# Data Project Part 3

Simran Kaur, Bhawanjot Mann, Maryellen Miyashita, Suhas Nagappala, Mannvir Singh

4/26/2021

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

1. Part 1

1.1. [2 marks] What is the problem your are addressing with these data? State the question you are trying to answer and let us know what type of question this is in terms of the PPDAC framework.

Problem: This problem is addressing causation. Are hospitals with more equipment, namely CT scanners, associated with lower patient stay lengths?

Plan: This is an observational study since we are analyzing already collected data from OECD. This study can also be classified as a retrospective study since we are analyzing data from the past.

Data: The data is from the OECD database which stands for the Organisation for Economic Cooperation and Development. This is a worldwide statistical organization that is known to be very reliable. We have the data from the years of 1990-2018. The data showcases the lengths of hospital stay associated with many countries of the OECD.

Analysis: We will analyze and visualize the data to see if there is a correlation between the average length of hospital stay and the average amount of equipment in various countries. We will use techniques learned in this course such as linear regression and scatter plots to visualize the relationship between variables.

Conclusion: As seen in our analysis, when hospitals have a higher number of CT scanners it is associated with lower patient stay lengths. Specifically, this is reflected in the scatter plot and regression line.

1.2. [2 marks] What is the target population for your project? Why was this target chosen i.e., what was your rationale for wanting to answer this question in this specific population?

As the dataset is for OECD countries, our target population would be OECD countries and countries that are similar. We chose this target population because OECD countries are typically more developed with many healthcare investments such as MRI units, with available data on them as well, so we may be able to show a causation between length of hospital stay and investments in equipment. This target population provides the best way to see this relationship, as data for non-OECD countries is not readily available.

To answer our question of if hospitals with more equipment have association with lower patient stay lengths we needed a population that has different amounts of hospital equipment amongst it. In other words, the only way to investigate our question was to compare countries with hospitals that have various average amounts of hospital equipment, and OECD countries meet this criteria.

1.3. [2 marks] What is the sampling frame used to collect the data you are using? Describe why you think this sampling strategy is appropriate for your question. To what group(s) would you feel comfortable generalizing the findings of your study and why.

The sampling frame is OECD countries. This can be seen as a convenience sample since data was only collected about OECD countries likely because data for them was readily available. Convenience sampling is appropriate for our question as we can only assess countries we have data on, which is OECD countries.

We may be able to generalize our findings to non-OECD countries as well. We can generalize to countries that are similarly developed, which can be determined with metrics such as GDP per capita, life expectancy, etc. as patient needs may be more similar amongst such nations. Countries we generalize to should have a significant number of patients with medical conditions that require this equipment, otherwise more equipment wouldn't have any effect on stay lengths.

1.4. [2 marks] Write a brief description (1-4 sentences) of the source and contents of your dataset. Provide a URL to the original data source if applicable. If not (e.g., the data came from your internship), provide 1-2 sentences saying where the data came from. If you completed a web form to access the data and selected a subset, describe these steps (including any options you selected) and the date you accessed the data.

This data was obtained from the OECD database that has data for each OECD country's healthcare investments collected from 1990-2018. OECD, the Organization for Economic Cooperation and Development, was founded in 1961 by member countries to stimulate economic progress, build better policies, and foster world trade. The data shows countries in the OECD and the average number of MRI units, CT scanners, and hospital beds in their hospitals, which is their healthcare investments. It also lists the average hospital stay length for each country in a given year.

