Correlation

Put your name here

Put the date here

Introduction

In this assignment we will learn how to run a correlation analysis. Correlation measures the strength of the linear relationship between two numerical variables.

Instructions

Presumably, you have already created a new project and downloaded this file into it. Please knit the document and work back and forth between this R Markdown file and the PDF output as you work through this module.

When you are finished with the assignment, knit to PDF one last time, proofread the PDF file **carefully**, export the PDF file to your computer, and then submit your assignment.

Sometimes you will be asked to add your own R code. That will appear in this document as a code chunk with a request for you to add your own code, like so:

```
## Add code here to [do some task]...
```

Be sure to remove the line `## Add code here to [do some task]...` when you have added your own code.

Sometimes you will be asked to type up your thoughts. That will appear in the document as follows:

Please write up your answer here.

Again, please be sure to remove the line "Please write up your answer here" when you have written up your answer. In these areas of the assignment, please use contextually meaningful full sentences/paragraphs (unless otherwise indicated) and proper spelling, grammar, punctuation, etc. This is not R code, but rather a free response section where you talk about your analysis and conclusions. If you need to use some R code as well, you can use inline R code inside the block between `\begin{answer}` and `\end{answer}`, or if you need an R code chunk, please go outside the `answer` block and start a new code chunk.

Load Packages

We load the standard `mosaic` package as well as the `reshape2` package for the `tips` data and the `OIdata` package for the `state` data. The `broom` package gives us tidy output.

```r
library(reshape2)
library(OIdata)
data(state)
library(broom)
library(mosaic)
```

We set the seed to make our results reproducible.

```r
set.seed(11)
```

Research question

Is there is a correlation between the size of a restaurant bill and the size of the tip?

Correlation

The word correlation describes a linear relationship between two numerical variables. As long as certain conditions are met, we can calculate a statistic called the Pearson correlation coefficient, denoted R. This value will be some number between -1 and 1. Coefficients close to zero indicate little or no correlation, coefficients close to 1 indicate strong positive correlations, and coefficients close to -1 indicate strong negative correlations. In between, we often use words like weak, moderately weak, moderate, and moderately strong. There are no exact cutoffs for when such words apply. You must learn from experience how to judge scatterplots and R values to make such determinations.

Let's examine a data set called `tips` from the `reshape2` package. Since there is also a `tips` data set in the `openintro` package, we'll use a trick to make sure we get the right one. The double colon is placed between the name of the package and the name of the data frame:

```r
tips <- reshape2::tips
```

These 244 observations were collected by one waiter over a period of a few months working in a restaurant. Our research question asks us to consider the variables `total_bill` and `tip`.

If all we wanted was the value of R, we could find it by using the `cor` command.

```r
cor(tip ~ total_bill, data = tips)
```

```
## [1] 0.6757341
```

This sample correlation R is an estimate of the true population correlation, called ρ, the Greek letter "rho". A typical null hypothesis is that there is no correlation between the two variables—in other words, $\rho = 0$. Under that assumption, the sampling distribution is somewhat complicated. Although the sample correlations don't follow a simple distribution, if we calculate

$$t = \frac{R - \rho}{\sqrt{\frac{1-R^2}{n-2}}} = \frac{R}{\sqrt{\frac{1-R^2}{n-2}}},$$

then the values of t follow a Student t distribution with $n - 2$ degrees of freedom. (The last step above takes into account the fact that the null value for ρ is zero.)

We can verify this with a basic simulation. First, we shuffle the values of `total_bill` to remove any association with tips to simulate the assumption of the null hypothesis. Here are a few examples:

```r
cor(tip ~ shuffle(total_bill), data = tips)
```

```
## [1] 0.005853122
```

```
cor(tip ~ shuffle(total_bill), data = tips)
```

```
## [1] 0.007911685
```

```
cor(tip ~ shuffle(total_bill), data = tips)
```

```
## [1] 0.1027245
```

We use the `do` command to do this a bunch of times.

