Statistical Report on Hospital Branches

Amanda Waldron

Fri Mar 11 2022

Contents

You were just hired as an applied behavioral scientist at a hospital. After some complaints from patients, the hospital board has brought up concerns that doctors at a remote branch of the hospital have been acting rudely toward patients. Your first assignment as the resident expert (see what I did there?) in behavioral science is to compare the branch receiving complaints with the main branch to determine if there could be a problem with doctor rudeness.

This hospital system already collects bedside manner ratings from patients, and stores them from a database. You decide that comparing the average bedside manner ratings of the patients at the remote branch with the main branch might be a good first-pass in investigating this rudeness issue. You extract the data from the database, and are left with the following variables in your dataset:

- `patient.ID`: a chronological count of patients who have filled out the bedside-manner questionnaire.
- `location`: a variable that indicates which of the two branches of the hospital the patient ratings came from. dist.1 is the main branch, and dist.2 is the remote branch (i.e. the branch whose doctors are supposedly more rude).
- Bedside manner ratings. This was a three-item scale that form a composite scale of perceived doctor bedside manner when averaged together. The questionnaire lists three behaviors and asks patients to indicate how often their doctor does the behaviors. The responses are on a 5-point Likert scale with 1 = *not at all often* and 5 = *very often.* The questions were:
  - "My doctor ignores me." (called `m.1` in your dataset). Important: this item is *reverse-scored*, such that 1 indicates better bedside manner and 5 indicates worse bedside manner. The responses in the data set are the raw responses, meaning you will have to recode this variable.
  - "My doctor listens to what I have to say." (called `m.2` in your dataset).
  - "My doctor shows interest in how I am feeling." (called `m.3` in your dataset).

Your goal is to compare the average bedside manner between the two hospital branches.

1 Prepare R Session For Analysis (Section Total: 1pt)

1.1 Load Packages

(1pt). Load the following packages into the library. Install packages that you don't already have. Show your code.

- dplyr
- DescTools
- lmtest
- jtools
- ggplot2

```
library(dplyr)
library(DescTools)
library(lmtest)
library(jtools)
library(ggplot2)
library(psych)
```

2 Inspect Data (Section Total: 13pt)

2.1 Loading Data

(1pt). We have a csv file called "HW3.data" (available on Blackboard). Read this file into R Studio.

```
Slytherin <- read.csv("HW3.data.csv")
```

2.2 Cleaning Data

(2pt). Check to see if there are problems with your variables. Show the code you used to find the problems.

```
#Slytherin$patient.ID %>% unique()

Slytherin$location %>% unique()
```

```
## [1] "dist.1" "dist.2"
```

```
Slytherin$m.1 %>% unique()
```

```
## [1] "2"      "4"      "3"      "1"      "5"      "dist.1"
```

```
Slytherin$m.2 %>% unique()
```

```
## [1] 3 4 2 5 1
```

```
Slytherin$m.3 %>% unique()
```

```
## [1]  4  2  3  5  1 33
```

There are no problems with location. `m.1` *is a factor, because there is a cell with "dist.1" in it.* `m.3` *has one value that is out of range for the scale (33).*

(2pt). Fix the problems you found. Show your code.

```
Slytherin1 <- Slytherin %>% filter(m.1 != 'dist.1')

Slytherin1 <- Slytherin1 %>% mutate(m.1= as.numeric(as.character(m.1)))


Slytherin1 <- Slytherin1 %>% filter(m.3 <= 6)

class(Slytherin1$m.1)
```

```
## [1] "numeric"
```

2.3 Scoring and Recoding Variables

(1pt). Use a combination of `mutate()` and `ifelse()` to recode the values in `location` to read as "main" and "remote" instead of "dist.1" and "dist.2", respectively. Name this new variable, which contains the recoded values, to `branch`. Use `branch` instead of `location` from now on.

```
Slytherin1 <- Slytherin1 %>%
  mutate(branch = ifelse(Slytherin1$location == "dist.1","main","remote"))

Slytherin1 %>% head
```

```
##   patient.ID location m.1 m.2 m.3 branch
## 1          1   dist.1   2   3   4   main
## 2          2   dist.2   2   3   4 remote
## 3          3   dist.2   4   3   2 remote
## 4          4   dist.2   2   4   3 remote
## 5          5   dist.2   3   3   3 remote
## 6          6   dist.1   3   3   3   main
```

(1pt). Show the code and output that checks what class the `branch` variable is. This variable needs to be factor. Is it? If not, also write some code that will change it to factor.

```
class(Slytherin1$branch)
```

```
## [1] "character"
```

```
Slytherin1$branch <- as.factor(Slytherin1$branch)

Slytherin1 %>% head
```

```
##   patient.ID location m.1 m.2 m.3 branch
## 1          1   dist.1   2   3   4   main
## 2          2   dist.2   2   3   4 remote
## 3          3   dist.2   4   3   2 remote
## 4          4   dist.2   2   4   3 remote
## 5          5   dist.2   3   3   3 remote
## 6          6   dist.1   3   3   3   main
```

(2pt). Create an item called `m.1.reversed` that is the reverse-scored version of `m.1`.

```
class(Slytherin$m.1)
```

```
## [1] "character"
```

```
Slytherin1 <- Slytherin1 %>%
  mutate(m.1.reversed = 6 - m.1)

Slytherin1 %>% head
```

```
##   patient.ID location m.1 m.2 m.3 branch m.1.reversed
## 1          1   dist.1   2   3   4   main            4
## 2          2   dist.2   2   3   4 remote            4
## 3          3   dist.2   4   3   2 remote            2
## 4          4   dist.2   2   4   3 remote            4
## 5          5   dist.2   3   3   3 remote            3
## 6          6   dist.1   3   3   3   main            3
```