The url is below: https://www.kaggle.com/babyoda/healthcare-investments-and-length-of-hospital-stay

1.5. [1 mark] Write code below to import your data into R. Assign your dataset to an object.

```
library(readr)
HealthHospitaldata <- read_csv("/home/jovyan/ph142-sp21/project/HealthHospitaldata.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Location = col_character(),
##   Time = col_double(),
##   Hospital_Stay = col_double(),
##   MRI_Units = col_double(),
##   CT_Scanners = col_double(),
##   Hospital_Beds = col_double()
## )
```

1.6. [3 marks] Use code in R to answer the following questions:

i) What are the dimensions of the dataset?

```
dim(HealthHospitaldata)
```

```
## [1] 518   6
```

The dimensions are 518 rows x 6 columns.

ii) Provide a list of variable names.

```
names(HealthHospitaldata)
```

```
## [1] "Location"      "Time"          "Hospital_Stay" "MRI_Units"
## [5] "CT_Scanners"   "Hospital_Beds"
```

iii) Print the first six rows of the dataset.

```
head(HealthHospitaldata)
```

```
## # A tibble: 6 x 6
##   Location  Time Hospital_Stay MRI_Units CT_Scanners Hospital_Beds
##   <chr>    <dbl>         <dbl>     <dbl>       <dbl>         <dbl>
## 1 AUS       1992           6.6      1.43        16.7          1.43
## 2 AUS       1994           6.4      2.36        18.5          2.36
## 3 AUS       1995           6.5      2.89        20.6          2.89
## 4 AUS       1996           6.4      2.96        22.0          2.96
## 5 AUS       1997           6.2      3.53        23.3          3.53
## 6 AUS       1998           6.1      4.51        24.2          4.51
```

1.7. [4 marks] Use the data to demonstrate a statistical concept from Part I of the course. Describe the concept that you are demonstrating and interpret the findings. This should be a combination of code and written explanation.

```
library(dplyr)
```

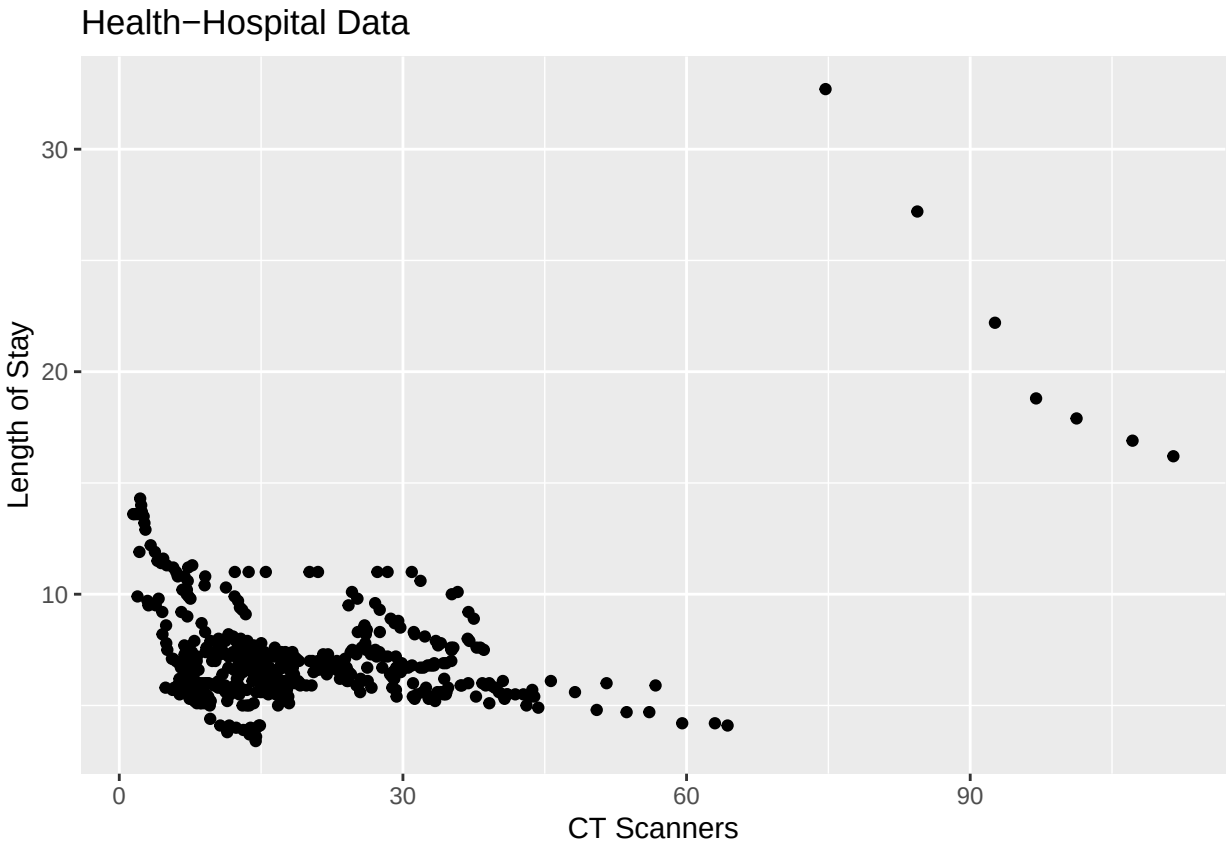
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