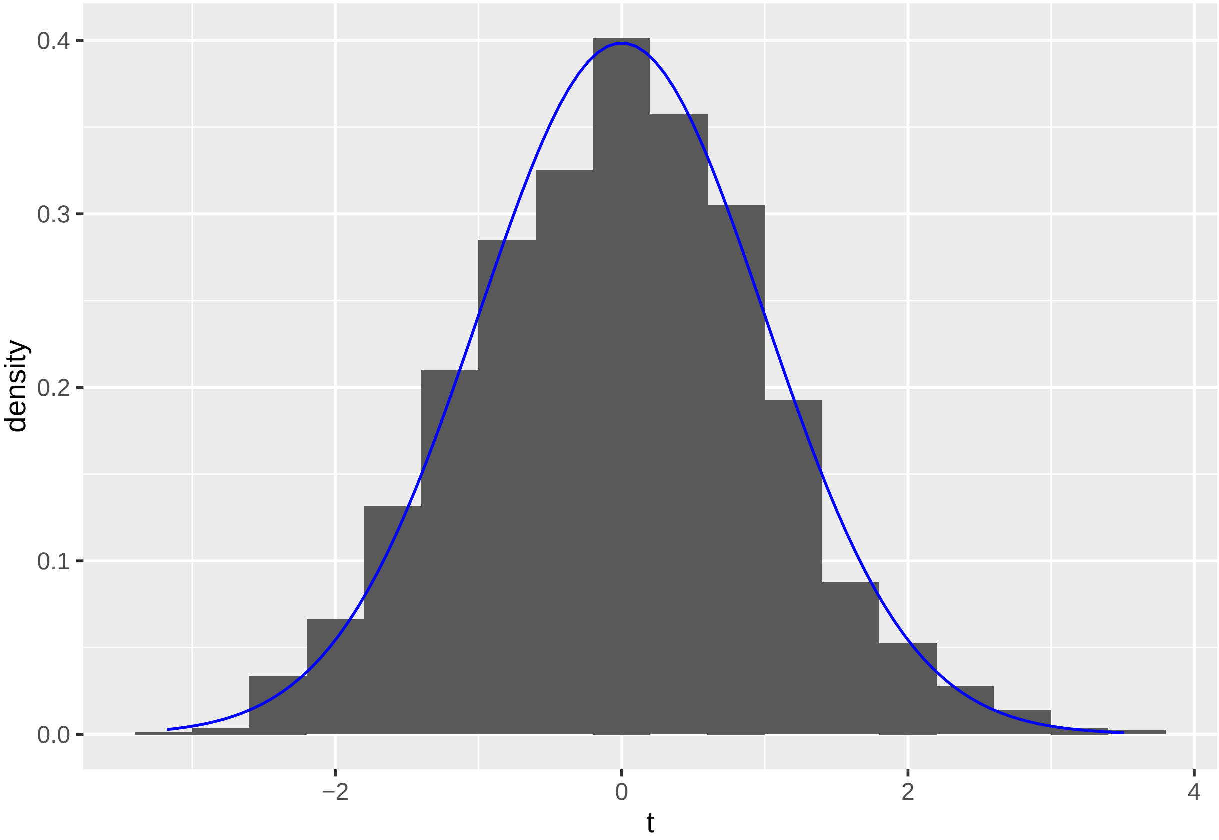
```
sims <- do(2000) * cor(tip ~ shuffle(total_bill), data = tips)
```

The t scores follow a Student t distribution, not the correlations themselves, so we have to calculate the t scores. We use the `mutate` command to compute the t score for each row of `sims`. The number 242 is $n - 2$.

```
sims <- mutate(sims, t = cor/(sqrt((1 - cor^2)/242)))
```

Now we can graph the simulated values. We superimpose the t distribution with $df = 242$ to show that it's a pretty good fit.

```
ggplot(sims, aes(x = t)) +
    geom_histogram(aes(y = ..density..), binwidth = 0.4) +
    stat_function(fun = "dt", args = list(df = 242), color = "blue")
```



Inference for correlation

Calculating a correlation coefficient blindly can be dangerous. Without meeting certain conditions, the value of R could be incredibly misleading. R will gladly compute the correlation coefficient for any data, whether appropriate or not. Therefore, we will follow our inferential rubric to decide if there is a statistically significant relationship between the bill and the corresponding tip. In truth, the entire inferential rubric is probably overkill for such a simple question. Nevertheless, the rubric does ensure that we take care to identify our hypotheses and check conditions.