(1pt). Rename the bedside manner items to something that better describes what the variables are.

```
Slytherin1 <- Slytherin1 %>%
  rename(bedside.1 = m.1.reversed,bedside.2 = m.2,bedside.3 = m.3)
```

(2pt). Create a new column that contains a composite measure (i.e. mean) of bedside manner questionnaire items for each patient. Name this variable `bedside.comp`.

```
Slytherin1 <- Slytherin1 %>%
  mutate(bedside.comp = rowMeans(dplyr::select(.,bedside.1,bedside.2,bedside.3)))
```

(1pt). Create a dataframe that contains the mean and standard deviation of the bedside.comp variable for each branch separately using `group_by()` and `summarize()`. You do not have to save the output.

```
Slytherin1descriptives <- Slytherin1 %>%
  group_by(branch) %>%
  summarize(mean.bedside.comp = mean(bedside.comp, na.rm = TRUE),
            sd.bedside.comp = sd(bedside.comp, na.rm = TRUE))
```

3 Conduct Statistical Analysis - Regression (Section Total: 25pt)

(1pt). Review the prompt at the beginning. What is the dependent variable? What is/are the independent variable(s)?

The dependent variable is the bedside manner ratings. The different hospital branches is the independent variable.

(1pt). What level of measurement is the dependent variable?

Interval and continuous

(1pt). What level of measurement is/are the independent variable(s)?

Interval and continuous

(1pt). What two equivalent statistical models can you use?

ANOVA & T tests & Regression (Since you demonstrated in class that they all return the same values)

3.1 Identify Assumptions - Regression

(2pt). What are the assumptions of simple linear regression with a binary independent variable.?

Linearity, Independence of Errors(No autocorrelation), Normality(Normally Distributed residuals with a mean of 0), and Homoscedasticity.

3.2 Inspect for Violations of Assumptions

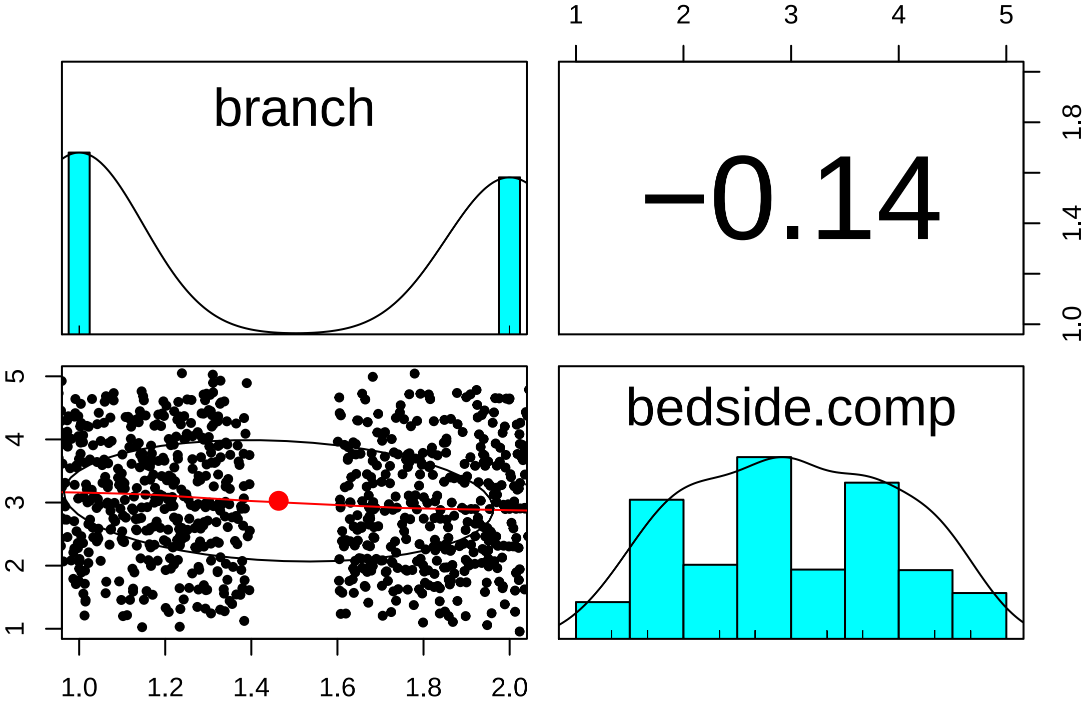
(1pt). Conduct a simple linear regression where you test whether there is a difference in bedside manner between hospital branches. Save this with the name `my.reg`.

```
my.reg <- lm(bedside.comp ~ branch, data = Slytherin1)
```

(8pt). Check the assumptions of the regression model, and describe what you see for each one.

```
#Linearity(In the case of a linear regression with a binary independent variable:it will ALWAYS be line

Slytherin1 %>% select(branch, bedside.comp) %>% pairs.panels(jiggle=TRUE)
```



They appear to be more or less linear thus this assumption holds up to be true and would be okay with moving forward with this test. Also, being that our independent variable is binary, then it will always be linear.

