MATH 8050 Homework 2

Parampreet Singh

Due February 23 at 11:59pm

Instructions: For the problems that follow, please show all of your work. For problems that you solve using software, please provide the code you used either with the problem or in an appendix at the end of your homework document. To submit your answers for homework, please upload a **single** PDF to Canvas.

Problem 1

A simple random sample of size $n = 19$ is taken from a Normal population. The sample values are listed below.

$$x_1, \ldots, x_n = \{11.3, 12.0, 10.1, 10.3, 13.5, 10.8, 13.7, 12.8, 11.9, 10.7, 11.0, 11.6, 10.1, 11.7, 10.6, 12.0, 11.7, 13.1, 11.3\}$$

(a) What is the maximum likelihood estimate for the population mean for this data?
(b) Provide a 92% confidence interval for the mean of the population. How do you interpret this interval?
(c) Perform an appropriate hypothesis test on this sample to determine if there is evidence that the true mean is different from 11. Use a significance level of $\alpha = 0.05$, and be sure to provide your hypotheses, test statistic, and decision rule.
(d) Provide a 90% confidence interval for the population variance σ^2 and interpret the interval.

Problem 2

Assume that we are sampling from a Normal population with unknown mean and known variance (equal to $\sigma^2 = 1.5625$). We want to create a (1- α) $\times$ 100% confidence interval for the unknown mean that has a particular interval width. Recall that the interval width is determined by the margin of error (ME).

(a) Derive a formula for n as a function of the margin of error (ME). That is, for a fixed/desired ME, find the minimum sample size needed to obtain a (1- α) $\times$ 100% confidence interval with the given ME.

(b) For $\alpha = 0.05$, what is the minimum sample size needed to produce a confidence interval with a margin of error equal to 0.25?

Problem 3

Suppose we want to compare the test scores of students from two different classes. To get the data for the class' test scores, download class_data.csv from Canvas. To read the data into R, you can do the following (you will need to replace ~/ with the file path of the location where you saved the data):

```
class_data <- read.csv("~/class_data.csv", header=TRUE)
class1 <- na.omit(class_data$class1)  # remove NA values since class 1 has fewer observations
class2 <- class_data$class2
```

(a) Plot the data for each class, and find the average test score for each.
(b) For a significance level of $\alpha = 0.05$, perform an appropriate test to determine whether the two classes have the same average test scores. Be sure to state the hypotheses and specify which type of test you used, even if you used software to perform the test.

(c) Using the same type of test as in part (b), do you make a different conclusion if you use a significance level of $\alpha = 0.01$?

Problem 4

We will study the the `trees` data set contained in R. The `trees` data set contains three measurements taken on $n = 31$ cherry trees; we will study how the `Height` and `Volume` variables are related to each other. To load this data into R, run `data("trees")` in your R console.

(a) Treating `Height` as the dependent variable, plot an appropriate scatterplot of `Volume` versus `Height`.
(b) Fit a simple linear regression model to the data, treating `Height` as the dependent variable. Plot the line from this model over your scatterplot.

(c) What are the values of the least squares estimates of β_0 and β_1? Interpret these estimates.
(d) What is the value of the least squares estimate for σ^2 for this regression fit?
(e) What is the value of the maximum likelihood estimate for σ^2 for this regression fit?

MATH 8050 HOMEWORK-2

Problem 1

Given,

$n = 19$

$x_1, \ldots x_n = \{11.3, 12.0, 10.1, 10.3, 13.5, \ldots\ldots 11.3\}$

(a) MLE for given sample set is

$$MLE = \bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$MLE = \frac{11.3 + 12.0 + 10.1 + \cdots\cdots + 13.1 + 11.3}{19}$$

$$\boxed{MLE = 11.59}$$

(b) $(100-\alpha)\%$ CI $= \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Now, σ is unknown but we can find s i.e sample std. Therefore we can use

