Homework Assignment 14: Applied Probabilistic Models

Central Limit Theorem

5273

1 Introduction

Central Limit Theorem (CLT) is an approximation one can use when the population to study is quite big (it would take a long time to gather data about each individual) and identifying its characteristics is desired. In statistical terms, one collects samples from a population and by combining the information from the samples, conclusions can be drawn about the population. In a nutshell, the approach of the CLT could be:

- Draw multiple samples sufficient in size.
- Calculate the individual mean of these samples.
- Calculate the mean of these sample means, and this value will give the approximate mean of the studied variable.
- Additionally, the histogram of the sample means will resemble a bell curve or normal distribution.

2 Applications

In this section, it is seen how the CLT can be used in real-world problems and how to apply it. It helps to solve problems where the population is not normal.

2.1 Manufacturing

Let assume a pipe manufacturing organization produces a different kind of pipes and the monthly data of the wall thickness of certain types of pipes are given. The organization wants to analyze the data by constructing confidence intervals to implement some strategies in the future and the challenge is that the distribution of the data is not normal. Data is simulated in R software [3]. Figure 1 shows a histogram of all the observations of the data. This graph denotes with a vertical red line the population mean, which is 12.802 and one can see that the population is not normal.

Therefore, to apply the CLT, it is needed to draw sufficient samples of different sizes and compute their means (known as sample means). Figure 2 shows this experiment, where sufficient samples are drawn, increasing its sizes. Means are calculated and are plotted in R. It is known that the minimum sample size taken should be 30, but even with samples of size 10, (see Figure 2a) nice bell-shaped curves are evidenced. The sampling distribution should approach a normal distribution as the sample sizes increase. Therefore, one can consider the sampling distributions as normal and the pipe manufacturing organization can use these distributions for further analysis.

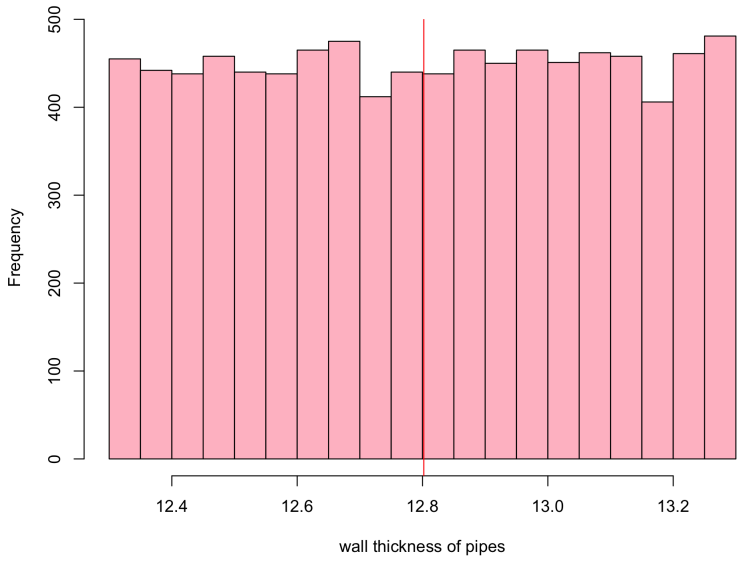


Figure 1: Histogram for Wall Thickness

2.2 Baseball

Another application can be found by Anderson and Bay [1], where CLT found an interesting application as hypothesis testing. This author answers the following question using the CLT: *Is there such thing as home-field advantage in Major League Baseball?*

Concerning this problem, the null hypothesis and the alternative hypothesis are:

$$H_0\text{: There is no home-field advantage,}$$
$$H_1\text{: There is a home-field advantage.}$$

To test this notion, in a Major League Baseball (MLB) season 2431 games are played. Let us take the 2013 MLB season, were 1308 of those games were won at home, therefore the observed value $\hat{p} = 0.5381$. To test the hypothesis, our null hypothesis will be 0.5, that is 50% of the MLB games are won at home and the other half on the road, hence, there is no home-field advantage. One method is using a confidence interval. For testing,

