Capstone Project:

A Comparative Analysis of Inpatient Length of Stay and Race in San Diego and San Mateo Counties

Brian Chang

Loma Linda University

Table of Contents

Section I. Abstract…………………………………………………………………………. 3

Section II. Deliverables……………………………………………………………………4

Section III. Literature Search…………………………………………………………..5

Section IV. Hypothesis…………………………………………………………………….6

Section V & VI. Database Selection & Data Collection………………………..6

Section VII. Validity Testing………………………………………………………...…..7

Section VIII. Conclusion….......................................................................................14

**Section I. Abstract**

The purpose of this project is to analyze two datasets containing frequency tables of inpatient hospital discharges in San Diego and San Mateo counties. The objective is to find any correlation between the race of patients and their length of stay in the two counties. Furthermore, the findings between race and length of stay will be compared between the two counties to see if there is a geographical phenomenon. More specifically, hypothesis testing will be conducted in determining these correlations. Hypothesis testing methods will include; chi-square tests, two sample t-tests, and descriptive statistics.

**Section II. Deliverables**

Data Analytics will be the main focus of this Capstone Project. The key objective for this data analysis project is to prove proficiency in data mining, data design, and hypothesis testing methods. Several courses in Loma Linda University’s Masters of Health Informatics program have contributed as preparation for this analysis project. Thorough research will be done on inpatient hospital discharges records, particularly patient’s Length of Stay (LOS) and race. Geographically, this study will be conducted on datasets from Northern California and Southern California.

The first county of analysis will be San Mateo in Northern California. The assumption is that San Mateo’s geographic location in the north would contribute to varying differences in the demographic of race, lifestyle, and various other factors that would make for an interesting study. An additional analysis will be performed on San Diego County. The geographical differences between the two counties were essential in their selection to be in this study. San Diego County is the southern most county in California, whereas San Mateo is one of the most northern. Several hypothesis testing methods will be used including a chi-square tests and a Z-test. The use of graphs and charts will also be utilized to illustrate any findings from the analysis of the two counties.

**Section III. Literature Search**

1. <http://www.ncbi.nlm.nih.gov/pubmed/25111874>
	1. This article theorizes the health patterns of immigrants who have resided in the United States for a long period of time. Their researchers believe that the longer an immigrant lives in the United States, their health declines due to poorer health behaviors and risks epitomized by the American lifestyle.
	2. The theory that immigrants may have declining health issues due to being Americanized for a longer period of time can directly contribute to this project’s analysis of LOS in inpatient hospitals. One county may have a higher immigrant percentage in which can contribute to an increased inpatient population in hospitals throughout that county.
2. <http://www.ncbi.nlm.nih.gov/pubmed/25801362>
	1. The purpose of this article is to discover if after-hours rehabilitation at inpatient facilities decreases length of stay and improve the overall functional outcome and performance of patients. The study was performed on adults and they concluded that an increase in rehabilitation may improve activities of daily living, but does not seem to affect LOS.
	2. This article disapproves an increase of rehabilitation increases physical activity but does not affect LOS. This article can be used as a precursor to several other theories that may affect LOS.
3. <http://www.ncbi.nlm.nih.gov/pubmed/23932602>
	1. The purpose of the article is to examine the risk associated between race and gender when pertaining to pediatric surgery. The study concluded that race and gender significantly affect postoperative results after surgery. They found that black patients are at a higher risk when it comes to postoperative mortality.
	2. Findings from this resource support the alternative hypothesis that race effects patient LOS. If race and gender have an affect on mortality rates as this study has found, the conclusion may be that the alternative hypothesis can be proven to be true.
4. <http://www.ncbi.nlm.nih.gov/pubmed/24973128>
	1. This article outlines race-based disparities in operative morbidity and mortality for several different procedures. This study in particular focuses on Pancreatoduodenectomy (PD) and race-based differences in the LoS among patients who are undergoing this procedure. The study found that mortality and LoS after PD were greater among black and Hispanic patients.
	2. This article can be used to support the alternative hypothesis that race does affect LOS.
5. <http://www.ncbi.nlm.nih.gov/pubmed/25187970>
	1. This resource studies the association between socioeconomic status, race, and ethnicity with outcomes of patients who undergo thyroid surgery. The study found that there were significant socioeconomic and racial disparities. Low-volume centers and surgeons had a much longer LoS and risk of complication than high others. They also found that there were inequalities existing in access to these high volume hospitals and surgeons.
	2. This article also shows evidence that the alternative hypothesis in which there is a relationship between patient race and LOS. Socioeconomic status can also correlate to high volume enters and race, in which can directly affect LOS.
6. <http://www.ncbi.nlm.nih.gov/pubmed/24313090>
	1. The focus of this article is to analyze the association between language, serious adverse events, and LOS among hospitalized children. The study found that hospitalized children that from Spanish speaking families had a significantly longer LOS in association with an adverse event. This study proves that patient care and quality can be improved if communication barriers were resolved.
	2. Language barriers are associated with race in which this article found that it has a direct correlation to a longer LOS, particularly with Spanish speaking patients. This can prove to be relevant when analyzing the data from San Diego county patients, in which has a high Spanish speaking population.