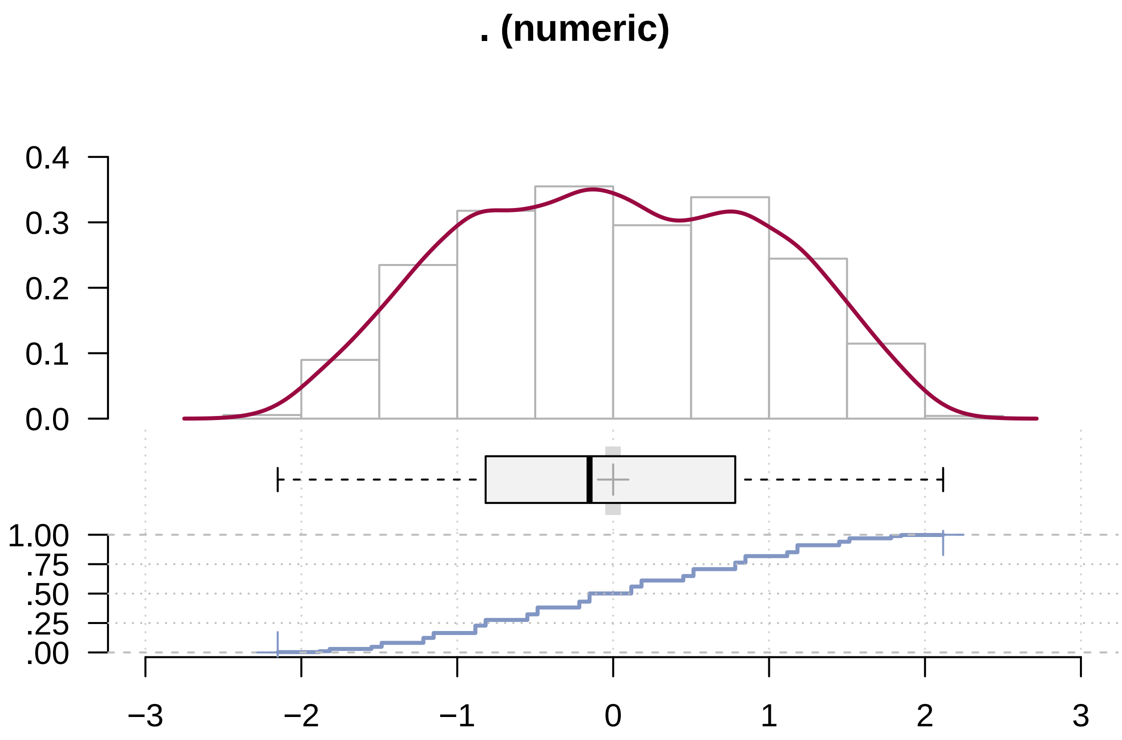
```
#Independence of Errors:No autocorrelation
DurbinWatsonTest(my.reg)
```

```
## 
##  Durbin-Watson test
## 
## data:  my.reg
## DW = 2.0177, p-value = 0.632
## alternative hypothesis: true autocorrelation is greater than 0
```

The Durbin Watson test indicates that there is no autocorrelation meaning that the residuals are independent from eachother. Our p value is not significant which overall shows that we cannot reject our null hypothesis of no autocorrelation which means that the residuals are found to be independent from eachother.Also,our DW value is less then 2.5 and more than 1.5 therefore we can determine the residulals are independent.This assumption of the linear regression model is found to be true so we can move forward with this test since there is no violation of the assumption of no autocorrelation.

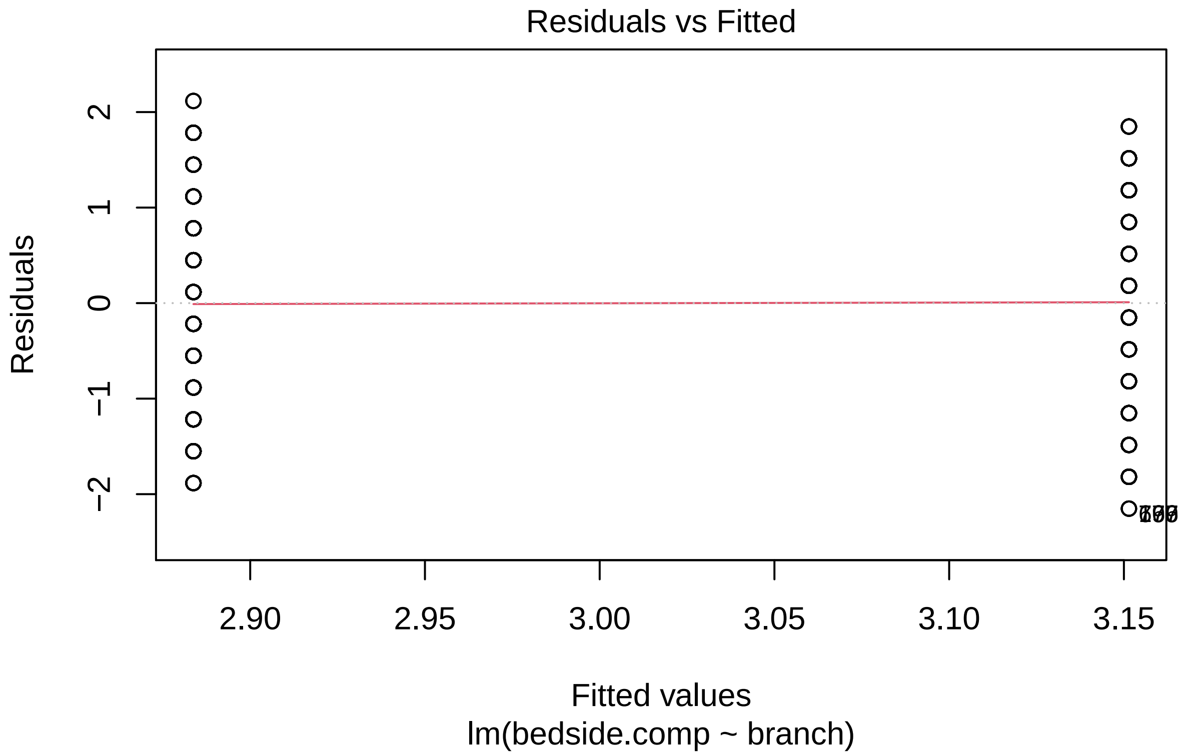
```
#Normality (Normally Distributed residuals with a mean of 0)
resid(my.reg) %>% Desc
```

```
## ------------------------------------------------------------------------------
## . (numeric)
##
##       length           n            NAs      unique         0s        mean’
##        1’448       1’448              0          28          0    0.0000000
##                   100.0%           0.0%                   0.0%
##
##          .05         .10            .25      median        .75          .90
##   -1.4847705  -1.2170889     -0.8181038  -0.1514372  0.7829111    1.1818962
##
##        range          sd          vcoef         mad        IQR         skew
##    4.2676816   0.9523318  2.6315074e+16   1.0857353  1.6010149   -0.0033962
##
##       meanCI
##   -0.0490925
##    0.0490925
##
##          .95
##    1.5152295
##
##         kurt
##   -0.9141961
##
## lowest : -2.1514372 (4), -1.8837556 (11), -1.8181038 (28), -1.5504223 (26), -1.4847705 (49)
## highest: 1.4495777 (43), 1.5152295 (42), 1.7829111 (28), 1.8485628 (13), 2.1162444 (3)
##
## heap(?): remarkable frequency (7.0%) for the mode(s) (= -0.151437151437151)
##
## ’ 95%-CI (classic)
```



