

Statistics : Final Exam

Baptiste Rouger

December 7, 2016

Exercise 1

Question 1

The population is the cones in the eye. The variable is the number of the L type of cone. In this case, the sample is the N cones observed.

Question 2

The population is the humans, the variable is the percentage of cone, and the sample here is the 10 people we observe.

The correlation test show, for $M_{\%}$ and $L_{\%}$ (FIGURE 1 & 6), a correlation coefficient of -0.96 with a p-value of $1.0620 \cdot 10^{-5}$. As the correlation coefficient is near from -1 and the p-value far under 0.05 , this shows that there is actually a correlation between $M_{\%}$ and $L_{\%}$, as the p-value is far lower than 0.05 .

For $M_{\%} - L_{\%}$ and $S_{\%}$ (FIGURE 2 & 7), the correlation coefficient is 0.192 and the p-value 0.5951 . As the correlation coefficient is near to 0 and the p-value is far above 0.05 , this test does not show a correlation between $M_{\%} - L_{\%}$ and $S_{\%}$ for this sample.

We can be pretty confident in the first correlation test, as the p-value is really low. Though, for the second test, it seems pretty unlikely that there actually is a correlation, as the correlation coefficient is far from 1 and the p-value far above 0.05 . This sample could be repeated to check this.

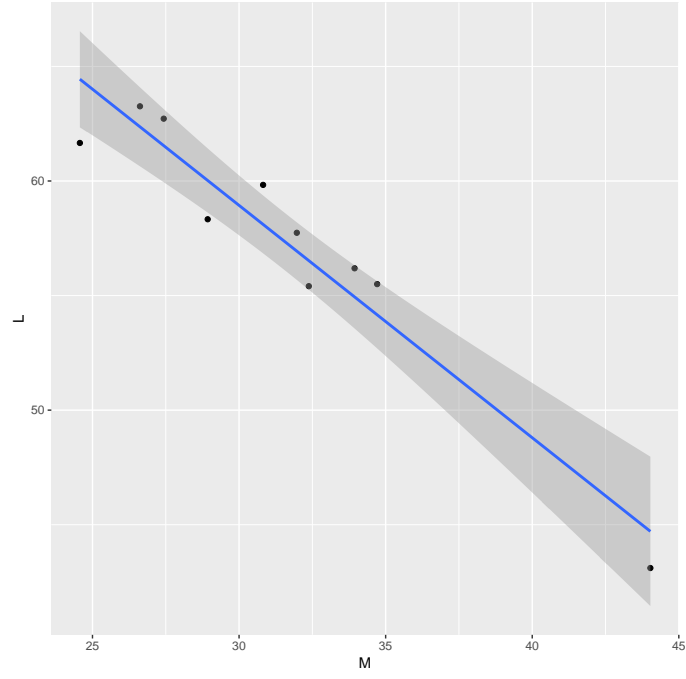


Figure 1: Plot of $M_{\%}$ against $L_{\%}$

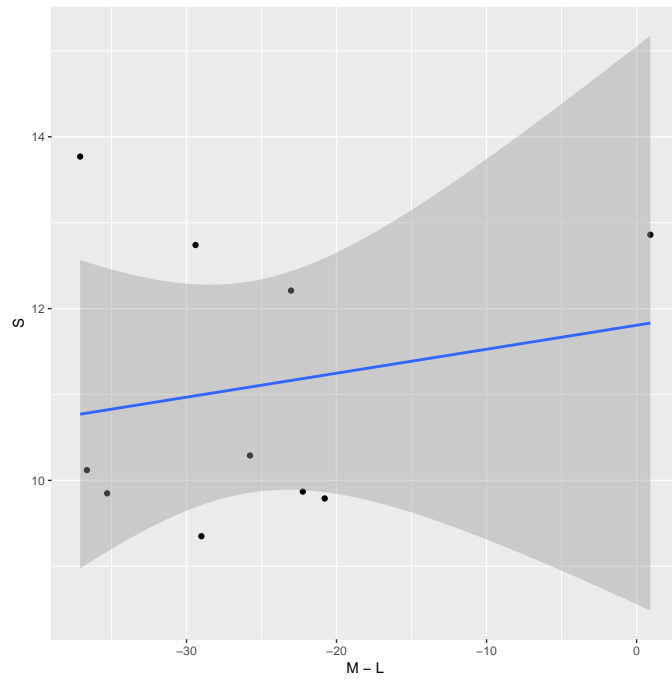


Figure 2: Plot of $M_{\%} - L_{\%}$ against $S_{\%}$

Exercise 2

Question 1

Here, we want to see if one of the sample is not equal to the other (i.e the brand put less powder than the other). Thus, our null hypothesis H_0 will be :

H_0 : the two brands tuned the machine to put the same weight of powder.

