Statistical Inference
Universidad Carlos III
October 2019
Rafaela Becerra & Marta Cortés

Absenteeism at Work

1. *Introduction*

   According to Gangai, K. (2014), absenteeism is a common topic that even though it has been studied by many authors, it is not fully understood by companies and needs to be treated with greater importance. The cost that involves the time of absences it is estimated to be high and can be drown if its behavior is understood and if there is a knowledge of the common causes of it, in order to work from a preventive point of view. Forbes (2013) estimated that the annual cost of absenteeism can vary from industry to industry and by occupation, so this variations can be imputed to some causes that are more common in some categories, in particularly, in the US, there is a considerable difference in the private professional sector where the losses are bigger and the time of absence is too.

   In Latin American there are fewer studies that treat this topic, since 2012 there are more studies but still there is a need of more information to comprehend it. From this group of countries, Brazil is the one that has published more investigations related (Tatamuez-Tarapues et al., 2018).

   This project aims to present an inference analysis about abseentism in one Brazilian enterprise, in order to compare some results with international index and throw some conclusions about it. The used database was compilated by Martiniano, A. et al. (2012) with records of absenteeism at work from July 2007 to July 2010 at a courier company. The total number of records is 716 which have been grouped by 36 individuals who worked in this company and presented absenteeism.

   The population objective are the potential employees that could have worked in this company, in the other hand, the sample corresponds to the data of the absences from these 36 employees from July 2007 to July 2010, this can be treated as a simple random sample.

   The present study is mainly concern in mean workload of the individuals and the proportion of which is the total time of the absenteeism, in order to relate this percentage against some standards that have been measured in other studies. Additionally, it is desired to know if there is any evidence of family obligations derived by children are causing greater absence at work.

   The complete data set presents 19 variables from which it has been selected 3 variables:

"Workload average": this variable corresponds to the average of hours worked by an employee per 30 days in each moment of the absenteeism, so it has been computed as the average for all the records of each individual. This variable can take a continuous value.

"Sons": this is a numerical value of the number of children that the employee has. This can take integer values.

"Absenteeism time in hours": this variable corresponds to the sum of the total hours that the employee missed work as a consequence of the absences. This is a continuous variable.

2. *Model selection*

It has been selected the Work Load Average per month per each employee from this data set as the continuous random variable for the next analysis.

Even though the F distribution of the variable is not known, but by the form of the cdf derived of the sample data, it is seen that F could belong to the Normal Family Distribution $\{\mathcal{N}(\mu, \sigma^2) : \mu \in R, \sigma^2 \in R^+\}$, as the plot bellow shows.

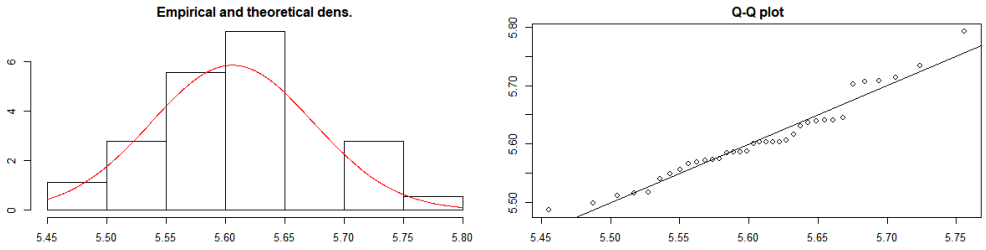

Figure 1. Histogram and Q-Q Plot Average Workload.

The histogram of the density (Figure 1) of the variable is fitted by the form of a theoretical normal distribution given by the red line. Moreover, the qq-plot which is the graphical visualization of the relation between the sample data (axis x), sort it in ascending order, and the theoretical quantiles calculated from a normal distribution (axis y), shows that the deviations of the points given by each observation from the straight line are minimal, therefore, they came from a normal distribution.

2.2. Estimate the model parameters by the method of moments or by maximum likelihood.

The method of moments can be used to find the unknown parameters of a Normal distribution through the sample moments.

Considering the population variable whose distribution depends on two unknown parameters (μ, σ^2), then, this population moments are functions of the unknown parameters. That is:

$$\alpha_{r=2} = a_{r=2}(\theta_1, \theta_2) = E\left[X^2\right]$$

The sample moments don't depend on (θ_1, θ_2) but the population moments do. That is why the method of moment is good for finding the parameters (θ_1, θ_2) such the method perfectly equate $\alpha_{r=2} = a_{r=2}$ when $n \to \infty$.

