Regional Analysis of Hospital Length of Stay Report

ANOVA

I. Introduction

Hospitals are intimidating for all. Whether you are in for a quick check-up or having life-altering surgery, hospitals will always be terrifying for all. Given this, choosing a hospital is essential for a smooth visit. In this report, we aim to investigate how hospitals in different geographical regions will affect the length of a patient's stay at the hospital.

The dataset we will be using is the ‘SENIC’ dataset which records the data of 100 patients' length of stay at a hospital and the region of the hospital. The explanatory variable in this study will be the region in which the hospital was located, there are 4 different ones: NC (North Central), NE (North East), W (West), and S (South). The response variable will be the length of stay the patient had at the hospital, which was measured in days.

To analyze the effect of hospital region on length of stay, we will perform diagnostics on the data for visualization and conduct a single-factor ANOVA test while ensuring all necessary assumptions are met and address any outliers as needed. Additionally, if we were to reject the null hypothesis, a power test will be conducted to see the possibility of us making a type I error. Furthermore, we will examine pairwise confidence intervals to identify which regions differ the most in terms of patient stay duration.

II. Summary of the Data

Before starting to investigate the data, we visualize it to better understand its trends. This involves examining the means, standard deviations, and sample sizes while also using visual representations to assess the data’s distribution and identify potential outliers.

	NC	NE	S	W	Total
Means	9.6834	11.1943	9.2030	8.2186	9.6371
Std. Dev	1.1929	2.9367	1.2695	1.0280	1.9261
Sample Size	32	21	33	14	100

Table 1: Table of sample means, standard deviations, and sample sizes based on the length of stay in each region

Table 1 shows the means, standard deviations, and sample sizes of each region to provide a clearer understanding of the data set. The table reveals that there is a skewed distribution of the samples as some regions have more data collected than others. Notably, the hospitals in the West have a much smaller sample size, which may create an imbalance in the data and potentially influence the results and interpretation of the analysis. This is important to keep in mind for future data analysis.

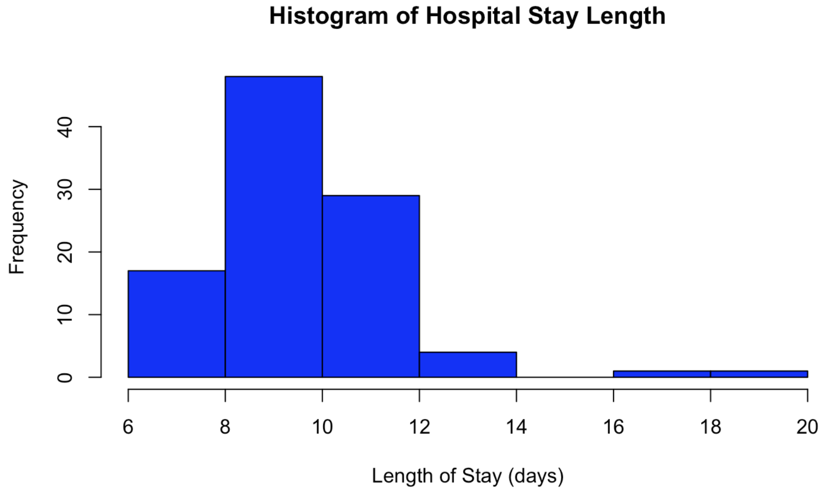

Figure 1. Histogram of Length of Hospital Stay for all data points

Next, to better visualize the trends in the data, a histogram was created to explore the distribution of the length of hospital stays, ignoring how the region may affect this. As shown in Figure 1, the data has a left skew which indicates there was a larger number of patients who had shorter hospital stays. It can be seen that most patients had a hospital length of stay between 8 to 10 days. In contrast, fewer data points are in the upper end of the distribution. This may suggest the presence of potential outliers that may influence the data analysis.

