Class Example 1 – Confidence intervals and Hypothesis test for large samples

Consider the data file, *class4_2.csv* containing 100 observations.

(a) Load the data file and carry out a descriptive statistic summary. What is the sample mean? What is the sample standard deviation?

Solution:
To read in the dataset we use `read.csv`. Having looked at the dataset we note that it does not contain headers, and so we need to tell R this.

```
> dataset1 <- read.csv("class4_2.csv",header=FALSE)
```

We then obtain the summary statistics (five number summary, mean, standard deviation and number of points) with the following commands

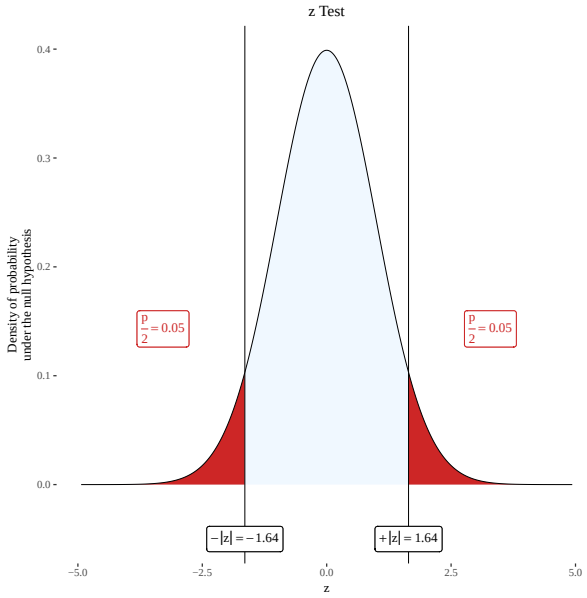
```
> summary(dataset1$V1) # mean and five number summary

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  163.4   186.5   197.9   201.0   215.8   252.6

> sd(dataset1$V1) # standard deviation

[1] 19.20648

> nrow(dataset1) # number of points

[1] 100
```

(b) Determine (large sample) confidence intervals of confidence levels: 90%, 95% and 99%.

Solution: As we are using a large sample (100 data points) we use the Normal distribution for the data. Recall that the calculation for a confidence interval of the mean is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

Usually we would looking up the relevant values of z^* for each confidence level using the standard tables, but for speed we will use the `qnorm` function to obtain the quantiles of the normal distribution. Remember as we are calculating two sided confidence intervals we need to split the remaining percentages to each end of the distribution. For example for a 90% confidence interval we would have



So we can calculate the confidence intervals using R in the following way:

```
> # define variables so we don't have to keep typing them
> mean1 <- mean(dataset1$V1)
> sd1 <- sd(dataset1$V1)
> n1 <- nrow(dataset1)
>
> # 90% confidence
> mean1 + qnorm(c(0.05,0.95))*sd1/sqrt(n1)

[1] 197.8719 204.1903

> # 95% confidence
> mean1 + qnorm(c(0.025,0.975))*sd1/sqrt(n1)

[1] 197.2667 204.7955

> # 99% confidence
> mean1 + qnorm(c(0.005,0.995))*sd1/sqrt(n1)

[1] 196.0838 205.9783
```

As we can see, as the confidence level increases the confidence interval increases in width.

(c) You now wish to test the Hypothesis: $H_0 : \mu = 203$ vs. $H_1 : \mu \neq 203$. Carry out a hypothesis test (z-test), and find the associated p-value. What are your conclusions with $\alpha = 0.1; 0.05; 0.01; 0.005$?

Solution: To calculate a hypothesis test we need to first standardise our random variable. We do this by

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Letting R do the heavy lifting we get

```
> z1 <- (mean1-203)/(sd1/sqrt(n1))
> print(z1)

[1] -1.025133
```

We now can use this to calculate the probability of $P(|Z| > z)$ which relates to our two sided test. We can either look up this value with the standard normal table or use R to calculate it for us. Note as we are doing a two sided test and the z is negative we simply use `pnorm` and multiply by two.

```
> 2*pnorm(z1)

[1] 0.3053004
```

So our p-value= $P(|Z| > z) = 0.3053$. As this is greater than all the α we **retain** the null hypothesis and conclude that the mean is not significantly different from 203.

Class Example 2 – Confidence intervals and Hypothesis test for small samples

Consider the data file, *class4_3.csv* containing 10 observations.

(a) Load the data file and carry out a descriptive statistic summary. What is the sample mean? What is the sample standard deviation?

Solution:
As in Class Example 1 we read in the data which does not contain headers into R and obtain the summary statistics.

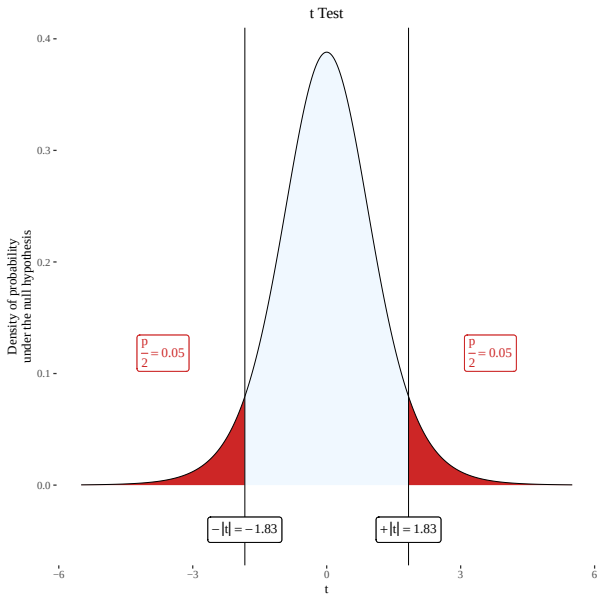
```
> dataset2 <- read.csv("class4_3.csv",header=FALSE)
> summary(dataset2$V1)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28.24   28.96   29.64   29.76   30.71   31.15

> sd(dataset2$V1)

[1] 1.077613

> nrow(dataset2)

[1] 10
```

(b) Determine confidence intervals of confidence levels: 90%, 95% and 99%.

