Faculty of Sciences and Bio-Engineering Sciences
Arthur Chomé - 0529279

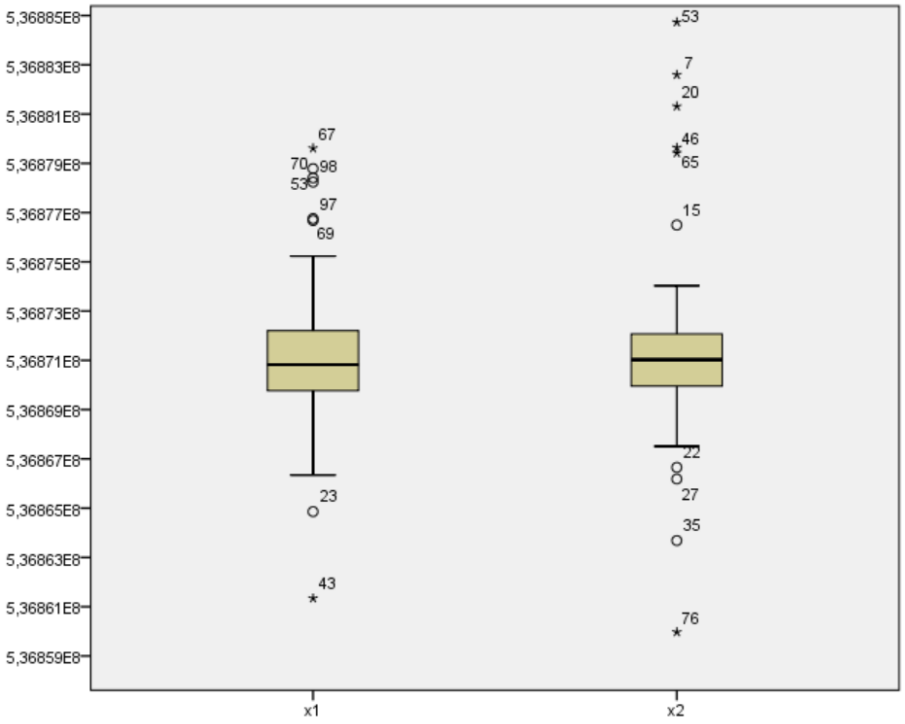
Assignment 2: Statistics

1 Question 1

Given our student number 0529279 and the online calculator[1], were assigned the sets 4, 1, 4 and 3. We used IBM's SPSS Statistics package version 23[2] for the first 3 exercises and R[3] for the question 4.
The used dataset –set 4 for the first question– was gathered by way of an experiment measuring the resonance frequencies of two types of crystal oscillators (variables x1 and x2) needed to provide timing information for high performance hardware.

1.1 Make an appropriate plot to show how the data are distributed. Explain why you used this type of plot (2+2pts)

The use of boxplots is a standardized way of displaying the distribution of data based on the five number summary.



For boxplots, between the first and third quartile lies 50 percent of all data

[1] https://ai.vub.ac.be/ bart/statsnumbers.html
[2] https://en.wikipedia.org/wiki/SPSS
[3] https://www.r-project.org/

and between the maximum and minimum lies about 99.75 percent of all data. Datasets that have a normal distribution have symmetrical boxplots with the means being in the center. This is clearly the case for both datasets. Set x1 however has a wider distribution of its data with the maximum and minimum of the data being further away from the median than in dataset x2.
The use of boxplots is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum. Using boxplots gives us a simple way of detecting symmetry and distribution of datasets and comparing them between each other.

1.2 Test whether the resonance frequencies of the two types could be identical (2 pts.)

Based on the experiment's description, we want to see if the same variable applied on two different populations (e.g. different groups of oscillator crystals) makes for a difference of the means of variables. Because the variance of both datasets differs significantly, a normal T-test would not be reliable in doing so. This is why we settle for Welch's T-test because it is robust even when the sample size or the variance is unequal. In this case, an independent T-test is the best option. Our null hypothesis in this case is that the means of both data collections do not differ.

ANOVA

frequency

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	186642792.2	1	186642792.2	1,508	,221
Within Groups	2.450E+10	198	123750522.7		
Total	2.469E+10	199			

Robust Tests of Equality of Means

frequency

	Statistic[a]	df1	df2	Sig.
Welch	1,508	1	135,500	,222
Brown-Forsythe	1,508	1	135,500	,222

a. Asymptotically F distributed.

