcars$transmission, mtcars$high_performance)
perf_table
```

```
            High Performance Standard
  Manual                   2       11
  Automatic               12        7
```

```r
# Proportion of high-performance by transmission type
prop.table(perf_table, margin = 1)
```

```
            High Performance  Standard
  Manual           0.1538462 0.8461538
  Automatic        0.6315789 0.3684211
```

```r
# Two-sample proportion test
prop.test(perf_table, alternative = "greater")  # automatic < manual
```

```
    2-sample test for equality of proportions with continuity correction

data:  perf_table
X-squared = 5.3488, df = 1, p-value = 0.9896
alternative hypothesis: greater
95 percent confidence interval:
 -0.7879206  1.0000000
sample estimates:
   prop 1    prop 2
0.1538462 0.6315789
```

The p-value is large (greater than 0.05). We fail to reject H_0. There is no evidence in this sample that manual-transmission cars are more likely to be high-performance. In fact the evidence may suggest the opposite.