The estimator of θ produced by the method of moments is simply referred as the moment estimator of θ and is denoted as $\hat{\theta}_{MM}$.

For estimating that two parameters, we need two equations. We compute the first two moments of the r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$. The first one is:

$$\alpha_1\left(\mu, \sigma^2\right) = E[X] = \mu$$

.

The second order moment issue from the variance σ^2:

$$\alpha_2\left(\mu, \sigma^2\right) = E\left[X^2\right] = Var[X] - E[X]^2 = \sigma^2 + \mu^2$$

On the other hand, the first two sample moments are given by:

$$a_1 = \bar{X}, \quad a_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Then, the equations to solve are:

$$\begin{cases} \alpha_1(\mu, \sigma^2) & = \bar{X} \\ \alpha_2(\mu, \sigma^2) & = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{cases}$$

Calculating the sample moments by the expressions, then, we have:

$$\hat{\mu}_{\text{MM}} = \bar{X} = 272.60$$
$$\hat{\sigma}^2_{\text{MM}} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = S_n^2 = 361.96$$

In consequence, the sample random variable has a distribution $X \sim \mathcal{N}(272.60, 361.96)$.

3. *One-sample inference*
3.1. Estimate the population mean of the continuous variable by two different estimators. Mention the good and bad properties of each estimator in your application. Which one is unbiased? Which one is more efficient?

There are several different estimators for calculating the population mean, two of them are the sample mean and the Trimmed mean.

The estimator sample mean $\bar{X}$ of the parameter population mean μ is a statistic with range in the parameter space .

The sample mean $\bar{X}$ is defined by:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 272.60$$

The trimmed mean, is the calculation of the mean that conserves only a part of the data from the center, assuming the extraction of a define percentage of the data distributed in the tails.

Due to the assumption of the normal form of the distribution and its symmetricity, both estimators of the population mean are unbiased. In the other hand, the trimmed mean is less efficient that the sample mean because of the increased the variation but, presents robustness for data that contains small contamination.

There is an outlier that could be identify through the graphical analysis shown above (Figure 2). Both the box plot and the histogram indicate a singular observation to the right of the variable which deviates from the median in more than 2 standard deviations.

outliers.png outliers.png

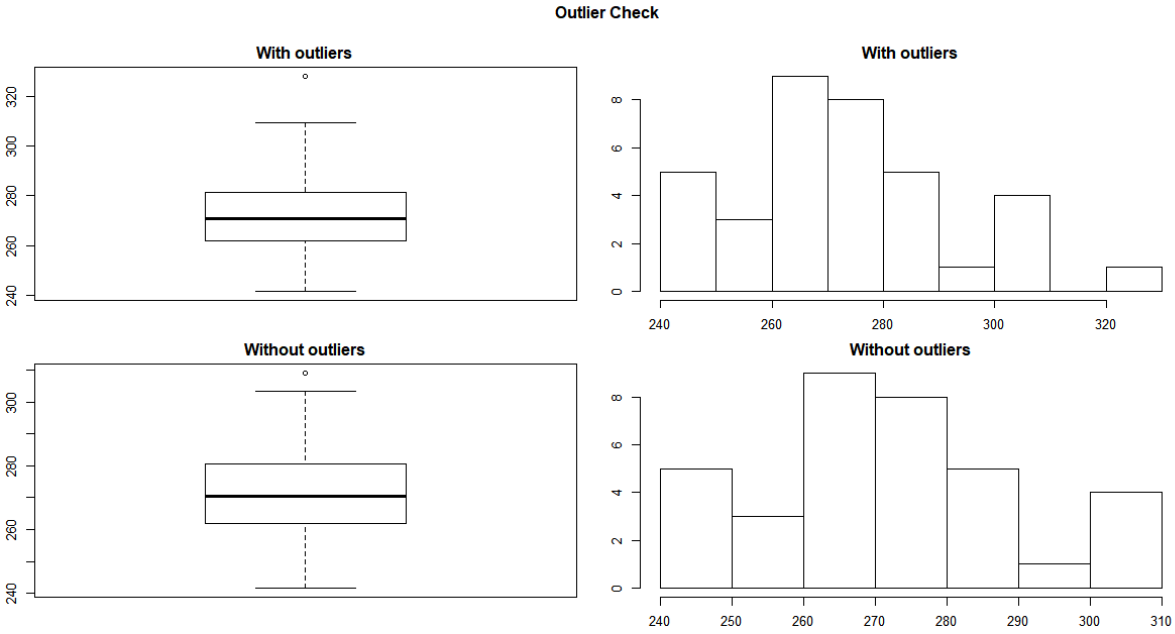

Figure 2. Outliers plot.