$$(100-\alpha)\% \text{ CI} = \bar{X} \pm t_{\alpha/2, n-1} \left(\frac{s}{\sqrt{n}}\right)$$

$$S = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$$

$$S = \sqrt{\frac{(11.3-11.59)^2 + (12-11.59)^2 + \cdots + (11.3-11.59)^2}{19-1}}$$

$$S = 1.088$$

$$92\%.CI = 11.59 \pm t_{0.04,18}\left(\frac{1.088}{\sqrt{19}}\right)$$

From t-table, $t_{0.04,18} = 1.855$

$$92\%.CI = 11.59 \pm \left(\frac{1.855 \times 1.088}{4.359}\right)$$

$$= 11.59 \pm 0.463$$

$$92\%.CI = [11.13, 12.05]$$

(c) Given,

$n = 19$, $S = 1.088$

$\bar{x} = 11.59$, $\alpha = 0.05$

Hypothesis,

$H_0 : \mu = 11$

$H_a : \mu \neq 11$

Test statistic → $\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ ~ t distribution
(t^*)

Rejection Region → $t^* > t_{0.025,18}$

or

$t^* < -t_{0.025,18}$

i.e $t^* > |t_{0.025,18}|$

From t table, $t_{0.025,18} = 2.101$

RR → $\{ t^* > |2.101| \}$

Calculate test statistic

$$t^* = \frac{11.59 - 11}{1.088/\sqrt{19}} = 2.364 > |2.101|$$

Decision →
Therefore, $t^* > |t_{0.025,18}|$, so we reject the null hypothesis.

Conclusion →
Yes there is sufficient evidence to state that the true mean is different from 11.

(d) 90% CI for pop. variance

$$90\%\ CI = \left[\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} , \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \right]$$

we use χ^2 distribution

$\alpha = 0.1$ $n = 19$ $s = 1.088$

From χ^2 distribution,

$$\chi^2_{0.05,18} = 28.87$$

$$\chi^2_{0.95,18} = 9.390$$

$$90\%\ CI = \left[\frac{18 \times (1.088)^2}{28.87} , \frac{18 \times (1.088)^2}{9.390} \right]$$

$$90\%\ CI = [0.74 , 2.27]$$

Problem 2

Given,

$$\sigma^2 = 1.5625$$

(a) Formula for n as a function of Margin of Error (ME)

$$ME = Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

where,

$ME \rightarrow$ margin of error

$Z_{\alpha/2} \rightarrow$ Z value for a given α

$\sigma \rightarrow$ population variance

$n \rightarrow$ sample size

Solving for n

$$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{ME}\right)^2$$

(b) $\alpha = 0.05$ $\quad ME = 0.25$

$$n* > \left(\frac{Z_{\alpha/2} \cdot \sigma}{ME}\right)^2$$

$$z_{0.025} = 1.96$$

$$n > \left(\frac{1.96}{0.25}\right)^2 \times 1.5625$$

$$n > 96.04$$

Minimum sample size needed is 97.