In addition to the standard "Random" and "10%" conditions, we introduce two new conditions. First, we need to know that the association is linear. Nonlinear relationships can exist, but the R value makes no sense for such situations. Finally, we need to check for outliers. These two conditions should be checked by looking at a scatterplot.

Exploratory data analysis

Use data documentaton (help files, code books, Google, etc.), the str command, and other summary functions to understand the data.

[Type `library(reshape2)` then `?tips` at the Console to read the help file.]

```
str(tips)
```

```
## 'data.frame':    244 obs. of  7 variables:
##  $ total_bill: num  17 10.3 21 23.7 24.6 ...
##  $ tip       : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...
##  $ sex       : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 2 2 2 2 2 ...
##  $ smoker    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ day       : Factor w/ 4 levels "Fri","Sat","Sun",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ time      : Factor w/ 2 levels "Dinner","Lunch": 1 1 1 1 1 1 1 1 1 1 ...
##  $ size      : int  2 3 3 2 4 4 2 4 2 2 ...
```

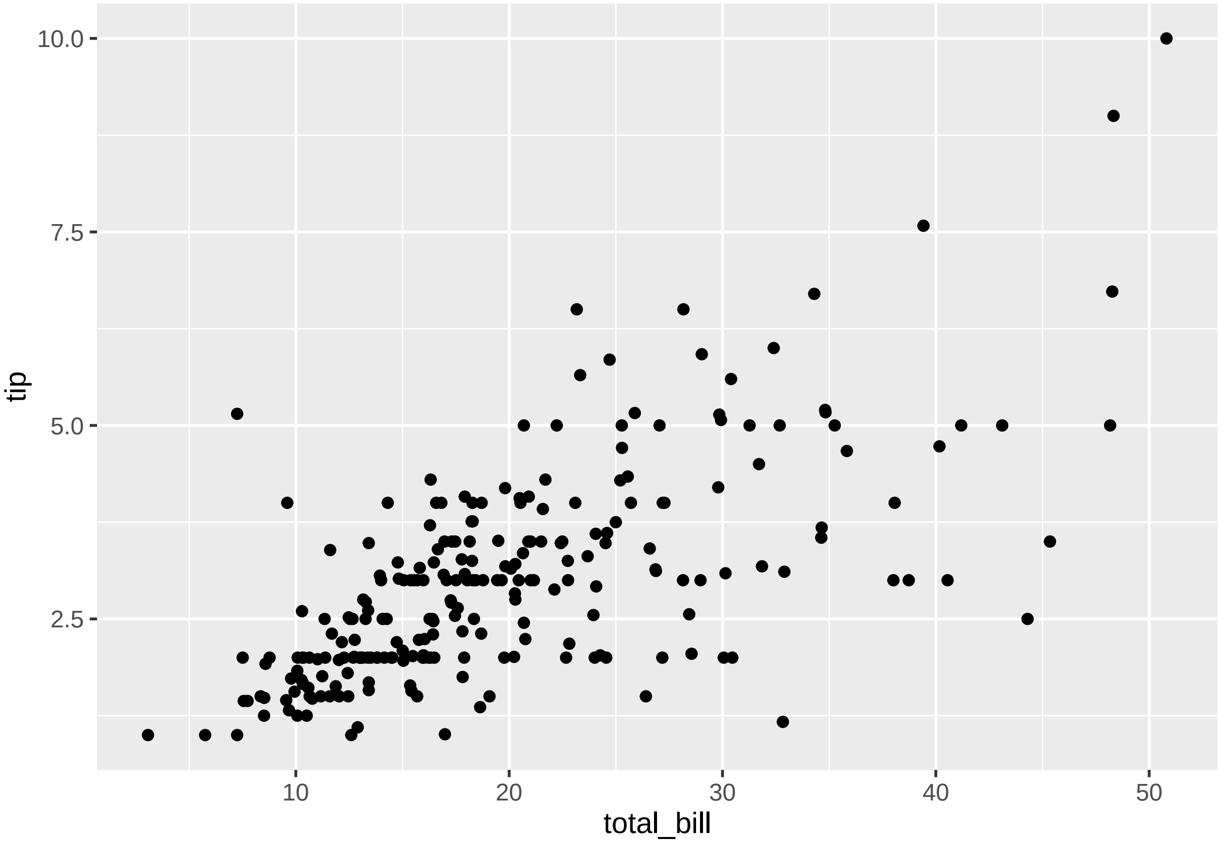
Prepare the data for analysis.

