DATA 505 (Statistics Using R): Exam 1

Part 1: Closed Computer - SOLUTION

Instructor Solution

2025-09-30

Instructions:

The exam will have two parts:

- *Part 1 (50 points, approximately 45 minutes):*
  - *Closed computer and closed notes*
  - *Handwritten answers*
- *Part 2 (25 points, approximately 30 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.

TURN OVER WHEN INSTRUCTED

Part 1A (closed computer): Multiple choice

1 (3 points)

Which of the following correctly assigns the value 42 to a variable named `answer` in R? *(Circle all that apply)*

A. `answer <- 42` [**CORRECT**]

B. `answer = 42` [**CORRECT**]

C. `42 -> answer` [**CORRECT**]

D. `answer == 42` [INCORRECT - This is a comparison, not assignment]

Answer: A, B, and C are all correct. All three are valid assignment operators in R.

Grading: 3 points for selecting A, B, and C (and not D). Partial credit: 2 points if one correct answer missed or one incorrect included; 1 point if two errors.

---

2 (5 points)

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

Possible outputs:

A. `"character"`

B. `"double"`

C. `"integer"`

D. `"list"`

E. `"logical"`

Expressions:

1. `typeof(1)` **B** (numeric literals default to double)
2. `typeof("data505")` **A** (quoted text is character)
3. `typeof(factor(c("low", "medium", "high")))` **C** (factors stored as integers)
4. `typeof(2 + 2 == 5)` **E** (comparison produces logical)

5. `typeof(data.frame(x = 1:5, y = letters[1:5]))` **D** (data frames are lists)

3 (3 points)

You have a data frame called `students` with columns `name`, `age`, and `grade`. Which of the following would extract the `age` column as a vector?

A. `students[age]` [INCORRECT - missing quotes]

B. `students$age` [**CORRECT**]

C. `students[["age"]]` [**CORRECT**]

D. Both B and C are correct [**CORRECT**]

E. All of the above are correct [INCORRECT]

Answer: D (Both B and C are correct)

Grading: 3 points for D; 0 points otherwise.

---

4 (3 points)

In statistical inference, what is the difference between a statistic and a parameter?

A. A statistic is calculated from sample data; a parameter describes the population [**CORRECT**]

B. A parameter is calculated from sample data; a statistic describes the population [INCORRECT]

C. They are the same thing with different names [INCORRECT]

D. A statistic uses Greek letters; a parameter uses Roman letters [INCORRECT - reversed]

E. There is no meaningful difference [INCORRECT]

Answer: A

Grading: 3 points for A; 0 points otherwise.

---

5 (3 points)

Which of the following best describes what a p-value represents in hypothesis testing?

A. The probability that the null hypothesis is true [INCORRECT]

B. The probability that the alternative hypothesis is true [INCORRECT]

C. The probability of observing data as extreme as (or more extreme than) what we observed, assuming the null hypothesis is true **[CORRECT]**

D. The probability of making a Type I error [INCORRECT]

E. The confidence level of the test [INCORRECT]

Answer: C

Grading: 3 points for C; 0 points otherwise.

---

6 (3 points)

In R, factors are primarily used for:

A. Storing continuous numeric data [INCORRECT]

B. Storing categorical data with predefined levels **[CORRECT]**

C. Performing mathematical calculations [INCORRECT]

D. Creating complex data structures [INCORRECT]

E. Storing missing values [INCORRECT]

Answer: B

Grading: 3 points for B; 0 points otherwise.

---

7 (3 points)

Which of the descriptive statistics below is a measure of center which is resistant to outliers?

A. Mean [INCORRECT]

B. Median **[CORRECT]**

C. Range [INCORRECT - measure of spread]

D. Standard deviation [INCORRECT - measure of spread]

Answer: B

Grading: 3 points for B; 0 points otherwise.