Since there is a contamination it could be better to use a much robust estimator that can be the trimmed mean, which its define by:

$$\hat{T}\alpha = \frac{1}{n - 2m(\alpha)} \sum_{i=m(\alpha)+1}^{n-m(\alpha)} X_{(i)}$$

Where,

$m(\alpha) = \lfloor n \cdot \alpha \rfloor$is the number of the trimmed observations at the extremes.

Applied to this case, and considering that the outlier is just one observation, an exact α should stand in 0.0278, this would allow to neat the observations in both sides by one, the final result of T_α is 271.88, which is really similar to the sample mean.

Considering the variable which average the average of work load and reviewing the raw data for this individual, it is almost improbable that there is a mistake in the

measurement so it is not consider necessary to use a estimator that removes observations, and worst yet to eliminate observations of the extremes for the database. Consequently, the sample mean is the best estimator for the population mean, contemplating that it is unbiased and more efficient.

3.2 Give measures of the quality of your estimators of the mean; concretely, estimate the coefficient of variation (CV) of your estimators.

The Coefficient of Variation of the sample mean ($\hat{X}$) is given by the next relation:

$$C\hat{V}_{(\bar{X})} = \frac{\sqrt{Var[\bar{X}]}}{\mu} * 100\%$$

$$Var[\bar{X}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{1}{n^2} * n * \sigma^2 = \frac{\sigma^2}{n}$$

By proving consistency for the quasivariance, it is known that:

$$S'^2 \xrightarrow{\text{P}} \sigma^2$$

Consequently,

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{n}{n-1} S^2 = 372.3$$

And, by Law of Large Numbers:

$$\bar{X} \xrightarrow{\text{P}} \mu$$

$$C\hat{V}_{(\bar{X})} = \frac{\sqrt{S'^2}}{\bar{X} * \sqrt{n}} * 100\% = \frac{19.29}{272.6 * \sqrt{36}} * 100\% = 1.18\%$$

In order to obtain the CV coefficient for the estimator of the trimmed mean a Bootstrapping procedure has been perform in R, by modeling 1000 independent random samples with parameters $X \sim \mathcal{N}(272.60, 361.96)$, and calculating the trimmed mean for each one considering an α of 0.0278.

The same procedure was performed for the estimator of the sample mean so that this two measures can be compared.

The results are the following:

$$C\hat{V}_{(\bar{X})} = \frac{\sqrt{Var[\bar{X}]}}{\mu} * 100\% = \frac{\sqrt{Var[\bar{X}]}}{\bar{X}} * 100\% = 1.1716\%$$

$$C\hat{V}_{(\bar{T})} = \frac{\sqrt{Var[\bar{T}]}}{\mu} * 100\% = \frac{\sqrt{Var[\bar{T}]}}{\bar{X}} * 100\% = 1.1918\%$$

In consequence, there is slightly difference of coefficient of variation of almost 0.02% positive for the trimmed mean, as expected as said before. This lost in efficiency can be seen in the following plot as a extended box plot with higher limits and more deviated observations for the trimmed mean.

Figure 3. Comparation between results from Bootstraping for Sample mean and Trimmed Mean.

3.3. Calculate a 95% confidence interval for the population mean.

Furthermore, since the population parameters (μ, σ^2) are not known, the interval for μ can be estimated using the estimation of σ^2 given by the quasi-variance S'^2 Then, we can calculate the CI of the population mean (μ) by applying the pivotal quantity method with T, which is a pivot with distribution T, t_{n-1}:

$$T = \frac{\bar{X} - \mu}{S'/\sqrt{n}}$$

Since Student's t is symmetric, it can be defined two constants as the critical values $c_1 = t_{n-1;1-\alpha/2}$ and $c_2 = t_{n-1;\alpha/2}$.