HW2

Parampreet Singh

2024-02-20

Problem 3

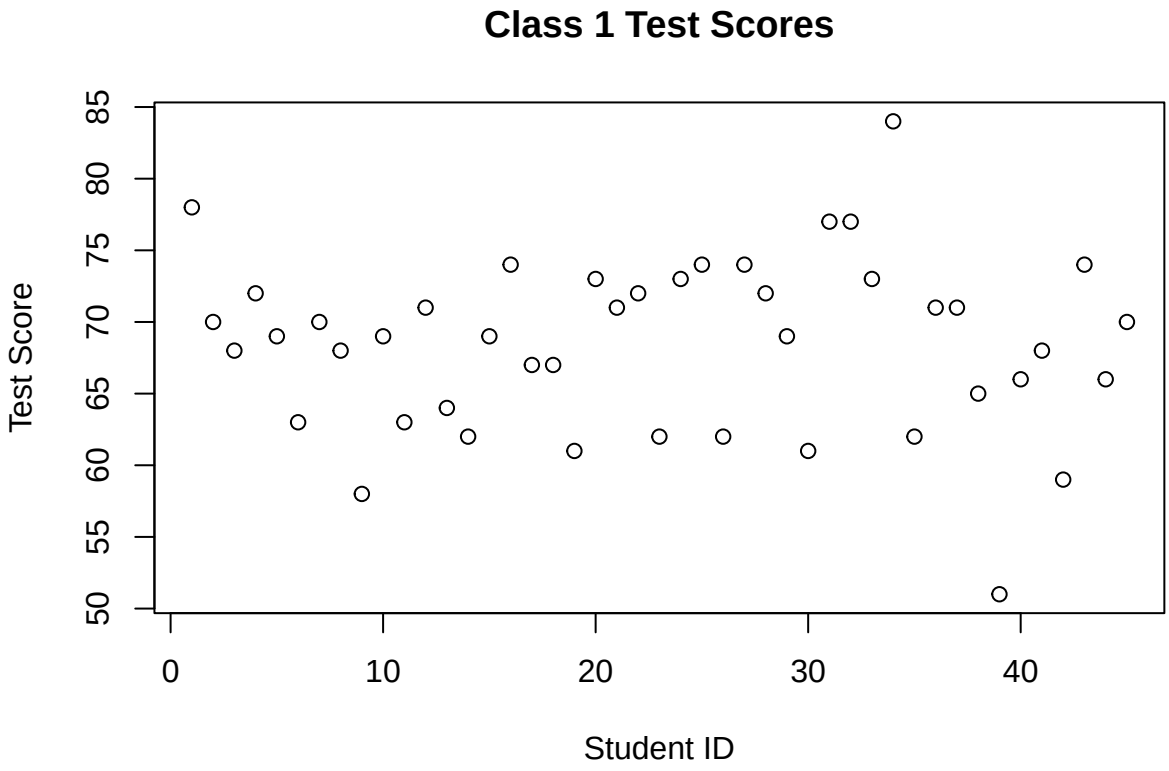
```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.2.3
```

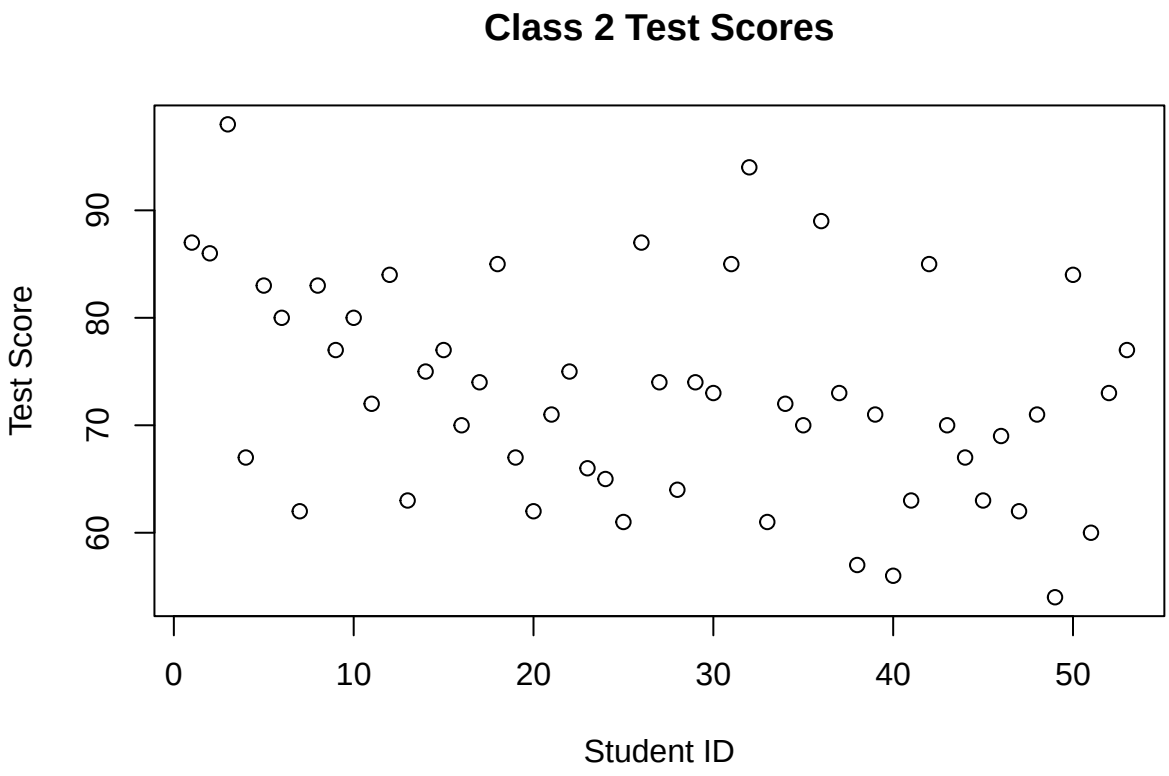
```
# Load the data from the CSV file
class_data <- read_csv("C:\\Coursework\\Spring'24\\Data Analysis\\class_data.csv")

# Remove rows with missing data (NA values)
class1 <- na.omit(class_data$class1) # Clean class1 data
class2 <- class_data$class2
```