This is considered to pretty much be a normal distribution; this assumption is true.This graph indicates an extremely close to normal distribution. The mean is not perefectly 0 but it is really close.The violations seem very mild therefore I feel comfortable with moving forward without worrying much about violations of normality in the residuals negatively effecting my results.

```
#Homoscedascity
plot(my.reg,1)
```



```
lmtest::bptest(my.reg)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  my.reg
## BP = 0.15125, df = 1, p-value = 0.6973
```

There is a pattern here so the homoscedasticity assumption appears to be violated but it is displayed like this due to having only these two variables. Also, the horizontal line in the fitted vs residuals plot indicates little-to-no violation of homoscedasticity. The BP test also demonstrates that we failed to reject the null hypothesis thus indicating little-to-no violation of homoscedasticity. Therefore,I wouldn't anticipate that my results would be any less trustworthy because of a violation of the assumption of homoscedasticity.

3.3 Interpret Results of Statistical Analysis - Regression

```
# Run this code:
my.reg %>% summ(confint = TRUE)
```

```
## MODEL INFO:
## Observations: 1448
## Dependent Variable: bedside.comp
## Type: OLS linear regression
```

```
##
## MODEL FIT:
## F(1,1446) = 28.43, p = 0.00
## R² = 0.02
## Adj. R² = 0.02
##
## Standard errors: OLS
## ----------------------------------------------------------
##                          Est.    2.5%   97.5%   t val.      p
## ------------------ ------- ------- ------- -------- ------
## (Intercept)              3.15    3.08    3.22    92.21   0.00
## branchremote            -0.27   -0.37   -0.17    -5.33   0.00
## ----------------------------------------------------------
```

(1pt). What percentage of the variability of bedside manner is explained by hospital branch?

The percentage of the variability of the bedside manner explained by hospital branch is 0.02% or 2%.

(1pt). How large of an effect size is this?

It could depend on the context but for the sake of predicting bedside manner(from hospital branch). The effect size is much smaller than I would like to see.

(2pt). What is the "reference level" in this regression (i.e. what level of the `branch` variable does the intercept represent)? What does the number of the intercept mean?

The level of the 'branch' variable that the intercept represents is the 'main' level. The number of the intercept is also the mean of our branch main group in the t test.

(1pt). Complete the following equation. What does the number mean?

3.15 - 0.27= 2.88; this number returns back the mean of our remote group in the t-test.

(1pt). What does the estimate (i.e. slope) of the branchremote tell you?

The slope(-0.27) also tells us the slope of the line tells us the rate of change of y relative to x. It appears to also tell us the difference between the means of the branch remote and branch main being that those means subtracted from eachother returned back a value of 0.27.

3.4 Model Inference - Regression

(3pt). Give a precise and accurate definition of the p-value for the slope of the branchremote term.

branchremote: the probability of observing a relationship between hospital branch and bedside manner this large or larger just due to sampling variability if the null were true.

(1pt). What is the null hypothesis that goes with the p-value from the last question?

There is no relationship between hospital branch and bedside manner at the population level.

4 Conduct Statistical Analysis - Independent Samples T-Test (Section Total: 12pt)

4.1 Identify Assumptions - Independent Samples T-Test

(3pt). What are the assumptions of the independent samples t-test?

Independence, Normality, and Homogeneity of variance(variances of dependent variables should be equal).

