Question 1: Prescribed clot-dissolving drugs

The variable **‘clotsolv’** indicates the type of clot dissolving drug prescribed to patients admitted to hospital for suspected myocardial infarction. Using R, produce the relevant graph and table(s) to summarise the **‘clotsolv’** variable. In the style of report writing, write a paragraph explaining the key features of the data by choosing the correct analysis.

The distribution for “clotsolv” is shown in Figure 1.

In this sample of 393 for the types of clot-dissolving drugs prescribed to patients, 5.85% is reported being Streptokinase, 45.04% being Reteplase, 43.77% being Alteplase and the remainder were not prescribed to the patients.

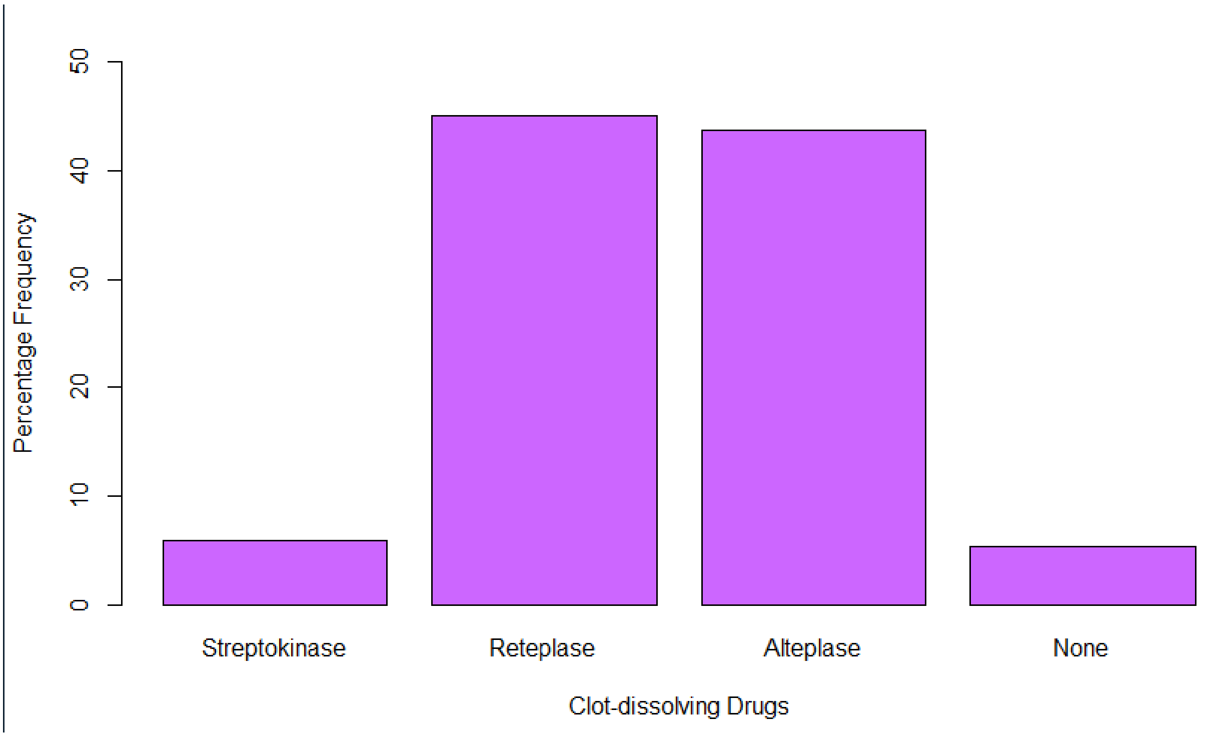
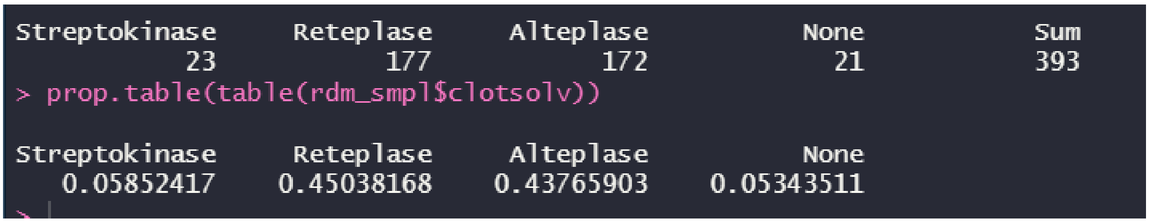


Figure 1: Distribution of clot-dissolving drugs.