In consequence,

$$P\left(c_1 \leq T \leq c_2\right) = 1 - \alpha$$

$$1 - \alpha = P\left(-t_{n-1;\alpha/2} \leq \frac{\bar{X} - \mu}{S'/\sqrt{n}} \leq t_{n-1;\alpha/2}\right)$$

A transformation that removes μ is performed, so that:

$$1 - \alpha = P\left(\bar{X} - t_{n-1;\alpha/2} S'/\sqrt{n} \leq \mu \leq \bar{X} + t_{n-1;\alpha/2} S'/\sqrt{n}\right)$$

Finally,

$$\mathrm{CI}_{1-\alpha}(\mu) = \bar{X} \mp t_{n-1;\alpha/2} \frac{S'}{\sqrt{n}}$$

Therefore, a confidence interval for μ at $\alpha = 95$ % confidence, is given by:

$$\mathrm{CI}_{0.95}(\mu) = \bar{X} \mp t_{35,0.025}\frac{S'}{\sqrt{n}} = 272.6 \mp 2.03\frac{19.29}{\sqrt{36}} = [266.07, 279.13]$$

3.4. Select a qualitative random variable (or discrete with few possible values) and one of the categories of this variable. Estimate the proportion of units in the population that belong to this selected category.

Since the sample mean for the variable of work load is 272.60, the average of hours worked in a day for the sample is 9.08, if the number of labor days in Brazil are consider as 256, the total of hours worked in a year sum 2324.48. If you missed 10 days in each year of work that means that the average percentage of absenteeism is 3.90%. Therefore, a new discrete random variable based on the total hours of absenteeism named Y is introduced. This variable categorized the sample with 1 if the total sum is higher than the sample mean of the average of the average of work load which gives a total of absence days of 30 in the 3 years, and zero if it is less or equal to this value (272.60), it is known that:

$$Y \sim Bin(n, p)$$

Where, p es the probability of success and n the number of observations, and $X_i \sim Ber(p), i = 1, \ldots, n.$ with $Ber(p) = Bin(1, p)X_1, \ldots, X_n$ so that,

$$Y = n\bar{X} = \sum_{i=1}^{n} X_i$$

And,

$$E[X_i] = p, \quad Var[X_i] = p(1-p) < \infty$$

By the CLT,

$$\bar{X} = \frac{Y}{n} \cong \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Consequently, the sample proportion which is the estimator of the population proportion is define by:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = 0.1944$$

This means that 19.44% of the observations of the sample have presented more than 272,6 hours of absenteeism in the three years and that most of the employees of the sample have a rate of absence lower than 3.90%.

Comparing a study carried by the European Foundation for the Improvement of Living and Working Conditions in 2010 and the results throw by this analysis from this company, can be told that there is moderate level of absence, since there are

countries which presented lower ranges like 0.8% Italy and much higher like 7% Norway. It must be said that this study consists in a compilation of surveys from different types that depend in different variables designated by each country, but despite that, it is a good measure for comparing the rates and the efficiency of this company in handling absenteeism because it shows tendencies that remain consistent over the years. Specifically, this ranges of absenteeism have been seen in other studies carried by other authors in 1997, (Gründemann, 1997) with a range from 3.5% to 8%, Barmby et al. (2002) from 1.8% to 6.3%.

3.5. Estimate the variability of your estimator of proportion.

The variability of the estimator of the proportion ($\hat{p}$) is given by:

$$Var[\hat{p}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{1}{n^2} * n * p(1-p) =$$

$$= \frac{p(1-p)}{n} = \frac{0.1944 * (1 - 0.1944)}{36} = 0.0043$$

Then, the variability of the estimator with a level of 0.43% which is consider as low.

3.6. Calculate a 95% confidence interval for the true proportion. As said before,

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

By the CLT,

$$U_n = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} \mathcal{N}(0,1)$$

And considering the LLN:

$$\hat{p} \xrightarrow{\text{P}} p$$

$$\hat{p}(1-\hat{p}) \xrightarrow{\text{P}} p(1-p)$$

Can be said that:

$$W_n = \sqrt{\frac{\hat{p}(1-\hat{p})}{p(1-p)}} \xrightarrow{P} 1$$

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} = U_n/W_n \xrightarrow{d} \mathcal{N}(0,1)$$

Considering n=36, as a large sample since it is greater than 30, it can be obtained an asymptotic confidence interval at level $1-\alpha$ for the population proportion , given:

$$Z(p) = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Consequently,

$$1-\alpha = P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \ p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$\mathrm{CI}_{0.95}(p) = \hat{p} \mp z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.1944 \mp 1.96\sqrt{\frac{0.1944 * 0.8056}{36}} = [0.0594, 0.3294]$$

Which mean that in 95% of all of the intervals the proportion of employees that have an absenteeism more than 3.90% lies between 5.94% and 32.94%.