4.2 Interpret Results of Statistical Analysis - Independent Samples T-Test Vs Regression

Run the chunks below to show the results of a t.test with the results of your regression. Then answer the questions that follow. (Note that for the t-test, you will have to replace `my.data` with whatever your dataset is called.)

```
t.test(bedside.comp ~ branch, Slytherin1)
```

```
##
##  Welch Two Sample t-test
##
## data:  bedside.comp by branch
## t = 5.3359, df = 1419.6, p-value = 1.106e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1692732 0.3660900
## sample estimates:
##   mean in group main mean in group remote
##             3.151437             2.883756
```

```
my.reg %>% summ(confint = TRUE, digits = 10)
```

```
## MODEL INFO:
## Observations: 1448
## Dependent Variable: bedside.comp
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,1446) = 28.4272682748, p = 0.0000001127
## R² = 0.0192802106
## Adj. R² = 0.0186019812
##
## Standard errors: OLS
## ------------------------------------------------------------------
##                               Est.            2.5%           97.5%
## ------------------ ---------------- ---------------- ----------------
## (Intercept)           3.1514371514    3.0843962824    3.2184780205
## branchremote         -0.2676815628   -0.3661648772   -0.1691982483
## ------------------------------------------------------------------
##
## -------------------------------------------------
##                             t val.              p
## ------------------ ---------------- --------------
## (Intercept)          92.2105847868   0.0000000000
## branchremote         -5.3317228243   0.0000001127
## -------------------------------------------------
```

- **(3pt). Give a precise and accurate definition of the p-value from the independent samples t-test.**

The probability of observing a relationship between hospital branch and bedside manner this large or larger just due to sampling variability if the null were true.If the p-value is very low (< alpha level), you reject the *null hypothesis and conclude that there's a statistically significant difference.In this way, T and P are linked. They are different ways to quantify the "extremeness" of our results under the null hypothesis. You can't change the value of one without changing the other in a t-test. The larger the absolute value of the t-value, the smaller the p-value, and the greater the evidence against the null hypothesis.*

- **(0.5pt). Where in the output from the t-test can you find the intercept from your regression?**

We can find the regression intercept in the t test as the mean in group main which is 3.15.

- **(0.5pt). Where in the output from the t-test can you find the slope for the branchremote regression term?**

The slope for the branchremote regression term can be found in the T-test when subtracting the means for remote and main groups but it just comes out positive.

- **(1pt). What does the 95% confidence interval pertain to in the t-test output? Why is this confidence interval the same as in the regression table for the branchremote term?**

The 95% CI in the t-test pertains to the exact same CI's in the regression and within this interval lies our slope which we can then come to the conclusion that our confidenece intervals in both the t test and regression not only are the same BUT do capture the true relationship(the slope= .27,-.27) between bedside manner and hospital branch. The CI is the same in the regression table because we are testing the same thing, same data, same relationship etc but through two equal measures that are both null hypothesis tests. (F=t anyway)

- **(0.5pt). What do you notice that's interesting about the p-value from the t-test?**

The p-value from the t-test is again exactly the same as the one from the regression model;regression= p = 0.0000001127 & T-test = p = 1.106e-07. Again demonstrating the equivalence of these two tests.

- **(0.5pt). What do you notice that's interesting about the t-value from the t-test?**

The interesting thing about the t-value from the t test is that the t-value in the branchremote portion of the regression output is the exact same number:5.33 except just negative. This only further supports how equal these tests are to eachother because all the outputs of numbers are the same but in different places.

(1pt). Create a version of you dataset called `my.data.remote` that only contains data from the remote branch, and another version called `my.data.main` that contains only values from the main branch.

```r
library(tidyr)
my.data.remote <- Slytherin1 %>% filter(branch == 'remote')
my.data.main <- Slytherin1 %>% filter(branch == 'main')
my.data.remote %>% head
```

```
##   patient.ID location m.1 bedside.2 bedside.3 branch bedside.1 bedside.comp
## 1          2   dist.2   2         3         4 remote         4     3.666667
## 2          3   dist.2   4         3         2 remote         2     2.333333
## 3          4   dist.2   2         4         3 remote         4     3.666667
## 4          5   dist.2   3         3         3 remote         3     3.000000
## 5          7   dist.2   3         2         2 remote         3     2.333333
## 6         12   dist.2   2         5         5 remote         4     4.666667
```

```
my.data.main %>% head
```

```
##   patient.ID location m.1 bedside.2 bedside.3 branch bedside.1 bedside.comp
## 1          1   dist.1   2         3         4   main         4     3.666667
## 2          6   dist.1   3         3         3   main         3     3.000000
## 3          8   dist.1   1         5         5   main         5     5.000000
## 4          9   dist.1   3         3         2   main         3     2.666667
## 5         10   dist.1   2         4         4   main         4     4.000000
## 6         11   dist.1   4         2         2   main         2     2.000000
```

(1pt). Use ?DescTools (i.e. the R documentation) to find a command in the DescTools package that can calculate Cohen's D measure of effect size. Calculate Cohen's D and show the output.

```
?DescTools

CohenD(x= my.data.remote$bedside.comp, y= my.data.main$bedside.comp)
```

```
## [1] -0.280983
## attr(,"magnitude")
## [1] "small"
```

(1pt). Are you surprised by this effect size, given what you learned from running this test as a regression before? Why or why not?

This effect size is bigger than the effect size given by the regression but this value at this same time is equivalent to our slope(or rather very very close -0.27) that is given by branchremote in our regression model. Maybe, I should not be surprised then because this returned our slope which was also the difference between our means in the t test. 0.2(in our case -0.28) is considered a small effect size which is on par with the effect size that was calculated by the regression being quite small too:.02. One can conclude that our slope is also representative of our effect size and this number also falls within our CI so it confirms a true relationship but just a small effect. (I should also not be surprised anyway because the formula for cohen's d is M1-M2/spooled).

