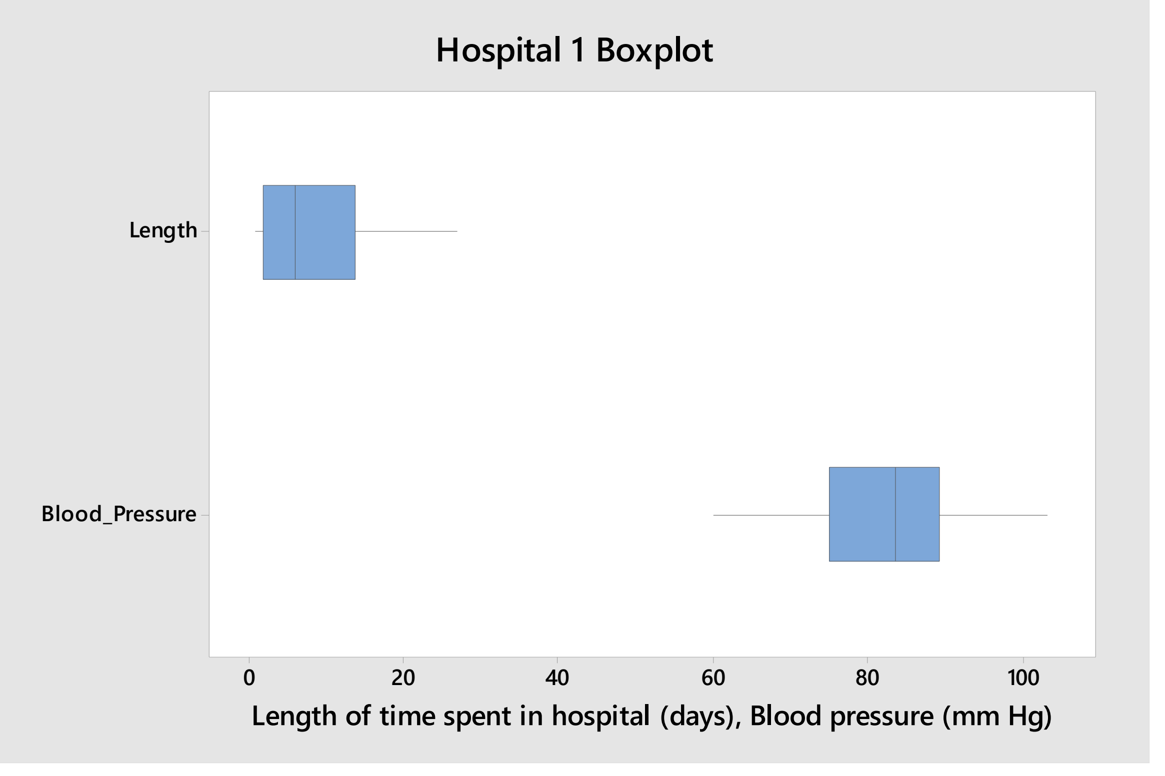
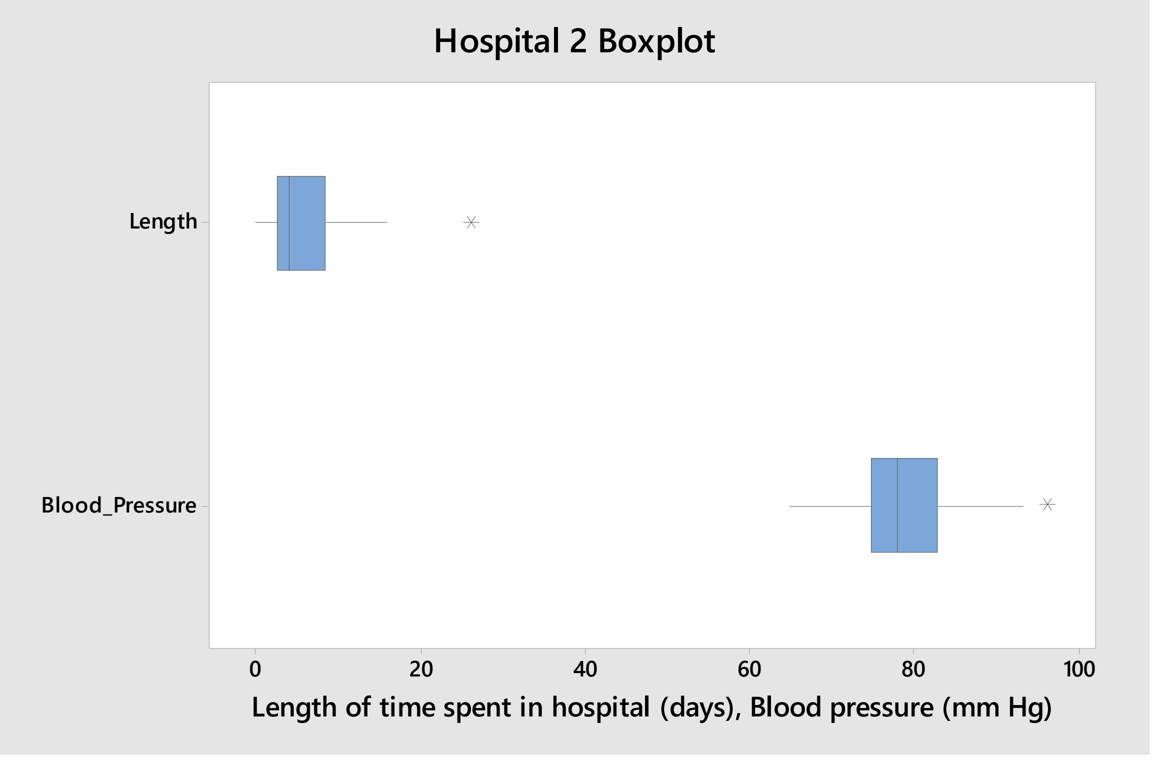
Up930537