Include your R input and output for this question here.

addmargins(table(rdm_smpl$clotsolv))
prop.table(table(rdm_smpl$clotsolv))



```
Streptokinase     Reteplase     Alteplase          None          Sum
           23           177           172            21          393
> prop.table(table(rdm_smpl$clotsolv))

Streptokinase     Reteplase     Alteplase          None
   0.05852417    0.45038168    0.43765903    0.05343511
>
```

Question 2: Patient age

The variable '**age**' indicates the age of patients in years. Using R, produce the relevant graph and table(s) to summarise the '**age**' variable. In the style of report writing, write a paragraph explaining the key features of the data by choosing the correct analysis.

The distribution of the age of patients in years in a sample of 391 patients is displayed in Figure 2. The distribution is approximately symmetric, and the average age of the patients was 63.73 years old. Typically, the age of the patients was between 58 to 69 years old. The 95% confidence interval indicates that the average age of patients was between 62.95 and 64.51 years old. There is no outlier.

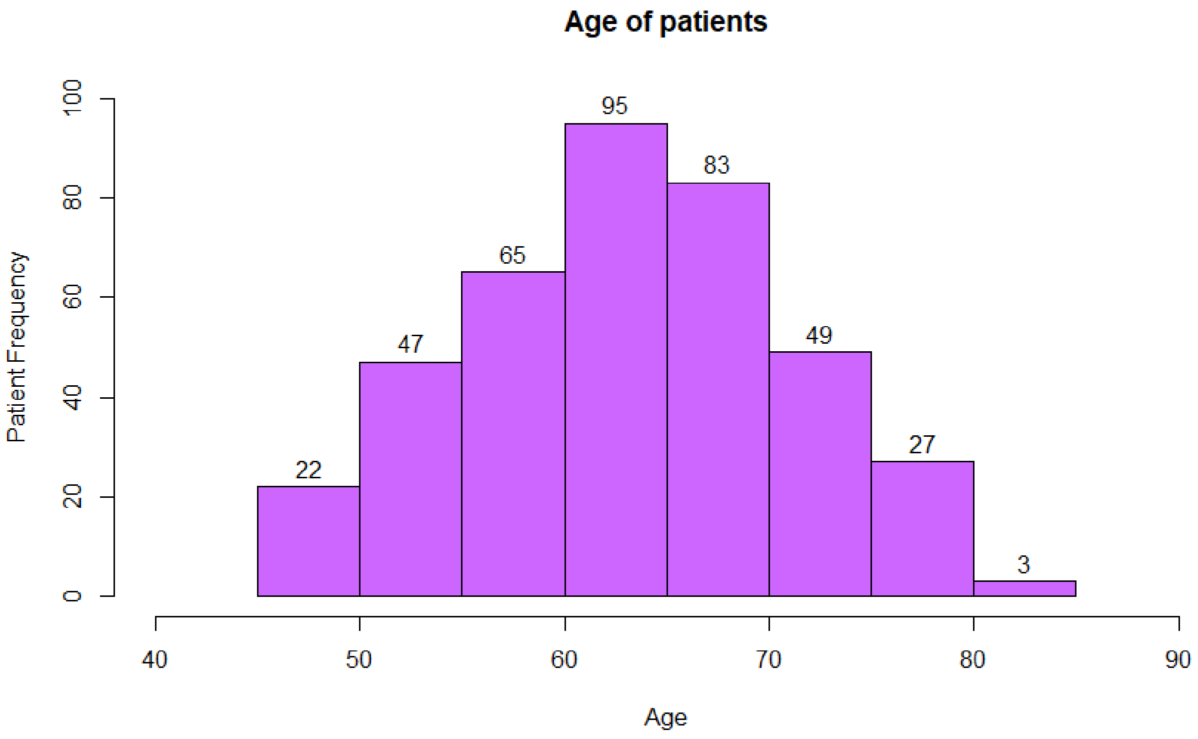
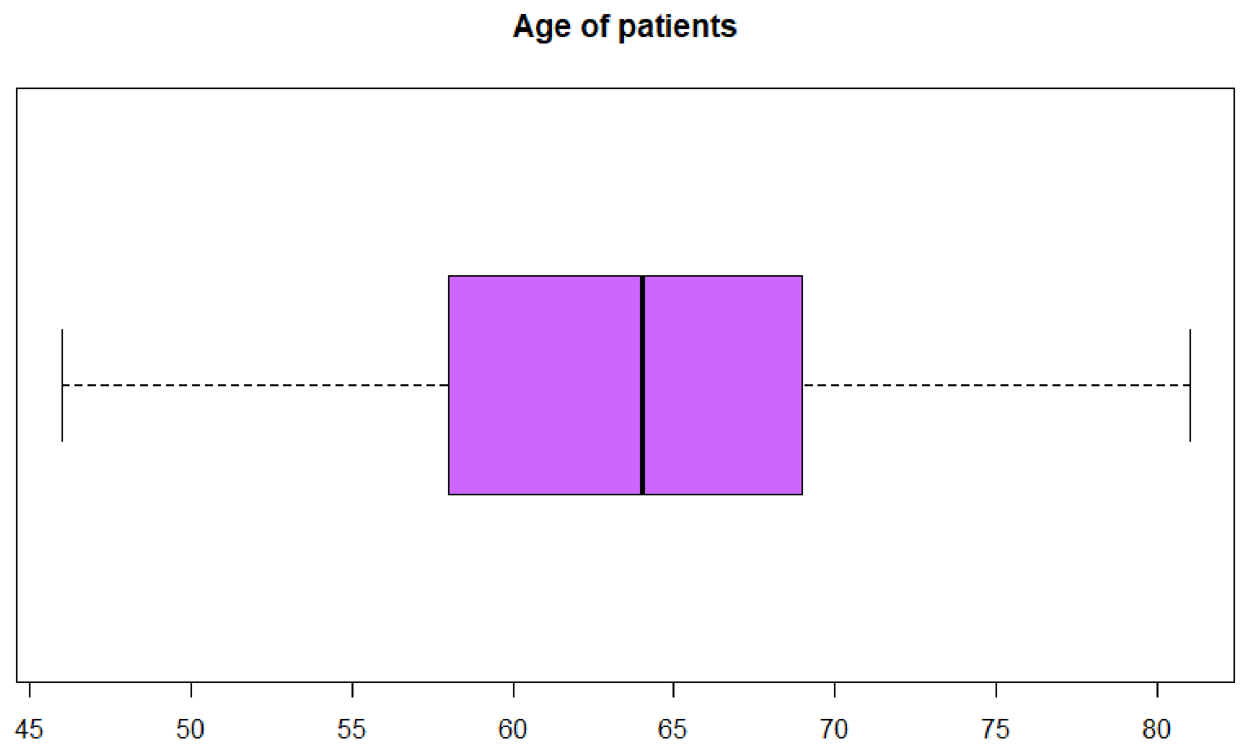