4. *Inference with more than one sample*

4.1. Select another qualitative variable (or discrete with few possible values), which divides the population in subgroups. Estimate the population mean of the continuous variable for each of those subgroups. Estimate the CV of your estimators.

The new qualitative random variable chosen in this case is Z, based on the number of sons of the employees, categorized with 0 if the employee does not have sons, with 1 if the employee has only one son and with 2, if the number of sons is two or greater.

It can be estimated the mean of the continuous random variable of these subgroups with the same moment parameter of section 2.2. ($\hat{\mu}$).

Then, it is calculated the Coefficient of Variation of the mean estimator ($\hat{\mu}$) by the next relation:

$$\hat{CV}(\bar{X}) = \frac{\sqrt{Var[\bar{X}]}}{\mu} * 100\% = \frac{\sqrt{Var[\bar{X}]}}{\bar{X}} * 100\%$$

Where the standard deviation is the squared root of the variance of the parameter estimator:

$$Var[\bar{X}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{1}{n^2} * n * Var[\bar{X}]$$

$$\hat{CV}(\bar{X}_1) = \frac{\sqrt{Var[\bar{X}_1]}}{\bar{X}_1 * \sqrt{n_1}} * 100\% = \frac{22.94}{278.26 * \sqrt{12}} * 100\% = 2.38\%$$

The results obtained for the rest of the estimators are summarised in the next table:

Table 1: Population mean estimator parameters and its coefficients of variation

Subgroup	Estimator	Sample Mean	CV
0 sons	$\hat{\mu}_0$	278.261	2.38%
1 son	$\hat{\mu}_1$	267.79	1.36%
2 or more sons	$\hat{\mu}_2$	271.18	2.01%

4.2. Estimate now the proportion of Section 3.4 for each population subgroup considered in 4.1. Estimate the MSE of your estimators.

The estimator of the proportion ($\hat{p}$) of the variable Y regarding the subgroups of the variable Z can be calculated by adding all of the elements that are 1 (if the sum of total hours of absenteeism is higher than 272.6 hours which is the workload mean of 30 days, which means that the employee has missed 10 days per year, considered as a representative value) in the subgroup, and divided it by the total elements of the subgroup.
Then, the proportion of the element with less total absenteeism hours than 272.6 is, $(1 - \hat{p})$. The summarise of the proportion for all the subgroups are detailed in the table below.

On the other hand, the Mean Squared Error (MSE) can be calculated by the next relation:

$$MSE\left(\hat{\theta}_n\right) = B^2\left(\hat{\theta}_n\right) + V\left(\hat{\theta}_n\right)$$

Where $B^2\left(\hat{\theta}_n\right)$ is the Bias of the estimator and $V\left(\hat{\theta}_n\right)$ is its variance.

As the sample proportion is a unbiased estimator of the population proportion, the Bias in the last relation is zero, then, the MSE is equal the variance.

The distribution followed in this case is a Bernoulli one, because it is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability (1-p).

The variance for the estimators is calculated by:

$$Var[\hat{p}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{1}{n^2} * n * p(1-p) = \frac{p(1-p)}{n} =$$

$$= \frac{0.167 * (1 - 0.167)}{12} = 0.01158$$

The next table shows the summary of the proportions estimators and their MSE:

Table 2: Proportion estimators and its MSE

Subgroup	Absenteeism hours group	Proportion($\hat{p}$)	MSE
0	1	0.167	0.01158
1	1	0.200	0.01600
2	1	0.124	0.00777

4.3. Select two of the population subgroups considered in 4.1. Compare the means of the continuous variable for the two subgroups by means of a confidence interval for the difference of means.

For this section, it has been selected the subgroups: 0 (none sons) and 2 (two sons or more).

To compare the sample means by the difference of means it is needed to know the confidence interval assuming a high value of confidence, for instance, 95%, regarding the difference of means.