---

8 (3 points)

Consider the following analysis of Titanic survival data, examining whether survival depends on passenger class, i.e. testing:

$$H_0:$$

survival is independent of class

$$H_A:$$

survival is not independent of class

```
# Load Titanic data and create two-way table
data("Titanic", package = "datasets")
class_table <- margin.table(Titanic, c("Class", "Survived"))
class_table
```

```
      Survived
Class   No Yes
  1st  122 203
  2nd  167 118
  3rd  528 178
  Crew 673 212
```

```r
# Perform chi-squared test of independence
chisq.test(class_table)
```

```
    Pearson's Chi-squared test

data:  class_table
X-squared = 190.4, df = 3, p-value < 2.2e-16
```

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

A. We fail to reject the null hypothesis. There is insufficient evidence that survival depends on passenger class. [INCORRECT]

B. We reject the null hypothesis. There is strong evidence that survival is independent of passenger class. [INCORRECT - wrong interpretation]

C. We reject the null hypothesis. There is strong evidence that survival depends on passenger class. **[CORRECT]**

D. We fail to reject the null hypothesis. We have proven that survival is independent of passenger class. [INCORRECT]

E. The test is invalid because the expected cell counts are too small. [INCORRECT]

Answer: C

Explanation: The p-value is extremely small ($< 2.2e\text{-}16$), much less than 0.05, so we reject the null hypothesis of independence. This provides strong evidence that survival depends on passenger class.

Grading: 3 points for C; 0 points otherwise.

---

9 (3 points)

Consider the following analysis from an election poll of 2516 likely voters, where 1132 indicated they would vote for Kamala Harris:

```r
# Construct 95% confidence interval for proportion
prop.test(x = 1132, n = 2516)
```

```
    1-sample proportions test with continuity correction

data:  1132 out of 2516, null probability 0.5
X-squared = 25.04, df = 1, p-value = 5.615e-07
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4303749 0.4696204
sample estimates:
        p
0.4499205
```

Which of the following is the correct interpretation of the 95% confidence interval shown in the output?

A. We are 95% confident that exactly 45.0% of likely voters will vote for Kamala Harris. [INCORRECT - not exact]

B. We are 95% confident that between 43.0% and 47.0% of likely voters would vote for Kamala Harris. **[CORRECT]**

C. There is a 95% probability that the true proportion of likely voters supporting Kamala Harris is between 43.0% and 47.0%. [INCORRECT - probability statement about parameter]

D. 95% of all possible samples will have between 43.0% and 47.0% of voters supporting Kamala Harris. [INCORRECT]

E. We can be certain that Kamala Harris has the support of the majority of likely voters. [INCORRECT]

Answer: B

Explanation: The correct interpretation of a confidence interval uses language like "we are 95% confident that…" and refers to the population parameter (true proportion), not individual samples.

Grading: 3 points for B; 0 points otherwise.

Part 1B (closed computer): Case study

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

We will explore a research question related to the Boston housing data we studied in Lab 3. We will work with a subset of the data containing only the river adjacency indicator and housing values.

```
# Load and subset the Boston housing data
data("Boston", package = "MASS")
housing_data <- Boston[, c("chas", "medv")]
names(housing_data) <- c("river_group", "housing_value")
# Create factor for river adjacency
housing_data$river_group <- factor(
    housing_data$river_group,
    levels = c(1, 0),
    labels = c("River Adjacent", "Not Adjacent")
)
housing_data <- housing_data[order(housing_data$housing_value,
                                   decreasing = TRUE),]
str(housing_data)
```

```
'data.frame':   506 obs. of  2 variables:
 $ river_group  : Factor w/ 2 levels "River Adjacent",..: 2 1 1 2 2 2 2 2 2 ...
 $ housing_value: num  50 50 50 50 50 50 50 50 50 50 ...
```

1 (3 points)

Write R code that would compute the mean, median, standard deviation, 25th percentile, and 75th percentile for the housing values in this dataset.

