DATA 505 (Statistics Using R): Final Exam

Part 1 Practice Exam - SOLUTION KEY

Instructor Solution

2025-12-09

Instructions:

This is the solution key for the Part 1 practice exam.

PRACTICE EXAM - PART 1 SOLUTIONS

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)` ← **CORRECT**

D. `a <- "hello"`

Answer: C

Explanation: Option C would result in an error because the object `w` has not been defined. R would return an error message like "object 'w' not found". The other options would all execute successfully.

2

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

A. `scores[scores > 85]` ← **CORRECT**

B. `scores > 85`

C. `subset(scores > 85)`

D. `filter(scores, > 85)`

Answer: A

Explanation: Option A uses logical indexing to extract values. `scores > 85` creates a logical vector, and `scores[scores > 85]` uses it to subset. Option B would only return the logical vector (TRUE/FALSE values), not the actual values. Options C and D have incorrect syntax.

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 ← **CORRECT**

C. Accept the alternative hypothesis as proven

D. The test is invalid

Answer: B

Explanation: Since the p-value (0.03) is less than the significance level α (0.05), we reject the null hypothesis. We never "accept" or "prove" hypotheses in statistical testing - we only reject or fail to reject the null hypothesis.

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()` ← **CORRECT**

D. `lm()`

Answer: C

Explanation: `prop.test()` is used to test hypotheses about proportions. This scenario involves testing whether the true proportion of heads equals 0.5. The syntax would be `prop.test(60, 100, p = 0.5)`. A t-test is for means, chi-squared test is typically for categorical associations, and `lm()` is for regression.

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 ← **CORRECT**

D. 35

Answer: C

Explanation: Substitute $X = 5$ into the equation: $\hat{Y} = 10 + 3(5) = 10 + 15 = 25$.

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 ← **CORRECT**

C. The model explains 4.2% of the variance

D. Students study an average of 4.2 hours

Answer: B

Explanation: The slope coefficient (4.2) represents the average change in the response variable (test scores) for a one-unit increase in the predictor (hours studied). The small p-value ($0.001 < 0.05$) indicates this relationship is statistically significant.

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 ← **CORRECT**

D. Boxplot of the residuals

Answer: C

Explanation: The residuals vs. fitted values plot is used to check for constant variance (homoscedasticity). If the variance is constant, the vertical spread of residuals should be roughly the same across all fitted values. A funnel shape would indicate non-constant variance.

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 ← **CORRECT**

D. A data frame with mean and sd columns

Answer: C

Explanation: The function uses `list()` to create a list with two named elements: `mean` and `sd`. When called, it returns a list structure containing both values. For example, `summarize_data(c(1,2,3))` would return `list(mean = 2, sd = 1)`.

9

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

A. `family = binomial` ← **CORRECT**

B. `family = gaussian`

C. `family = normal`

D. `family = logistic`

Answer: A

Explanation: Logistic regression uses the binomial family. The syntax is `glm(y ~ x, family = binomial, data = mydata)`. The gaussian family is used for normal linear regression, and there is no "normal" or "logistic" family option in R.

10

In logistic regression, predictions from the model represent:

A. The exact outcome (0 or 1)

B. The probability that the outcome equals 1 ← **CORRECT**

C. The log-odds of the outcome

D. The residual error

Answer: B

Explanation: When using `predict(model, type = "response")` on a logistic regression model, the predictions are probabilities between 0 and 1, representing the probability that $Y = 1$. The linear predictor gives log-odds, but predictions are typically given as probabilities.

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?

SOLUTION:

Hypotheses:

- H_0: The mean blood pressure reduction is the same for the drug group and placebo group (i.e., $\mu_{drug} - \mu_{placebo} = 0$)
- H_A: The mean blood pressure reduction differs between the drug group and placebo group (i.e., $\mu_{drug} - \mu_{placebo} \neq 0$)

Conclusion:

The p-value is 0.0011, which is less than $\alpha = 0.05$, so we reject the null hypothesis. There is statistically significant evidence that the drug produces a different blood pressure reduction than the placebo. Specifically, the drug group had a mean reduction of 15.2 mmHg compared to 7.7 mmHg for the placebo group, a difference of 7.5 mmHg. The 95% confidence interval for the difference (3.2 to 11.8 mmHg) does not contain 0, confirming the drug is effective at reducing blood pressure more than the placebo.

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?

SOLUTION:

The slope coefficient for `occupants` is 83.43 kWh. This means that for each additional household occupant, monthly electricity usage is estimated to increase by about 83.43 kWh on average.

Yes, this relationship is statistically significant at the $\alpha = 0.05$ level. The p-value for the occupants coefficient is 1.68e-10 (which is 0.000000000167639585816), which is much less than 0.05. This provides very strong evidence that the number of occupants is a significant predictor of monthly electricity usage.

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.

SOLUTION:

From the regression output, the equation is:

$$\widehat{\text{electricity}} = 122.42 + 83.43 \times \text{occupants}$$

For a household with 7 occupants:

$$\begin{aligned}\widehat{\text{electricity}} &= 122.42 + 83.43 \times 7 \\ &= 122.42 + 584 \\ &= 706.41 \text{ kWh}\end{aligned}$$

The predicted monthly electricity usage for a household with 7 occupants is approximately 706 kWh.

Note: This prediction is an extrapolation beyond the observed data (which ranged from 1 to 6 occupants), so it should be interpreted with some caution.

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"`

SOLUTION:

```r
grade_exam <- function(score) {
  if (score >= 90) {
    return("A")
  } else if (score >= 80) {
    return("B")
  } else if (score >= 70) {
    return("C")
  } else if (score >= 60) {
    return("D")
  } else {
    return("F")
  }
}
```

Testing the function:

```
# Test cases
grade_exam(95)  # Should return "A"
```

```
[1] "A"
```

```
grade_exam(85)  # Should return "B"
```

```
[1] "B"
```

```
grade_exam(75)  # Should return "C"
```

```
[1] "C"
```

```
grade_exam(65)  # Should return "D"
```

```
[1] "D"
```

```
grade_exam(55)  # Should return "F"
```

```
[1] "F"
```

Alternative solution using `ifelse()` or `case_when()`:

```
# Using nested ifelse
grade_exam_v2 <- function(score) {
  ifelse(score >= 90, "A",
         ifelse(score >= 80, "B",
                ifelse(score >= 70, "C",
                       ifelse(score >= 60, "D", "F"))))
}

# Using dplyr::case_when (if dplyr is loaded)
library(dplyr)
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
grade_exam_v3 <- function(score) {
  case_when(
    score >= 90 ~ "A",
    score >= 80 ~ "B",
    score >= 70 ~ "C",
    score >= 60 ~ "D",
    TRUE ~ "F"
  )
}
```