5 Visualize Results (Section Total: 5pt)

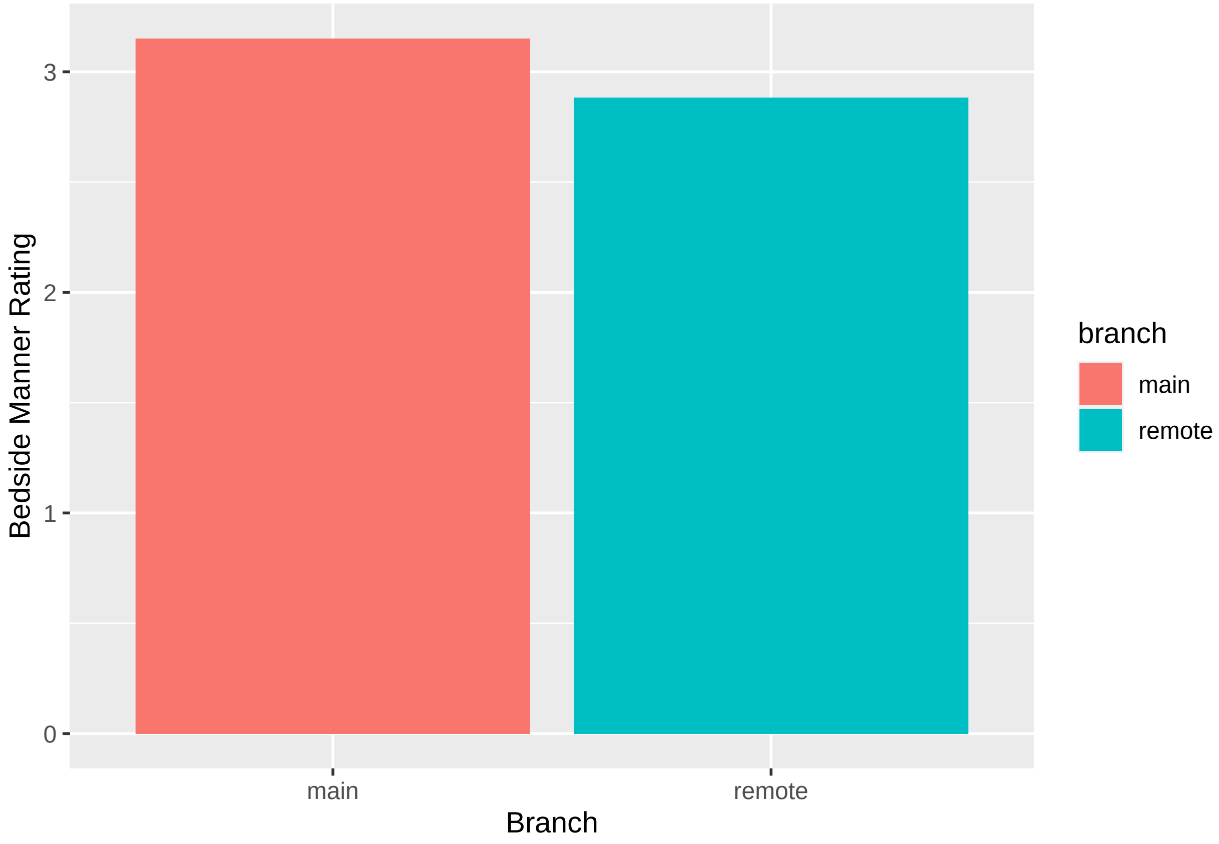
In the chunks below, I've used `ggplot2` to create a bar graph, a density plot, and a box-and-whisker plot. Run the code in each chunk, and take a look at each plot. Note that you will have to uncomment the code and replace `my.data` with whatever you named your dataset.

```
#bar plot
my.bar.plot <- ggplot(Slytherin1, aes(x = branch, y = bedside.comp, fill = branch)) +
   geom_bar(stat = "summary", fun.y = "mean") +
   labs(x = "Branch", y = "Bedside Manner Rating")
```

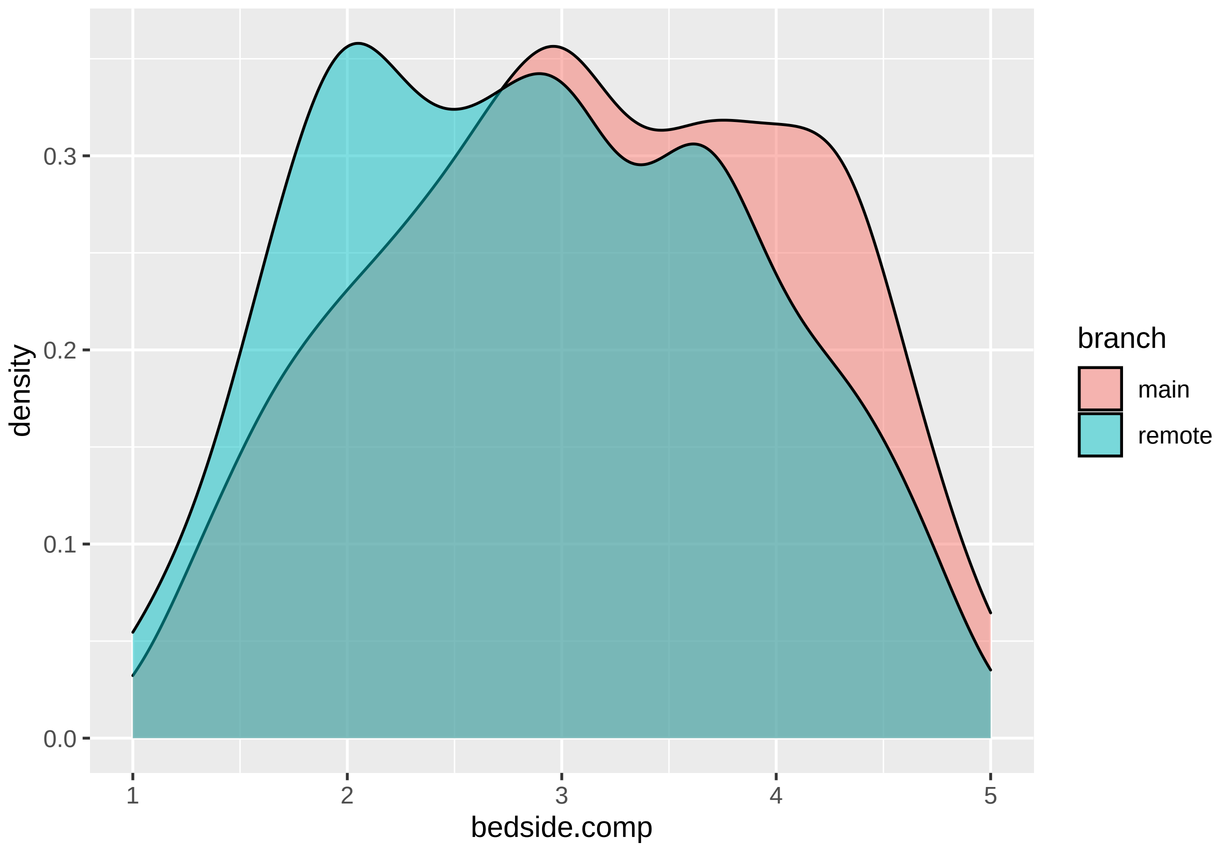
```
## Warning: Ignoring unknown parameters: fun.y
```

```
 my.bar.plot
```