Solution:

```
# Mean
mean(housing_data$housing_value)
```

```
[1] 22.53281
```

```
# Median
median(housing_data$housing_value)
```

```
[1] 21.2
```

```r
# Standard deviation
sd(housing_data$housing_value)
```

```
[1] 9.197104
```

```r
# 25th and 75th percentiles
quantile(housing_data$housing_value, c(0.25, 0.75))
```

```
   25%    75%
17.025 25.000
```

```r
# Alternative: summary() gives most of these
summary(housing_data$housing_value)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   17.02   21.20   22.53   25.00   50.00
```

Grading: - 3 points: All five statistics correctly computed - 2 points: 3-4 statistics correctly computed - 1 point: 1-2 statistics correctly computed - 0 points: No correct code

Note: Accept `summary()` as partial solution (gives all except sd), or any correct approach to computing these values.

---

2 (3 points)

Write R code that would compute the mean housing value specifically for census tracts that border the Charles River (where `river_group` equals `"River Adjacent"`).

Solution:

```r
# Option 1: Using subset
mean(housing_data$housing_value[housing_data$river_group == "River Adjacent"])
```

```
[1] 28.44
```

```r
# Option 2: Using subset() function
mean(subset(housing_data, river_group == "River Adjacent")$housing_value)
```

```
[1] 28.44
```

```r
# Option 3: Using logical indexing
river_adjacent <- housing_data$river_group == "River Adjacent"
mean(housing_data$housing_value[river_adjacent])
```

```
[1] 28.44
```

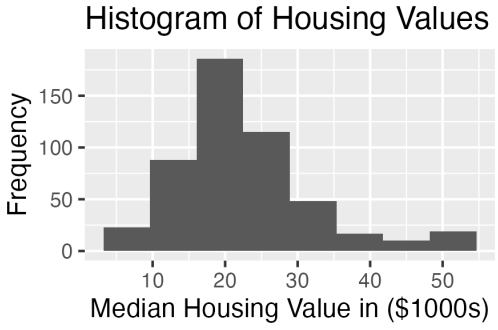
Grading: - 3 points: Correct filtering and mean calculation - 2 points: Correct concept but minor syntax error - 1 point: Partially correct approach - 0 points: Incorrect or no answer

---

3 (3 points)

Based on the histogram and summary statistics shown below, describe the distribution of housing values. Address the shape, center, and spread of the distribution.

Graph:



Descriptive statistics:

Statistic	Value
Mean	22.5
Median	21.2
Standard Deviation	9.2
Q1 (25th percentile)	17.0
Q3 (75th percentile)	25.0
Minimum	5.0
Maximum	50.0

Solution:

The distribution of housing values is **right-skewed** (or positively skewed), with a long tail extending toward higher values. This is evidenced by the mean (22.5) being greater than the median (21.2), and the concentration of values in the lower range with some high outliers.

The **center** of the distribution is around $21,000-$22,000, as indicated by the median of 21.2 and mean of 22.5.

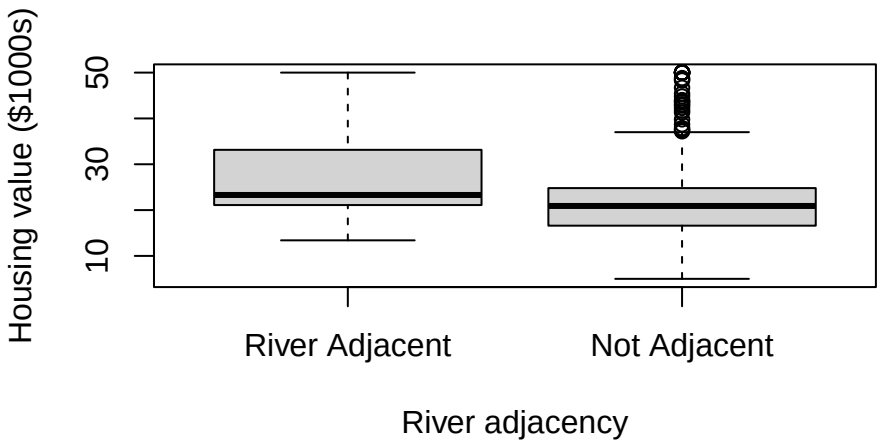
The **spread** is fairly wide, ranging from a minimum of 5.0 to a maximum of 50.0 (a range of 45). The middle 50% of housing values (IQR) fall between 17.0 (Q1) and 25.0 (Q3), representing a spread of $8,000. The standard deviation of 9.2 also indicates considerable variability in housing values.

