Average gestation duration of women relative to smoking

1 Abstract

Using the babies' dataset of R, we check whether the average pregnancy duration of the women smokers is 40 weeks. Also, we check whether there is some difference in the average pregnancy duration for both smokers and non-smokers mothers. The final results indicate that the 40 weeks are not the average pregnancy duration and there is a difference between smokers and non-smokers mothers in their average pregnancy duration.

2 Introduction

Below we will present the data that we will process. We will show that they do not meet the assumption of normality. We will assume that the average pregnancy duration of women who smoke is 40 weeks and we will show by various statistical tests that our data do not confirm this hypothesis. However, considering that the average duration of gestation for women who smoke and for women who do not smoke is the same, we will show that there is a deviation between the average values.

3 Data presentation

The dataset contains 1236 recorded pregnancies observations. The variables that we interested in and which we will analyze are gestation and smoke, which indicate the duration of pregnancy in days and the mother's status on smoking, respectively. The results we extract in the process of mothers' variables during their pregnancy in smoking are: in the 10 cases we do not know the woman's behavior towards smoking, the 484 cases are for women who smoke and the 742 cases are for women who do not smoke. In some of these cases, however, the recorded pregnancy duration far exceeds what the laws of nature define. Thus, pregnancy intervals over 300 days are considered outliers and excluded from the statistical processing. And so the results are as follows: the 456 cases are for women smokers and the 684 cases are for non-smoking women.

4 Statistical data analysis

4a. Women who smoke

We will start analyzing the data by studying the normality or non-normality of our distribution in the mothers who smoked.

Table 4.1: Descriptive statistical measures for the mothers who smoked.

	Value	Standard Error
Mean	276.212719 0.6134	
Median	278	
Variance	171.574431 16.014971	
Q1	270	
Q3	285	
Asymmetry	-1.085816	0.114708
Kurtosis	1.97294	0.229416
Shapiro-Wilk, p-τιμή	0	

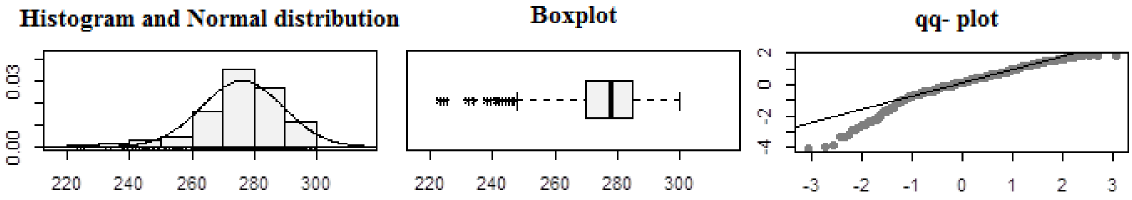

Figure 4.1: Graphical plots for the mothers who smoked.

Table 4.1 and Figures 4.1 indicate that the distribution of the data is not normal. In particular, this is shown in detail as follows: the distribution curve appears to have asymmetry on the left with the Pearson skewness coefficient at -1.08581, that is, below 0 and standard asymmetry indicator at 9.4659, that is, above the threshold 2, so the conclusion is that asymmetry does affect the normality of the data. In addition, the Q-Q plot indicates that the sample points do not fall on the straight line. The positive kurtosis indicator at 1.97294 means that we have a leptokurtic distribution and the standard kurtosis indicator at 8.5998, that is, above the threshold 2, we conclude that kurtosis is the deviation problem from normality.

In the project, we make the following assumptions: if we consider that the average pregnancy duration of mothers who smoke is 40 weeks, namely, we have the hypothesis H_0: μ=280(40 weeks * 7 days), versus of the hypothesis H_1: μ ≠280, and we apply the function t.test() to see if the assumption μ =280 is correct. The implementation of the function indicates that 95% for the average pregnancy duration is from 275.0073 to 277.4182, so the 280 value is not included in this interval, and we reject the H_0 hypothesis. Furthermore, the hypothesis H_0: μ = 280 is rejected at significance level α = 0.05, because p-value = 2.2 * 10-16 <0.05. As we have already seen, because our distribution is not normal we apply a Box-Cox transformation for normality and stabilization of the variance. With the function boxcox() of the library MASS we can find the appropriate λ for the normality. So for λ = 7.4 we transform the data and apply the stats.d() function and we get the results shown in Table 4.2:

Table 4.2: Descriptive statistical measures in a normalized sample for the mothers who smoked.

	Value	Standard Error
Mean	164767159192960608	2367694174255544
Median	164702340239965856	
Variance	2.55632492047846e+33	1.70254815813486e+32
Asymmetry	0.000416	0.114708
Kurtosis	0.0227	0.229416
Shapiro-Wilk, p-τιμή	0.022883	

From Table 4.2 we can see that for the transformed data we reject the assumption of normality again but the asymmetry has improved. We calculate the corresponding of the 280 for the transformed data and do the hypothesis H_0: μ=1.736752*10^{17}, versus H_1: $\mu \neq$1. 736752*10^{17} with the function t.test().The results indicate the p-value = 0.0001904, which is a much lower value than the case of non-transformed data. Even with normalization of the sample, however, the average duration of pregnancy is still not 40 weeks as the 95% is from 1.601142*10^{17} to 1.694201*10^{17}, so this interval does not include the μ value. Because our sample is not normally distributed, we can simply assume it symmetrical and apply another (non-parametric) test for the average value of gestational days. The non-parametric Wilcoxon test results that p-value=7.389 *10^{-7} and the hypothesis H_0: μ = 280, it is not true. Therefore, all tests confirm that the average pregnancy duration of women who smoke is not 280 days but less.

