D207 Performance Assessment

September 04, 2021

**Part A.**

1. A question that could be asked by this organization would be to see if Rural, Suburban, or Urban patients spend the most amount of time in the hospital and if they pay more.
2. The stakeholders could benefit from this by learning about what type of origin patients are coming from and if there are factors related to the costs and area type which may indirectly say if they have more or less problems that could be observed in the hospital.
3. The data I will be using is the area column to determine if the patient is rural, suburban, or urban. I will also be using the “Initial\_days”, “TotalCharge”, and “Additional\_charges“ to calculate how much each patient spent while in the hospital along with how long they were in the hospital.

**Part B.**

1. For my analysis, I used a Student t-test in order to find if there was any sort of significant difference between the 3 groups. I paired them up in Rural/Suburban, Suburban/Urban, and Urban/Rural to check the t value and p value.
2. The results were:

Rural/Suburban: t-value: -0.58; p-value: 0.56

Suburban/Urban: t-value: -0.55; p-value: 0.58

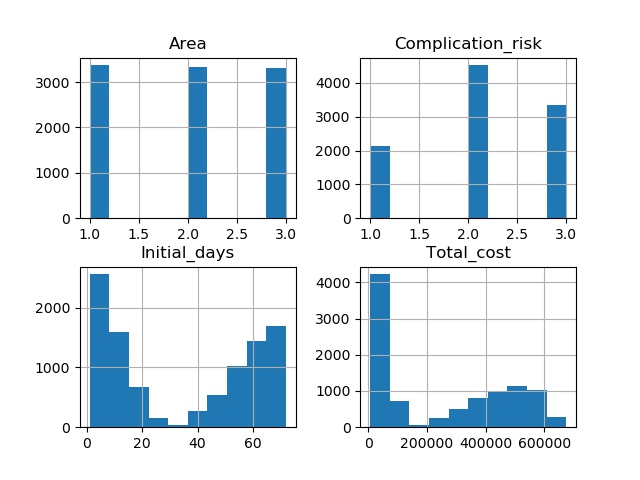
Urban/Rural: t-value: 1.13; p-value: 0.26

These values look bad initially, but they are good since it shows that there isn’t any real significant difference between any of the living areas. The only comparison that’s not like the others, its the Urban/Rural divide, but with a p-value of 0.26, the results are have incredibly low confidence still. This could be easily interpreted to show that no meaningful difference is noticed within the area divides. Code is in attached document.

1. When choosing to use a t-test, I had to eliminate the options of chi-square and ANOVA. I eliminated chi-square test since it needed 2 categorical variables for analysis and my analysis didn’t have multiple categories to compare since I was looking at a categorical and a continuous variable for cost and area of residence so the chi-square test wouldn’t work. ANOVA test was eliminated as an option since it’s a test to see which factors of a sample impacts a result the most if the variable is being changed. Since the ANOVA test would just give us a “Well there’s something significant in there” result without being exact, I rejected the use of the test.

**Part C.**

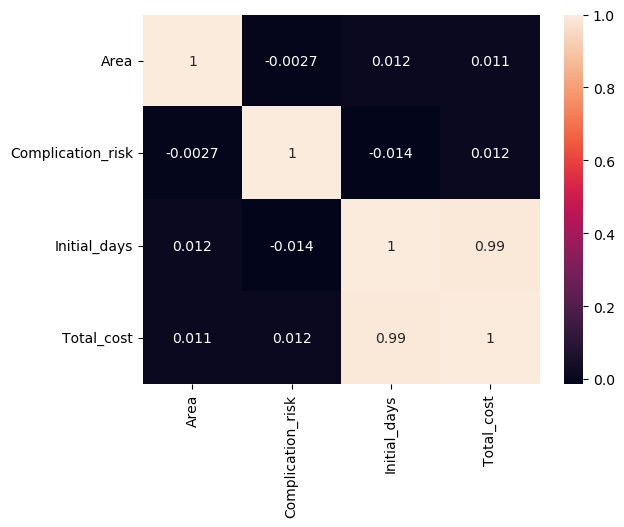
1. The 2 continuous variables I chose to analyze were the number of days in the hospital, and the number days in the hospital, while the 2 categorical variables area and complication risk. I chose to put these into a histogram to help compare all the variables to themselves in an univariate analysis. Below is the histogram representation as well as an attached image.



You can see that the area is balanced between rural, urban, and suburban areas. The complication risk shows that most patients are at a moderate risk of complications between all areas. The number of days in the hospital are interesting since it’s a reverse bell curve. Patients seem to spend either a lot of time, or very little time in the hospital based on the “Initial\_days” inverse bell curve. The total cost is also interesting as it shows almost 2 bell curves, one around $5,000 and another around $50,000.

**Part D.**

1. I chose to use the same variables for a bivariate analysis and compare them all to one another in the form of a heatmap to see how related each variable is to each other. Below is the image generated from the code attached.



The heatmap above is comparing the relationships between all the components and the closer the number is to 1, the more related the 2 factors are. As you can see, the only factors that seem to be relatable are the days a patient stay in the hospital and the total cost, which makes sense. The more time a person is in the hospital, the more it will cost them, this seems like a rational linear relationship.

**Part E.**

1. The original hypothesis that was being tested was to see if patients from different areas showed any kind of significant difference existed between them in prices they paid. After doing a full analysis, it appears that there isn’t any significant difference between rural, suburban, or urban patients in terms of cost. The R values are 0.56, 0.58, and 0.26 which is such a massive uncertainty that no reasonable person would be able to conclude that there’s any variation between the area people are from, and the cost they incur. Since these values are so far off, I can conclude that the null hypothesis was correct, in that the factors of area and cost seem to be completely unrelated to each other. The alternative hypothesis was incorrect because if it was correct, then we would see a differential between the area of residency and cost incurred at the hospital, but since my null hypothesis was correct, the alternative hypothesis is incorrect.
2. A couple limitations from the dataset that I used is the fact that I ended up adding the costs all together to get a total cost. This could cause a few issues, one being that perhaps if I analyzed and compared rural, suburban, and urban patients and the amount of time in the hospital or additional charges that were added onto their bill, we might be able to find some sort of discrepancy, but this analysis ignores that go into the full picture. Another issue that could have occurred is the bivariate analysis between area, complication risk, initial days, and total cost. By using a t-test, it appears that there only appears to be a relationship between initial days and total cost, but because of how far out the R and P values were, it doesn’t show a high level of confidence in the result, especially since the Urban/Rural split doesn’t appear like the others. I think doing a chi-square test would have been able to help sort out the anomalies that would be present in our alternative hypothesis, but this t-test analysis shows that our null hypothesis was correct.
3. I think the best course of action with this data would be to look further into amount of days a patient stays in the hospital, the amount they pay per day, and the additional charges independently in order to determine if the total cost happens to average out because of all those factors, or if there really is no discrepancy between areas on any factors involving costs.

**Part F.**

Panopto video: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=2400cf42-ba62-4ed4-8ce7-ada200f6b266>

**Part G.**

References

<https://www.youtube.com/watch?v=pTmLQvMM-1M> Bozeman Science, Student’s t-test

<https://www.youtube.com/watch?v=WXPBoFDqNVk> Bozeman Science, Chi-squared Test

<https://www.youtube.com/watch?v=oOuu8IBd-yo> CrashCourse, ANOVA: Crash Course Statistics #33

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html> SciPy, ttest function

<https://seaborn.pydata.org/> Seaborn documentation

<https://matplotlib.org/stable/index.html> matplotlib documentation

<https://www.statisticshowto.com/> general statistics guide