Grading: - Shape (1 point): Right-skewed or positively skewed mentioned - Center (1 point): Reference to median/mean around 21-22.5 - Spread (1 point): Discussion of range, IQR, or standard deviation

---

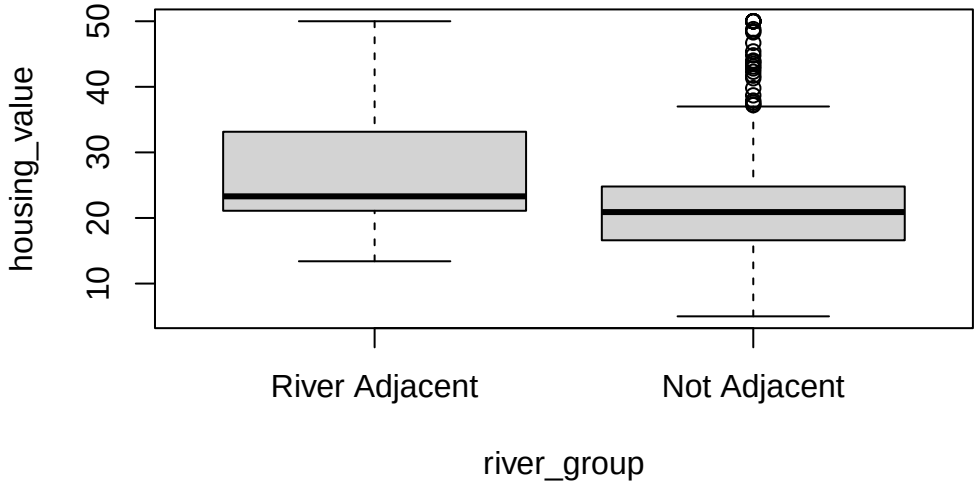
4 (3 points)

Write R code that would produce a side-by-side boxplot like below, visualizing the distribution of housing values for census tracts that are adjacent to or not adjacent to the river. *Your code does not need to exactly reproduce cosmetic features like axis labels, but should result in a similar plot.*