HealthHospital_scatter <- ggplot(data = HealthHospitaldata, aes (x = CT_Scanners, , y = Hospital_Stay))
labs(x = "CT Scanners", y = "Length of Stay", title = "Health-Hospital Data")
HealthHospital_scatter
```



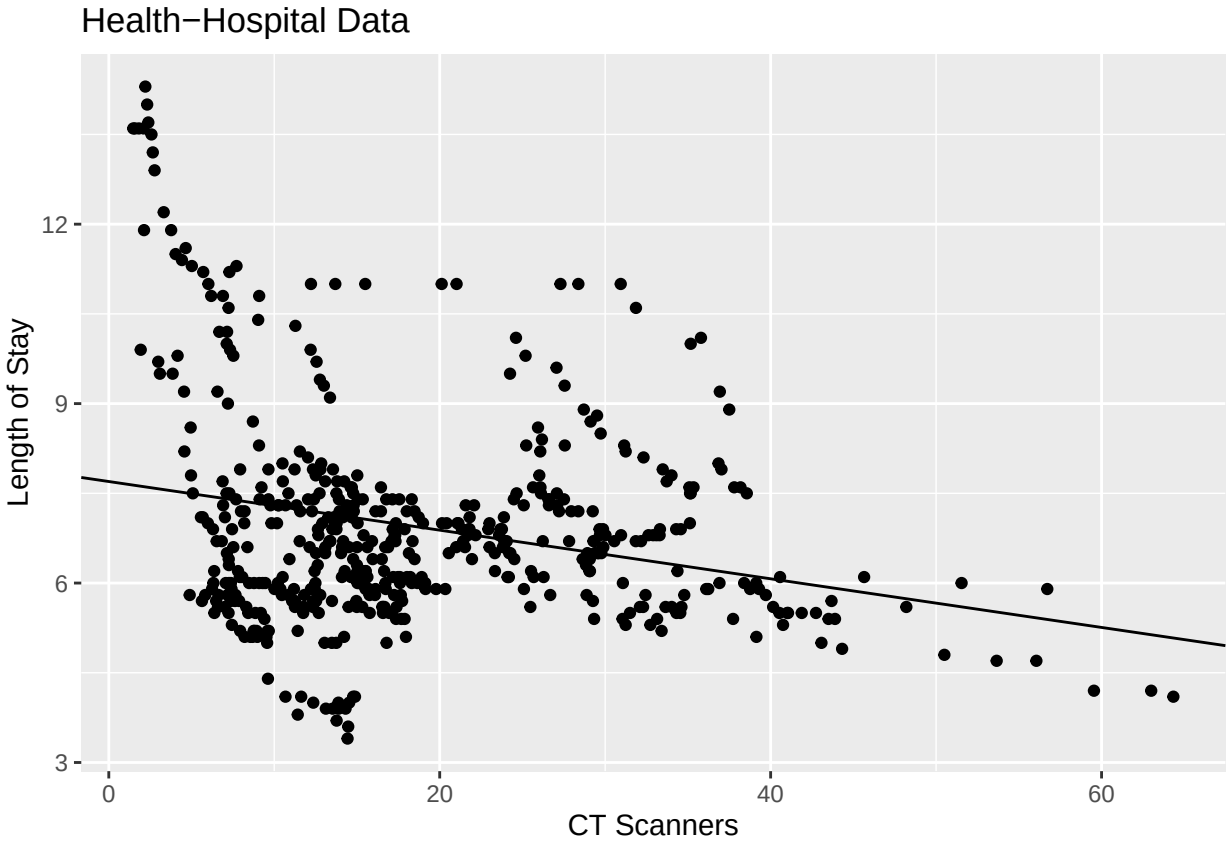

There were two groups of data and both groups of data had a negative correlation. From here, we saw there were outliers, namely points where the CT scanner count was greater than approximately 65. They each shared the same downward pattern.

For the group of the data where the CT scanner count is less than 65, which is the majority of the data, we generated a scatterplot and the regression line. To start, we manipulated the dataset so that only data where CT scanner count less than 65 is present so we can investigate it on its own.