Figure 2: Age of patients.



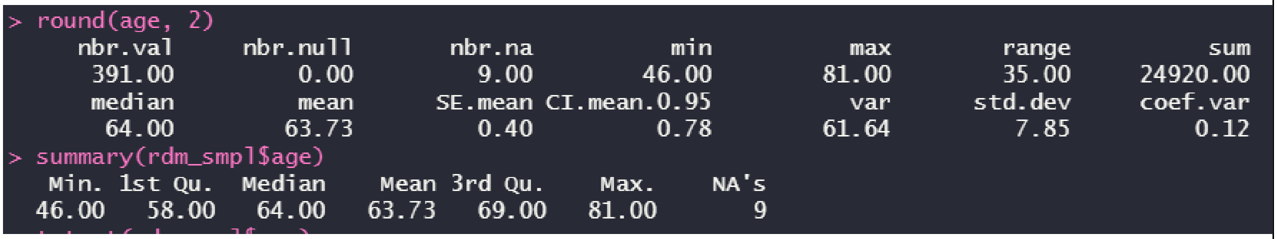
Include your R input and output for this question here.

age<-stat.desc(rdm_smpl$age)
round(age, 2)

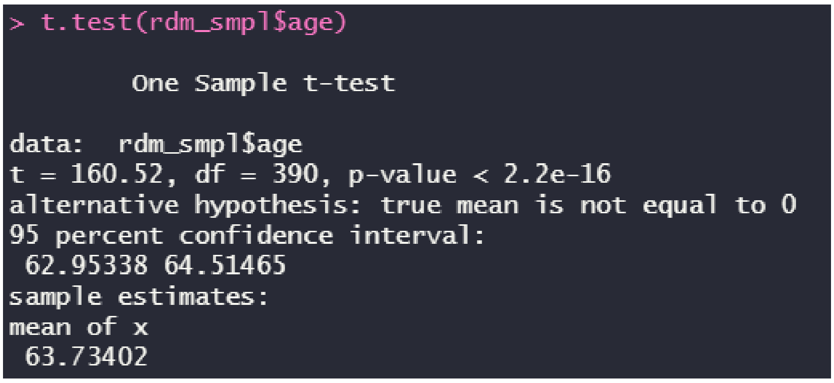
summary(rdm_smpl$age)

t.test(rdm_smpl$age)

boxplot.stats(rdm_smpl$age)$out



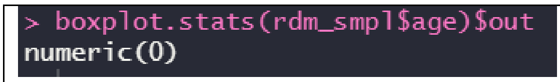
```
> round(age, 2)
     nbr.val      nbr.null       nbr.na          min          max        range          sum
      391.00          0.00         9.00        46.00        81.00        35.00     24920.00
      median          mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
       64.00         63.73         0.40         0.78        61.64         7.85         0.12
> summary(rdm_smpl$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  46.00   58.00   64.00   63.73   69.00   81.00       9 
```



```
> t.test(rdm_smpl$age)

	One Sample t-test

data:  rdm_smpl$age
t = 160.52, df = 390, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 62.95338 64.51465
sample estimates:
mean of x 
 63.73402 
```



```
> boxplot.stats(rdm_smpl$age)$out
numeric(0)
```

Question 3: Length of stay in the hospital

The variable **'los'** indicates the number of days that patients spent in the hospital. Using R, produce the relevant graph and table(s) to summarise the **'los'** variable. In the style of report writing, write a paragraph explaining the key features of the data by choosing the correct analysis.

The distribution of the number of days patients spent in the hospital in a sample of 392 patients is displayed in Figure 3. The distribution is positively skewed, and the average number of days patients spent in the hospital is 7.22 days. Typically, the number of days patients spend in the hospital was between 5 to 9 days. 3 patients spent an exceptionally high number of days in the hospital which is 18 days.

