Practice 6: Statistical Tests

Mayra Cristina Berrones Reyes 6291

October 13, 2020

1 Activity

Please answer the next questions, taking into consideration the shared links.

a) Describe the relationship between hypothesis testing and statistical tests.

To understand the relationship between these two concepts, first we describe them separately.

A statistical test gives us a mechanism to be able to make quantitative decisions about processes. The goal of this decisions is to determine if there is enough evidence to "reject" a conjecture or hypothesis about our process. In this case, the conjecture is called null hypothesis [8]. A null hypothesis proposes that no significant difference exists in a set of given observations [1].

A classical use of the statistical tests occurs in process control studies.

Hypothesis testing is a way we can test if the results of an experiment really have meaningful results. This means that we test to see if the results we have are valid by checking out the odds that said results could have happened by chance [6].

It is very common in statistics to estimate a parameter from sample data. For example, a sample of the mean of the data is then used as the point estimate of the population mean. An hypothesis test addresses the uncertainty of the sample estimate, and instead of providing an interval, it attempts to refute a specific claim about a population parameter based on the sample data we took for the experiment [8].

A common format for hypothesis test is shown in Listing 1.

- H_0: Null hypothesis.
- H_a: Alternative hypothesis.
- Statistic test: This is based on the specific hypothesis test.
- Significance level: Commonly known as α, defines the sensitivity of the test.
- Critical value: This encompasses the values of the statistic test that lead to a rejection of the null hypothesis.

Table 1: Names and types of errors of the significance level.

	Truth about population	
Decision based on sample	H_0 is true	H_0 is false
Fail to reject H_0	Correct decision (Probability = $1-\alpha$)	**Type II error** Fail to reject H_0 when it is false (Probability = β)
Reject H_0	**Type I error** Rejecting H_0 when is true (Probability = α)	Correct decision (Probability = $1-\beta$)

To answer the original question, seeing the common format for a hypothesis test, we can now say that they are both needed for a good analysis of the experimentation, because we should not just trust that the sample we took for our statistical analysis is the correct one, so we should combine engineering judgement with statistical analysis.

b) What would indicate to reject the null hypothesis?

To reject a null hypothesis we perform a statistical test, and then we compare its results with the critical value. If it is greater than the critical value, the hypothesis is rejected. A great explanation Jonathan Christensen [3] says "In a theoretical underpinning, hypothesis tests are based on the notion of critical regions: the null hypothesis is rejected if the test statistic falls in the critical region. The critical values are the boundaries of the critical region. If the test is one sides, then there will only be one critical value, but in other cases there will be two."

So we reject the null hypothesis if this falls in the critical value.

c) How is the output of a statistical test interpreted?

In this case, the output of the statistical test can be that we accept the null hypothesis, or we reject it. If ithe null hypothesis falls in the critical value, we reject it. The null hypothesis is a statement about belief [8], so the test procedure is done in a way that the risk of rejecting the null hypothesis is small when it is in fact true.

d) How to select the alpha value?

The risk we mentioned in previous questions, α, is often referred as value significance level of the test. When we use a small value of α it is often said that it actually proves something when the null hypothesis is rejected.

There is no magic significance level that will give us a 100% accuracy results. The most common α values are 0.05 and 0.01, but they are mainly used based on tradition. Because this test are based in probability, there is always a chance of a wrong conclusion.

When doing this experiments, there are 2 types of errors, which are related and determined to the level of significance and the test used. So we should determine which error has worst concecuences for the experiment, before defining the risks.

In Table 1 we name these type of errors.

Table 2: Examples of different parametric and non parametric procedures for the same type of analysis.

Analysis Type	**Parametric**	**Non parametric**
Compare means between 2 distinct/independent groups	Two sample t-test	Wilcoxon rank sum test
Compare 2 quantitative mesurements taken from the same individual	Paired t-test	Wilcoxon signed rank test
Compare means between 3 or more distinct/independent groups	Analysis of variance (ANOVA)	Kruskal-Wallis test
Estimate the degree of association between two quantitative variables	Pearson coefficient of correlation	Spearman rank correlation

e) What are the most frequent misinterpretations of the p-value?

The p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. It is the measure pf the probability that an observed difference may have occurred by random [2].

The incorrect interpretation of the p-value is very common. One of the most frequent is to interpret it as the probability of making a mistake by rejecting the true null hypothesis, which as we can see in Table 1 is a Type I error.