```
HealthHospitaldata_filtered <- HealthHospitaldata %>% filter( CT_Scanners < 65 )
HealthHospital_filtered_lm <- lm(Hospital_Stay ~ CT_Scanners, data = HealthHospitaldata_filtered)
HealthHospital_filtered_lm
```

```
##
## Call:
## lm(formula = Hospital_Stay ~ CT_Scanners, data = HealthHospitaldata_filtered)
##
## Coefficients:
## (Intercept)  CT_Scanners
##     7.69723     -0.04066
```

```
HealthHospital_filtered_scatter <- ggplot(data = HealthHospitaldata_filtered, aes(x = CT_Scanners, y = L
labs(x = "CT Scanners", y = "Length of Stay", title = "Health-Hospital Data") +
geom_abline(intercept = 7.69723, slope = -0.04066)
HealthHospital_filtered_scatter
```




Overall, a weak negative correlation exists between the number of CT scanners and length of patient stays. Based on the slope, an increase in one CT scanner was associated with reduction in stay length of approximately 0.04 days, which equates to approximately 1 hour.The data is scattered around the regression line, but overall when hospitals have a higher number of CT Scanners they seem to have lower patient stay lengths.

2. [2 marks] Describe a quantity you will estimate as an outcome in your problem using probability notation. Are you planning to calculate marginal probabilities? Conditional probabilities?

In our problem, we are looking into how hospitals with more equipment, particularly CT scanners, will relate to lower patient stay lengths. The probability notation that would be used in this particular problem would be P(A|B), where A would represent the patient stay length while B would represent the average amount of equipment. Among the hospitals that have more equipment what percent resulted in lower patient lengths would be a probability we are trying to solve. From this example, we would be calculating conditional probability.

3. [3 marks] Describe the type of theoretical distribution that is relevant for your data. What type of variable(s) are you investigating (continuous, categorical, ordinal, etc)? What theoretical distribution that we have talked about would potentially be appropriate to use with these data (Normal, Binomial, Poisson...) Why is this an appropriate model for the data you are studying?

The variables that are relevant in our data would be continuous and discrete variables. The continuous variable would be the length the patient stays in the hospital and the discrete variable would be the total number of equipment at each hospital. The theoretical distribution that we have talked about that could apply to this data is a normal distribution. The normal distribution would likely be best suited for this data, since the variables we are testing (length of patient stay and total number of equipment) are likely to be evenly distributed, with the mean and median equivalent to each other.

4. [4 marks] Use the data you have to demonstrate a statistical concept from Part II of the course. Describe the concept that you are demonstrating and interpret the findings. This may include code in R, a visual of some kind and text interpretation.

I want to demonstrate calculating the probability and percent of the amount of MRI Units in the hospitals. First thing we did was calculate the mean and standard deviation of the amount of MRI Units across the hospitals by year.

In order to calculate the mean and standard deviation, we used the following code:

```
HealthHospitaldata %>% summarize(MRI_units = mean(MRI_Units), standard_deviation_MRI_Units = sd(MRI_Uni
```

```
## # A tibble: 1 x 2
##   MRI_units standard_deviation_MRI_Units
##       <dbl>                        <dbl>
## 1      10.6                         8.69
```

The mean of the data was 10.5655 MRI units and the standard deviation was 8.68557. This data is useful in order to find out probabilities based on a normal curve. Using the functions pnorm() and qnorm() we can determine areas under the curve which can tell us about probability. If we want to figure out the probability of having less than 15 MRI units, assuming an approximately normal distribution, we would do:

```
pnorm(q= 15, mean = 10.57, sd = 8.68)
```

```
## [1] 0.6951034
```

This shows that 70% of the hospitals by year have MRI units under 15. Then, to show how many hospitals over the course of the years are over that amount would be about 30%. In relation to pnorm(), we can use qnorm() for percentiles. To find the third quartile, which would be above 75% of the data, we use q = 0.75:

```
qnorm(p = 0.75, mean = 10.57, sd = 8.68)
```

```
## [1] 16.42457
```

This function shows that, assuming the data is approximately normally distributed, 75% of the data falls below 16.42 MRI Units, and so 16.42 is the 75th percentile.

Part 3

5a. [2 marks] Identify a statistical test to apply to your data (must be a concept we covered in part III of the course). In plain language, write the question you are trying to answer.

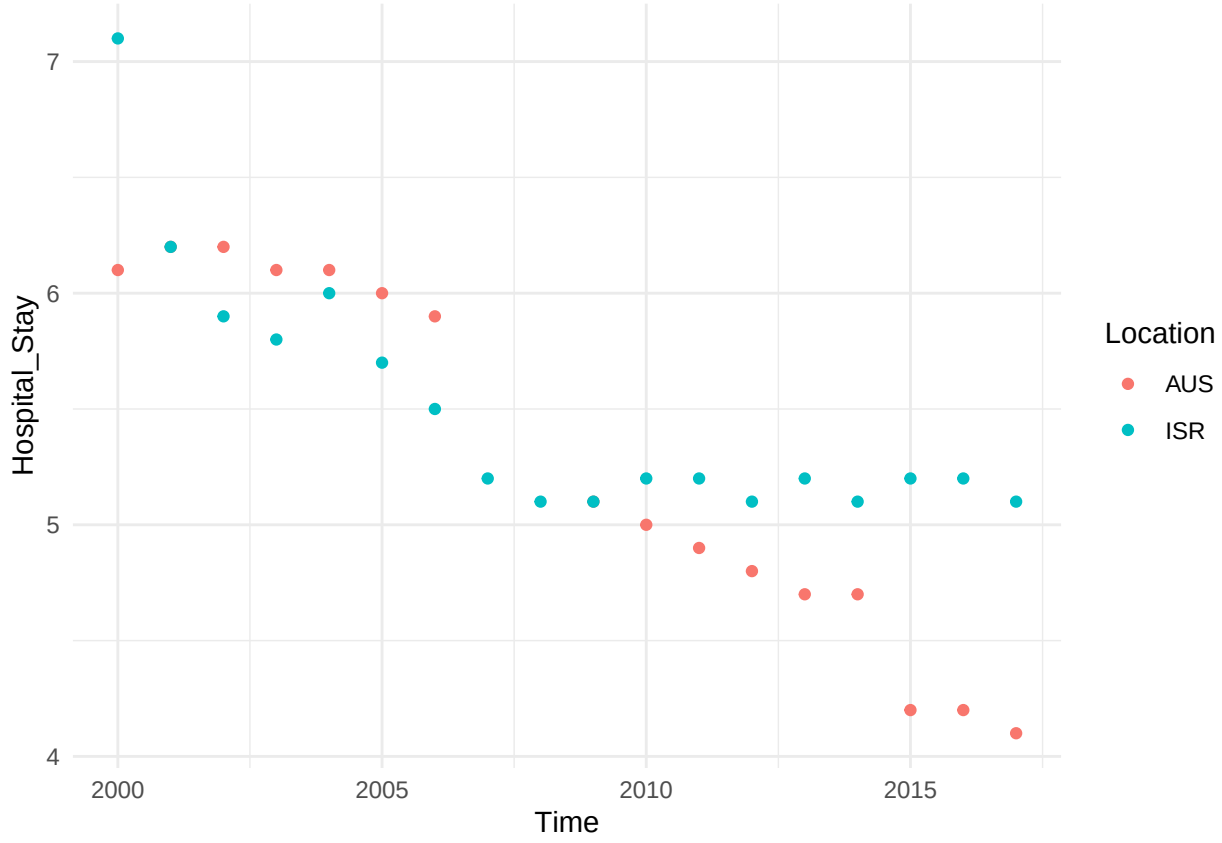
We will use a Wilcoxon Sign Rank paired 2 sample t-test to apply to our data. We are trying to determine, if over time, the mean hospital stay has been different between Israel and Austria over the years using sample data from 2000-2017. We chose these two countries because they had hospital stay data for every year from 2000-2017, while other countries did not have data for a couple years within this range.

5b. [2 marks] What assumptions are required by the method you chose in 5.a)? Show how you assessed whether these assumptions are met by your dataset.

This assumes that the distributions have the same general shape but assumes nothing about the shape. By looking at the plots of Israel and Austria hospital stay lengths below, one can see that the two samples follow approximately the same "downward" trend shape of numerical values.