We then perform a two sample Student analysis with R to see if the two means are equal (FIGURE 8), and we get a statistic result of 2.2942 and a p-value of 0.03481. We can see here that we refute the null hypothesis as the statistic result is above 1.96 and the p-value is under 0.05. Thus, we can say that the brand A and B tuned the machine in different ways.

We perform an other two sample Student analysis with the null hypothesis being

H_0 : the brand A put more powder than B.

We then get from the t-test (FIGURE 8) a statistic result of 2.2942 as above and a p-value of 0.0174. We reject the null hypothesis as the statistic result is under 3.92 and the p-value under 0.05. Thus, A puts significantly less powder than B.

Question 2

We want to check if the two brands actually sell 100mg of powder. For both brands, our null hypothesis will then be :

H_0 : the brand sell doses of 100mg.

We perform a one sample Student analysis for both brand samples (FIGURE 10). We get for the brand A a statistic result of -2.095 and a p-value of 0.06565. This shows us that we can not refute the null hypothesis as the statistic result is under -1.96 and the p-value higher than 0.05. Thus we cannot tell that the brand A sells, in mean, less or more than 100mg of powder, and would need more samples to be determine this.

For the brand B, after performing the same test (FIGURE 10), we get a statistic result of -6.4272 and a p-value of 0.0001213. This statistic result is far under -1.96 and the p-value is far under 0.05, thus we do not refute the null hypothesis and can say that the brand B sells, in mean, 100mg of powder.

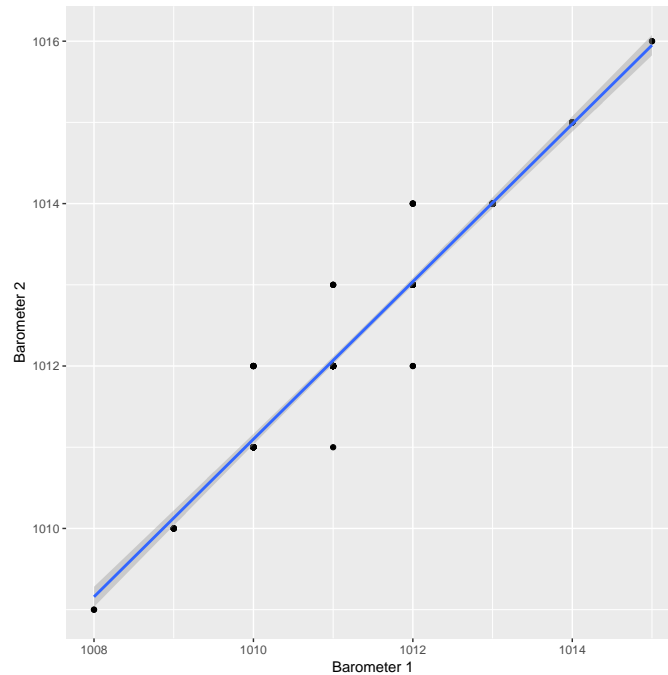


Figure 3: Plot of the data of the second barometer against the first one

Exercise 3

Question 1

Using visual inspection first (FIGURE 3), we observe a pretty good correlation between the two barometers. As we can see, all the points are around a line (here the linear model regression in blue, with confidence interval of 95%).

We then use the correlation function of R (FIGURE 11) to see if this correlation exists (with the Spearman option). We obtain a correlation coefficient of 0.96 which is close to 1, and a p-value of 0 (it is likely that the actual value is higher, but it is the value returned in the matrix by R. It seems this value is under $1 \cdot 10^{-5}$). We can conclude that there is a significant correlation between the two barometers.

Question 2

Using the built-in R function *lm* (FIGURE 12), we find an equation of the linear model such as $y = 0.97035x + 31.04732$. We then use the summary function on the linear model variable, and see that the significance level is $< 2.2 \cdot 10^{-16}$. Thus, we can say that our linear model is significant.

Question 3

Using the built-in R function for the Kolmogorov-Smirnov test, I try the null hypothesis

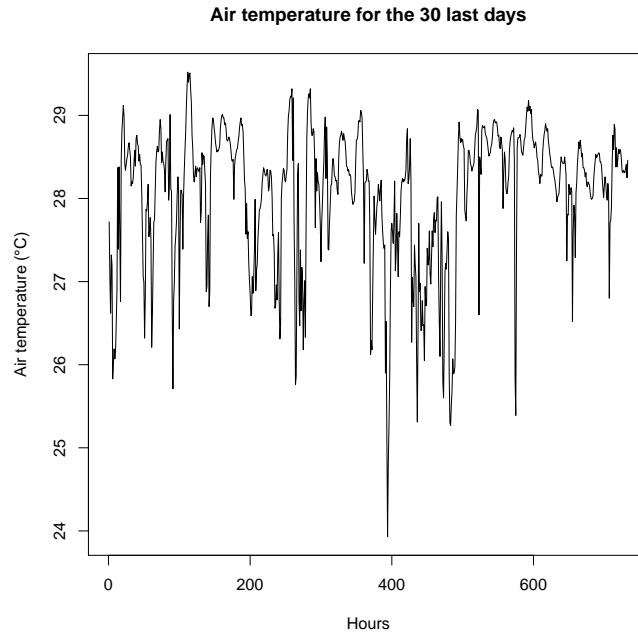


Figure 4: Plot of the temperature the last 30 days

H_0 : The two samples have the same distribution.