4b. Women who not smoke

To make a comparison between the means, we first examine the distribution of the sample of non-smoking women.

Table 4.3: Descriptive statistical measures for the mothers who non-smoked.

	Value	Standard Error
Mean	277.913743 0.552436	
Median	280	
Variance	208.746575 31.418738	
Q1	273	
Q3	287	
Asymmetry	-2.440384	0.093659
Kurtosis	13.495117	0.187317
Shapiro-Wilk, p-τιμή	0	

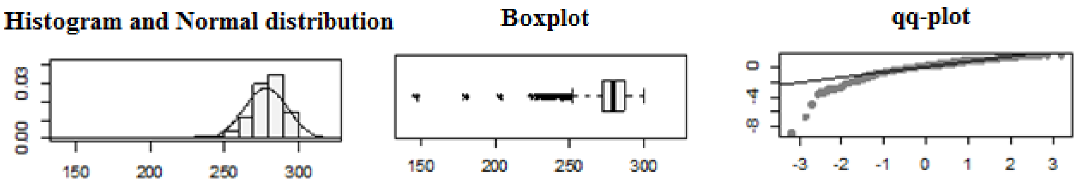

Figure 4.2: Graphical plots for the mothers who non-smoked.

From Table 4.3 and Figure 4.2 we can see that the sample is not normally distributed. The additional feature in this sample is that we have pregnancies of less than 150 days. In this case, since the probability of survival of an infant less than 150 days is impossible, we consider these values as outliers and remove them from the sample. Figure 4.3 presents the histograms of our samples for smoking and non-smoking mothers.

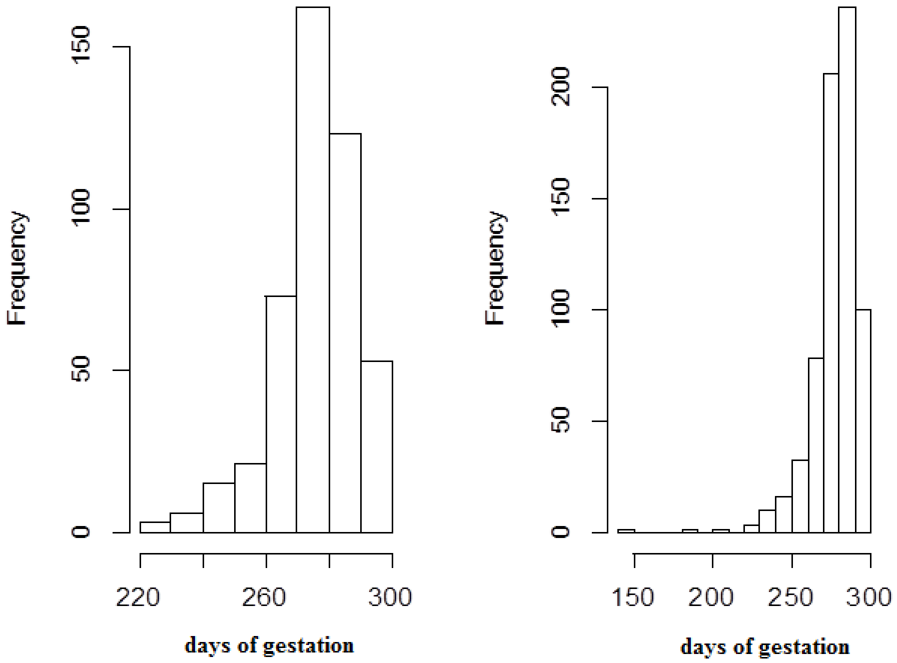

Figure 4.3: Histograms for smoking and non-smoking mothers.

Initially, we will check whether the variances of the two samples are equal. The hypothesis is H_0: $\frac{\sigma_1^2}{\sigma_2^2}$ =1 versus H_1: $\frac{\sigma_1^2}{\sigma_2^2}$ $\neq$1. By implementing the function var.test() indicates that the hypothesis H_0 is not valid, but because the p-value = 0.409 we cannot reject equality. Therefore we will check the equality of the means by considering the variances of the samples as unknown but equal. Thus, using the function t.test() we check whether the hypothesis H_0: μ_1 =μ_2 versus H_1: $\mu_1 \neq \mu_2$ is valid. The results indicate that 95% of the means difference is from 3.479545 to -0.302923. The zero value does not belong to this interval. Also, p-value= 0.01965, so based on the above, hypothesis H_0 is rejected. With probability 0.05 to be wrong, the true means of the samples are different. Because the samples do not follow a normal distribution, we perform a non-parametric Wilcoxon test. Using the function ks.test() we obtain p-value=0.1984, so we cannot reject the equality of the means, but the implementation of wilcox.test rejects the equality of means and because p-value=0.001544, this difference is considered statistically significant.

5 Conclusions

The statistical analysis of the babies dataset regarding the pregnancy duration and the women's attitude towards smoking shows that women who smoke have an average pregnancy duration of less than 40 weeks, as well as less than women who do not smoke.