Final Project

Ariana Schindler
PHP1510: Principles of Biostatistics and Data Analysis
Dr. Shira Dunsinger

(Dated: December 14, 2021)

Research Question: Test whether one of the two insurance companies has been more successful at minimizing the length of hospitalization (and therefore the respective healthcare cost).

I. METHODS

A. VARIABLE SELECTION

To test if one of the insurance companies is more successful at minimizing length of stay, the variables `company` and `los` will be the guiding variables of my analysis. I will also explore the variables `type` and `hosp.id`, as insurance company could possibly impact whether a patient seeks care at a private or public hospital, and following from that, the associated hospital IDs. I also want to look at `type` and `hosp.id` in relation to `los` to see if either variable significantly affects how long a patient is admitted for. The last variable I want to prioritize is `complic` because length of stay could possibly be extended due to a complication, and complications could possibly arise from insufficient insurance coverage.

The very first thing I did was run basic summary statistics (see FIG. 1) and checked the variable types for the given dataset. I transformed `los` into a new categorical variable to easily group the observations by length of stay, and I did the same transformation for `hosp.id`.

The final part of my initial variable exploration was visualizing the distribution of the length of stay for all observations (see FIG. 2) and the count of observations for each `los` by `company` (see FIG. 3).

B. EXPLORING RELATIONSHIPS WITH MAIN VARIABLE FOCUS

I first looked into the proportions of each company for all hospitals by creating a new dataframe of the proportion table for the variable `company`. After creating the dataframe, I visualized the proportion table for `company` with a barplot (see FIG. 4). We were given the information that there were 392 patients insured with Insurer A and 395 patients insured by Insurance B. After confirming that the proportions between the two companies were similar for all hospitals, I wanted to visualize the proportions of each company for all hospitals for each day in `los`. I created another dataframe grouped by the variables `company` and `los` and then visualized this with a grouped barplot for each day in length of stay for the proportions of company (see FIG. 5). From this figure, it looked like the distribution of company for each day was not the same, so this prompted the first hypothesis test I ran to check whether the means of the length of stay were the same for each company.

The variable `company` is a bivariate categorical variable, so to look at the relationship between `company` and the numerical variable `los`, I used a two-sample t-test on the original dataframe with the following hypotheses:

$$H_0 = \textit{The means of the length of stay for each insurance company are the same.}$$
$$H_1 = \textit{The means of the length of stay for each insurance company are different.}$$

The results of this test were significant with a p-value of $p = 2.032e-09$ at the significance level $\alpha = 0.05$. The mean length of stay for Insurer A is 2.30 days and the mean length of stay for Insurer B is 2.89 days. Therefore, we can reject the null hypothesis that the means of the length of stay of each insurance company are the same. Next I used a boxplot to visualize the difference in means of the length of stay between the two companies (see FIG. 6).

Since the proportion of companies varied when grouped by length of stay, the next patient-characteristic I wanted to test was the significance of difference between the expected and observed frequencies of complications and insurance company. I used a chi-squared test with the following hypotheses for testing since both of these variables are categorical:

$$H_0 = \textit{The expected frequencies for company and complications are the same as observed frequencies.}$$
$$H_1 = \textit{The expected frequencies for company and complications are different than observed frequencies.}$$

This test resulted in a p-value of $p = 0.8058$, which is not significant at the significance level $\alpha = 0.05$, so we fail to reject the null hypothesis.

With an insignificant relationship between the variables `company` and `complic`, I used a two-sample t-test to check if a patient with a complication affected my other main variable of focus, length of stay. The hypotheses used for this test are:

$$H_0 = \textit{The means of the length of stay for whether a complication occurred or not are the same.}$$
$$H_1 = \textit{The means of the length of stay for whether a complication occurred or not are different.}$$

This test also failed to reject the null hypothesis that the mean length of stay was the same for each group in

`complic` at the significance level $\alpha = 0.05$ with a p-value of $p = 0.2202$.