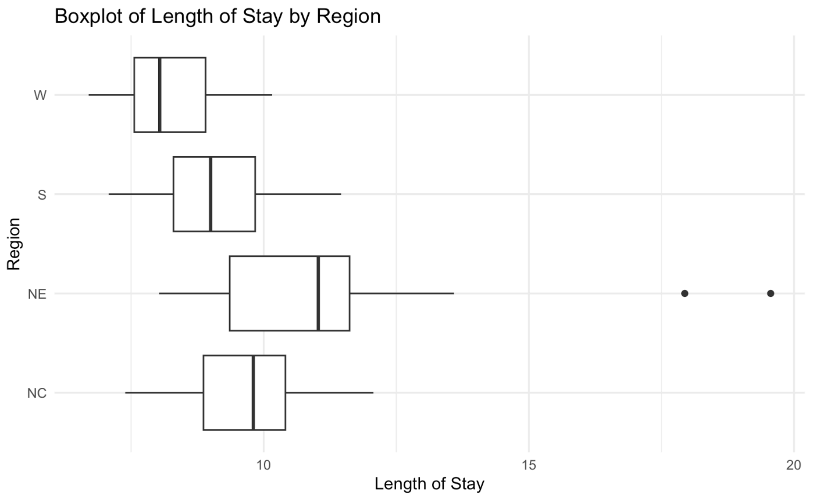

Figure 2. Boxplot of Length of Stay based on the Region of the Hospital

To further investigate this, boxplots were created to look at the distribution of data based on the different regions and identify any potential outliers in the data. In Figure 2, it can be seen that the hospital in the West had the lowest median which indicates that patients in that area had shorter hospital stays compared to other regions. In contrast, the Northeast region had the longest median stay as well as longest stays in general. Additionally, it can be seen that there are two outliers observed in the Northeast region which will be examined and potentially removed to ensure a cleaner dataset as it might violate assumptions for testing later on.

III. Diagnostics

When looking at test statistics and confidence intervals, we need to verify that all our assumptions can be made. This includes that all Y_{ij} were randomly sampled, and $\varepsilon_{ij} \sim N(0, \sigma_{\epsilon}^2)$ for all i, j. We will assume that all Y_{ij} were randomly sampled and all groups are independent. Thus, we must check that our data is normally distributed and has a constant variance.

III-1. Removing Outliers

To make sure that our data is normally distributed and has constant variance, we must check to see if any outliers can cause non-normality and non-constant variance. Thus, we first performed outlier detection and used the recommended cutoff value in R. Based on this criteria, we identified and removed the following outliers: Observation 67 (Length of Stay = 19.56 days, Northeast) and Observation 90 (Length of Stay = 17.94 days, Northeast). These outliers were also visible in Figure 2 and their removal helps prevent extreme values from influencing our model.

III-2. Assessing Normality

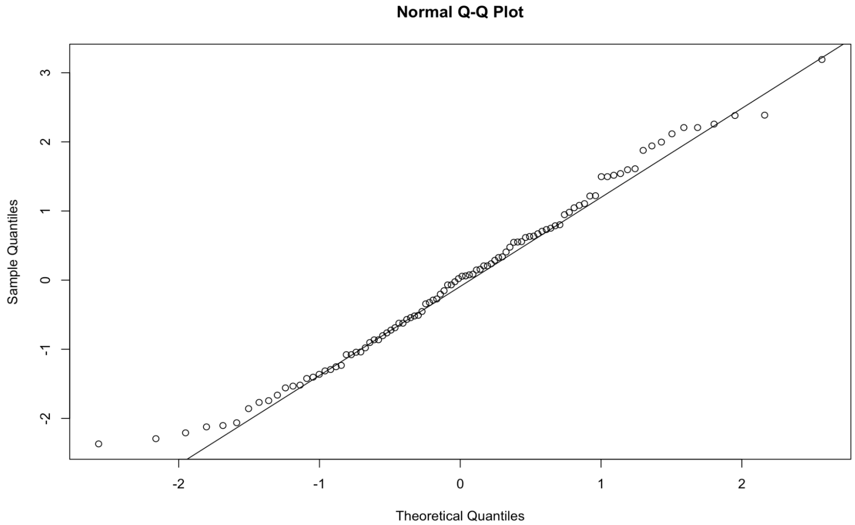

Figure 3. Q-Q Plot for assessing normality

Next, we assessed key model assumptions which include normality and constant variance. A Q-Q plot was created in order to examine normality. A Q-Q plot calculates the actual centered percentiles of the data and compares them to what they should be if the data was normal. In Figure 3, we can see that the data seems approximately normal with most of the data points on the 45-degree reference line. Deviation from this line would suggest skewness which there is a bit in our plot. Since identifying normality can be subjective when looking at a plot, we must conduct the Shapiro-Wilk test to assess whether the data follows a normal distribution in a more systematic and objective manner. The hypothesis for this test is as follows:

$$H_0 : The\ error\ is\ normally\ distributed$$

$$H_A : The\ error\ is\ not\ normally\ distributed$$

Computing this in R, we found that at W = 0.98494, p = 0.3278 which is larger than α = 0.05 which means that we fail to reject the null hypothesis, therefore, we have sufficient evidence to conclude that the values e_{ij} are normal.