Misinterpreting the p-value as the error rate creates the illusion that there is more evidence against the null hypothesis. But if we base the whole experiment on a study of p-value of 0.05, the difference observed in the sample may not exist in the whole population [4].

f) What is the statistical power and what is it for?

The statistical power of an experiment refers to how likely it is to distinguish an actual effect from one chance. It is the likelihood that the test we perform is correctly rejecting the null hypothesis, which looking at Table 1 is the probability of avoiding Type II error [11].

A high statistical power means that the results we have are likely valid, and as the power increases, the probability of making a Type II error decreases. It can be used as a tool to estimate the sample size required in order to detect an effect in an experiment [5].

g) Examples of parametric and non-parametric statistical tests.

The definition of non parametric is very convoluted, and better explained by examples. In Table 2 we see examples of the type of analysis and the parametric and non parametric procedure.

h) Summarize THE GUIDE to find the statistical test you are looking for.

It is importan to select correctly the type of statistical test we are going to use to analyze our data. There is a very helpful flowchart [9] shown in Figure 1 that helps to find the most suitable statistical test, depending on the type of data to analyze.

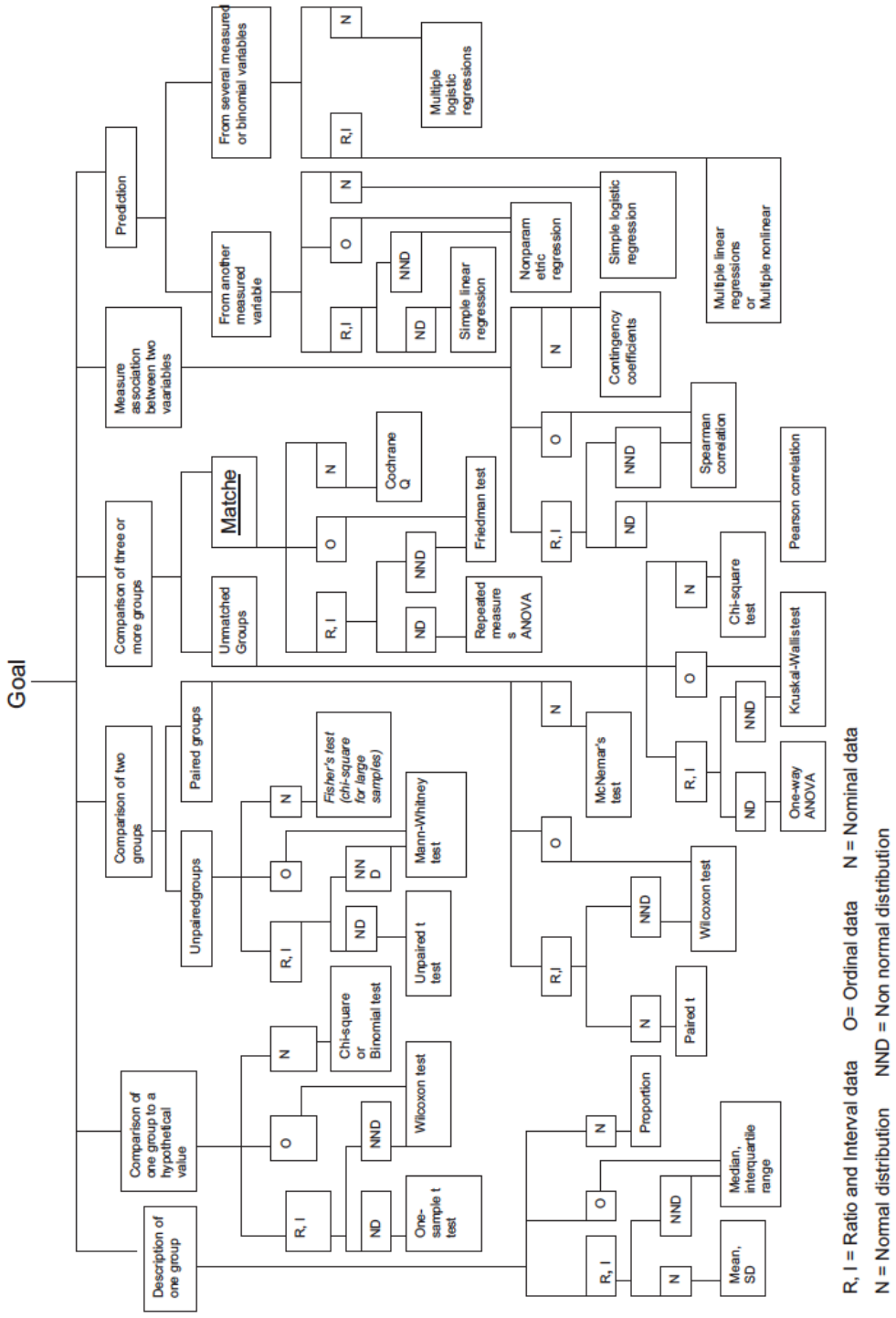


Figure 1: Flowchart for the selection of appropriate statistical tests

i) What are the assumptions to apply parametric techniques?