Next, I focused on `company` and the hospital characteristic `type`. In the summary statistics for `type`, 655 patients sought care at a private hospital, and 132 sought care at a public hospital. A 1-sample proportions test on `type` resulted in a p-value of $p = 2.2e-16$ and a confidence interval of $(0.8039, 0.8573)$, so we can reject the null hypothesis that the two types of hospitals have the same probability of 0.5 at our significance level $\alpha = 0.05$. Since `company` and `type` are both categorical variables, I ran a chi-squared test with the hypothesis:

H_0 = *The expected frequencies for company and type are the same as observed frequencies.*
H_1 = *The expected frequencies for company and type are different than observed frequencies.*

This test resulted in a p-value of $p = 7.52e-15$, which is significant at the significance level $\alpha = 0.05$. Therefore, we can reject the null hypothesis that the expected frequencies of each company for type of hospital are the same as the observed frequencies. At this point, my main variables of focus became `company`, `los`, and `type`.

C. NEW DATAFRAMES FOR HOSPITAL TYPE

Since we rejected the null hypothesis that the mean length of stay is the same for each insurance company and we also rejected the hypothesis that the expected frequencies for `company` and `type` are the same as observed frequencies, next I explored the alternative hypotheses. I began by creating two new dataframes: One for private hospitals and one for public. After creating the dataframes, I performed the same analysis on the two new dataframes as I did on the original dataframe including company proportions, company proportions by length of stay, and testing if the means for length of stay were the same for each company.

II. RESULTS

1. HOSPITAL TYPE: PRIVATE

The private dataframe consisted of 24 unique hospital IDs and 655 patients were observed, so I first used a 1-sample proportions test to see if the proportions of `company` were the same for the private dataframe. This test resulted in a p-value of $p = 0.0023$, so we can reject the null hypothesis that the probabilities of each insurance company are the same for private hospitals at the significance level $\alpha = 0.05$. I used a barplot to visualize this result (see FIG. 7). Next I created a new dataframe of private hospitals grouped by `company` and `los` and then visualized this with a grouped barplot for each day in length of stay for the proportions of each company for each day (see FIG. 8). From the figure, it appeared that patients insured with Insurer A had a greater probability of having a shorter duration of length of stay, so next I wanted to test if the means of the length of stay for each company were the same. I tested the hypothesis:

H_0 = *The means of the length of stay for each insurance company are the same for private hospitals.*
H_1 = *The means of the length of stay for each insurance company are different for private hospitals.*

A two-sample t-test resulted in a p-value of $p = 1.661e-09$ which is less than our level of significance, $\alpha = 0.05$. The mean length of stay for a patient with Insurer A is 2.30 days and the mean length of stay for a patient with Insurer B is 2.96 days for private hospitals. So for private hospitals, we can reject the null hypothesis that the means of `los` are the same for each insurer in `company`. I used a boxplot to visualize this difference in means of the length of stay between the two companies in private hospitals (see FIG. 9). From the boxplot, we can also see a few outliers in duration for both companies.

2. HOSPITAL TYPE: PUBLIC

The public dataframe consisted of 5 hospitals and 132 patients were observed. Again, I used a 1-sample proportions test to see if the proportions of `company` were the same for the public dataframe and this resulted in a p-value of $p = 1.787e-12$. We can reject the null hypothesis that the probabilities of each insurance company are the same for public hospitals at the significance level $\alpha = 0.05$. To visualize this significance, I used a barplot (see FIG. 10). The proportions of company for public hospitals were drastic, with the proportion of Insurer A at 19% and the proportion of Insurer B at 81%. I then created a new dataframe grouped by `company` and `los` and used a grouped barplot to visualize the proportions of company for each day in length of stay (see FIG. 11). This grouped barplot reflected the drastic difference in proportions of company over all days, as 5 out of the 9 days in length of stay only had patients with Insurer B.

I ran a two-sample t-test to test if the means of length of stay were the same for each insurance type in public hospitals. I tested the same hypothesis for public hospitals:

H_0 = *The means of the length of stay for each insurance company are the same for public hospitals.*
H_1 = *The means of the length of stay for each insurance company are different for public hospitals.*

This test resulted in a p-value of $p = 0.1192$ which is greater than our level of significance, $\alpha = 0.05$. The mean length of stay for a patient with Insurer A is 2.28 days and the mean length of stay for a patient with Insurer B is 2.71 days. For public hospitals in this study, we fail to reject the null hypothesis that the means of `los` are the same for each insurer in `company`. A boxplot of the insurance companies by length of stay shows

the similarity in means for each company (see FIG. 12). However, this boxplot has more outliers than both the original dataframe boxplot and the boxplot created from the dataframe containing only private hospitals.