Statistical Theory and Methods I Coursework U30214

1)

a)





b)

This is not confirmed by the boxplots. The median length of time patients spent in hospital 1 was 6 days, and 4 days in hospital 2.

c)

The length is right skewed. Blood pressure is slightly left skewed.

d)

It would be expected that the mean length of stay is greater than the median because the boxplot shows it is right-skewed, so the median would be to the left of the mean on a distribution.

e)

It signifies an outlier. This was decided by Minitab by multiplying the interquartile range by 1.5, then subtracting from Q1, and adding to Q3. Anything outside this range is an outlier.

$(IQR)(1.5) = (6)(1.5) = 9$

$Q3 + 9 = 8.5 + 9 = 17.5$

$26 > 17.5$, so that value was an outlier.

2)

a)

Null hypothesis: The median difference in recorded speeds from the detectors is equal to zero.

Alternative hypothesis: The median difference in recorded speeds from the detectors is not equal to zero.

Where "difference" = "Detector A" – "Detector B"

Wilcoxon Signed Rank Test: A-B

```
Test of median = 0.000000 versus median ≠ 0.000000

        N for   Wilcoxon          Estimated
      N  Test  Statistic      P     Median
A-B  20    20       29.0  0.005   -0.07500
```

Wilcoxon Signed Rank CI: A-B

```
                                  Confidence
        Estimated    Achieved      Interval
      N    Median  Confidence   Lower   Upper
A-B  20    -0.075        95.0  -0.170  -0.030
```

Conclusion:

The p-value is less than 0.05, 0.005<0.05. There is sufficient evidence to reject the null hypothesis that there is no significant difference in results between the speed detectors at the 5% level of significance.

The confidence interval does not include zero, which suggests a difference in results between the speed detectors. Therefore, there is sufficient evidence to reject the null hypothesis at the 5% level.

b)

The p-value equals 0.005. 0.005 is less than the level of significance, so there is sufficient evidence to reject the null hypothesis at the 5% level. We can conclude there is a significant difference between results of the two types of speed detector, and the data implies “detector A” recorded the cars at lower speeds than “detector B”.

c)

A parametric test is a statistical test where assumptions about the parameters are made. For example, in a T-test, the sample population is assumed to be normally distributed.

The parametric equivalent to the Wilcoxon Signed Rank Test would be the paired T-test, because the two samples are dependent of each other.