Solution: Given we have a small sample (10 observation) we use the Student t-distribution with $n-1=9$ degrees of freedom to obtain our standard values. The confidence interval calculation now uses

$$\bar{x} \pm t_9^* \frac{s}{\sqrt{n}}$$

Note that this uses the sample standard deviation, however is of a similar form to the Normal Confidence interval. Here we use `qt(t,df)` to obtain the quantiles of the t distribution. Remember the percentage left after the confidence interval is again split to each end.



In R we obtain the confidence intervals as follows:

```
> # Again defining variables as we are lazy
> mean2 <- mean(dataset2$V1)
> sd2 <- sd(dataset2$V1)
> n2 <- nrow(dataset2)
>
> df2 <- n2-1
>
> # 90% confidence
> mean2 + qt(c(0.05,0.95),df2)*sd2/sqrt(n2)

[1] 29.13933 30.38867

> # 95% confidence
> mean2 + qt(c(0.025,0.975),df2)*sd2/sqrt(n2)

[1] 28.99312 30.53488

> # 99% confidence
> mean2 + qt(c(0.005,0.995),df2)*sd2/sqrt(n2)

[1] 28.65655 30.87145
```

Once again we can see as the confidence level increases the range of the interval increases.

(c) You now wish to test the Hypothesis: $H_0 : \mu = 30$ vs. $H_1 : \mu < 30$. Carry out a hypothesis test (t-test). What are your conclusions with $\alpha = 0.1, 0.05, 0.01, 0.005$?

Solution: Here we are conducting a one-sided test and so will be looking for the probability of $P(X < 30)$. However as our tables are a standard t-distribution we have to first standardise our random variable. We do this by

$$t_9 = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

In R we get:

```
> t2 <- (mean2-30)/(sd2/sqrt(n2))
> print(t2)

[1] -0.6925472
```

Now we can look up this value in our table. As the table is a $P(T > t)$ table we can simply remove the negative sign and look where the test statistics falls. When we do this we find that $p > 0.25$. Using `pt(t,df)` in R we find that $P(T < t) = 0.2530342$. Once again this is above all the values of α and so we again retain the null hypothesis. We conclude that the mean is not significantly different from 30.

We can check our calculation using the inbuilt `t.test` function of R. The output we obtain using the options `mu=30` to set our hypothesised mean and `alternative="less"` to set the alternative hypothesis is:

```
> t.test(dataset2$V1,mu=30,alternative="less")


One Sample t-test

data:  dataset2$V1
t = -0.69255, df = 9, p-value = 0.253
alternative hypothesis: true mean is less than 30
95 percent confidence interval:
     -Inf 30.38867
sample estimates:
mean of x
   29.764
```

Here we see that the output agrees with our test statistic, degrees of freedom and p-value. It also calculates a one-sided confidence interval (not cover in the course) and the sample mean.

Class Example 3 – Previous Examination Question

An engineer is designing a kayak that holds a single person together with a bag. It is known that kayak users have a mean (person) weight of $\mu_1 = 78$ and a standard deviation of $\sigma_1 = 12.3$ (all units in kilograms). Denote the (unknown) population mean and population standard deviation of bag weights, by μ_2 and σ_2 respectively.

The engineer measures bags of 16 randomly selected users. A sample mean of $\bar{x} = 25.2$ and a sample standard deviation of $s = 4.2$ are obtained.

Previous manufacturers have assumed a bag weight of 30. However the data collected indicate a potentially lower weight. Hence, the engineer wishes to claim that bag weight assumptions were too conservative. She decides to test the hypothesis,$H_0 : \mu_2 = 30$ vs. $H_1 : \mu_2 < 30$, setting $\alpha = 10\%$. What does the engineer conclude? Carry out the hypothesis test.

Solution:

Method 1 - Critical Value

Given that $\alpha = 0.1$ we look up in the t-table the t^* for $P(T > t) = 0.1$, given our table is a greater than table. As the hypothesis test is for a less than alternate hypothesis, we simply add a negative sign to get our t value. This give a value of $t = -1.341$. Now reversing the standardisation to get to the distribution of the weight of the bags we perform the following calculation.

$$\begin{aligned} -1.341 &= \frac{x - 30}{\frac{4.2}{\sqrt{16}}} \\ -1.277 &= x - 30 \\ 28.723 &= x \end{aligned}$$

As $25.3 < 28.723$ we conclude that the weight of the bags is less than 30 kg as 25.3 is within the rejection region.

Method 2 - P-value

Here we need to calculate the probability of the mean bag weight being less than 30 kg. First we need to calculate the test statistic,

$$\begin{aligned} P(X < 30) &= P\left(T < \frac{25.2 - 30}{\frac{4.2}{\sqrt{16}}}\right) \\ &= P(T < -4.4761905) \end{aligned}$$

So $t = -4.476$ and so we look this up in the t-table with $16 - 1 = 15$ degrees of freedom. As our table is a greater than table and we have a negative t we can simply remove the negative sign and look this value up to get the probability directly. Thus $0.0001 < p < 0.0005$. This gives strong evidence to support the alternative hypothesis so we conclude that the weight of a bag is less than 30 kg.