I get the result (with the built-in R function : FIGURE 13) such as $D = 0.60174$ and a p-value such as $2.2e - 16$. As D is lower than 1.96 which corresponds to the 5% confidence interval value in the tables and the p-value far under 0.05; we can say that these two samples are the statistically the same.

Question 4

Please find the data presentation in the FIGURE 4. The graphical presentation of the correlation is presented in FIGURE 5.

We use the correlation function of R to see if a correlation exists (FIGURE 14). We obtain a correlation coefficient of 0.04 wich is far below 1, and a p-value of 0.3259. Thus, we cannot conclude that there is a significant correlation between the two barometers.

Question 5

Using the built-in R function *lm* (FIGURE 15), we find a equation of the linear model such as $y = 7.958e - 02x + 1.007e + 03$. We then use the summary function on the linear model, and see that the signification level is 0.3259. Thus, we cannot say that our linear model is relevant.

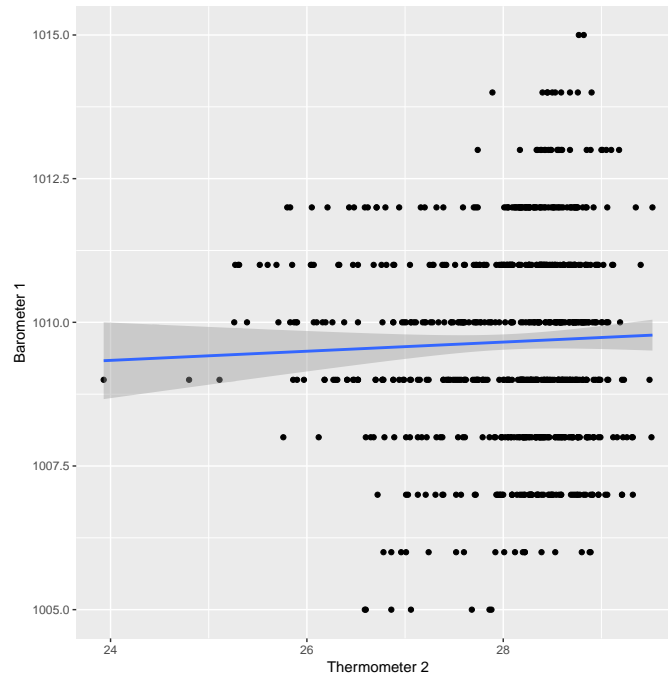


Figure 5: Plot of the data of the first barometer for against the air temperature for the last 30 days

Question 6

Even though the two data are different on the physical approach, I think it could make sense to test if they have the same distribution. Indeed, the statistical distribution of the values it takes are not influenced by the physical nature of these parameters which could follow any distribution. It could then be interesting to do a one sample Kolmogorov-Smirnov test. Though, I don't see any point to compare one variable to the other with the same test.

Appendix

```
ml <- rcorr(M,L)
print("ML")
print(ml)
```

Figure 6: Computation of the correlation between M and L

```
mls <- rcorr(M-L, S)
print("MLS")
print(mls)
```

Figure 7: Computation of the correlation between M-L and S

```
tab <- t.test(A,B)
print("TAB")
print(tab)
```

Figure 8: Computation of the Student test between A and B

```

tabg <- t.test(A,B, alternative="greater")
print("TABg")
print(tabg)

```

Figure 9: Computation of the Student test between A and B for A is greater than B

```

ta <- t.test(A, mu=100)
print("TA")
print(ta)

```

```

tb <- t.test(B, mu=100)
print("TB")
print(tb)

```

Figure 10: Computation of the Student test for mean of A and mean of B equal 100

```

bar7 <- rcorr(Bars$Bar1, Bars$Bar2, type="spearman")
print("Bars7")
print(bar7)

```

Figure 11: Computation of the correlation between the two barometers

```

lmbar7 <- lm(Bars$Bar2 ~ Bars$Bar1)
summary(lmbar7)

```

Figure 12: Computation of the linear model of the correlation between the two barometers

```

t723 <- ks.test(Bar23, Bars$Bar1)
print(t723)

```

Figure 13: Computation of the Kolmogorov-Smirnov test between the 7 days and 23 days samples

```

ap <- rcorr(Bar130[,2], Air2[,2])
print("AP")
print(ap)

```

Figure 14: Computation of the correlation between the barometer and the thermometer for 30 days


```
lmAP <- lm( Bar130 [ ,2] ~ Air2 [ ,2] )  
summary(lmAP)
```

Figure 15: Computation of the linear model of the correlation between the barometer and the thermometer data for 30 days