Name: Ventsislav Radev
Lecturer: Rebecca Killick

# 104 Long Coursework Assessment

23/04/18

## q1

Retrieve your individual data set from Moodle using the "Long coursework data" quiz link below the link to this file. You should read the data into R by double clicking the file downloaded from Moodle. This will make the data object marks available in your R session.

Use the marks data provided in that file to answer the questions below. You are strongly encouraged to use R to compute the answers and include the R code in your solutions. However you calculate the answers, remember to ensure that you keep a large number of decimal places for quantities such as the Standard Error and only curtail to 3 decimal places for the final answer.

a. All students have received different datasets. To ensure a match between the data set you have used and the data set in the mark scheme, please type marks[1] into R and write down the returned value.

### Answer

*Code: marks[1]*

26.487 *[3dp]*

## q2

The lecturer of a course was unfortunately absent for a large portion of the teaching of a module in 2018. They want to investigate whether the marks of the students were affected by the absence. The marks from the 2018 students are located in the object marks.

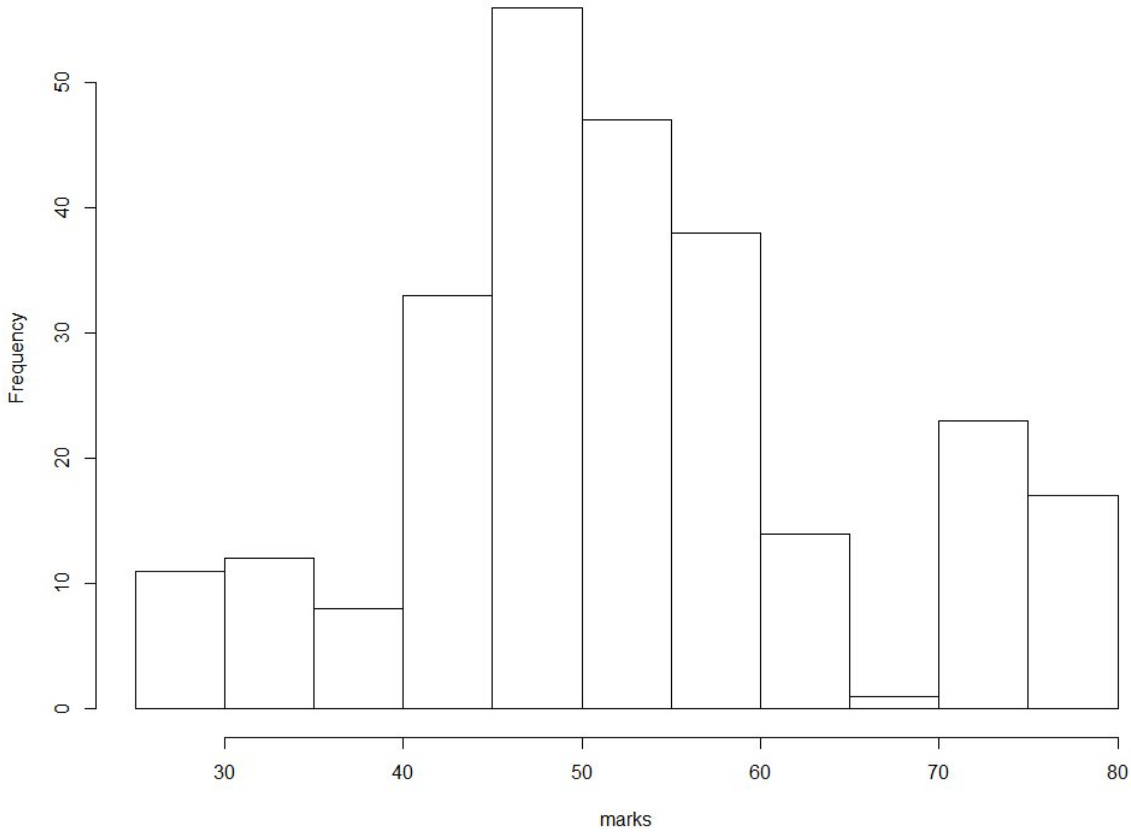
Draw (recommended to use R and print) a histogram of the 2018 marks data.

a. Describe what features are present. [6]
b. Calculate the mean of the marks data. [2]
c. Calculate the variance of the marks data. [2]
d. Calculate the proportion of students achieving a first class mark in the module (70 marks or more). [2]
e. Calculate the proportion of students failing the module (less than 45 marks). [2]

### Answer

q2a

*Code: hist(marks, breaks=10)*



There is an unusual dip between the frequency of students who got between 60 and 70 marks. Except for that dip, the histogram closely resembles a normal distribution. The histogram is also unimodal.

