DATA 505 (Statistics Using R): Final Exam

Part 1 Practice Exam

Name: _____

2025-12-09

Instructions:

This is a practice exam for Part 1 of the final. The actual exam will have two parts:

- ***Part 1 (closed computer):*** *Multiple choice and short answer questions. You may not use your computer or any notes. You will write answers on this exam sheet.*
- ***Part 2 (open computer):*** *Coding and analysis problems. You may use your computer and any online resources that are "not alive" (i.e., you may freely use online resources, but you may not communicate with any other person). You will submit a completed Quarto document.*

PRACTICE EXAM - PART 1

Part 1A: Multiple choice

1

Which of the following R expressions would result in an error?

A. `x <- c(1, 2, 3)`

B. `y <- x + 5`

C. `z <- mean(w)`

D. `a <- "hello"`

2

You have a vector `scores <- c(85, 92, 78, 95, 88)`. Which command extracts the values greater than 85?

A. `scores[scores > 85]`

B. `scores > 85`

C. `subset(scores > 85)`

D. `filter(scores, > 85)`

3

You conduct a hypothesis test and obtain a p-value of 0.03. Using $\alpha = 0.05$, what is the appropriate conclusion?

A. Fail to reject the null hypothesis

B. Reject the null hypothesis

C. Accept the alternative hypothesis as proven

D. The test is invalid

4

Which R function would you use to test whether a coin is fair (i.e., the probability of heads equals 0.5) based on observing 60 heads in 100 flips?

A. `t.test()`

B. `chisq.test()`

C. `prop.test()`

D. `lm()`

5

In a regression equation $\hat{Y} = 10 + 3X$, what is the predicted value of Y when $X = 5$?

A. 15

B. 18

C. 25

D. 35

6

You fit a linear regression model to predict student test scores from hours studied. The slope coefficient is 4.2 with a p-value of 0.001. What does this tell you?

A. Hours studied does not significantly predict test scores

B. For each additional hour studied, test scores increase by about 4.2 points on average, and this relationship is statistically significant

C. The model explains 4.2% of the variance

D. Students study an average of 4.2 hours

7

Which diagnostic plot would best help you assess whether the constant variance assumption of linear regression is satisfied?

A. Histogram of the response variable

B. Scatterplot of predictor vs response

C. Residuals vs. fitted values plot

D. Boxplot of the residuals

8

Consider this R function:

```r
summarize_data <- function(x) {
  list(mean = mean(x), sd = sd(x))
}
```

What does this function return when called with a numeric vector?

A. Only the mean

B. Only the standard deviation

C. A list containing both the mean and standard deviation

D. A data frame with mean and sd columns

9

To use logistic regression in R, which family argument should be specified in `glm()`?

A. `family = binomial`

B. `family = gaussian`

C. `family = normal`

D. `family = logistic`

10

In logistic regression, predictions from the model represent:

A. The exact outcome (0 or 1)

B. The probability that the outcome equals 1

C. The log-odds of the outcome

D. The residual error

Part 1B: Short answer

Write your answers in complete sentences where appropriate.

11

Hypothesis Testing: A pharmaceutical company tests a new drug's effect on blood pressure reduction. They measure blood pressure reduction (in mmHg) for 30 patients who took the drug and 30 patients who took a placebo.

The output of a two-sample t-test is shown below:

```
    Welch Two Sample t-test

data:  reduction_drug and reduction_placebo
t = 3.45, df = 57.2, p-value = 0.0011
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.2  11.8
sample estimates:
mean of x mean of y
    15.2       7.7
```

What are the null and alternative hypotheses being tested? Based on the output and using $\alpha = 0.05$, what is your conclusion about the drug's effectiveness?

12

Regression Analysis: A researcher fits a linear regression model to predict monthly electricity usage (in kWh) based on the number of occupants in a household.

```
Call:
lm(formula = electricity ~ occupants, data = usage_data)

Residuals:
    Min      1Q  Median      3Q     Max
-31.545 -17.842  -0.967  13.923  52.621

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  122.416     14.982   8.171 3.03e-06 ***
occupants     83.428      4.238  19.687 1.68e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.35 on 12 degrees of freedom
Multiple R-squared:   0.97, Adjusted R-squared:  0.9675
F-statistic: 387.6 on 1 and 12 DF,  p-value: 1.676e-10
```

Based on the regression output above, what is the estimated increase in monthly electricity usage for each additional household occupant? Is this relationship statistically significant at the $\alpha = 0.05$ level?

13

Making Predictions: Using the regression model from the previous question, predict the monthly electricity usage for a household with 7 occupants. Show your calculation.

14

Writing Functions: Write an R function named `grade_exam` that takes a single numeric argument `score` representing an exam score out of 100. - If the score is 90 or above, return `"A"` - If the score is 80-89, return `"B"`
- If the score is 70-79, return `"C"` - If the score is 60-69, return `"D"` - If the score is below 60, return `"F"`