3. FILTERING THE TYPE DATAFRAMES

Recalling the grouped barplots for the original, private, and public dataframes, some days in `los` had only one of the insurance companies for all observations. I searched through `hosp.id` to check if there were specific hospitals that had all of their patients with the same insurer. I found that 4 private hospitals had observations with only one company: Insurer A, and 2 public hospitals had observations with only one company: Insurer B. I then filtered these hospitals out of the private, public, and original dataframes to see if the effect of company on length of stay was significant or not when I tested the hypothesis with only hospitals which took both insurers.

Filtering all single-company hospitals out of the main dataframe and running a two-sample t-test again with the same hypothesis for `company` and `los`, we get a p-value of $p = 7.404e-08$ with a mean length of stay for a patient with Insurer A at 2.33 days and a mean length of stay for a patient with Insurer B at 2.91 days. As with the hypothesis test without filtering single-company hospitals, we still reject the null hypothesis that the means of length of stay are the same for each insurer in `company` for our filtered dataset with both private and public hospitals. I visualized the difference in means of the length of stay for each company in the filtered dataset with a boxplot (see FIG. 13). Comparing this boxplot of only hospitals containing observations with both insurers with FIG. 6 (the boxplot containing all observations for length of stay and company), we can see that the filtered boxplot actually has more outliers than the unfiltered boxplot.

For the private and public hospital dataframes, I ran the same two-sample t-tests again with the same hypothesis for `company` and `los` on the new filtered dataframes. The filtered private hospital dataframe returned a p-value of $p = 5.248e-08$, so we still reject the null hypothesis that the means of length of stay for each insurance company are the same at the significance level $\alpha = 0.05$. I used a boxplot to visualize the mean for the length of stay per company for the filtered private hospitals (see FIG. 14) and compared it with the boxplot for the unfiltered private dataframe. As with the filtered dataframe with both private and public hospitals, the filtered private hospital boxplot also contained more outliers than the unfiltered boxplot.

For the filtered public hospital dataframe, running the same two-sample t-test with the same hypothesis returned a p-value of $p = 0.2024$, so as with the unfiltered public hospital dataframe, we still fail to reject the null hypothesis that the means of length of stay for each insurance company are the same at the significance level $\alpha = 0.05$. I created yet another boxplot to visualize the difference in means for the length of stay for each company for the filteredd public hospitals (see FIG. 15). Unlike the filtered original and filtered private dataframes, the filtered public dataframe contained less outliers than the unfiltered public dataframe.

Filtering out the hospitals which only take one insurance did not result in a different conclusion regarding the significance of the relationship between `los` and `company` for either `type` or for the original dataframe containing both types of hospitals. Since the proportions of hospital type are significantly different and the results of analysis on `company` and `los` vary depending on `type`, it did not make sense to perform further analysis on specific hospital IDs at this time.

III. DISCUSSION

While the dataset consisted of roughly the same number of patients with Insurer A and Insurer B, the significance of the effect of `company` on `los` in this dataset depends on whether the patient sought medical treatment at a private or publicly-owned hospital. While there are > 30 observations in the public hospital dataframe, the observations are coming from only 5 different hospitals, and when the single-company hospitals are filtered out, the public dataframe consists of observations from only 3 hospitals. There are also > 30 observations in the private hospital dataframe, but the observations are only from 24 different hospitals, and when the single-company hospitals are filtered out, the observations are only coming from 20 different hospitals.

To better answer the research question, I would want observations from more hospitals, ideally more than 30 private hospitals and more than 30 public hospitals. Since `type` influences significance of `company` and length of stay in this analysis, the amount of observations being evenly distributed between private and public hospitals would also increase the accuracy of the results. More data would also illuminate whether it is company, hospital type, or specific hospital ID that has the greatest effect on the length of stay.

If more data were provided as described above, the patient-characteristics that were not explored in this analysis could provide insight into whatever results that data would produce. After analyzing whether company has a significant effect on length of stay, possible questions to further explore could include whether a company is more successful at minimizing length of stay for a particular age, race, or sex. Because of the limitations of the data, I did not explore these patient-characteristic variables in this analysis because even if a test resulted in a significant finding, more data would still be required to substantiate the result.

