# Statistical tests

G. Palafox

October 12, 2020

**Abstract**

Elementary facts of statistical tests are presented. Common statistical tests are performed on real data sets.

# 1 Introduction

In this report, basic facts about statistical tests are presented, along with examples of tests performed on real data sets of Mexico City subway system. The data was downloaded directly from INEGI's website [4], and includes kilometers traveled, passengers transported, and energy consumed by Mexico City's subway in 2018 and 2019. A fragment of this data is shown in Table 1. The work was coded in R version 4.0.0 [7] on a Jupyter [5] notebook[1].

# 2 Statistical hypothesis testing

The following is a brief summary of the theory regarding statistical hypothesis testing. It is not original work by the author. All presented here, and further topics, can be found in excellent books such as the one by Ross [8], Casella and Berger [2], or Navarro's online electronic book [6]. A statistical hypothesis is a statement about the nature of data, usually in terms of some statistical parameter [8]. To test the hypothesis, it must be decided whether a data sample appears to be consistent with said hypothesis For example, it may be the case where the real mean of some data is unknown, and a sample is tested to see if it is consistent with data having mean $m_0$. The hypothesis to be tested is called *null hypothesis*, denoted by $H_0$, and the alternative is called *alternative hypothesis*, denoted by $H_1$ or $H_a$. In the previous example, the test would be written as

$$H_0 : \mu = m_0, \quad H_1 : \mu \neq m_0, \tag{1}$$

where $\mu$ denotes the real mean of the population. The alternative hypothesis can also be written as $\mu > m_0$ or $\mu < m_0$. The null hypothesis can also be written as $\mu - m_0 = 0$, which would change the alternative hypothesis to $\mu - m_0 > 0$, and so on. The typical set up consists of specifying a small value $\alpha$ (called *significance level*, commonly set as 0.05) and then requiring the test to have the property that, whenever $H_0$ is true, its probability of being rejected is less than or equal to $\alpha$. Statistical tests often output a $p$-value: this is the smallest significance level at which we would allow for the rejection of the null hypothesis for the given data [6, 8].

## 2.1 Practical considerations

When performing statistical tests, certain questions arise, such as choosing significance levels, or interpreting the output of the test. We intend to address some of these questions here, showing examples of statistical tests performed in R.

[1] The notebook with the code containing our analysis, as well as this report, can be found in the Github Repository: `https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment6`

Table 1: Fragment of Mexico City subway data.

| Period | Km traveled (thousands) | Passengers Transported (millions) | Energy consumed (thousands of KWH) |
|---|---:|---:|---:|
| 2018 | | | |
| January | 3816.46 | 129.10 | 89340.53 |
| February | 3389.20 | 124.69 | 80122.84 |
| March | 3688.76 | 131.28 | 89370.53 |
| April | 3646.12 | 132.84 | 84142.34 |
| May | 3768 | 135.99 | 89340.53 |

Table 2: $p$-values from Shapiro-Wilk normality test on different data samples.

| Data | $p$-value |
|---|---:|
| Passengers | 0.245 |
| Passengers in 2018 | 0.346 |
| Passengers in 2019 | 0.927 |
| Kilometers traveled | 0.495 |
| Energy consumption | 0.112 |

### 2.1.1 Interpreting the output of a statistical test

If the $p$-value is less than the significance level $\alpha$, the null hypothesis is rejected and the alternative hypothesis is accepted. On the contrary, if $p \geq \alpha$, we do not reject $H_0$. Rejecting the null hypothesis should be taken as a strong indicative that it does not appear consistent with the observed data [8], while not rejecting $H_0$ is a weak indicative that $H_0$ is consistent with the data [8]. For example, many tests require an assumption of normality. Shapiro-Wilk normality test is a test whose null hypothesis is: *data comes from a normal distribution.* In R, it is easy to perform this test on a numeric array `passng` containing the (millions) of passengers transported by Mexico City subway each month. It is shown in Listing 1. This gives a $p$-value of 0.24. Taking a significance level $\alpha = 0.05$, the $p$-value is greater than $\alpha$, so we do not reject the hypothesis that data comes from a normal distribution, and accept it instead. The same test was performed on different data samples, whose outputs are shown in Table 2.

Listing 1: Normality test

```
shapiro.test(passng)
#         Shapiro-Wilk normality test

#data:   passng
#W = 0.94795, p-value = 0.2445
```

On the other hand, suppose it is of interest to see whether the number of passengers and the kilometers traveled had the same variance. Fisher's F test can be performed, as seen in Listing 2, and it outputs a $p$-value smaller than $2.2 \times 10^{-16}$. In this case, the null hypothesis is rejected, and the ratio of the variance is assumed to be distinct to one.