As the variances are unknown, it is needed to assume, at least, that they are equal for applying the relation with equal variances.

In order to be able to assume that evidence, it is necessary the calculation of the confidence interval (with confidence of 95%) of the ratio of variances to verify if this condition can be concluded. The equation to calculate the Confidence Interval for ratio of variances is:

$$Cl_{(95\%)}\left(\sigma_1^2/\sigma_2^2\right) = \left[\frac{S_1'^2/S_2'^2}{\mathcal{F}_{n_1-1,n_2-1,\alpha/2}}, \frac{S_1'^2/S_2'^2}{\mathcal{F}_{n_1-1,n_2-1,1-\alpha/2}}\right] = \left[\frac{22.936^2/20.392^2}{\mathcal{F}_{11,13,0.025}}, \frac{22.936^2/20.392^2}{\mathcal{F}_{11,13,0.975}}\right]$$

$$Cl_{(95\%)}\left(\sigma_1^2/\sigma_2^2\right) = (0.3956, 4.2910)$$

As the value 1 is inside the interval, there is not enough evidence against that the variances are not the same.

After this, it can be applied the relation of difference of means with equal variances:

$$Cl_{(95\%)}\left(\mu_1 - \mu_2\right) = \bar{X}_1 - \bar{X}_2 \mp t_{n_1+n_2-2,\alpha/2} S_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} =$$

$$= 278.261 - 271.177 \mp t_{24,0.025} * 21.595 * \sqrt{\frac{1}{12} + \frac{1}{14}}$$

Where S_c is:

$$S_c^2 = \frac{\left(n_1 - 1\right) S_1'^2 + \left(n_2 - 1\right) S_2'^2}{n_1 + n_2 - 2}$$

Then,

$$Cl_{(95\%)}\left(\mu_1 - \mu_2\right) = (-10.450, 24.618)$$

The zero is inside this interval, so, there is no enough statistic evidence to assume that the true mean depends on the belonging to a subgroup.

4.4. Compare the means of the continuous variable for the two subgroups by means of hypothesis testing of equality of means.

To apply the hypothesis testing by equality of means in order to check if it can be consider that the means are different for belonging to one concrete subgroup, it is used the two-side test, where the primal hypothesis (H_0) is consider equal means ($\mu_1 = \mu 2$) and the secondary hypothesis (H_1) is consider different means ($\mu_1 \neq \mu_2$). Then, we can reject H_0 if $|T| > t_{n_1+n_2-2;\alpha/2}$. Where:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_C\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{278.261 - 271.180}{21.595 * \sqrt{\frac{1}{12} + \frac{1}{14}}} = 0.834$$

On the other hand, $t_{n_1+n_2-2;\alpha/2} = 2.064$.

As $\{|T| < t_{n_1+n_2-2}\}$ we can not reject H_0, therefore, with a 95% of probability, we do not have enough evidence for assuming that one of the subgroups should has higher or lower mean of workload.

4.5. Compare the proportion of Section 3.4 for the two subgroups by means of a confidence interval for the difference of proportions.

Even though that the two groups have less than 30 observations, it is assume that by CLT, the intervals can be calculated with the data as if it fit a normal distribution. The relation to calculate the Confidence Interval (with a 95% of confidence) by difference of proportions is:

$$Cl_{95\%}\left(p_A - p_B\right) = \left(\hat{p}_A - \hat{p}_B\right) \mp Z_{\alpha/2}\sqrt{\frac{\hat{p}_A\left(1-\hat{p}_A\right)}{n_A} + \frac{\hat{p}_B\left(1-\hat{p}_B\right)}{n_B}} =$$

$$= (0.1667 - 0.1243) \mp Z_{0.025}\sqrt{\frac{0.1667(1-0.1667)}{12} + \frac{0.1243(1-0.1243)}{14}}$$

$$Cl_{95\%}\left(p_A - p_B\right) = (-0.2302, 0.315)$$

At the same time than section 4.3. the zero is inside this interval, so, we do not have enough statistic evidence to assume that the true proportion depends on the belonging to a subgroup.

4.6. Compare the proportion of Section 3.4 for the two subgroups by means of hypothesis testing of equality of proportions.