There are four important assumptions when it comes to the use of parametric tests in the analysis of data [10], which are listed in Listing 1.

- **Normal distribution data:** The p-value depends on a normal sampling distribution. If the sample size is big enough and the sample data point value are aproximately normally distributed, then the central limit ensures a normally distributed distribution.
- **Homogeneity of variance:** The variable in the population where we took the samples have been taken in similar variance of these populations.
- **Interval data:** The data point values should be for numerical variable and measured at this level.
- **Independence:** Data point values for variables for different groups should be independent of each other. In regression analysis, the errors should likewise be independent.

2 Experimentation

Using the same topic as practice 1, "Enviromental practices" from the page of INEGI [7] we attempt to replicate as many statistical test as possible. In this case, we still use the tables from the water section, but because the behavior of the previous data was to fractured and not compatible with any of the tests, we are focusing now on the quality of certain services on urban spaces. In this case, they give a 100 percent, and distributed in bad, regular and good quality.

In R we made a variable with each distribution, hoping to get a more normally distributed data than the one that we used before. In Figure 2 we see the distribution of each variable.

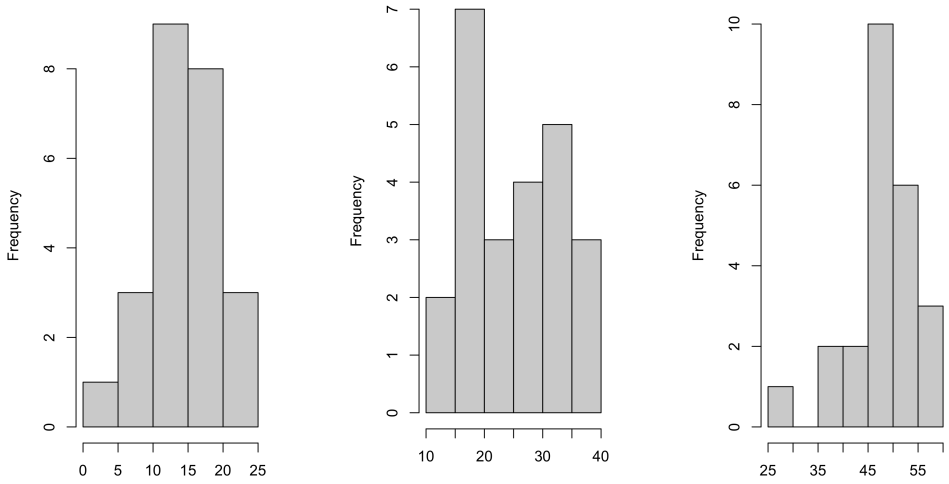
At plain sight we can speculate that Figure 2a and 2b are more or less resembling a normal distribution. But to be certain, we can perform some experimentation on the data.

2.1 One sample t-test

This is a parametric test to prove if the mean of a sample from a normal distribution could reasonably be a specific value. This test needs as values the name of the data, and a number that represents the posible mean. Seeing all three histograms, we set this parameter in each one as 15, 25, and 45, because is the most central in each one.

The results of the test are shown in Table 3.

Interpreting this results, we put our null hypothesis in each case that the mean of each data variable was 15, 25, and 45 for `bad`, `regular` and `good` respectively. Seeing the results on Table 3 we can see that in all three cases we accept the null hypothesis. In column mean we also can see that our first assumption to use the medium value of the histogram as mean was not so far off. The farthest one was the `good` variable, and it was off only by 2.51.



(a) Distribution of bad reviews (b) Distribution of regular reviews (c) Distribution of good reviews

Figure 2: Distribution of the public opinion about the water service of urban spaces.

Table 3: Output in R of the One sample t-test.