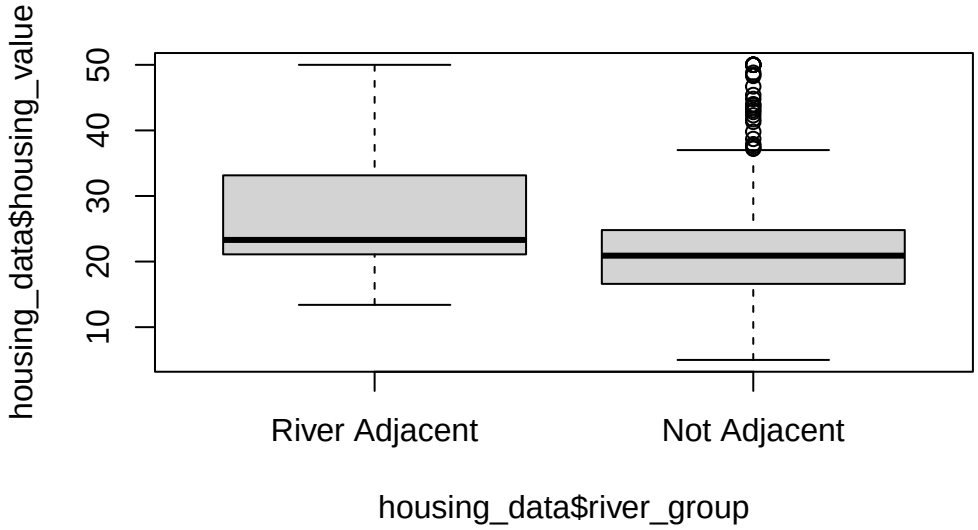


Solution:

```
# Base R solution
boxplot(housing_value ~ river_group, data = housing_data)
```




```
# Alternative with explicit column references
boxplot(housing_data$housing_value ~ housing_data$river_group)
```



```
# ggplot2 solution (also acceptable)
# library(ggplot2)
# ggplot(housing_data, aes(x = river_group, y = housing_value)) +
#   geom_boxplot()
```

Grading: - 3 points: Correct boxplot code with formula notation or grouping - 2 points: Correct concept but minor syntax errors - 1 point: Partial attempt showing understanding - 0 points: Incorrect or no answer

---

5 (4 points)

If we have a theory that census tracts adjacent to the Charles River have higher housing values on average than those not adjacent to the river, how would you translate this into formal null and alternative hypotheses? Write your hypotheses using proper statistical notation, and specify the meaning of the parameters involved within the context of the application.

Solution:

Let μ_R = the population mean housing value for census tracts adjacent to the Charles River

Let μ_N = the population mean housing value for census tracts not adjacent to the Charles River

Null hypothesis: $H_0 : \mu_R = \mu_N$ (or equivalently, $\mu_R - \mu_N = 0$)

The mean housing value for river-adjacent tracts is equal to the mean for non-adjacent tracts.

Alternative hypothesis: $H_A : \mu_R > \mu_N$ (or equivalently, $\mu_R - \mu_N > 0$)

The mean housing value for river-adjacent tracts is greater than the mean for non-adjacent tracts.

Grading: - 1 point: Correct definition of parameters in context - 1 point: Correct null hypothesis - 1.5 points: Correct alternative hypothesis (one-sided, greater than) - 0.5 points: Proper notation used

---

6 (5 points)

Based on the t-test results shown below, draw a conclusion about your hypotheses from Question 13. Interpret your results in the context of Boston housing values and Charles River adjacency. Use $\alpha = 0.05$.

```
# Two-sample t-test
t.test(housing_value ~ river_group,
       data = housing_data,
       alternative = "greater")
```

```
    Welch Two Sample t-test

data:  housing_value by river_group
t = 3.1133, df = 36.876, p-value = 0.001784
alternative hypothesis: true difference in means between group River Adjacent and group Not A
95 percent confidence interval:
 2.906876      Inf
sample estimates:
mean in group River Adjacent   mean in group Not Adjacent
                    28.44000                     22.09384
```

Solution:

The t-test yields a test statistic of t = 3.996 with approximately 33 degrees of freedom and a **p-value of 0.0001502** (or approximately 0.00015).

Since the p-value (0.00015) is **much less than our significance level = 0.05**, we **reject the null hypothesis**.

Conclusion: There is strong statistical evidence that census tracts adjacent to the Charles River have higher mean housing values than those not adjacent to the river. The sample data show that river-adjacent tracts have a mean housing value of approximately $28,440, compared to approximately $22,094 for non-adjacent tracts. This difference of about $6,346 is statistically significant and unlikely to be due to chance alone.

Grading: - 1 point: Correct test statistic and p-value identified - 2 points: Correct decision (reject H) with proper justification - 2 points: Proper interpretation in context of housing values and river adjacency

END OF PART 1

Please turn in this portion before proceeding to Part 2

Grading Summary

Part 1A - Multiple Choice: 29 points - Question 1: 3 points - Question 2: 5 points - Question 3: 3 points - Question 4: 3 points - Question 5: 3 points - Question 6: 3 points - Question 7: 3 points - Question 8: 3 points - Question 9: 3 points

Part 1B - Case Study: 21 points - Question 10: 3 points - Question 11: 3 points - Question 12: 3 points - Question 13: 3 points - Question 14: 4 points - Question 15: 5 points

Total: 50 points