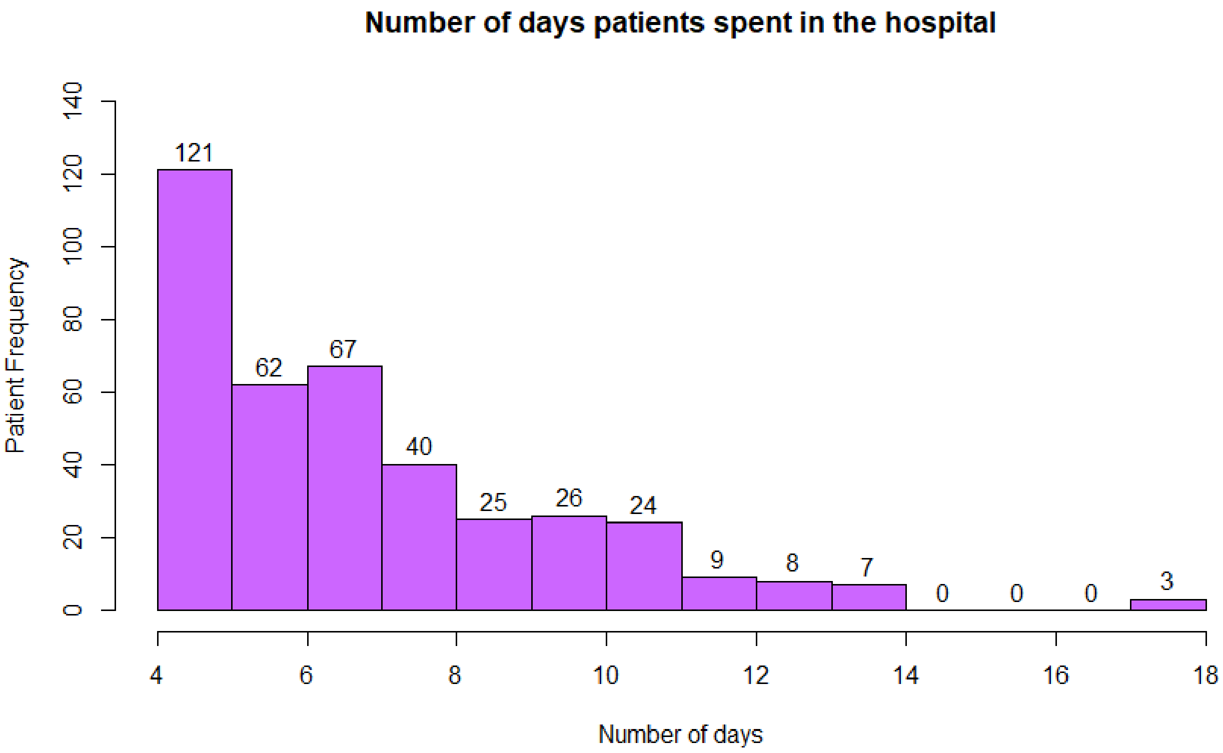
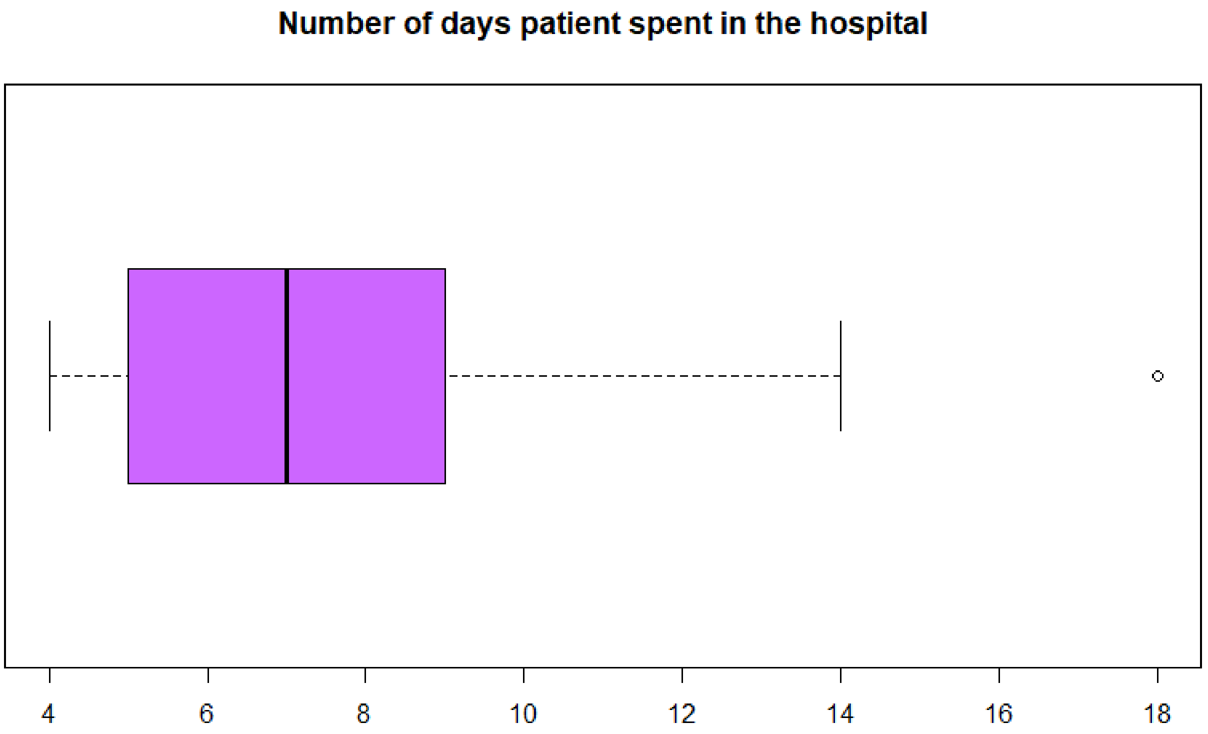


Figure 3: Number of days patients spent in the hospital.