**Section IV. Hypothesis**

Null hypothesis: There is no relationship between Length of stay San Mateo County.

Alternative hypothesis: There is a relationship between Length of stay in San Mateo County.

Null hypothesis: There should be no variability between San Mateo and San Diego counties length of stay, based on race.

Alternative: There is variability between San Mateo and San Diego counties length of stay, based on race.

**Section V & VI. Database selection and data collection**

1. <http://www.oshpd.ca.gov/HID/Products/PatDischargeData/FrequencyTables/SanMateo/12SanMateoCounty.pdf>

One of the databases of focus is from the Office of Statewide Health Planning & Development, or OSPHD. OSHPD contains an inpatient hospital discharge frequency tables by county database. The databases are categorized into county and made available from the year 2002-2013. This database in particular focuses on San Mateo County in 2012.

1. <http://www.oshpd.ca.gov/HID/Products/PatDischargeData/FrequencyTables/SanDiego/12SanDiegoCounty.pdf>

This is a similar dataset from the same source that focuses on San Diego County in 2012. These two datasets contain several frequency tables such as; Race group by Patient County of Residence, MS-DRG by Patient County of Residence, and Disposition by Patient County of Residence and Type of Care. These frequency table variables will be used for statistical analysis and hypothesis testing techniques to prove or disapprove the hypothesis.

**Section VII. Validity Testing**

 The first validity test will determine if the null hypothesis, which states; there is no difference in patient’s LOS in San Diego County compared to San Mateo County. A two sample or two tailed T-test method was used for this validity test. First, two tables were created to show patients race and their average LOS. A third variable or column was added so that a Chi-Square test may be implemented for additional validity testing. This third column shows whether or not the patient’s average LOS is lower or higher than the county wide average LOS. The table is shown below: (Page 8)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Group 1 |   | Group 2 |   |
|  | San Mateo (7.1 avg LOS) | San Diego (5.1 avg LOS) |
| RACE | AVG LOS |   | AVG LOS |   |
| Unknown | 5.00 | lower | 6.50 | higher |
| White | 6.00 | lower | 5.30 | higher |
| Black | 11.00 | higher | 5.20 | higher |
| Hispanic | 12.00 | higher | 4.70 | lower |
| Asian | 5.40 | lower | 5.00 | lower |
| Native Am. | 4.00 | lower | 5.00 | lower |
| Other | 4.80 | lower | 4.50 | lower |

The average LOS in San Mateo County is 7.1, versus 5.1 in San Diego County. The races shown in the table are Unknown, White, Black, Hispanic, Asian, Native American, and Other. The process of calculating the two sample T-test was done by inputting all of the patients LOS in San Mateo County separately from the patients in San Diego County into a data analysis tool provided by Microsoft Excel. The “Comparing means (T-test)” tool in Excel calculated a descriptive statistics chart, in which the information provided will determine whether or not the null hypothesis will be rejected. The descriptive statistics chart is depicted below:

|  |
| --- |
| **Comparing Means [ t-test assuming equal variances (homoscedastic) ]** |
| *Descriptive Statistics* |
| *VAR* | *Sample size* | *Mean* | *Variance* |
|  | 7 | 6.88571 | 10.38476 |
|  | 7 | 5.17143 | 0.41905 |
|  |  |  |  |
| *Summary* |
| *Degrees Of Freedom* | 12 | *Hypothesized Mean Difference* | 0.E+0 |
| *Test Statistics* | 1.37989 | *Pooled Variance* | 5.4019 |
|  |  |  |  |
| *Two-tailed distribution* |
| *p-level* | 0.19279 | *t Critical Value (5%)* | 2.17881 |
|  |  |  |  |
| *One-tailed distribution* |
| *p-level* | 0.09639 | *t Critical Value (5%)* | 1.78229 |
|   |   |   |   |
| *G-criterion* |
| *Test Statistics* | 0.34286 | *p-level* | 0.05131 |
| *Critical Value (5%)* | 0.459 |   |   |
|  |  |  |  |
|  |  |  |  |
| *Pagurova criterion* |
| *Test Statistics* | 1.37989 | *p-level* | 0.78555 |
| *Ratio of variances parameter* | 0.96121 | *Critical Value (5%)* | 0.02607 |

The data shows that the critical value is 2.17881. The critical value is then used to determine the range in which will be used to either reject the null or accept the alternative hypothesis. If the test statistic shown on the chart is in between a range of -2.17881 and 2.17881, then the null hypothesis can be accepted. Through this t-test, it is determined that the test statistic for the two-tailed distribution is 1.37989. Because 1.37989 falls in between the critical value range, the null hypothesis is accepted and the alternative hypothesis is rejected. In conclusion, the two-sample t-test determined that there is no evidence to prove that there is a difference in the LOS in San Diego and San Mateo County.

 The next validity testing method that was utilized was the Chi Square Test. Initial thoughts of using the Chi square test proved to be challenge. Because this type of test is performed on categorical variables, the collected data sets alone did not meet this criterion. In order to be able to perform a Chi square test, the same table of data that was used in the t-test was used for this testing method as well.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Average LOS of Patients by Race |  |  |
|  |  |  |  |  |
|  | Group 1 |   | Group 2 |   |
|  | San Mateo (7.1 avg LOS) | San Diego (5.1 avg LOS) |
| RACE | AVG LOS |   | AVG LOS |   |
| Unknown | 5.00 | lower | 6.50 | higher |
| White | 6.00 | lower | 5.30 | higher |
| Black | 11.00 | higher | 5.20 | higher |
| Hispanic | 12.00 | higher | 4.70 | lower |
| Asian | 5.40 | lower | 5.00 | lower |
| Native Am. | 4.00 | lower | 5.00 | lower |
| Other | 4.80 | lower | 4.50 | lower |

The data sets that were collected contained averaged variables. In order to create categorical variables, the average LOS of all races were categorized as “lower” or “higher” compared to the county wide average LOS. Group 1 represents San Mateo County, with an average LOS of 7.1 and Group 2 represents San Diego County, with an average LOS of 5.1. The average LOS of each race according to their county is also depicted in the table. The next step in performing the Chi Square Test is to determine the observed and expected frequencies. The observed frequencies are found by inserting “1” in the table cell that depicts “lower” or “higher” LOS according to each race. The row and column total was computed to complete the observed frequencies table shown below (Page 10):

|  |  |  |  |
| --- | --- | --- | --- |
|  | San Mateo (OBSERVED) |  |  |
|  |  |  |  |
| Race | Lower | Higher | Row total |
| Unknown | 1.00 |   | 1.00 |
| White | 1.00 |   | 1.00 |
| Black |   | 1.00 | 1.00 |
| Hispanic |   | 1.00 | 1.00 |
| Asian | 1.00 |   | 1.00 |
| Native Am. | 1.00 |   | 1.00 |
| Other | 1.00 |   | 1.00 |
| column total | 5.00 | 2.00 | 7.00 |

 In order to find the expected cell frequencies, the following formula was computed: Row Total x Column Total / Grand Total of All Cells. This formula was repeated for every cell in observed frequency table for both counties. If done correctly, the row total and the column total should equate to the observed cell frequency tables. Both expected frequency tables for San Mateo and San Diego counties are shown below:

|  |  |  |
| --- | --- | --- |
|  | San Mateo (EXPECTED) |  |
|  |  |  |  |
| Race | Lower | Higher | Row total |
| Unknown | 0.71 | 0.29 | 1.00 |
| White | 0.71 | 0.29 | 1.00 |
| Black | 0.71 | 0.29 | 1.00 |
| Hispanic | 0.71 | 0.29 | 1.00 |
| Asian | 0.71 | 0.29 | 1.00 |
| Native Am. | 0.71 | 0.29 | 1.00 |
| Other | 0.71 | 0.29 | 1.00 |
| column total | 5.00 | 2.00 | 7.00 |

With both tables completed, the Chi-Square test was conducted using the formula below. The “O” variable is the observed cell in the observed frequency tables. The “E” variable is the expected cell in the expected cell frequency table.