Null & Alternative Hypothesis

For this T-test, our null hypothesis is that no significant difference in means exists between the datasets x1 and x2 of the oscillator crystals. The alternative would be that a significant difference in means exists between the two. All

hypothesis tests make use of a p-value[4] to weigh the strength of the evidence. If the value is smaller than 0.05, we say there is strong evidence against the null hypothesis so it gets rejected.

Conclusion
Because the p-value in this case is greater than 0.05 being 0.222, we cannot conclude that a significant difference exists and we accept our nullhypothesis: there is no significant difference in the means of both datasets.

1.3 Also test whether they are different from the target frequency of 536 870 912 Hz. Explain why you used this test. (2 pts.)

This test is used to see whether a dataset's mean is equal to a given average value. It can only be applied on a random sample with a normal distribution with the variance being –optionally– unknown. Previous boxplots show our datasets to be normally distributed, we assume them to be randome samples.

Null & Alternative Hypothesis
Our null hypothesis here is that –for both crystal x1 and x2– their mean does not differ significantly from the average target frequency of 536 870 912 Hz. If our p-value is smaller than 0.05, we would reject the null hypothesis and accept the alternative hypothesis: the difference is significant.

T-test Dataset x1
After executing the T-test on dataset x1 of the crystal oscillators, we get the following tables. Our null hypothesis is that there is no significant difference between dataset x1's mean and given dataset average.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
x1	100	536869604.5	14415,165	1441,516

One-Sample Test

	Test Value = 536870912					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
x1	-,907	99	,367	-1307,530	-4167,81	1552,75

[4] https://en.wikipedia.org/wiki/P-value

Conclusion
Because the p-value 0.367 is bigger than 0.05 we conclude there is no significant evidence against our null hypothesis and accept our claim that x1's mean is equal to 536 870 912 Hz.

T-test Dataset x2
After executing the T-test on dataset x2 of the crystal oscillators, we get the following tables. Our null hypothesis is that there is no significant difference between dataset x2's mean and given the given average.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
x2	100	536871536.5	6301,11703	630,11170

One-Sample Test

	Test Value = 536870912					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
x2	,991	99	,324	624,53000	-625,7483	1874,8083

Conclusion
Because the p-value 0.324 is bigger than 0.05 we conclude there is no significant evidence against our null hypothesis and accept our claim that x2's mean does not differ significantly from 536 870 912 Hz.

1.4 Would you trust these crystals in your application, and why (not)? (2 pts.)

Because both dataset means do not differ significantly, are normally distributed and don't differ significantly from the average 536 870 912 Hz, we can trust both crystal types x1 and x2.

2 Question 2

We were assigned set 1 for this question, the data represents the number of times visitors bought something or simply left on two variants of a website (site1 and site2). The files whose name ends in a contain the results after 200 visits to each variant.

buy * site Crosstabulation

Count

		site		
		,00	1,00	Total
buy	,00	200	199	399
	1,00	0	1	1
Total		200	200	400

Figure 1: The data seems to be a contigency table with the row representing the number of times whether something got bought or not in the two sites.

2.1 Do an appropriate test to see whether these data indicate a difference between the two sites. Explain why you used this test and give the effect size. (2+1+2 pts.)

Because dataset A of our sites clearly represents a contingency table, we could opt to use either Fisher's Exact Test or Pearson's Chi-Squared Test. However, Fisher's test is more appropriate for small sample size and because the size of our test data (400 site visitors) is fairly low, we opt for Pearson Chi-Squared Test.

Null & Alternative Hypothesis
For this experiment, our null hypothesis is that there is no significant difference in buying behavior of the two sites. The alternative would be that there is significant difference between the two.

Conclusion
For both the one-sided and two-sided variant of Fisher's Exact Test, we get p-value of 1,000 and 0,500 respectively. This means that we accept the null hypothesis that there is no significant buying behavior between the two sites.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	1,003[a]	1	,317	1,000	,500	
Continuity Correction[b]	,000	1	1,000			
Likelihood Ratio	1,389	1	,239	1,000	,500	
Fisher's Exact Test				1,000	,500	
Linear-by-Linear Association	1,000[c]	1	,317	1,000	,500	,500
N of Valid Cases	400					

a. 2 cells (50,0%) have expected count less than 5. The minimum expected count is ,50.

b. Computed only for a 2x2 table

c. The standardized statistic is 1,000.