The two variables of interest are coded correctly as numerical variables in the `tips` data frame, so we don't need to do anything for this step.

Make tables or plots to explore the data visually.

The appropriate plot for two numerical variables is a scatterplot. We are thinking of `total_bill` as the explanatory variable and `tip` as the response variable.

```
ggplot(tips, aes(x = total_bill,  y = tip)) +
    geom_point()
```



There does appear to be a moderate, positive association between these variables.

Exercise

There appears to be a pattern of unusual bunching in the lower left part of the graph. Can you explain this pattern? (Hint: use `View(tips)` at the Console and sort the `tip` variable.)

Please write up your answer here.

Hypotheses

Identify the sample (or samples) and a reasonable population (or populations) of interest.

The sample consists of 244 meals a waiter served over the course of several months working at a restaurant. The population is presumably all meals this waiter might ever serve at this restaurant. (It would not make sense to include other servers or other restaurants in this population as bills and tips vary widely from person to person and restaurant to restaurant.)

Express the null and alternative hypotheses as contextually meaningful full sentences.

H_0: There is no correlation between the total bill and the tip.

H_A: There is a correlation between the total bill and the tip.

Express the null and alternative hypotheses in symbols.

$H_0 : \rho = 0$

$H_A : \rho \neq 0$.

Commentary: We are performing a two-sided test here. One could perform a one-sided test if the question of interest was about a positive or a negative correlation specifically. Unless otherwise specified, though, the default is to run a two-sided test.

Model

Identify the sampling distribution model.

We use a t model with 242 degrees of freedom.

Check the relevant conditions to ensure that model assumptions are met.

- Random
  - This is not a random sample, but over several months, it seems reasonable that this is representative of this waiter's experiences at this restaurant.
- 10%
  - Assuming the waiter works at this restaurant for several years, 244 meals is probably less than 10% of all meals he will serve.
- Linear association
  - The scatterplot shows a reasonably linear pattern.
- Outliers
  - We don't see any significant outliers. There are a few dots here and there that are a little far from the main cloud, but nothing that worries us too much, especially given the large sample size.

Commentary: No data will ever line up in a perfect straight line. The "linear association" condition is meant to suggest that the "cloud of dots" should be more or less in a straight pattern moving across the plot. We are most concerned here with checking that the pattern does not curve substantially, and this does not appear to.

Mechanics

Compute the test statistic.

```r
cor_test <- tidy(cor.test(tips$total_bill, tips$tip))
t <- cor_test$statistic
```