```
StayLengths_Austria_Israel <- HealthHospitaldata %>% select(Location, Time, Hospital_Stay) %>% filter(L
```

```
ggplot( StayLengths_Austria_Israel , aes(x = Time , y= Hospital_Stay )) + geom_point(aes(col=Location))
```




5c. [2 marks] Explain why this test is appropriate for the data you have and the question you are trying to answer. Use at least one visualization technique and include both the output and the R code that generated it.

Since the corresponding year can affect the hospital stays due to population increases etc., data for each year may not be "fully independent" of each other per se and so we see them as paired. Furthermore, since we do not have a large sample size, a Wilcoxon Sign Rank test is appropriate for the data we have. This test assumes that the difference between the two sample's measures is 0 under its null hypothesis; thus, this test

is appropriate for investigating if the stay lengths are different amongst the two countries during this time period. Hospital stay lengths from the same year will be "paired" with each other.

5d. [2 marks] Clearly state the null and alternative hypotheses for this test.

Null Hypothesis: There is no difference between the mean hospital stay length of Austria and Israel.

Alternative Hypothesis: There is a difference between the mean hospital stay length of Austria and Israel.

6. [2 marks] Include the R code you used to generate your results. Annotate your code to help us follow your reasoning.

```
hospital_stay <- tribble( ~Aus, ~ISR,
                          6.1, 7.1,
                          6.2, 6.2,
                          6.2, 5.9,
                          6.1, 5.8,
                          6.1, 6.0,
                          6.1, 5.7,
                          6.0, 5.5,
                          5.9, 5.2,
                          5.1, 5.1,
                          5.0, 5.1,
                          4.9, 5.2,
                          4.8, 5.2,
                          4.7, 5.1,
                          4.7, 5.2,
                          4.2, 5.1,
                          4.2, 5.2,
                          4.1, 5.2,
                          4.1, 5.1)


wilcox.test(hospital_stay %>% pull(Aus), hospital_stay %>% pull(ISR), paired = T, correct = FALSE)
```

```
## Warning in wilcox.test.default(hospital_stay %>% pull(Aus), hospital_stay %>% :
## cannot compute exact p-value with ties
```

```
## Warning in wilcox.test.default(hospital_stay %>% pull(Aus), hospital_stay %>% :
## cannot compute exact p-value with zeroes
```