3)

a)

Null hypothesis: There is no association between region and number of sales.

Alternative hypothesis: There is an association between region and number of sales.

b)

Chi-Square Test for Association: Region, Recoded Number_of_Sales

```
Rows: Region   Columns: Recoded Number_of_Sales

          low  medium   high  All

A         111      56     33  200
        99.50   58.00  42.50

B          88      60     52  200
        99.50   58.00  42.50

All       199     116     85  400

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 7.043, DF = 2, P-Value = 0.030
Likelihood Ratio Chi-Square = 7.085, DF = 2, P-Value = 0.029
```

The expected value for the “low” category in “region A” is 99.50.

The contribution to the chi-squared statistic for the “medium” category in “region B” is approximately 0.069.

Expected value for the “high” category in “region B”:

$(row\ total)\left(\frac{column\ total}{grand\ total}\right)$=E

$(200)\left(\frac{85}{400}\right)$=42.5

Contribution to Chi-squared test:

$$\frac{(O-E)^2}{E} = \frac{(52-42.5)^2}{42.5} = \frac{361}{170} \approx 2.12$$

c)

$$v = (r-1)(c-1) = (3-1)(2-1) = (2)(1) = 2$$

d)

$$\chi^2 = 7.043$$

$$\chi^2_{5\%}(2) = 5.99$$

$$\chi^2 > \chi^2_{5\%}(2)$$

There is sufficient evidence to reject the null hypothesis that there is no association between region and number of sales at the 5% level of significance. We can conclude there is an association between region and the number of sales.

e)

A “type 1” error might have occurred because the conclusion was that there was sufficient evidence to reject the null hypothesis. If the null hypothesis were true, this would be a “type 1” error.

4)

a)

Probability Density Function

```
Continuous uniform on 0 to 3

       x    f( x )
0.898504  0.333333
0.399398  0.333333
0.688326  0.333333
0.654048  0.333333
0.160691  0.333333
0.785923  0.333333
0.986407  0.333333
0.948507  0.333333
0.052011  0.333333
0.547418  0.333333
```

b)

Correlation: x, x^4

```
Pearson correlation of x and x^4 = 0.861
P-Value = 0.000
```

Correlation: x, minuses

```
Pearson correlation of x and minuses = -1.000
P-Value = *
```

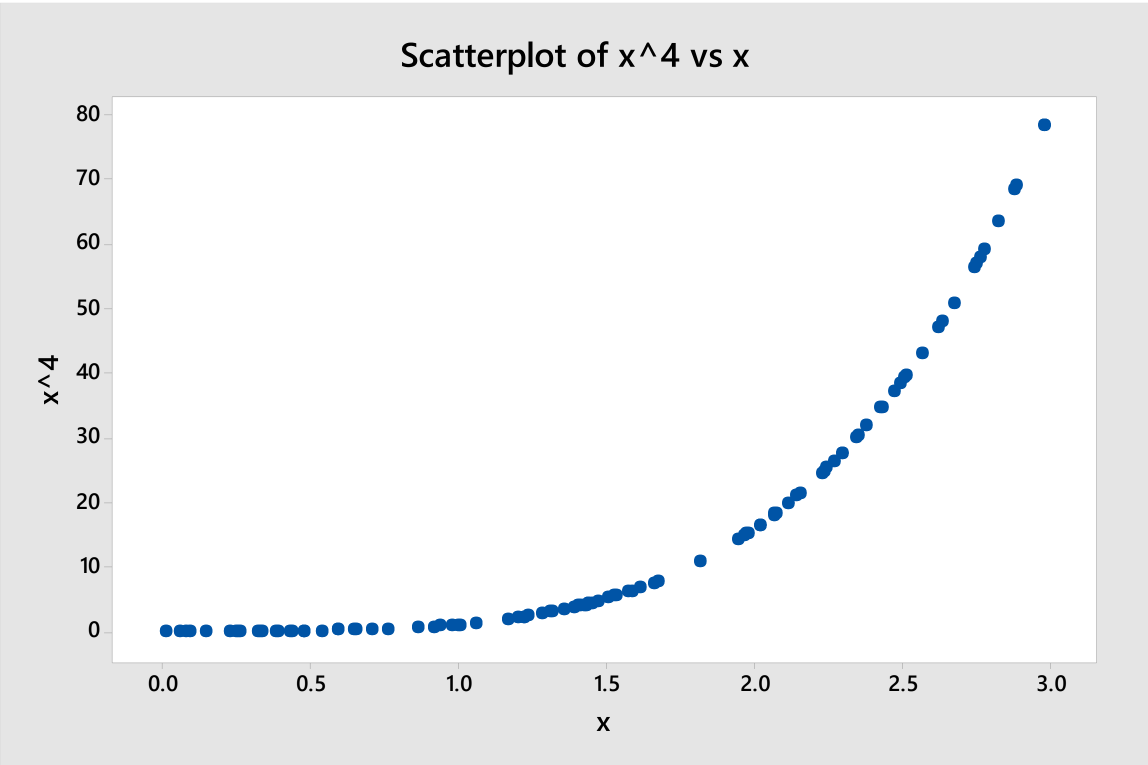
Spearman Rho: x, x^4

```
Spearman rho for x and x^4 = 1.000
P-Value = *
```