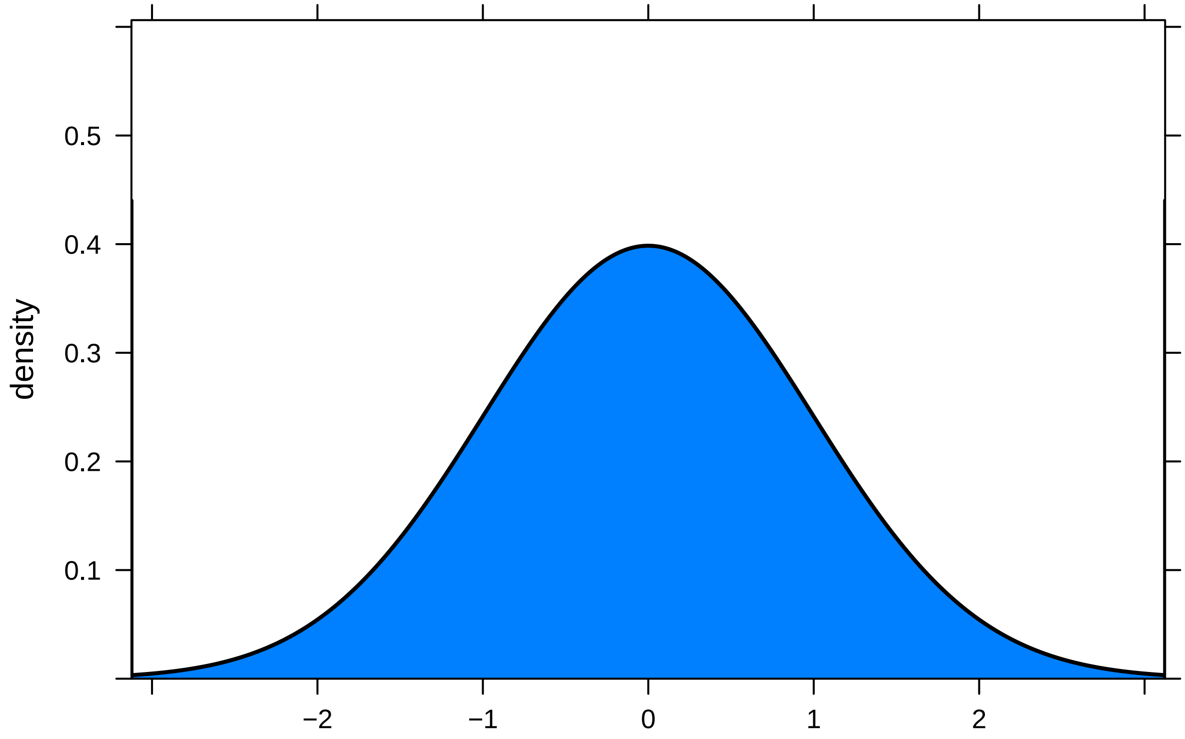
The t score is 14.260355.

Commentary: Unlike the other hypothesis tests we've done before, `cor.test` doesn't use the "tilde" notation. You just put in both variables separated by a comma. The order doesn't matter; correlation is symmetric, so the R value is the same independent of the choice of explanatory and response variables.

You may have noticed that this is an insanely large t score. This is typical of correlation tests. If there is enough visual evidence of a correlation in the scatterplot, the R value will be pretty far from 0. That's why the full rubric for inference is somewhat overkill for questions about correlation.

Plot the null distribution.

```
pdist("t", df = cor_test$parameter, q = c(-t, t))
```



```
## [1] 3.346235e-34 1.000000e+00
```

Calculate the P-value.

```
cor_test$p.value
```

```
## [1] 6.692471e-34
```

$P < 0.001$

Commentary: $P < 0.001$ is quite the understatement. The P-value is 0.000…00067 with 33 zeros after the decimal point!

Conclusion

State the statistical conclusion.

We reject the null hypothesis.

State (but do not overstate) a contextually meaningful conclusion.

There is sufficient evidence that there is a correlation between the total bill and the tip (for this waiter at this restaurant).

Identify the possibility of either a Type I or Type II error and state what making such an error means in the context of the hypotheses.

If we've made a Type I error, then there is actually no correlation between total bill and tips, but this data shows one.

Confidence interval

Conditions

All the conditions have already been checked.

Calculation

```
cor_test$conf.low
```

```
## [1] 0.6011647
```

```
cor_test$conf.high
```

```
## [1] 0.7386372
```

Conclusion

We are 95% confident that the true correlation between total bill and tips (for this waiter at this restaurant) is captured in the interval (0.6011647, 0.7386372).

Your turn

The `state` data from the `OIdata` package has a number of variables collected from various sources. There are 51 rows, representing the 50 states and the District of Columbia. Run a correlation test to determine if the median household income in each state is correlated with the percentage of the state's population that smokes.

There is something unusual about this example that you will need to consider in your answer. The sample is 50 states and the District of Columbia. The population is tricky though because these states don't represent some larger groups of states; we already have all the states in our data. One can think of this data, though, as a snapshot of what was true in each state at one point in time. Therefore, the population can be thought of as similar measurements taken at other times.

This also makes it difficult to check conditions. We do not have a random sample of states (as we have all of them), but remember that we're thinking of this as a random sample across a number of years in which we might have gathered this data. Having said that, I imagine that median income goes up every year with inflation, so it may or may not be representative of other years. The 10% condition also requires some thought.