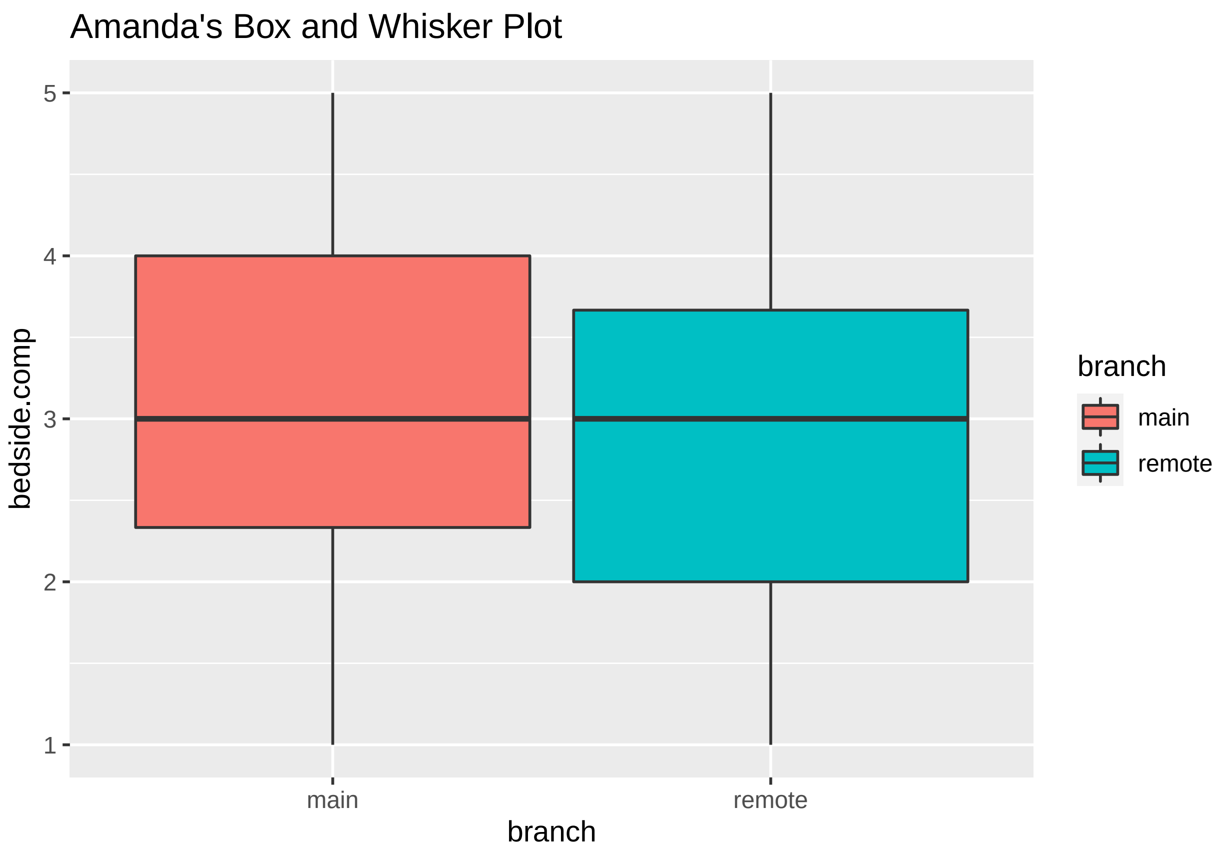
```
## No summary function supplied, defaulting to `mean_se()`
```



```
# Density plot
my.dens <-
   ggplot(Slytherin1, aes(x = bedside.comp, fill = branch)) +
   geom_density(alpha = 0.5)
 my.dens
```



```
# Box-and-whisker plot
my.box <-
   ggplot(Slytherin1, aes(x = branch, y = bedside.comp, fill = branch)) +
   geom_boxplot(width = 2) +
   labs(title = "Amanda's Box and Whisker Plot")
 my.box
```