A test statistic under H_0 has an asymptotic normal distribution, then it can be follow the same relation than the section 4.4., but for the estimator parameters of proportions:

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}} = \frac{0.167 - 0.124}{\sqrt{\frac{0.1667(1-0.1667)}{12} + \frac{0.1243(1-0.1243)}{14}}} = 0.309$$

The asymptotic critical region for difference of proportion is given by:

$$C_c = \{|Z| > z_{\alpha/2}\}, \quad where \quad z_{\alpha/2} = 1.96$$

As $\{|Z| < z_{\alpha/2}\}$ we can not reject H_0, therefore, with a 95% of probability, we do not have enough evidence for assuming that one of the subgroups should has higher or lower proportion of more than 272.6 hours of absenteeism.

5. *Conclusions*

   Summarize the most important conclusions of your analyses. Mention limitations and possible extensions of this project.

   As we have already discussed, the random continuous variable selected (workload of the employees) follows a normal distribution, this has a lot of sense as this variable is already the mean of the workload of the employees during all the days in these 3 years of study, and it helps to bring the data near to the mean.

   The population mean estimated by the first moment has a value of 272.6 per 30 days of work, which is higher that the normal legal hours of work in Brazil.

   On the other hand, our variance, 361.96, give us the value of the standard deviation of our sample ($\sqrt{361.96}$), then, the variation or dispersion of our set of values is 19.025, which is a bit high considering that there is people who worked 19 hours more in 30 days than the mean.

   There is evidence that the absenteeism is moderate in this company, considering that only 19.44% exceeds a mean rate of 3.90%, and that there are other international studies that have estimated a range of 0.08% to 7.7%.

   The estimation of the means by the sample mean and the trimmed mean has a difference between them of 0.2% which means both are good approximation of the population mean.

   The CV of the sample mean estimator is 0.02% less than the Trimmed mean. We can assume then that a better estimator is the sample mean, considering that the sample is small and that all the records are presumed measured without errors for that variable.

   We have selected the variable of number of sons in order to study how this reality influence in the workload and in the absenteeism cause, a prior, it seems to be related to. In fact, according to the study of the Eurofond (2010, p.9) the Hungarian survey showed that the highest rates observed in absence were presented among women aged between 28 and 37, which is consistent with high pressures from childcare responsibilities in this group. On the contrary, in the study, the results show there is not enough evidence for us to say that there is a difference between the absenteeism in employees and those with children.

   Therefore, it is recommendable that the data retrieved as samples must be categorized by gender. This missing distinction may be the reason why the results are not congruent with previous studies that showed that there are, in fact, differences between absenteeism among women with children, and men.

   Simple inferential analysis like the performed, can approximate something to the true and allows us to draw much appropriate conclusions about a certain case. In this case, it is commonly known that costs related to work absenteeism can be lowered with an understanding of it. This is the reason why it is considered convenient that continuous samples should be taken in each company to perform analysis that allow them to approximate the behavior of their workers and focus their actions to prevent absenteeism by creating new measures.

Bibliography

Eurofond (2010). Absence from work. Retrieved on October 19, from: *https* : *//www.eurofound.europa.eu/sites/default/files/effiles/docs/ewco/tn*0911039*s/ tn*0911039*s.pdf*

Gangai, K. (2014). ABSENTEEISM AT WORKPLACE: WHAT ARE THE FACTORS INFLUENCING TO IT? Retrieved on October 19, 2019 from: *https* : *//www.academia.edu/*24968588/$ABSENTEEISM_AT_WORKPLACE_WHAT_ARE_THE_FACTO$

Martiniano, A., Ferreira, R. P., Sassi, R. J., Affonso, C. (2012). Application of a neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp. 1-4). IEEE. Retrieved September 16th, 2019 from: *https* : *//archive.ics.uci.edu/ml/datasets/Absenteeism + at + work*

NA (2013). The Causes And Costs Of Absenteeism In The Workplace. FORBES. Retrieved on October 20, 2019 from: *https* : *//www.forbes.com/sites/investopedia/*2013/07/10/*the − causes − and − costs − of − absenteeism − in − the − workplace/*2*be*3*af*033*eb*6

Tatamuez-Tarap ues RA, Domínguez AM, Matabanchoy-Tulcán SM. (2018) Revisión sistemática: Factores asociados al ausentismo laboral en países de América Latina. Univ. Salud. 2019;21(1):100-112. DOI: *http* : *//dx.doi.org/*10.22267/*rus*.192101.143