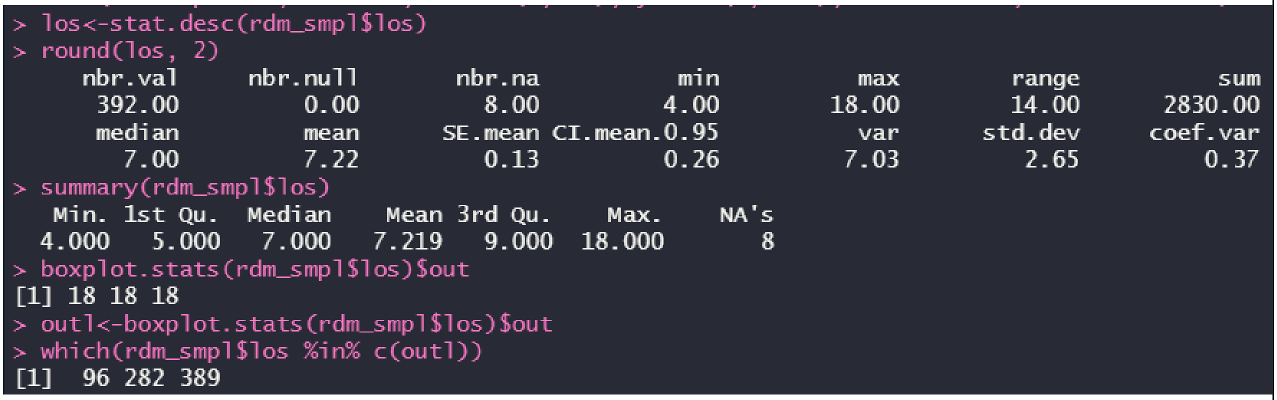


Include your R input and output for this question here.

```
los<-stat.desc(rdm_smpl$los)
round(los, 2)

summary(rdm_smpl$los)

boxplot.stats(rdm_smpl$los)$out
outl<-boxplot.stats(rdm_smpl$los)$out
which(rdm_smpl$los %in% c(outl))
```



```
> los<-stat.desc(rdm_smpl$los)
> round(los, 2)
     nbr.val     nbr.null       nbr.na          min          max        range          sum 
      392.00         0.00         8.00         4.00        18.00        14.00      2830.00 
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var 
        7.00         7.22         0.13         0.26         7.03         2.65         0.37 
> summary(rdm_smpl$los)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  4.000   5.000   7.000   7.219   9.000  18.000       8 
> boxplot.stats(rdm_smpl$los)$out
[1] 18 18 18
> outl<-boxplot.stats(rdm_smpl$los)$out
> which(rdm_smpl$los %in% c(outl))
[1]  96 282 389
```

Question 4: [does not require R]

Nick is a psychology student who participates in sprint triathlon competitions which consist of swimming, cycling, and running in one event. In the last competition, Nick completed the swimming race in 12 minutes and 46 seconds (766 seconds), cycling in 33 minutes and 52 seconds (2032 seconds), and the running race in 17 minutes and 3 seconds (1023 seconds).

Completion times for participants in Swimming are normally distributed with a mean of μ = 695 seconds and a standard deviation σ = 62 seconds. Completion times for participants in Cycling are normally distributed with a mean of μ = 2184 seconds and a standard deviation σ = 76 seconds. And completion times for participants in Running are normally distributed with a mean of μ = 1083 seconds and a standard deviation σ = 50 seconds.

In which competition race (swimming, cycling, or running) was Nick's performance better, relative to others on average who took part in that competition? Justify your answer, quoting relevant statistics as part of your explanation.

$$z = \frac{X - \mu}{\sigma}$$

Swimming:

$$z = \frac{766 - 695}{62}$$

$$z = \frac{71}{62}$$

z = 1.15

Cycling:

$$z = \frac{2032 - 2184}{76}$$

z = -2.00

Running:

$$z = \frac{1023 - 1083}{50}$$

$$z = -\frac{6}{5}$$

z = -1.20

Out of the 3 categories in the triathlon, Nick’s performance was the best in Cycling. His time for Cycling was two standard deviations lower than the mean. However, his performance was the worse in Swimming as the z-score for this was 1.15 which means that this time was 1.15 standard deviations more than the mean. As for his performance in Running, he performed better in this category as well, as his time was -1.2 standard deviations lower than the mean.