2.2 The files whose name ends in b contain data from a longer run (5000 visits) of the experiment. Do an appropriate test to see whether these data indicate a difference between the two sites. Explain why you used this test and give the effect size. (1+1+1 pts)

For set B, we have data of about 500 site visitors. Because we have a bigger sample size than for set B, we now opt to use the non-parametric chi-squared test.

sites * data Crosstabulation

Count

		data		Total
		,00	1,00	
sites	,00	4955	45	5000
	1,00	4885	115	5000
Total		9840	160	10000

Figure 2: 0,00 represents the times a site got visited and nothing got bought. 1,00 is the value for when a site got visited and something got bought.

Null & Alternative Hypothesis
Our null hypothesis is that there is no significant difference in buying behavior of the two sites (just a previous test).

Conclusion
We can see that our significance level (p-value) when performing the Pearson

Chi-Square test is equal to 0,000 hence less than 0.05. This is why we reject the null hypothesis that the buying behaviour of both sites is equal and we accept the alternate hypothesis that they are different.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	31,123[a]	1	,000		
Continuity Correction[b]	30,240	1	,000		
Likelihood Ratio	32,183	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	31,120	1	,000		
N of Valid Cases	10000				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 80,00.

b. Computed only for a 2x2 table

2.3 What did this teach you about the importance of sample size (a.k.a. "big data")? (2 pts)

Sample size is directly related to a statistic's margin of error, or how accurate a statistic can be calculated to be. The bigger the sample size, the less chance there is to wrongfully accept the null hypothesis.

3 Question 3

We are using set number 4 for the fourth question. It deals with the grade difference of a group of students that followed remedial lectures and a control group who did not.

3.1 Investigate whether the remedial lectures had any effect, and explain what you did and why you did it (2+2 pts.)

If the variances between the datasets would not differ significantly, we could have opted for a normal independent T-test. Because Levene's test for the equality of variances has a p-value of less than 0.05 (being 0.014), we conclude that both samples their variances differ significantly. Welch's test is a better alternative being robust for independent samples with different variances.

3.1.1 Null & Alternative Hypothesis

Our null hypothesis here is that –for both groups– the difference in study performance do not differ significantly from one another. If our p-value is smaller than 0.05, we would reject the null hypothesis and accept the alternative hypothesis: the difference is significant.

		Levene's Test for Equality of Variances	
		F	Sig.
data	Equal variances assumed	6,564	,014
	Equal variances not assumed		

Robust Tests of Equality of Means

data

	Statistic[a]	df1	df2	Sig.
Welch	1,718	1	38,810	,198
Brown-Forsythe	1,718	1	38,810	,198

a. Asymptotically F distributed.

3.1.2 Conclusion

Because Welch's test has a p-value of 0.198 (greater than 0.05), we accept the null hypothesis that the means of both dataset are equal.

3.2 What was the power of this experiment? What does this mean, and are you happy with this (and why)? (2+1 pts)

When it comes to hypothesis testing, we can make type I and type II errors. Type I errors occur when we would reject the null hypothesis when it is in fact true, type II errors occur when we would accept the null hypothesis when it is false. The power of a test is the probability of rejecting the null hypothesis, given it is false. It depends on several factors including the choice of Alpha and the sample size.

Discussion
Our total sample size for this experiment is of 50. With with a p-value of 0.196 and an effect size of 0.035, we can say that there is a 25 % chance that the detected difference in the test is really there. Hence, the probability of making a type II error would be 75 %.
Having a 75 % change of accepting the null hypothesis when it is false makes

Tests of Between-Subjects Effects

Dependent Variable: data

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	1,479[a]	1	1,479	1,718	,196	,035	1,718	,250
Intercept	13160,909	1	13160,909	15284,129	,000	,997	15284,129	1,000
groups	1,479	1	1,479	1,718	,196	,035	1,718	,250
Error	41,332	48	,861					
Total	13203,720	50						
Corrected Total	42,811	49						

a. R Squared = ,035 (Adjusted R Squared = ,014)

b. Computed using alpha = ,05

the experiment very unreliable. We must enlarge the sample size to get more reliable results.

3.3 How large would the groups need to be to achieve power 0.8? Explain. (2+1 pts)

For this experiment, we can use Cohen's d (= effect size) –expected difference between the means of the target values between the experimental group and the control group– as to determine the sample size using the power tables.
The variables at work here are the power level, the means of the control and experiment group, the pooled standard deviation of the 2 groups and our alternative hypothesis.