Listing 2: Fisher's F test

```
var.test(passng, km)

F test to compare two variances

#data:   passng and km
#F = 0.0015576, num df = 23, denom df = 23, p-value < 2.2e-16
#alternative hypothesis: true ratio of variances is not equal to 1
#95 percent confidence interval:
#0.0006738075 0.0036006134
#sample estimates:
#ratio of variances
#0.001557601
```

### 2.1.2 Meaning of rejecting the null hypothesis

It should be noted that the goal of a statistical test is not to determine whether $H_0$ is true or not, but to determine whether it being true is consistent with the given data [8]. Rejecting the null hypothesis should mean that our observed data is very unlikely if $H_0$ is true [8]. Consider the case of performing a One Sample $t$-Test to see whether the mean of passengers transported is `mu`= 100. Carrying this out in R (see Listing 3) gives a $p$-value smaller than $2.2 \times 10^{-16}$, which leads to rejecting the null hypothesis of the sample having mean equal to a hundred. This means that, if the true mean was 100, it would be very unlikely to observe data as the observed in `passng`.

Listing 3: One sample t-test.

```
 t.test(passng, mu = 100)
 #        One Sample t-test

 #data:   passng
 #t = 26.513, df = 23, p-value < 2.2e-16
 #alternative hypothesis: true mean is not equal to 100
 #95 percent confidence interval:
 #130.2191 135.3338
 #sample estimates:
 #mean of x
```

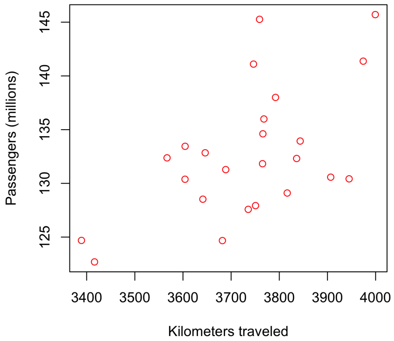

Figure 1: Plot of kilometers traveled against passengers transported.

```
#132.7765
```

A similar conclusion can be reached with Wilcoxon signed rank test, which unlike one sample $t$-test, does not assume a normal distribution for our data. The results are seen in Listing 4, which show a $p$-value of $1.192 \times 10^{-7} < 0.05$.

Listing 4: Wilcoxon signed rank test.

```
wilcox.test(passng, mu=100, conf.inf=TRUE)

#Wilcoxon signed rank exact test

#data: passng
#V = 300, p-value = 1.192e-07
#alternative hypothesis: true location is not equal to 100
```

### 2.1.3 Parametric and non-parametric tests. Assumptions and examples

Common parametric tests include Student's $t$-test, Pearson's correlation coefficient, linear regression and ANOVAs. Common assumptions for these tests is that observations are independent, and that the samples come from normal distributed populations with common variance. As seen in Listing 5, a correlation test can be used to see whether there is correlation between kilometers traveled and passengers transported. In this case, the null hypothesis is that true correlation is equal to zero. Given how a $p$-value of $0.003 < 0.05$ was obtained, the null hypothesis can be rejected, concluding there is some correlation between the variables. This can be further supported with the graphic in Figure 1, where kilometers traveled are plotted against passengers transported ($x$ and $y$ axis respectively).

Listing 5: Correlation test.

```
 cor.test(km, passng)
 #Pearson's product-moment correlation

# data: km and passng
# t = 3.3106, df = 22, p-value = 0.003181
#alternative hypothesis: true correlation is not equal to 0
#95 percent confidence interval:
#0.2257709 0.7950927
#sample estimates:
#cor
#0.5766491
```

Among the non-parametric tests one can find $\chi^2$ tests, Spearman-Kendall correlation coefficients, and Kruskal-Wallis tests. A $\chi^2$ test is performed in Listing 6 to see if two categorical variables are dependent, by means of a contingency table. The categorical variables are low and high kilometers traveled and energy consumption, defined as being below or above the median. The null hypothesis is that the variables are independent, which cannot be rejected with the obtained $p$-value (0.219). This is a weak indication of the variables being dependent.

Listing 6: Chi Squared test.

```
df2 <- data.frame(km, kwh)

df2$cat_x <- (df2$km < median(km))

df2$cat_y <- (df2$kwh < median(kwh))
```

```
chisq.test(table(df2$cat_x, df2$cat_y), correct = FALSE)

#Pearson's Chi-squared test

#data:  table(df2$cat_x, df2$cat_y)
#X-squared = 1.5105, df = 1, p-value = 0.2191
```

Other examples of statistical tests are shown next. In Listings 7 and 8, a null hypothesis of whether the number of passengers transported in 2018 has the same mean as the number of passengers transported in 2019 is tested. The difference between these two tests is that the $t$-test assumes the samples are drawn from a normal distribution, while Wilcoxon rank-sum test does not. In both tests, a $p$-value larger than 0.05 is obtained, so the null hypothesis cannot be rejected, suggesting both samples may have the same mean.