Spearman Rho: x, minuses

```
Spearman rho for x and minuses = -1.000
P-Value = *
```

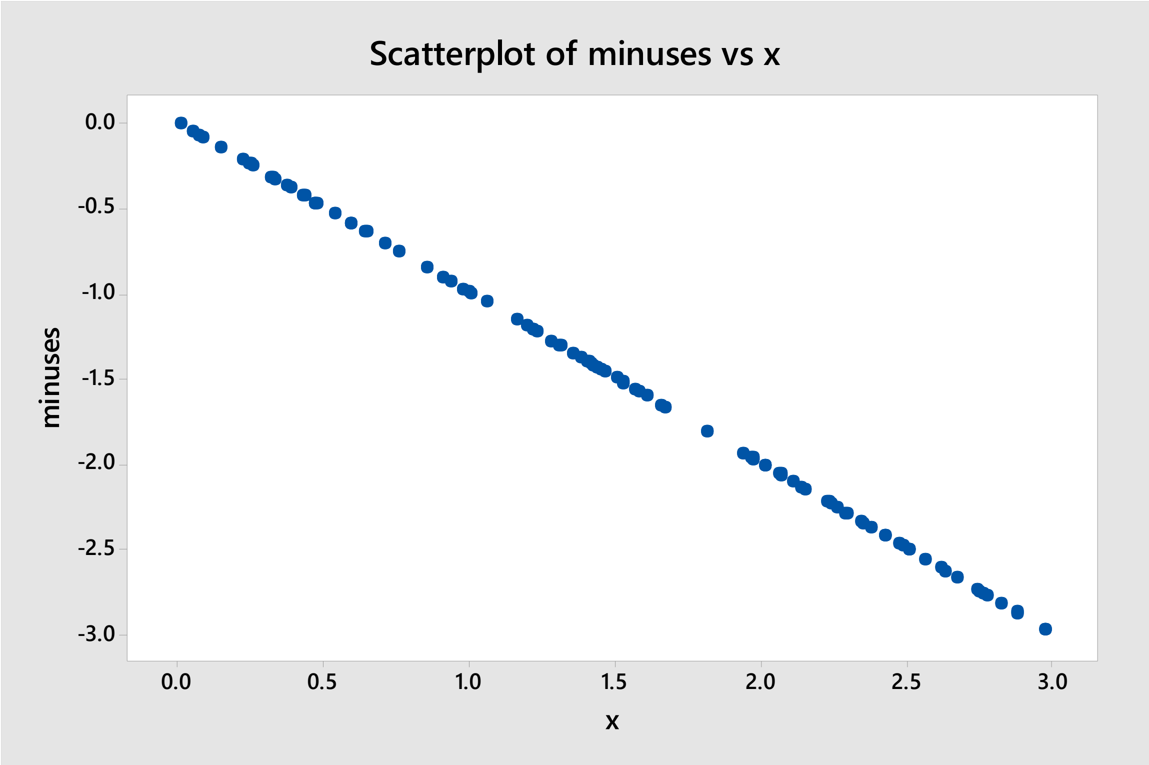
c)



For the scatterplot of x, x^4 the Pearson correlation coefficient and Spearman are not in agreement because the scatterplot is not a linear relationship.

d)

The Spearman and Pearson correlation are in agreement because the relationship is linear. We definitly know that “x” against “minuses” will have a strong negative correlation since it’s the same values on the y axis, except negative.