It is inconclusive to suggest that one insurance company does a better job at minimizing length of stay and healthcare cost because more data is needed to either validate or challenge the results of this analysis.

APPENDIX A: TABLES AND GRAPHS

```
> # basic summary statistics
> summary(hospital)
       id              age            sex                    race          los            hosp.id
 Min.   :  1.0   Min.   : 2.000   boy :503   African-American:275   Min.   :1.000   Min.   : 1.00
 1st Qu.:197.5   1st Qu.: 3.000   girl:284   Asian            : 49   1st Qu.:2.000   1st Qu.: 7.00
 Median :395.0   Median : 5.000              Hispanic         :207   Median :2.000   Median :22.00
 Mean   :394.7   Mean   : 6.713              other            : 38   Mean   :2.596   Mean   :25.19
 3rd Qu.:591.5   3rd Qu.:10.000              white            :218   3rd Qu.:3.000   3rd Qu.:45.00
 Max.   :789.0   Max.   :17.000                                     Max.   :9.000   Max.   :58.00
      company         beds              beds.grp        type     complic
 Insurer A:392   Min.   :  2.0   1-99 beds   : 14   private:655   No :772
 Insurer B:395   1st Qu.:179.0   100-249 beds:353   public :132   Yes: 15
                 Median :260.0   250-400 beds:306
                 Mean   :281.2   401-650 beds:114
                 3rd Qu.:359.5
                 Max.   :650.0
```