q2b

*Code: mean(marks)*

52.372 *[3dp]*

q2c

*Code: var(marks)*

159.884 *[3dp]*

q2d

*Code: sum(marks>70)/length(marks)*

0.154 *[3dp]*

q2e

*Code: sum(marks<45)/length(marks)*

0.246 *[3dp]*

## q3

The lecturer claims that the underlying true average mark is 65. The lecturer wants to test if the 2018 cohort marks support this hypothesis.

a. Write down the null and alternative hypothesis we would use to test the lecturers claim that the true average mark is 65. [4]
b. What are the assumptions for conducting a hypothesis test around this data? Are these satisfied? [6]
c. Calculate the p-value for your hypothesis. [11]
d. What is your conclusion when conducting a 95% hypothesis test? [2]

**Answer**

q3a

$H_0$: $\mu = 65$

$H_A$: $\mu \neq 65$

q3b

1. Each student's mark is independent from his classmates' marks: satisfied, when doing the test, students wouldn't know their peers' marks.
2. The data is not skewed - it is evident from the histogram.
3. Large sample size (>30) - length(marks) = 260, and 260>30.

q3c

```
Code:  pnorm((mean(marks)-65)/(sd(marks)/sqrt(260)))
```

1.196 *[3dp]* * 10^-58

q3d

As 1.196215 * 10^-58 < 0.05, we reject the null hypothesis. Therefore the 2018 cohort's true average is not 65, as the lecturer's claim would suggest.

## q4

To ensure fairness the lecturer administers the same final exam as in 2017 (note I won't be doing this!) and wants to use this to assess if any scaling of marks should be applied to the 2018 data. The average exam mark for the 100 students in 2017 was 63.45 and the standard deviation was 10.489.

a. What are the assumptions for calculating a confidence interval for the difference in means from the two cohorts? Are these satisfied? [6]
b. Calculate a 99% confidence interval for the difference in means from the two cohorts (2017 and 2018). [8]
c. Interpret your confidence interval. [2]
d. Would a 95% confidence interval be larger or smaller? State your reasoning. [2]

**Answer**

q4a

1. Each student's mark is independent from his classmates’' marks and from the marks of the other cohort. - satisfied, when doing the test, students wouldn't know their peers' marks.
2. Assume that the students' exam marks from 2017 can be approximated by a normal distribution.
3. The 2018 cohort's data is not skewed, but we have very little information about the 2017 cohort and thus can't determine if it's skewed or not.

q4b

*Code:*

```
63.45 - mean(marks)

qnorm(0.995)

sqrt( ((sd(marks)**2)/length(marks)) + ((10.489**2)/100) )

#First Bound
(63.45 - mean(marks)) - qnorm(0.995)*sqrt( ((sd(marks)**2)/length(marks)) +
((10.489**2)/100) )

 #Second Bound
(63.45 - mean(marks)) + qnorm(0.995)*sqrt( ((sd(marks)**2)/length(marks)) +
((10.489**2)/100) )
```

In conclusion, the 99% confidence interval for the difference between the means of the 2017 and 2018 cohorts is 7.705 and 14.452 *[3dp]*.

q4c

The mean of the 2018 cohort is vastly different from that of the 2017 cohort. As the 2017 cohort has a reasonably large sample size, there is little doubt regarding the accuracy of the mean.

q4d

A 95% confidence interval would be smaller than a 99% interval. Since with a 95% interval, you can make more mistakes, you can afford to be more specific and therefore your interval is narrower. Consider a 100% confidence interval; that interval would have to be the set of real numbers, because technically anything is possible; a height of 4m is possible, but very unlikely. Now, for a 1% confidence interval, we would only need 1% of our data to fall within it, so it would be very narrow. Similarly, the 95% confidence interval is smaller than the 99% confidence interval.

**q5**

The university want to deliver at least 25% of degrees as first class degrees. The lecturer decides to consider the proportion of first class marks in their class and test whether the true proportion is lower than the university expects.

a. What distribution would you use to model the proportion of first class marks? [1]
b. Calculate the method of moments estimator for the parameter of this distribution. [8]
c. Write down the method of moments estimator for the marks data. [1]

d. Write down the null and alternative hypothesis we would use to test the lecturers claim that the proportion is lower than 25%. [4]

e. What are the assumptions for conducting a hypothesis test around this data? Are these satisfied? [6]

f. Calculate the standard error of the estimate of the proportion of first class marks. [5]

g. Using your standard error, or otherwise, test your hypothesis at the 90% level [5]

h. Without conducting further calculations, what is your conclusion when conducting a 95% hypothesis test? [2]

i. Did your hypothesis test use the distribution you stated from (a)? If not, why could you use an alternative distributuion? [3]