Listing 7: Two sample t-test.

```
t.test(passng18, y=passng19)

#Welch Two Sample t-test

#data:  passng18 and passng19
#t = -0.087937, df = 21.716, p-value = 0.9307
#alternative hypothesis: true difference in means is not equal to 0
#95 percent confidence interval:
#-5.468207  5.023664
#sample estimates:
#mean of x mean of y
#132.6653  132.8876
```

Listing 8: Wilcoxon rank sum test.

```
 wilcox.test(passng18, passng19, alternative = "g")

 #Wilcoxon rank sum exact test

# data:  passng18 and passng19
#W = 68, p-value = 0.6006
#alternative hypothesis: true location shift is greater than 0
```

In Listing 9, a null hypothesis of whether the number of passengers transported in 2018 and in 2019 come from the same distribution is tested. A $p$-value of 0.99 is obtained, which suggests the data indeed comes from the same distribution.

Listing 9: Kolmogorov Smirnov test.

```
 ks.test(passng18, passng19)

#Two-sample Kolmogorov-Smirnov test

#data:  passng18 and passng19
# D = 0.16667, p-value = 0.9985
# alternative hypothesis: two-sided
```

### 2.1.4 Further considerations

The **choosing of significance level** $\alpha$ should depend on how dangerous is to reject $H_0$ when it is true, with lower $\alpha$ values associated with a lower risk tolerance [9]. Navarro mentions in her book [6] that a **common mistake** is thinking the $p$-value is the probability of the null hypothesis to be true. **Statistical power** refers to the probability of rejecting the null hypothesis when it is false [3]. It gives a method of discerning between competing tests of the same hypothesis, with the test with the higher power being preferred [3].

## 3 Conclusion

Statistical hypothesis tests give an important set of tools to use when working with data, as is usually done in many scientific domains. Assumptions and conclusions must be carefully looked at, to avoid reaching to erroneous conclusions.

## A Guide

Part of a guide for choosing the correct statistical test, created by the UCLA's Institute for Digital Research and Education [1], is shown in Table 3.

Table 3: Guide for choosing statistical tests.

| Number of dependent variables | Nature of Independent Variables (IVs) | Nature of Dependent Variable(s) | Test(s) |
|---|---|---|---|
| 1 | Zero IVs (1 population) | interval and normal | one-sample t-test |
| | | ordinal or interval | one-sample median |
| | | categorical (two categories) | binomial test |
| | | categorical | Chi-square goodness-of-fit |
| 1 | One IV with two levels (independent groups) | interval and normal | two independent sample t-test |
| | | ordinal or interval | Wilcoxon-Mann Whitney test |
| | | categorical | Chi-Square test, Fisher's exact test |
| 1 | One IV with two or more levels (independent groups) | interval and normal | one-way ANOVA |
| | | ordinal or interval | Kruskal-Wallis |
| | | categorical | Chi-square test |
| 1 | One IV with two levels (dependent / matched groups) | interval and normal | paired t-test |
| | | ordinal or interval | Wilcoxon signed ranks test |
| | | categorical | McNemar |
| 1 | Two or more IVs (independent groups) | interval and normal | factorial ANOVA |
| | | ordinal or interval | ordered logistic regression |
| | | categorical (two categories) | factorial logistic regression |
| 1 | One interval IV | interval and normal | correlation, simple linear regression |
| | | ordinal or interval | non-parametric correlation |
| | | categorical | simple logistic regression |

# References

[1] *Choosing the correct statistical test in SAS, STATA, SPSS and R.* `https://stats.idre.ucla.edu/other/mult-pkg/whatstat/`.

[2] G. Casella and R. L. Berger, *Statistical inference*, Thomson Learning, 2nd ed ed., 2002.

[3] B. Everitt, *The Cambridge dictionary of statistics*, Cambridge University Press, 2nd ed ed., 2002.

[4] Instituto Nacional de Estadística y Geografía, *Transporte Urbano de Pasajeros. Principales características del sistema de transporte colectivo metro de la Ciudad de México.* `https://www.inegi.org.mx/app/tabulados/?nc=100100042`.

[5] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al., *Jupyter notebooks—a publishing format for reproducible computational workflows*, in Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing, IOS Press, 2016, p. 87.

[6] D. Navarro, *Learning statistics with R.* `https://learningstatisticswithr.com/`.

[7] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[8] S. M. Ross, *Introductory statistics*, Academic Press/Elsevier, 3rd ed., 2010.

[9] M. R. Spiegel, R. Hernández Heredero, and L. Abellanas Rapun, *Estadística*, McGraw-Hill, 1991.