FIG. 1. Summary Statistics for Original Dataset

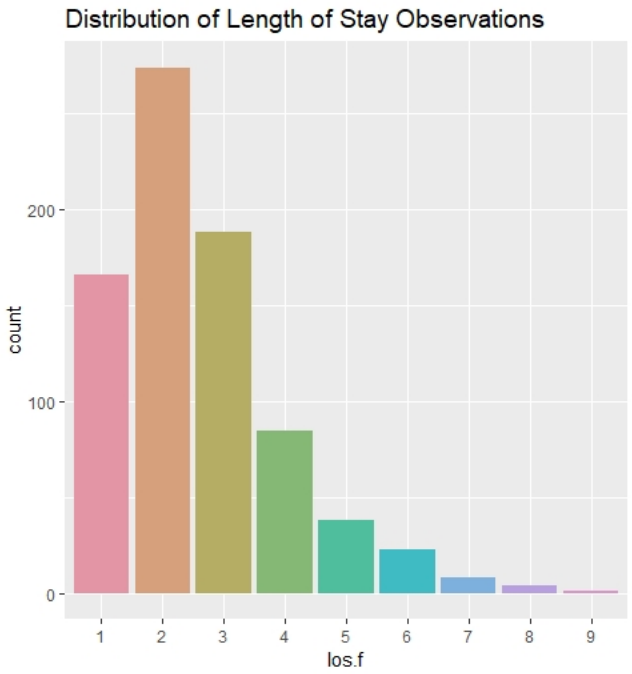

FIG. 2. Distribution of Length of Stay

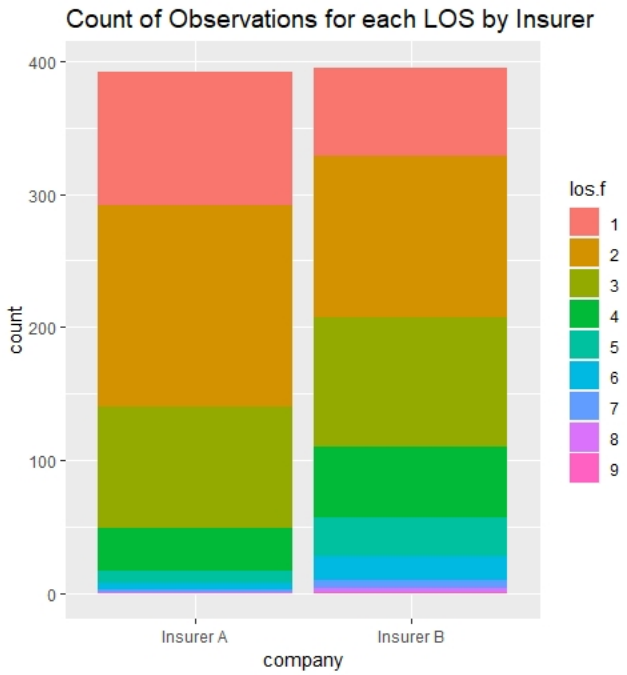

FIG. 3. Count of Observations for each LOS by Insurer

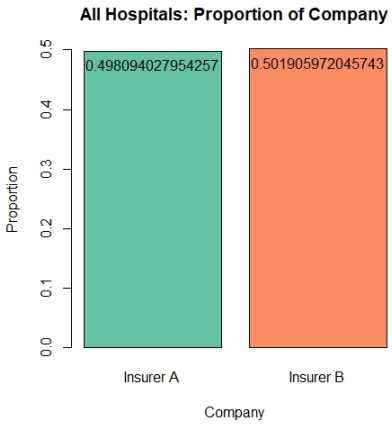

FIG. 4. Proportion of Company for All Hospitals

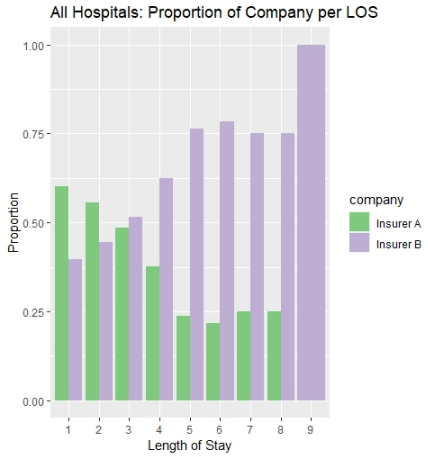

FIG. 5. Proportion of Company per LOS

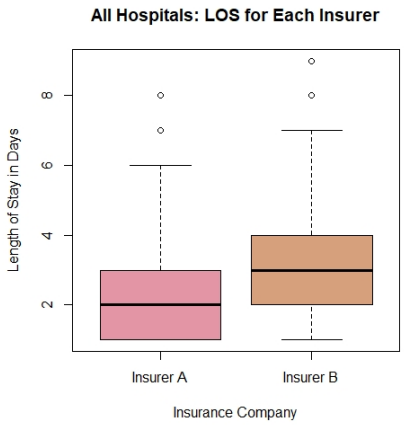

FIG. 6. All Hospitals: LOS for each Insurer

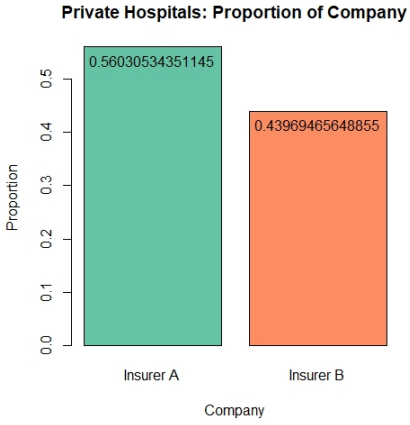

FIG. 7. Proportion of Company for Private Hospitals

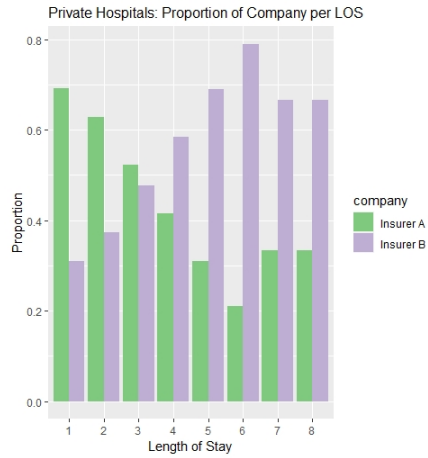

FIG. 8. Private Hospitals: Proportion of Company per LOS