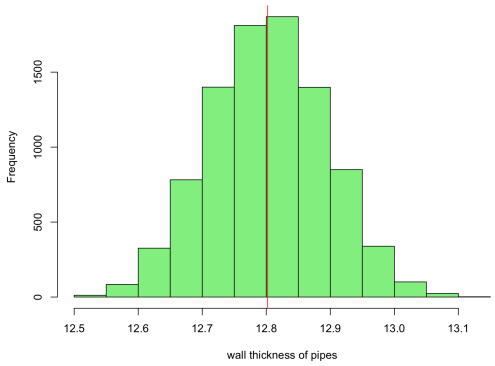
$$H_0\text{: } p = 0.50,$$
$$H_1\text{: } p > 0.50.$$

at the 0.05 level of significance a right-sided 95% confidence interval for p can be constructed. If our test statistic of $p = 0.5$ is in the interval, then we fail to reject H_0 at the 0.05 level of significance. If $p = 0.5$ is not in the interval, we reject H_0. The right-sided $100(1-\alpha)\%$ confidence interval for p for a large sample is given by

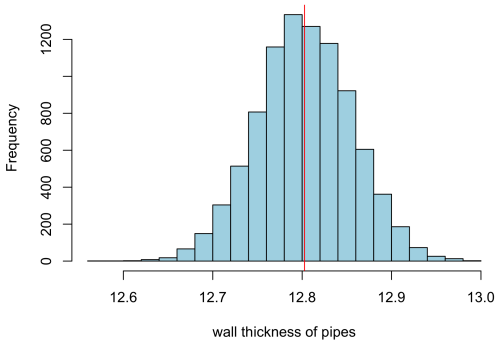
$$\hat{p} - z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p \leq 1,$$

where α is the level of significance.

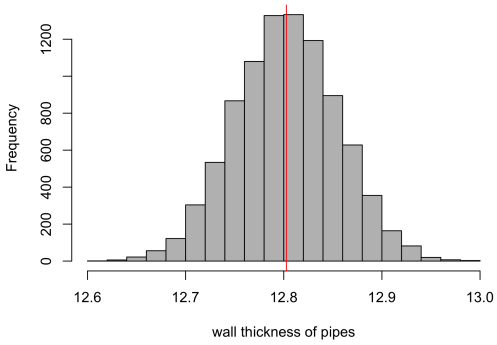
Since $n = 2431, \hat{p} = 0.5381, \alpha = 0.05$, and $z_{0.05} = 1.645$, obtained from the standard normal table [2], which is used to find the probability that a statistic is observed below, above, or between values on



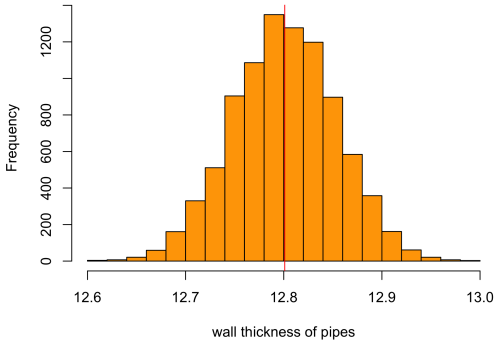
(a) Histogram of samples of size 10



(b) Histogram of samples of size 30



(c) Histogram of samples of size 50



(d) Histogram of samples of size 500

Figure 2: Histograms of the sample means

the standard normal distribution, and by extension, any normal distribution. Therefore a right-sided 95% confidence inerval for p is:

$$\begin{aligned} 0.5381 - 1.645\sqrt{\frac{(0.5381)(1-0.5381)}{2431}} < p \leq 1, \\ 0.5381 - 1.645(0.0101114) < p \leq 1, \\ 0.5215 < p \leq 1. \end{aligned}$$

Since $0.5 \notin (0.5215, 1]$, reject $H_0 : p = 0.5$ in favor of $H_1 : p > 0.5$ at the 0.05 level of significance, that is, there is enough evidence to support that there is a home-field advantage, and the home team wins more than 50% of the games played at home, hence there is such thing as home-field advantage in MLB.

Also, if one wants to see if there is a difference between the American League and the National League a similar analysis can be performed. For these separate leagues, a 99% confidence interval is used, and for the National League it is obtained that $0.5 \notin (0.5117, 1]$, therefore it is concluded that the National League has a home-field advantage. On the other hand, in the American League $0.50 \in (0.4978, 1]$, fail to reject $H_0 : p = 0.50$. That is, one does not have enough evidence to support that there is a home-field advantage in the American League based on the 2013 season.

References

[1] Nicole Anderson and Thunder Bay. Central limit theorem and its applications to baseball. 2014.

[2] F. James Rohlf, Robert R. Sokal, et al. *Statistical tables*. Macmillan, 1995.

[3] The R Foundation. The R Project for Statistical Computing. `https://www.r-project.org/`, 2020.