	One sample t-test variables				
Data	t	df	p-value	95% confidence interval	Mean
Bad	-0.4531	23	0.6547	12.4916 — 16.6067	14.54
Regular	0.0600	23	0.9527	21.8560 — 28.3318	25.09
Good	1.7898	23	0.0866	44.6074 — 50.4312	47.51

Table 4: Output in R of the Wilcoxon signed rank test.

	Wilcoxon variables			
Data	V	p-value	95% confidence interval	Pseudo median
Bad	141	0.8115	12.4177 — 16.8965	14.77
Regular	151	0.9888	21.6393 — 28.4672	25.15
Good	234	0.0150	45.7382 — 50.6462	48.25

Table 5: Output in R of the Shapiro test.

	Shapiro test variables	
Data	W	p-value
Bad	0.9664	0.5814
Regular	0.9485	0.2515
Good	0.8776	0.0074

2.2 Wilcoxon Signed Rank Test

We move to a experimentation to test the mean of a sample when normal distribution is not assumed of the data. In Table 4 we see the results of this test on each variable.

In this case, without the assumption of normality in our distribution the p-value, we reject the null hypothesis on the `good` variable. The difference in the media is also a bit larger in all three variables.

2.3 Shapiro test

This test is used to see if our sample follows a normal distribution. Same as the experiments before, we are going to check if all of our variables follow a normal distribution. Table 5 shows the results of this test.

As expected, variables `bad` and `regular` pass the test as normal distribution, but the `good` variable does not.

2.4 Kolmogorov and Smirnov test

This test helps find out if two samples follow the same distribution. From Figure 2 we can see that the distribution of our variables are different, but in Table 6 we prove it further than just analyzing a figure.

2.5 Fisher F-test

This test can be used to check if two samples have the same variance. We used the same parings as the Kolmogorov-Smirnov test. The results can be seen in Table 7. In this case, all three variables pass the null hypothesis, and the ratio of variances gives us also a positive result

Table 6: Output in R of the Kolmogorov-Smirnov test.

	Kolmogorov-Smirnov variables	
Data	D	p-value
Bad and Good	1.0000	6.195×10^{-14}
Regular and Good	0.9166	6.996×10^{-11}
Bad and Regular	0.6250	0.0001

Table 7: Output in R of the Fisher F-test.

	Fisher F-test variables					
Data	F	num df	denom df	p-value	95% confidence interval	Ratio of variances
Bad and Good	0.4993	23	123	0.1028	0.2159 — 1.1541	0.4992
Regular and Good	1.2364	23	23	0.6150	0.5348 — 2.8582	1.2364
Bad and Regular	0.4038	23	23	0.0343	0.1746 — 0.9334	0.4038

between them.

2.6 Correlation

This test gives us as a result the linear relationship of the two continuous variables. The null hypothesis in this case is that the true correlation between both variables is zero. The results of this tests are shown in Table 8.

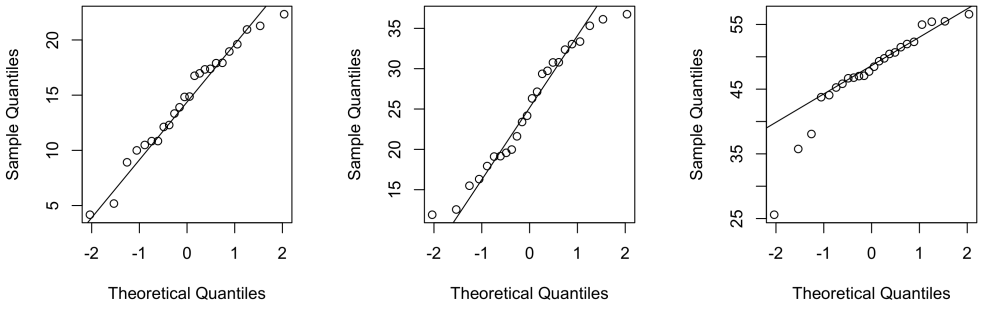
In this case the null hypothesis of the correlation could only be accomplished by the pair of `bad` and `good` variables.

3 Other experimentation

We have talked and experimented with different known statistical tests. And even if now when, after all the research they seem a lot more understandable, visual queues remain a favorite.

Table 8: Output in R of the correlation test.

	Correlation variables				
Data	t	df	p-value	95% confidence interval	Correlation
Bad and Good	1.5459	22	0.1364	-0.1034 — 0.6360	0.3130
Regular and Good	2.7261	22	0.0123	0.1242 — 0.7532	0.5024
Bad and Regular	2.9768	22	0.0069	0.1689 — 0.7723	0.5358



(a) Normal Q-Q plot of bad reviews (b) Normal Q-Q plot of regular reviews (c) Normal Q-Q plot of good reviews

Figure 3: Normal Q-Q plot of all the variables

So, after some searching, we found other ways to show our data. In Figure 3 we have the Q-Q plot, or more commonly known as the quantile-quantile plot. It helps to assess if a set of data came from some distribution such as normal or exponential.