a)Plot the data for each class, and find the average test score for each. ANS -

```
# Plot the data for each class
plot(class1, main = "Class 1 Test Scores", xlab = "Student ID", ylab = "Test Score")
```




```
plot(class2, main = "Class 2 Test Scores", xlab = "Student ID", ylab = "Test Score")
```




```
# Calculate and print the average test score for each class
average_class1 <- mean(class1)
average_class2 <- mean(class2)
print(paste("Average test score for Class 1:", average_class1))
```

```
## [1] "Average test score for Class 1: 68.4444444444444"
```

```
print(paste("Average test score for Class 2:", average_class2))
```

```
## [1] "Average test score for Class 2: 72.9811320754717"
```

b) For a significance level of 0.05, perform an appropriate test to determine whether the two classes have the same average test scores. Be sure to state the hypotheses and specify which type of test you used, even if you used software to perform the test.

ANS -

```
#Perform a hypothesis test to determine if there is a difference in means:
#H0: mean1 = mean2 (the two classes have the same average test scores)
#H1: mean1 != mean2 (the two classes do not have the same average test scores)

#We use a Two sample t-test for independent samples
```

```r
t_test_result <- t.test(class1, class2, alternative = "two.sided", conf.level = 0.95)

# Print the results of the t-test
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  class1 and class2
## t = -2.739, df = 87.111, p-value = 0.007473
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.828788 -1.244587
## sample estimates:
## mean of x mean of y
##  68.44444  72.98113
```

```r
# Check the conclusion based on the p-value for alpha = 0.05
if(t_test_result$p.value < 0.05) {
  print("Reject the null hypothesis: There is a significant difference in the average test scores.")
} else {
  print("Fail to reject the null hypothesis: There is no significant difference in the average test sco
}
```

```
## [1] "Reject the null hypothesis: There is a significant difference in the average test scores."
```

c) Using the same type of test as in part (b), do you make a different conclusion if you use a significance level of 0.01?

ANS - Repeat the same hypothesis test with alpha = 0.01

```r
# Check the conclusion based on the p-value for alpha = 0.01
if(t_test_result$p.value < 0.01) {
  print("Reject the null hypothesis at the 0.01 level: There is a significant difference in the average
} else {
  print("Fail to reject the null hypothesis at the 0.01 level: There is no significant difference in th
}
```

```
## [1] "Reject the null hypothesis at the 0.01 level: There is a significant difference in the average
```