III-3. Assessing Constant Variance

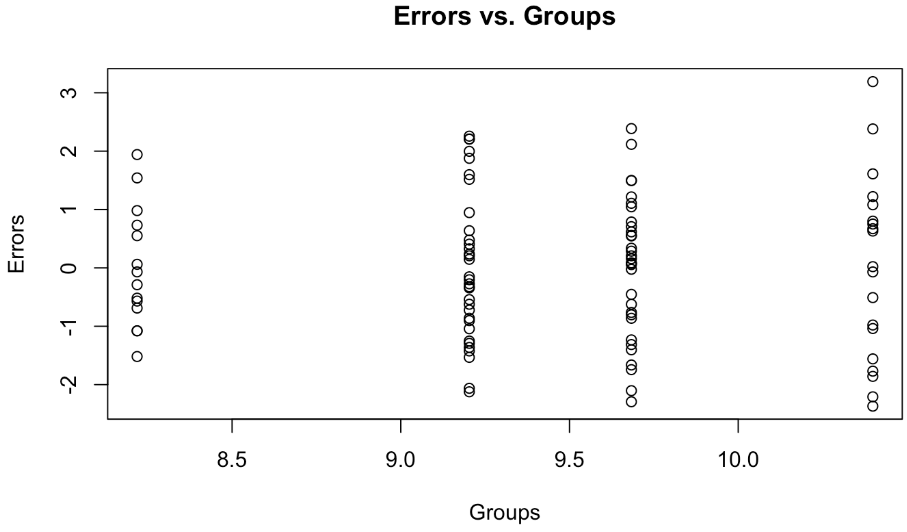

Figure 4. Residual Plot for assessing constant variance

Now that we know our data is normally distributed, we need to assess if the data has constant variance. We first use a plot of e_{ij} $vs.$ $\widehat{E}\{Y_{ij}\}$ which is plotting e_{ij} vs. the fitted values of $\widehat{E}\{Y_{ij}\} = \overline{Y}_{i.}$. This means that when we are plotting the residuals by group, we are assessing if there is an equal spread for all the groups. Looking at Figure 4, it can be seen that the plot may be approximately unequal since the data does not seem to be scattered in a manner that is constant across all groups. This suggests that there is a violation of homoscedasticity which means that the variances may not be equal across groups. In order to confirm this, the Brown-Forsythe test is conducted since the analysis of a plot is subjective. The hypothesis of the tests is as follows:

$$H_0 : \sigma_{NC}^2 = \sigma_{NE}^2 = \sigma_S^2 = \sigma_W^2$$

$$H_A : \ At\ least\ one\ \sigma_i^2\ is\ not\ equal$$

Computing this in R, we get the *p*-value of 0.2668604 which is larger than α = 0.05, we would fail to reject the null which means we can conclude that the group variances are equal.

In summary, by addressing the outliers and removing them, we have minimized their impact on the model. Our diagnostics helped confirm that our data is normally distributed and has constant variance which are assumptions that we need to make sure before conducting the ANOVA test and creating confidence intervals.

IV. Analysis & Interpretation

IV-1. Model Fit

Now that we know all assumptions have been met for a single factor ANOVA test, we will be able to conduct it. In this analysis, we will use the group mean model to determine whether there are significant differences between patients' average length of stay across different geological regions. The model is as follows:

$$Y_{ij} = \mu_i + \epsilon_{ij}, i = 1,\ldots,4, \epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$$

The main goal of this study is to determine whether the average length of stay for patients varies significantly among the four regions: North Central (NC), North East (NE), West (W), and South (S). Therefore, we want to determine if at least one region's average length of stay differs from the others. In this case, μ_i represents the mean length of stay for patients, which will be estimated using our ANOVA sample group averages for each region *i*, where *i* = NC, NE, W, S. In this model, ϵ_{ij} represents the individual error for any given jth value in the ith group.