The values for both cell frequencies were calculated and then summed to find the test statistical for both counties. The calculated test statistic for San Mateo County was calculated to be 2.87. The degree of freedom of 6 was determined by finding the difference between the sample size, “N” or 7 and 1. The test statistic was then compared against the critical value. The ideal alpha that was set to 0.95, which was the complement of 0.5. Using a table of upper-tail critical values of chi square distribution based on the degree of freedom from the National Institute of Standards and Technology, the critical value was determined to be 12.592. Because the test statistic of 2.86 is not greater than the critical value of 12.592, the null hypothesis was not rejected. In conclusion, the sample does not represent statistical evidence that race affects the LOS in San Mateo County.

 The same formula and process was applied to San Diego County in determining whether or not to reject or accept the null hypothesis. An observed frequency table was created for San Diego County using the same method that was applied to San Mateo County’s observed table. The expected frequency table was filled out by using the Row Total x Column Total / Grand Total of All Cells formula as mentioned previously.

|  |  |  |  |
| --- | --- | --- | --- |
|  | San Diego (OBSERVED) |  |  |
|  |  |  |  |
| Race | Lower | Higher | Row total |
| Unknown |   | 1.00 | 1.00 |
| White |   | 1.00 | 1.00 |
| Black |   | 1.00 | 1.00 |
| Hispanic | 1.00 |   | 1.00 |
| Asian | 1.00 |   | 1.00 |
| Native Am. | 1.00 |   | 1.00 |
| Other | 1.00 |   | 1.00 |
| column total | 4.00 | 3.00 | 7.00 |

|  |  |  |
| --- | --- | --- |
|  | San Diego (EXPECTED) |  |
|  |  |  |  |
| Race | Lower | Higher | Row total |
| Unknown | 0.57 | 0.43 | 1.00 |
| White | 0.57 | 0.43 | 1.00 |
| Black | 0.57 | 0.43 | 1.00 |
| Hispanic | 0.57 | 0.43 | 1.00 |
| Asian | 0.57 | 0.43 | 1.00 |
| Native Am. | 0.57 | 0.43 | 1.00 |
| Other | 0.57 | 0.43 | 1.00 |
| column total | 4.00 | 3.00 | 7.00 |

Using the Chi-Square test formula, the calculated test statistic for San Diego County was 9.31. The degree of freedom was already established to be 6 as well as the critical value of 12.592. Because the test statistic of 9.31 is not greater than the critical value of 12.592, the null hypothesis will not be rejected. This concludes that the sample from San Diego County does not represent statistical evidence that race affects LOS.

**Section VIII. Conclusion**

 The two sample t-test was used to determine if there was difference in a patient’s LOS in San Diego County compared to San Mateo County. The result of the test showed that there is no evidence to prove that the alternative hypothesis was true. The findings of this test were in direct contrast to initial assumptions that because San Mateo and San Diego counties were geographically apart, the racial diversity would have a factor on the LOS. This was not the case as the t-test proved that the two datasets that were used in this study did not show those findings. The Chi Square Test was used to determine if inpatient race had any affect on their LOS in both counties respectively. The test found that because the test statistics for both counties were not greater than the critical value, the null hypothesis was not rejected. This meant that the data collected did not show any correlation between inpatient race and LOS.

 There is some speculation to believe that because of the particular datasets used in this analysis, the findings were not completely accurate. The datasets used for both the t-test and the Chi Square test were averages of the race group for each county. If the datasets showed each individual patient in the entire county and contained variables of LOS, race, age, etc., the use of a pivot table could have more accurately depicted the population. The assumption is that the validity testing methods would have been more effective in rejecting or accepting the null hypotheses in this analysis.