5)

a)

A cofounding factor is a factor that influences both the dependent and independent variable.

An example is the time when cutting the above ground stems on day 3 for the random sample of 10. If some were slightly earlier or later than another, this would influence both variables. This was controlled by them all being cut at the same time.

b)

The seedlings that had their above ground stems cut on day 3 cannot be compared with the seedlings that were not cut. If they were it would not be a fair test, as this will likely affect the growth.

c)

Descriptive Statistics: Light

Variable	Total Count	Mean	SE Mean	StDev
Light	10	25.50	1.96	6.19

Descriptive Statistics: Dark

Variable	Total Count	Mean	SE Mean	StDev
Dark	10	23.50	1.91	6.04

d)
Null hypothesis: There is no significant difference between measurements of the seedlings from “group D” and “group L”.

Alternative hypothesis: There is a significant difference between the measurements of the seedlings from “group D” and “group L”.

Where “difference” = “group L” – “group D”

e)

Two-Sample T-Test and CI: Light, Dark

```
Two-sample T for Light vs Dark

        N   Mean  StDev  SE Mean
Light  10  25.50   6.19      2.0
Dark   10  23.50   6.04      1.9


Difference = μ (Light) - μ (Dark)
Estimate for difference:  2.00
95% CI for difference:  (-3.75, 7.75)
T-Test of difference = 0 (vs ≠): T-Value = 0.73  P-Value = 0.474  DF = 18
Both use Pooled StDev = 6.1146
```

Conclusion:

The P-value is greater than the level of significance, 0.474>0.05. There is insufficient evidence to reject the null hypothesis that there is a difference in measurement between the seedlings from “group D” and “group L” at the 5% level of significance.

The confidence interval includes zero, which suggest no difference. So, this suggests there is insufficient evidence to reject the null hypothesis at the 5% level of significance.

f)

Assumptions:

The data follows a normal distribution.

The two samples are independent.

The standard deviations of the two samples are equal.

g)

Null hypothesis: The proportion of mustard seeds that germinated is the same for the sample that were grown in the dark as there were in the light.

Alternate Hypothesis: The proportion of mustard seeds that germinated is not the same for the sample that were grown in the dark as there were in the light.

h)

Test and CI for Two Proportions: Group, Recoded Germinate?

```
Event = Light
```

```
Recoded
Germinate?   X   N  Sample p
0            9  24  0.375000
1           51  96  0.531250


Difference = p (0) - p (1)
Estimate for difference:  -0.15625
95% CI for difference:  (-0.374147, 0.0616467)
Test for difference = 0 (vs ≠ 0):  Z = -1.37  P-Value = 0.171

Fisher's exact test: P-Value = 0.254
```

Conclusion:

The P-value is greater than the level of significance, 0.171>0.05. There is insufficient evidence to reject the null hypothesis that the proportion of seeds that germinated is the same in each sample at the 5% level of significance.

The confidence interval includes zero, which suggests no difference, so there is insufficient evidence to reject the null hypothesis at the 5% level of significance.

From the results in this experiment, we conclude that light doesn't have a significant effect on the probability of mustard seeds germinating.

6)

a)
Null hypothesis: The mean difference between swimming times of male and female for the 200m pool equals zero.

Alternate hypothesis: The mean difference between swimming times of male and female for the 200m pool does not equal zero.