```
##
##  Wilcoxon signed rank test
##
## data:  hospital_stay %>% pull(Aus) and hospital_stay %>% pull(ISR)
## V = 36.5, p-value = 0.1027
## alternative hypothesis: true location shift is not equal to 0
```

```
# This was the R code used to determine if the mean hospital stay has been different
```

```
wilcox.test(hospital_stay %>% pull(Aus), hospital_stay %>% pull(ISR), paired = T, correct = FALSE)
```

```
## Warning in wilcox.test.default(hospital_stay %>% pull(Aus), hospital_stay %>% :
## cannot compute exact p-value with ties
```

```
## Warning in wilcox.test.default(hospital_stay %>% pull(Aus), hospital_stay %>% :
## cannot compute exact p-value with zeroes
```

```
##
##  Wilcoxon signed rank test
##
## data:  hospital_stay %>% pull(Aus) and hospital_stay %>% pull(ISR)
## V = 36.5, p-value = 0.1027
## alternative hypothesis: true location shift is not equal to 0
```

```
#This was the R code used to determine if the mean hospital stay has been different
```

7. [4 marks] Present your results in a clear summary. This should include both a text summary and a table or figure with appropriate labeling.

The table below summarizes the two groups of average hospital stays - AUS (Australia) and ISR (Israel). The values in the data chart are the average hospital stay lengths in Austria and Israel, starting from 2000 (denoted 1) to until 2017 (denoted 17).

Year AUS | ISR 2000 6.1 | 7.1 2001 6.2 | 6.2 2002 6.2 | 5.9 2003 6.1 | 5.8, 2004 6.1 | 6.0 2005 6.1 | 5.7 2006 6.0 | 5.5 2007 5.9 | 5.2 2008 5.1 | 5.1 2009 5.0 | 5.1 2010 4.9 | 5.2 2011 4.8 | 5.2 2012 4.7 | 5.1 2013 4.7 | 5.2 2014 4.2 | 5.1 2015 4.2 | 5.2 2016 4.1 | 5.2 2017 4.1 | 5.1

Here we have the results of the Wilcoxon signed rank test we conducted on this data, with data from AUS and ISR : data: hospital_stay %>% pull(Aus) and hospital_stay %>% pull(ISR) V = 36.5, p-value = 0.1027

The results above shows the V (36.5), which is the sum of ranks, and our p-value (0.1027) which is the probability of observing the difference we saw **under the null hypothesis** based on the test we used.

8. [4 marks] Interpret your findings. Include a statement about the evidence, your conclusions, and the generalizability of your findings.

The evidence from our statistical analysis shows that the p-value is greater than 0.05, which means that our P-value is not statistically significant. Under the null hypothesis, the probability of observing the difference in hospital stays we say would be 0.1027. Thus, we fail to reject the null hypothesis that there is no underlying difference between the overall mean hospital stay lengths of Austria and Israel. The generalizability of these findings to other countries may be limited and not be feasible, as larger countries with larger populations may have more chronic patients in hospitals that may change the outcome of these results. We only looked at Austria and Israel, which are European countries that may be more comparable to one another than other "pairs" of countries, so we cannot generalize our results to say hospital stay lengths are similar across all countries. Furthermore, our data was only from 2000-2017, so we likely cannot generalize to the present day either.

9. [2 marks] Create a statement of contribution.

Members Suhas Nagappala and Mannvir Singh had primary responsibility for questions in the Part I of the project (Q#1-7), **and** questions 5a, 5b, 5c, and 5d of Part III, and made some minor corrections to Part III. Suhas Nagappala compiled the answers into datahub and uploaded to gradescope. Bhawanjot Mann, Simran Kaur, and MaryEllen Miyashita completed Part II, questions 1-4. The remaining questions in Part III were the primary responsibilities of Bhawanjot Mann (Q #6), Simran Kaur (Q #8), and MaryEllen Miyashita (Q #9).