3.3.1 Significance, Power level & Alternative Hypothesis

These parameters for the sample size are the same for both the 2 groups of students. As is common practice in statistics, an experiment's significance level –chance of a type I error– is set at 0.05 which is the significance we will go for. As was specified in the description, our power level is of 0.8 (80 %).
For our null hypothesis, we agree that there is no difference between the means of the two groups. For our alternative hypothesis, we will go for the two-sided approach since this is the default used in Statistics. Whether we would go for a one or two-sided alternative hypothesis affects which table we would like to use.

3.3.2 Medians and Pooled Standard Deviation

The inter-individual variability is the variation among the experimental subjects, expressed as the standard deviation. We have to calculate the standard deviation for both groups.

Descriptive Statistics

	N	Mean	Std. Deviation
Control	25	16,052	1,1314
Experiment	25	16,40	,665
Valid N (listwise)	25		

Figure 3: We can see that the control group has a standard deviation of 0.9347 and the experimental group a standard deviation of 0.665.

Pooled Deviation

A problem with Cohen's D is that it is supposed to be applied to 2 groups that have the same standard deviation which is not the case here. In this case, we need to take the average of the standard deviations using following formula:

$$SD_{pooled} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{2}}$$

This gives a pooled deviation equal to: 0.8111.

3.3.3 Cohen's D

Using the means of 2 groups and the pooled deviation, we can compute Cohen's D and compare the results with a power table for a 2-sided alternative hypothesis. We use following formula:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\mathrm{SD}}$$

This gives the following result: 0.4291.

3.3.4 Conclusion

After searching the corresponding number in a power table[5] with power 0.8 with a two-tailed alternative hypothesis and significance of 0.05, we need a sample size of 99 participants.

4 Question 4

We are using set number 3 for the fourth question as the data of the performance of two algorithms for deep learning. We have to perform binary logistic regression as to estimate relationships between a dependent variable (algorithm type and parameters) and binary independent variable (probability of success). To do so, we have to built up a model for the regression specifying a target variable (binary) and the covariate variables that decide the target's value.

```
# Read a txt file, named "mtcars.txt"
question4 <- read.delim(file.choose(), header = T)
question4$type <- as.numeric(question4$type)
train = question4[sample(2, nrow(question4), replace = T, prob = c(1.0, 0.1))==1,]
# Fit the model: question4 is the training data.
mymodel = glm(success ~ type + parameters, data = train, family = 'binomial')
```

Figure 4: To set up our model, we have to load up the data .txt file, give it training data, and specify that our result success variable is in function of the parameters variable and type variable.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.334e+00  9.064e-01  -5.885 3.97e-09 ***
type        -6.678e-02  4.081e-01  -0.164     0.87
parameters   2.992e-04  6.234e-05   4.800 1.59e-06 ***
---
```

Figure 5: When looking at the p-values of the model, we can see the parameters variables is the most statistically significant for the outcome of the regression.

4.1 Can you figure out whether one of the algorithms is better, controlling for the number of parameters? Explain how you determined this (3+3 pts).

Binary Logistic regression was performed to measure the effect of parameters on successfully performing tasks for both algorithm types independently. For this, we changed all parameter entries to the median of each respective group and calculated the probability of success for each group/algorithm with the new parameter values.

[5] http://www.pilesofvariance.com/Chapter13/Cohen_Power_Tables.pdf

```
> predict(mymodel,  with(question4, data.frame(parameters=mean(parameters),
 type=factor(1:2))), type="response")
        1         2
0.4340497 0.4165832
> |
```

Figure 6: We use the same model for both experiments

Conclusion
In our regression, we see that the probability for success for both algorithms when using the mean amount of parameters for each. Algorithm 1 has a probability of success of 0.4340497 and algorithm 2 of 0.4165832. This makes algorithm 2 slightly more performant.

4.2 At which number of parameters does each of the algorithms have a 50 % chance of succeeding? Explain how you calculated this. (2+2 pts).

The model that we built for the logistic regression can be used again to calculate the probability of success in one specific case. For this purpose, we can use the predict function that takes in a model, training data (hence, all the data) and its return value (in this case response).

```
> predict(mymodel,  data.frame(type=1, parameters=18050), type="response")
       1
0.499966
> predict(mymodel,  data.frame(type=2, parameters=18274), type="response")
        1
0.5000283
```

Figure 7: Finding the right amount of parameters for the algorithms was a process of trial-and-error.

Conclusion
After trying different inputs, the closest we could get to a 50 % probability of success would be 18050 parameters for algorithm 1 and and 18274 parameters for algorithm 2.