Where "Difference" = "Female" - "Male"

Two-Sample T-Test and CI: 200M_pool, Recoded Sex

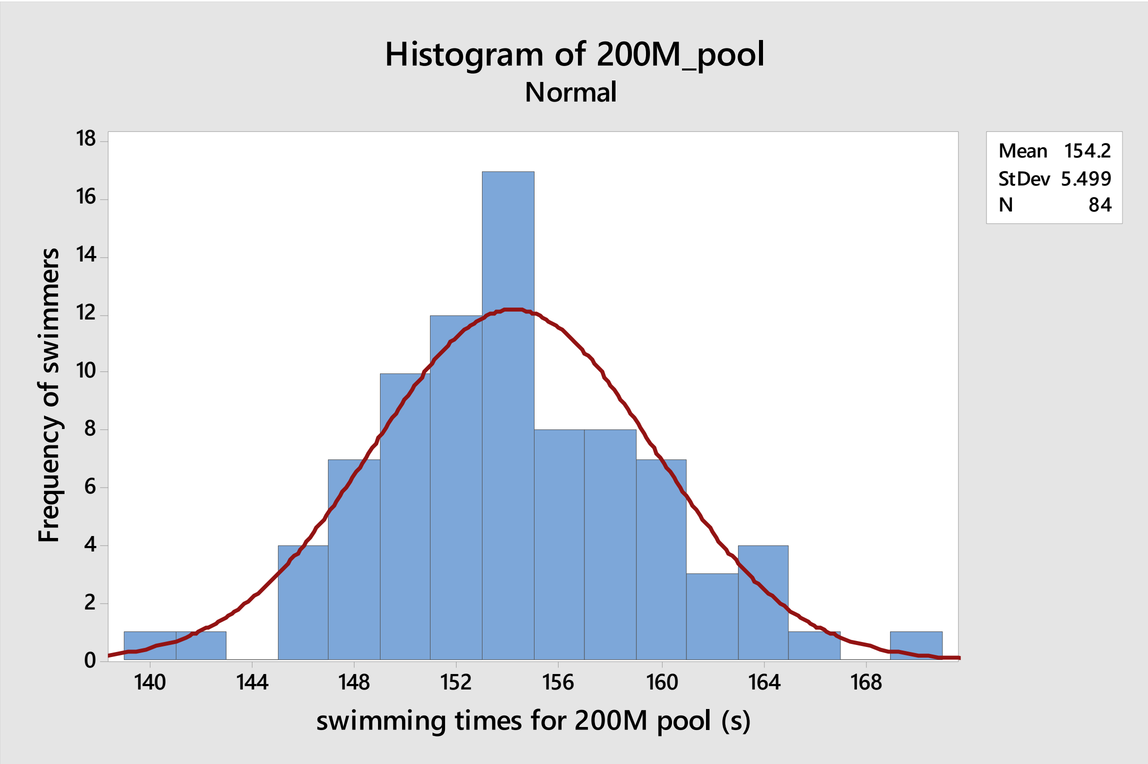
```
Two-sample T for 200M_pool

Recoded
Sex       N    Mean  StDev  SE Mean
1        31  156.31   7.09      1.3
2        53  152.98   3.88     0.53


Difference = μ (1) - μ (2)
Estimate for difference:  3.32
95% CI for difference:  (0.95, 5.70)
T-Test of difference = 0 (vs ≠): T-Value = 2.78  P-Value = 0.007  DF = 82
Both use Pooled StDev = 5.2892
```

Conclusion:

The two sample T test was used because the samples are not paired.

The P-value is less than the level of significance, 0.007<0.05. There is sufficient evidence to reject the null hypothesis that there is no difference between the male and female swimming times for the 200m pool at the 5% level of significance.

The confidence interval does not include zero, which suggests a difference, so there is sufficient evidence to reject the null hypothesis at the 5% level of significance. The confidence interval implies the female swimming times are higher than male.



b)

Null hypothesis: The mean difference between the swimming times in the pool and sea of males equals zero.

Alternate hypothesis: The mean difference between the swimming times in the pool and sea of males does not equal zero.