IV-2. Hypothesis Test

We conduct a hypothesis test to see if there are significant differences between the average length of stay for patients and their corresponding regions. Therefore, our null hypothesis claims that the average length of stay for regions NC, NE, W, and S, are all the same, and there is no difference. Alternatively, our alternative hypothesis would be that at least one of the average length of stay for the regions would not be the same, signifying that there would be a difference.

$$H_0 : \mu_{NC} = \mu_{NE} = \mu_W = \mu_S$$

$$H_A : At\ least\ one\ \mu_i\ differs\ from\ the\ others.$$

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Region	3	42.096	14.0319	7.5428	4.487e-05 ***
Residuals	94	154.399	1.6425		

Table 2. ANOVA Table

To determine whether to reject or fail to reject the null hypothesis, we conducted an ANOVA test. To compute the p-value, we first calculated the F-value using the formula $F_s = \frac{MSA}{MSE}$, where MSA is the

Mean Square among groups and MSE is the Mean Squared Error. Conducting this calculation in R, we yield an F-value of 7.5428 as seen in Table 2. Thus our p-value is 4.487e-05 (look at Table 2 for reference), which is smaller than $\alpha = 0.01$, meaning that we can reject the null hypothesis and have sufficient evidence to conclude that at least one of the average length of stay for patients differs from the other average length of stay across different four regions. If in reality there was no difference in the true average length of stay for patients across NC, NE, W, and S regions, we would observe our sample data or more extreme with a probability less than .001.

IV-3. Power Test

Since we rejected the null hypothesis, there is a possibility that we made a Type I error, which occurs when we have incorrectly rejected the true null hypothesis. In order to assess the reliability of our conclusion, we will perform a power calculation and determine the probability that we correctly rejected the null hypothesis when the alternative is true. This probability is given by $1 - \beta$, where β is the probability of making a type II error (failing to reject the null hypothesis when it is false). A high power would indicate a stronger test and reduce the likelihood of missing a true effect. The power does not directly measure the probability of making a Type I error, rather it helps evaluate whether our study had a sufficient sample size and effect size to detect a true relationship, thereby increasing the confidence of our findings.

The equation we will be using to calculate the power is as follows:

$$\phi = \frac{1}{\sigma_e}\sqrt{\frac{\sum_{i=1}^{n} \mu_i(\mu_i - \mu_.)^2}{a}}$$

Doing so in R, we get the power value is 0.9924617, which means that our test has a very high probability (99.25%) of correctly rejecting the null when the alternative hypothesis is true. This indicates that our study is well-powered, which reduces the likelihood of making a Type II error. With a power value close to 1, this suggests that our sample size and effect size were sufficient enough to detect a true effect, thus giving us strong confidence in our results.

Furthermore, to ensure that our sample size was large enough to reliably detect the true effect, we used R to calculate the minimum required sample size per group. In doing so, we determined that $k = a - 1 = 4 - 1 = 3$, where *k* represents the number of groups minus one. The effect size was 0.5113851, which was computed using our new dataset with the same F-statistic equation as before. With a significance level of $\alpha = 0.01$ and a power of 0.9 (rounding down the previously computed power value), our calculations indicated that the minimum required sample size per group should be at least 24 (specifically, 23.78621). This suggests that the sample sizes for the West and Northeast regions may be too small and might need to be increased to meet this minimum requirement. Overall, our power calculations confirm that our study is well-powered which ensures a low probability of Type II errors while highlighting the potential areas where our sample size may need to be adjusted for more reliable results.

IV-4. Confidence Intervals

We will also construct 95% confidence intervals for pairwise comparisons across different regions. The formula below helps us to compare the four region groups in pairs:

$$(\overline{Y}_{i.} - \overline{Y}_{i'.}) \ \pm t_{1-\alpha/2,\, n_T - a}\sqrt{\frac{MSE}{n_i}},\ i \ = NC,\ NE,\ W,\ S$$

Since we are interested in all pairwise confidence intervals for differences in means, so we use $1 - \alpha/2$ as a percentage of the T distribution, and $n_T - a$ as the degree of freedom to calculate the 95% confidence interval. The degree of freedom is 94 with the test statistic equal to 3.588.