This is the visual check of something that the t-test does, which is assume that the variable is normally distributed, so it is not a complete proof, but something to help visualize the data. In the case of Figure 3 we compare our results with Table 3, and we find similar results.

Another visual queue besides the histograms if Figure 2 to help find the mean of our data are the box plots. In Figure 4 we show the box plots of all the variables. Here we can see more clearly the mean of each variable, as well as the odd behavior of some of the data in variable `Good`.

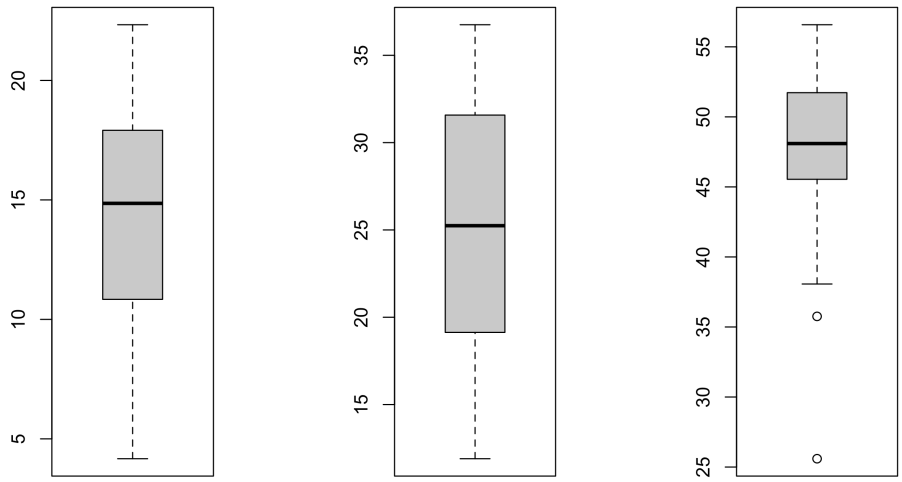
4 Conclusions

So far I was understanding fine the histograms and plots made in class, but in the last practice one of the mayor faults in my work was that I misinterpreted some data. With this practice the concepts of the tests and their results are clearer. They can help us move from just guessing our answers from the plots and actually proving some points, because many of the articles and books researched for this practice say that plots are only visual aids, and should require a bit more experimentation to back them up.

References

[1] Statistical tests: When to use which. `https://towardsdatascience.com/statistical-tests-when-to-use-which-704557554740`. Accessed: 2020-10-12.

[2] Brian Beers. P value definition. `https://www.investopedia.com/terms/p/p-value.asp`. Accessed: 2020-10-12.



(a) Box plot of bad reviews (b) Box plot of regular reviews (c) Box plot of good reviews

Figure 4: Box plots of all the variables

[3] Jonathan Christensen. What is a critical value in statistics. `https://math.stackexchange.com/questions/281940/what-is-a-critical-value-in-statistics`. Accessed: 2020-10-12.

[4] Minitab Blog editor. How to correctly interpret p values. `https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-correctly-interpret-p-values`. Accessed: 2020-10-12.

[5] Statistics for everyone. Statistical power: What it is, how to calculate it. `https://www.statisticshowto.com/statistical-power/`. Accessed: 2020-10-12.

[6] Stephanie Glen. Hypothesis testing. `https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/`. Accessed: 2020-10-12.

[7] INEGI. Medio ambiente. `https://www.inegi.org.mx/temas/practicas/default.html#Tabulados`. Accessed: 2020-10-12.

[8] Douglas C Montgomery, George C Runger, and Norma F Hubele. *Engineering statistics*. John Wiley & Sons, 2009.

[9] DISTRIBUTION OR NOT. How to select appropriate statistical test? *Journal of Pharmaceutical Negative Results| October*, 1(2):61, 2010.

[10] JP Verma and Abdel-Salam G Abdel-Salam. *Testing statistical assumptions in research*. John Wiley & Sons, 2019.

[11] Angela L.E. Walmsley. What is power? https://www.statisticsteacher.org/2017/09/15/what-is-power/. Accessed: 2020-10-12.