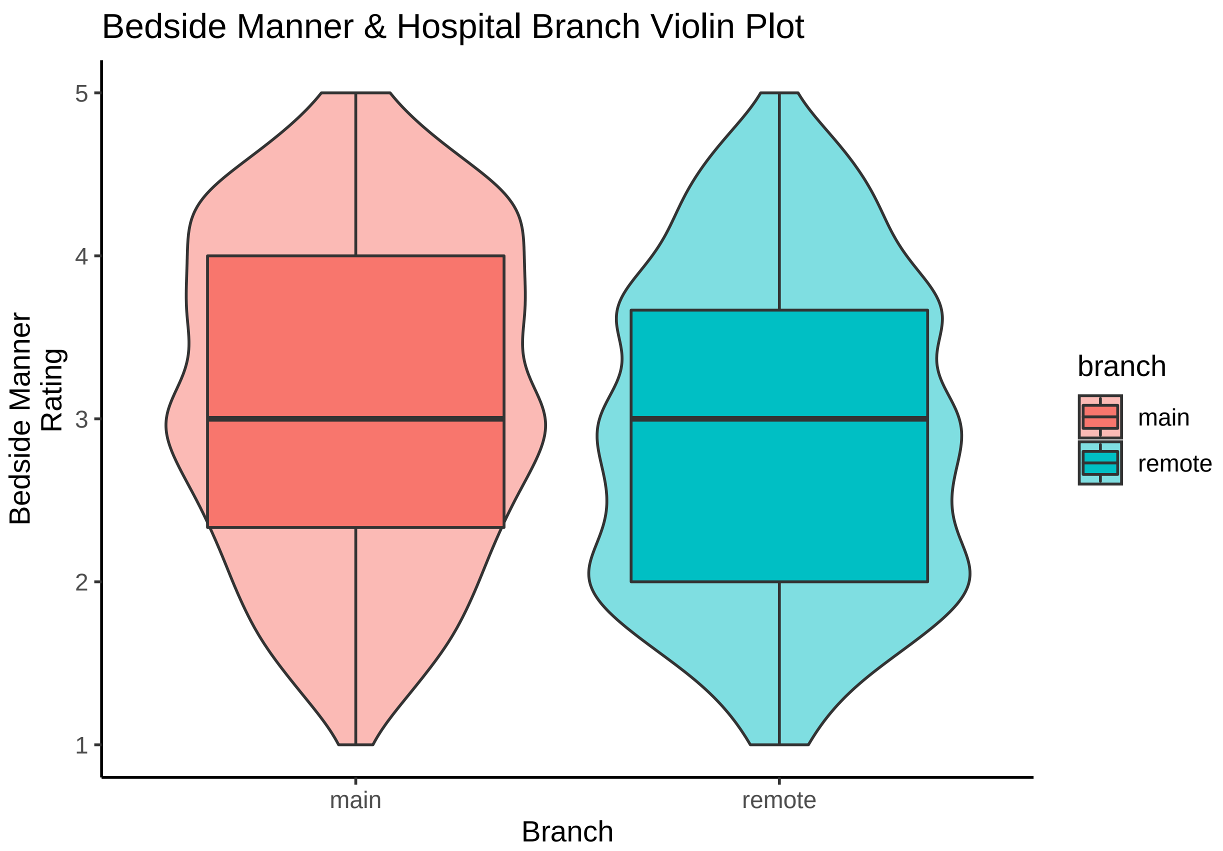


(5pt). Now, create a violin plot using ggplot. Call it "my.violin". Use the clues I've given you to construct a violin plot. Do the following to your plot:

- Make the colors correspond to the branch.
- Add a boxplot overlay, and set the width to make it look nice on top of the violin plot.
- Make the violins a bit see-through.
- Add an appropriate title to the plot, as well titles for the x-axis and y-axis.
- Add a theme.

```
# your violin plot
my.violin <- ggplot(Slytherin1, aes(x=branch, y=bedside.comp, fill=branch))+
  geom_violin(alpha=0.5)+
  geom_boxplot(width=0.7)+
  labs(title="Bedside Manner & Hospital Branch Violin Plot", x= "Branch",y="Bedside Manner
       Rating")+
  theme_classic()

my.violin
```



6 Summarizing Your Findings (Section Total: 4pt)

(4pt). Based on what you found, give a plain-english explanation of your findings that your statistics-illiterate hospital board members will be able to understand, and provide a firm explanation - as well as any caveats - concerning the relative levels of rudeness of doctors at the two hospital branches. Provide a recommendation about whether you think the board should intervene in some way to decrease rudeness (or increase bedside manner) at the remote branch of the hospital.

The relationship between hospital branch and bedside manner is considerably weak but positive.(This relationship is true though as confirmed by our slope value falling wihin confidence intervals and the p-values still being significant) Yes - I would say that bedside manner is predicted by hospital branch but on a very small scale that might not be translatable to our hospitals nor Doctors/nurses etc.Also, only 2% of bedside manner is explained by hospital branch. That means that 98% of bedside manner ratings are still not being accounted for. We did find that the effect size as a result of the CohenD test revealed a stronger but still small negative effect size -0.28 which was equivalent to the amount of change and difference between our average scores for the remote and main branches. I would recommend considering and looking at other possible variables that contribute to or rather could also explain bedside manner ratings. In simpler terms, we should evaluate what other factors could contribute to the bedside manner ratings for each branch(especially understanding why the remote branch has lower ratings) thus posibly revealing a stronger relationship than what we found. I believe that it is time to intervene in some way to increase bedside manner at our remote hospital branch and then we can reevaluate if those changes implemented have positively impacted the bedside manner ratings for the remote branch.