**Answer**

q5a

A binomial distribution. More specifically, X ~ Bin(260, 0.25)

q5b

$$\mathrm{n} \;= \tfrac{1}{\mathrm{n}}\textstyle\sum_{i=1}^{n} xi$$
$$= \tfrac{1}{\mathrm{n}^2}\textstyle\sum_{i=1}^{n} xi$$

q5c

$$= \frac{1}{\mathrm{n}^2}\sum_{i=1}^{n} xi = \frac{1}{\mathrm{n}} * \frac{\sum_{i=1}^{n} xi}{n} = \frac{1}{n} * mean(marks) = \frac{mean(marks)}{n}$$

*Code:*
```
mean(marks)/length(marks)
```

0.201 *[3dp]*

q5d

$H_0$: p = 0.25

$H_A$: p < 0.25

q5e

1. Each student's mark is independent from his classmates' marks. - satisfied, when doing the test, students wouldn't know their peers' marks.
2. The data is not skewed - it is evident from the histogram.
3. Large sample size (>30) - length(marks) = 260, and 260>30.

q5f

*Code:*
```
p1 <- sum(marks>70)/length(marks)

SE1 <- sqrt((p1*(1-p1))/length(marks))

sqrt((p1*(1-p1))/length(marks))
```

0.022 *[3dp]*

q5g

```
Code:  PE1 <- p1

       z1 <- (PE1 - 0.25)/SE1

       pnorm(z1)
```

8.649 *[3dp]* * 10^-6

8.649 *[3dp]* * 10^-6 is smaller than 0.1, therefore we reject $H_0$. In other words, the proportion of students who achieved first class is lower than 25%.

q5h

8.649 *[3dp]* * 10^-6) is also smaller than 0.05, therefore for a 95% hypothesis test we would still reject $H_0$. In other words, the proportion of students who achieved first class will still be lower than 25% for a 95% hypothesis test.

q5i

No. My hypothesis test used a Normal Distribution. This is because for a large enough sample size (roughly >30), any distribution approximates a normal distribution accurately.

**q6**

The university lecturer also wants to ensure that not too many people fail their course (obtain a mark less than 45). They decide to test if the proportion failing their course in 2018 is the same as in 2017. In 2017 15 out of 100 of students failed.

a. Write down the null and alternative hypothesis we would use to test the lecturers claim that the proportion failing their class is the same in as 2017. [4]
b. What are the assumptions for conducting a hypothesis test around this data? Are these satisfied? [6]
c. Calculate the standard error of the estimate of the difference in proportion of failing marks. [5]
d. Test your hypothesis at the 95% level [5]
e. Without conducting further calculations, would a 95% confidence interval give the same conclusion? State your reasoning. [2]

**Answer**

q6a

$H_0$: p17 - p18 = 0

$H_A$: p17 - p18 ≠ 0

## q6b

1. Each student's mark is independent from his calssmates' marks. - satisfied, when doing the test, students wouldn't know their peers' marks.
2. The data is not skewed - it is evident from the histogram.
3. Large sample size (>30) - length(marks) = 260, and 260>30.

## q6c

```
Code:  p2 <- (sum(marks<45) + 15) / (length(marks) + 100)

        #Standard Error = sqrt( (p(1-p)/n1) + (p(1-p)/n2) )

       SE2 <- sqrt( (p2*(1-p2)/length(marks)) + (p2*(1-p2)/100) )

       SE2
```

0.049 *[3dp]*

## q6d

```
Code:  #Point Estiamte = p17 - p18

       PE2 <- 15/100 - sum(marks<45)/length(marks)

       #Null value = 0

       #SE2 from previous question,

        Z2 <- (PE2 - 0)/SE2

       pnorm(Z2)
```

0.0242 *[4dp] (because 0.024 is very close to 0.025)*

As this is a 95% level, two-tailed test, we compare our result to 0.025. 0.0242 < 0.025, therefore we reject the null hypothesis $H_0$. In other words, the proportion of failed students in 2018 is different to the proportion in 2017.

## q6e

Yes. For the 95% hypothesis test we reject any results outside of the central section of the bell curve with area 0.95 (i.e. results in the outside tails with area 0.025). Similarly, the 95% confidence interval is the interval such that any number outside of it has a probability less than 0.025. In other words, the confidence interval's bounds are just the points on the x-axis that form the central section of the bell curve with area 0.95. Therefore both methods give the same conclusion.

**q7**

**Answer**

Based on your analyses, would you conclude that the students marks in 2018 were affected by the lecturers absence and should be scaled? State your reasoning [3]

q7

Most, if not all, of our calculations above imply that the marks of the 2018 cohort are abnormally low. In other words, the results were affected by the lecturer’s absence, and need to be scaled. The most striking number is the answer to question 3c: 1.196 *[3dp]* * 10^-58. This number is immensely smaller than one would expect it to be and shows the extent to which the results have been affected.