Where “difference” = “swimming times in pool” – “swimming times in sea”

Paired T-Test and CI: 200M_pool, 200M_sea

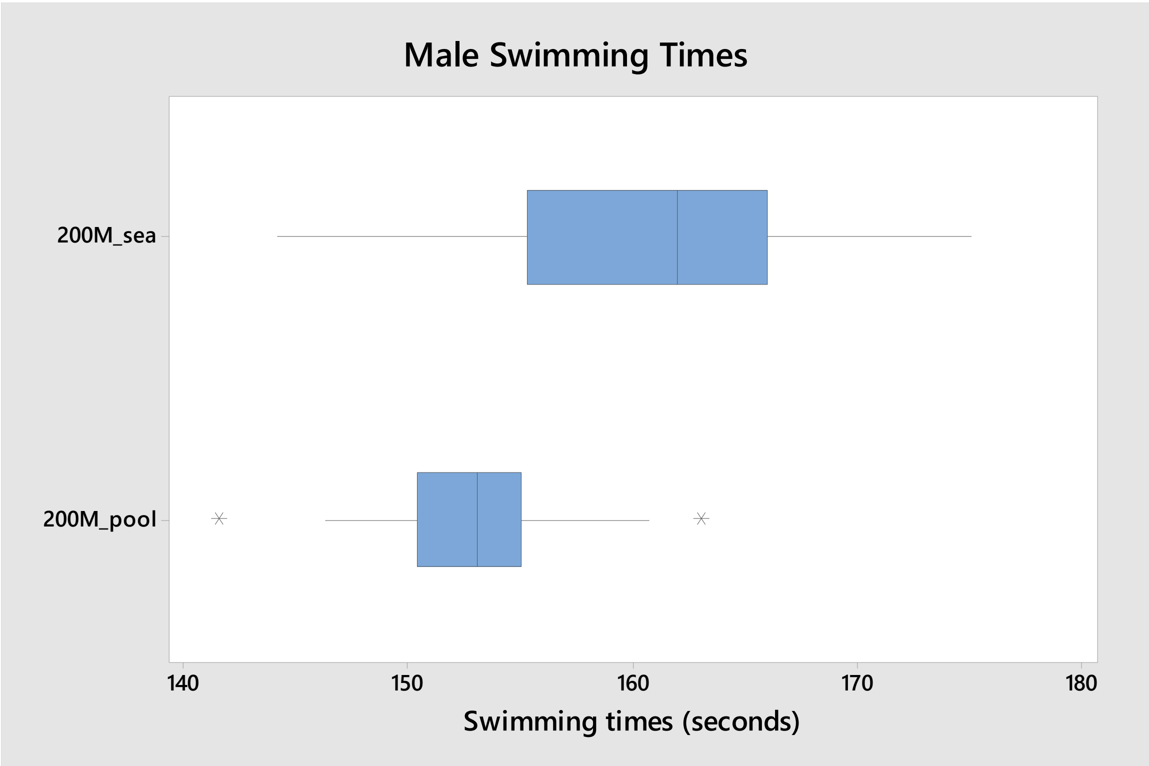
```
Paired T for 200M_pool - 200M_sea

              N      Mean  StDev  SE Mean
200M_pool    53   152.985  3.884    0.534
200M_sea     53   160.802  7.117    0.978
Difference   53    -7.817  4.441    0.610


95% CI for mean difference: (-9.041, -6.593)
T-Test of mean difference = 0 (vs ≠ 0): T-Value = -12.81  P-Value = 0.000
```

Conclusion:

This test was used because the samples are paired. The same men swam in the pool and the sea.

The P-value is approximately zero, which is less than the level of significance, 0<0.05. There is sufficient evidence to reject the null hypothesis that there is no mean difference between the male swimming times in the pool and sea at the 5% level of significance.

The confidence interval does not include zero, which suggests no difference. It implies the times in swimming times were higher in the sea than in the pool.



c)

Null hypothesis: The mean difference between national and local club swimming times of males in the pool is equal to zero.

Alternate hypothesis: The mean swimming times of the males from the local club are less than the national times.

Where "difference" = "club swimming times" – "national swimming times"

One-Sample Z: 200M_pool

```
Test of μ = 155 vs < 155
The assumed standard deviation = 10


Variable     N    Mean  StDev  SE Mean  95% Upper Bound      Z      P
200M_pool   53  152.98   3.88     1.37           155.24  -1.47  0.071
```

Conclusion:

A one sample Z test was used because the population standard deviation is known.

The P-value is greater than the level of significance, 0.071>0.05. There is insufficient evidence to reject the null hypothesis that the mean difference in swimming times is equal to zero at the 5% level of significance.

The club cannot claim that on average the men are faster than the national average at the 5% level of significance.