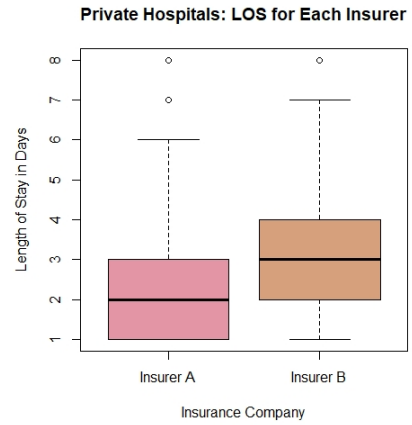

FIG. 9. Private Hospitals: LOS for each Insurer

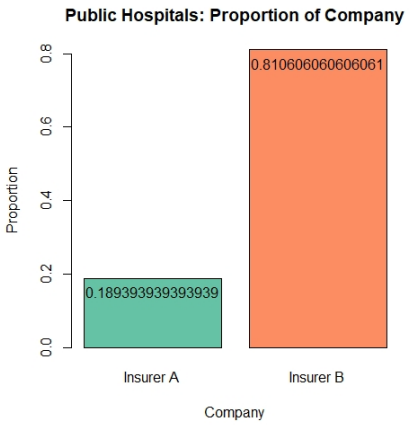

FIG. 10. Proportion of Company for Public Hospitals

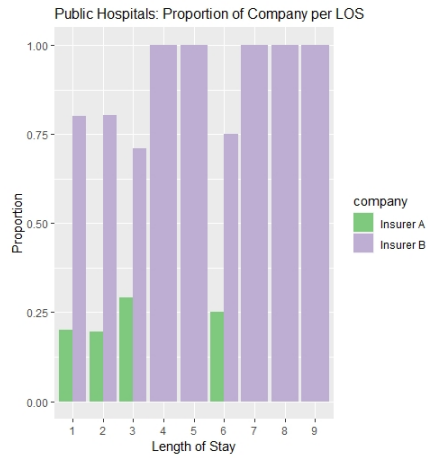

FIG. 11. Public Hospitals: Proportion of Company per LOS

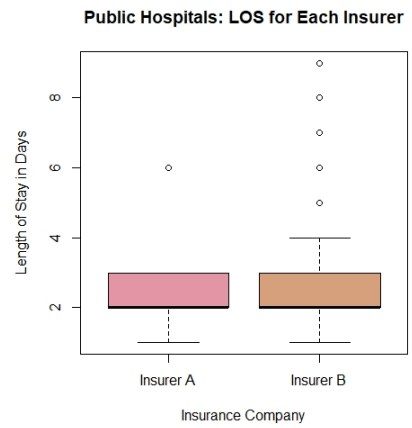

FIG. 12. Public Hospitals: LOS for each Insurer

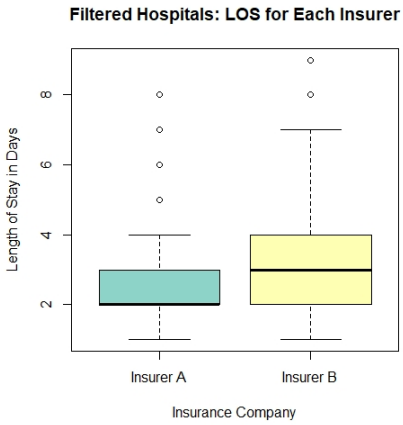

FIG. 13. Filtered Hospitals: LOS for each Insurer

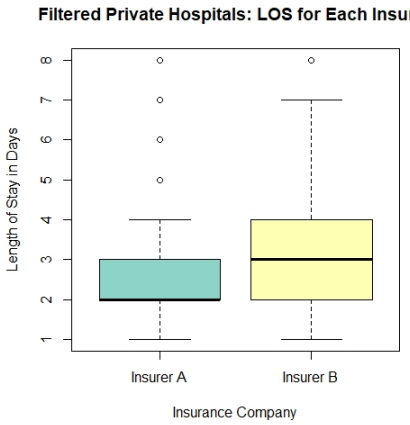

FIG. 14. Filtered Private Hospitals: LOS for each Insurer

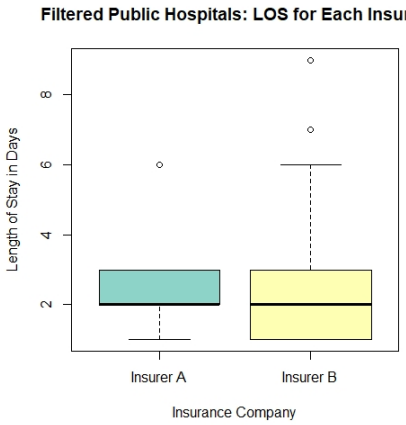

FIG. 15. Filtered Public Hospitals: LOS for each Insurer