Question 5: [does not require R]

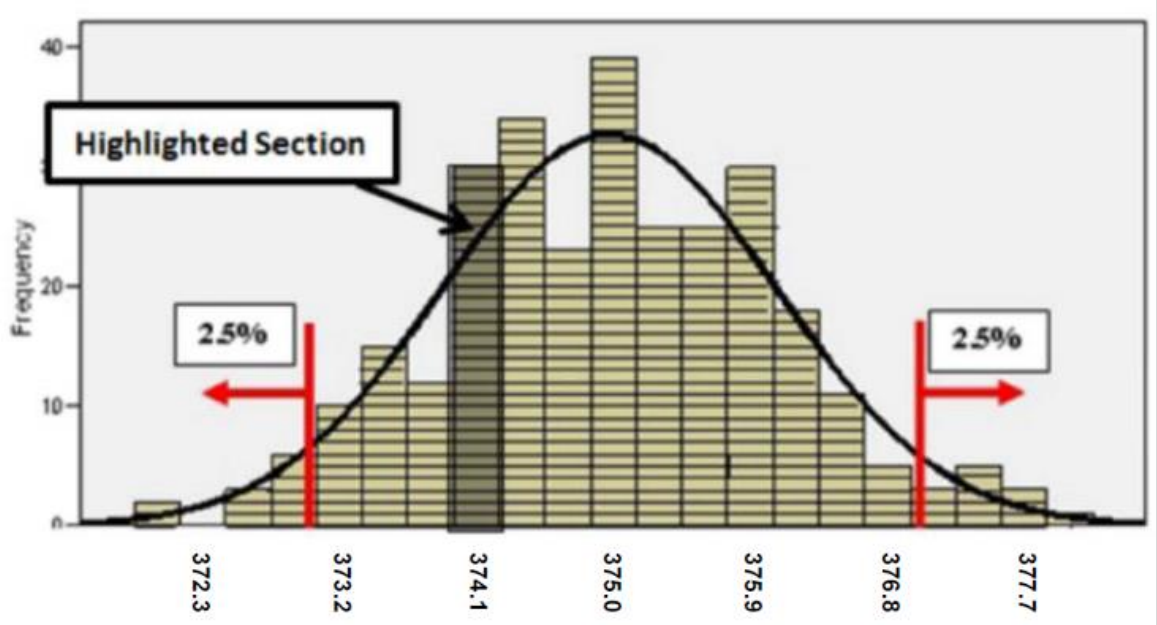
Vitality Creams produce a moisturising cream, which has been selling extremely well – particularly the 375ml container [σ = 20ml]. Some stockists, however, have recently complained that these containers appear to contain less than 375ml of moisturising cream.

The production line manager thinks that the machinery filling the containers might not be working properly, so he takes a random sample of 500 moisturising cream containers produced during January to check the contents.

a. What is the population we can draw conclusions about in this study?

All the moisturising cream containers produced in January by Vitality Creams.

We have produced a sampling distribution using 200 samples of size 500, taken from a population where the mean is 375ml [σ = 20ml]. The sampling distribution is displayed below in *Figure 1*.



b. What does the highlighted section of the sampling distribution in *Figure 1* represent?

All the sample of 500 moisturising cream containers with the mean of 374.1 ml.

c. The random sample of 500 containers of moisturising cream taken by the production line manager turn out to have a mean of 377.4ml.

Does this sample look like it belongs to the sampling distribution displayed in *Figure 1*? Justify your answer.

It does not belong to the sampling distribution because it is in the rejection region of the sampling distribution.

d. Given that the sample was randomly selected from all containers of moisturising cream produced by Vitality Creams during January and the amount of moisturising cream in each container was measured accurately. Based on part (c),

(i) what conclusion can we reach for the stockists' concern? Explain.

(ii) can we conclude that the mean weight of all containers of moisturising cream produced by Vitality Creams during January is specifically 377.4 ml? Explain.

(iii) can we confidently make the conclusion about the mean weight of containers at other general times of year is specifically 377.4 ml? Explain.

(i) The amount of moisturising cream in each container that was produced by Vitality Creams in January does not correlate to the stockist' concern as the mean weight was recorded to be more than 377.4 ml.

(ii) The mean weight of all the containers of moisturising cream produced in January by Vitality Creams is significantly higher than 377.4 ml therefore it cannot specifically be 377.4 ml as it 2 standard deviations above the mean.

(iii) We cannot confidently make the conclusion about the mean weight of the containers at other general times of year because this sample was done for the month of January only which might be biased therefore no conclusion can be made for the other genera times of year.