Using R, we calculated the pairwise confidence intervals for all of the contrasts $\mu_i's$ above. The results are as follows:

Pairwise Comparison	Pairwise Confidence Interval
$\mu_{NC} - \mu_{NE}$	$(-1.45250834, 0.02148861)$
$\mu_{NC} - \mu_{S}$	$(-0.1509251, 1.1117395)$
$\mu_{NC} - \mu_{W}$	$(0.6494614, 2.2802708)$
$\mu_{NE} - \mu_{S}$	$(0.4630905, 1.9287436)$
$\mu_{NE} - \mu_{W}$	$(1.284084, 3.076668)$
$\mu_{S} - \mu_{W}$	$(0.1728230, 1.7960947)$

IV-4-1. Confidence Interval for $\mu_{NC} - \mu_{NE}$

We are 95% confident that there is no significant difference between the average lengths of stay for the North Central and North East regions since the confidence interval contains zero, ranging from -1.4525 to 0.0215 days. The true difference in mean lengths of stay in NC and NE regions lies between -1.4525 days and 0.0215 days. This means that the length of stay in the NC region could be less than, equal to, or more than the length of stay in the NE region. Furthermore, since 0 is included in the interval, it indicates that there is insufficient evidence to conclude a difference in the average length of stay between these regions.

IV-4-2. Confidence Interval for $\mu_{NC} - \mu_{S}$

We are 95% confident that the true average length of stay in the hospital between the North Central and South regions is between -0.1509 and 1.1117 days. Similar to the previous confidence interval, since 0 is included in the interval, it indicates that there is insufficient evidence to conclude a difference in the average length of stay between these regions.

IV-4-3. Confidence Interval for $\mu_{NC} - \mu_{W}$

We are 95% confident that the true average length of stay for the hospital in the North Central region is greater than the West region by between 0.6495 and 2.2803 days. This indicates that there is likely a difference in the means of the North Central and West region.

IV-4-4. Confidence Interval for $\mu_{NE} - \mu_{S}$

We are 95% confident that the true average length of stay for the hospital in the Northeast region is greater than the South region by between 0.4631 and 1.9287 days. This indicates that there is likely a difference in the means of the Northeast and South region.

IV-4-5. Confidence Interval for $\mu_{NE} - \mu_{W}$

We are 95% confident that the true average length of stay for the hospital in the Northeast region is greater than the West region by between 1.2841 and 3.0767 days. This indicates that there is likely a difference in the means of the Northeast and West region,

IV-4-6. Confidence Interval for $\mu_{S} - \mu_{W}$

We are 95% confident that the true average length of stay for the hospital in the South region is greater than the West region by between 0.1728 and 1.7961 days. This indicates that there is likely a difference in the means of the South and West region

Based on our confidence intervals we can see there is insufficient evidence of a significant difference between the North Central regions in comparison to the Northeast and South regions. However, there is sufficient evidence of significant differences between other regions. Most notable is that hospitals in the

Northeast regions have a longer average stay in comparison to the West and South regions. These findings further support the idea that hospital stays vary depending on geographical region.

V. Conclusion

In conclusion, through our data analysis, we can conclude at least one region's average length of stay differs from the others. Through looking at the confidence intervals, we further concluded that the hospitals in the Northeast region had significantly longer stays on average than those in the West and South; however, the North Central region has insufficient evidence to show a difference when compared to the Northeast and South regions. However, since we rejected the null, we could possibly have made a type I error which led us to do a power calculation which indicated that the sample sizes of West and North East might be too small and may have affected our results. However, it indicated that our test is well-powered and is unlikely we committed a Type II error.

This report suggests that there is a regional difference in hospital stays, this means that the healthcare system may benefit from looking at how hospitals in the North Central, West, and South may differ from the hospital in the North East so that their patients have shorter hospital stays. Longer hospital stays may be indicative of longer recovery times therefore it may be helpful to allocate resources to reduce longer times. However, it is important to acknowledge the limitations of this study. Notably, the small sample sizes for the West and North East regions may affect the reliability of the findings. Additionally, this study only considers two variables, and further research is necessary to investigate potential confounding factors that could influence the observed results. Overall, this analysis provides a basic insight into the differences in hospital length of stay based on regional differences.