Bivariate Hypothesis Testing

Dr. Sarah Hunter

7/21/2020

Introduction to Bivariate Hypothesis Testing

In this week's classes, we have studied how to find a statistical relationship between two variables. You can do this in the form of a *bivariate hypothesis test*. In this document, we will discuss three basic tests: tabular analysis, difference of means, and a correlation coefficient. These can be done quickly and easily in R. This document shows step by step how to do each test.

The first step for this lab is to load the data used for all three tests. This follows the same structure as the previous documents.

```
#Setting the working directory
setwd("/Users/sarahhunter/Desktop/Data")

#Loading CSV data
mydata<-read.csv("nes2004subset3.csv")

#Getting summary statistics
summary(mydata)
```

```
##        X             religion          bush           female
##  Min.   :   1.0   Min.   :1.000   Min.   :0.000   Min.   :0.0000
##  1st Qu.: 302.5   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000
##  Median : 607.0   Median :1.000   Median :1.000   Median :1.0000
##  Mean   : 606.1   Mean   :2.311   Mean   :0.508   Mean   :0.5319
##  3rd Qu.: 908.5   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :1212.0   Max.   :7.000   Max.   :1.000   Max.   :1.0000
##                   NA's   :15      NA's   :396
##    unionhouse        partyid         eval_WoT         eval_HoE
##  Min.   :0.0000   Min.   :0.000   Min.   :-2.000   Min.   :-2.0000
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:-2.000   1st Qu.:-2.0000
##  Median :0.0000   Median :3.000   Median : 1.000   Median :-1.0000
##  Mean   :0.1724   Mean   :2.872   Mean   : 0.152   Mean   :-0.3467
##  3rd Qu.:0.0000   3rd Qu.:5.000   3rd Qu.: 2.000   3rd Qu.: 2.0000
##  Max.   :1.0000   Max.   :6.000   Max.   : 2.000   Max.   : 2.0000
##  NA's   :6        NA's   :17      NA's   :29       NA's   :36
##     ideology       bush_therm       education         income
##  Min.   :1.000   Min.   :  0.00   Min.   :0.000   Min.   : 1.00
##  1st Qu.:3.000   1st Qu.: 30.00   1st Qu.:3.000   1st Qu.:11.00
##  Median :4.000   Median : 60.00   Median :4.000   Median :16.00
##  Mean   :4.268   Mean   : 54.94   Mean   :4.305   Mean   :14.97
##  3rd Qu.:6.000   3rd Qu.: 85.00   3rd Qu.:6.000   3rd Qu.:20.00
##  Max.   :7.000   Max.   :100.00   Max.   :7.000   Max.   :23.00
##  NA's   :288                                      NA's   :141
```

```
#Get the first 5 lines of data

head(mydata)
```

```
##   X religion bush female unionhouse partyid eval_WoT eval_HoE ideology
## 1 1        7    0      0          1       3       NA        0        4
## 2 2        1    0      0          0       2       -1       -1        4
## 3 3        1    1      1          0       6        2        2        6
## 4 4        1   NA      0          0       3       -2       -1        4
## 5 5        1    1      1          0       6        2        2        6
## 6 6        1   NA      0          0       3       -2       -2        6
##   bush_therm education income
## 1         70         7     17
## 2         40         4     19
## 3        100         6     23
## 4         50         2      3
## 5        100         3     12
## 6         60         7     12
```

Tabular Analysis

Tabular analysis is a method of finding relationships between two *categorical* variables. This method is also frequently called "Cross Tabs". Tabular Analysis starts by looking at a table of categorical variables. The table basically breaks down how many (and/or what percent) of observations are in each two categories. For example:

Policy	Republican	Democrat
Support	285	60
Oppose	208	432

In the table above, you see that 285 Republicans and 60 Democrats support a certain policy while 208 Republicans and 432 Democrats oppose a certain policy. We use tabular analysis to see if there is a pattern in that table.

In R, it is easy to get the above table (called a frequency table) using the `table` command. For the rest of this section, we will use the above data set. The variables used for this example are bush (if the respondent voted for George W. Bush in 2004) and unionhouse (whether or not a member of the household is a member of a union). The following code demonstrates how to get a frequency table from R:

```
table(mydata$bush, mydata$unionhouse)
```

```
##
##       0   1
##   0 300  97
##   1 356  54
```

The rows of the tables refer the first variable entered in the table command. Therefore, the first row is when `bush`=0, while the first column is when `unionhouse`=0. So, the upper left cell of the table contains the number of observations for which the respondent did not vote for George W. Bush and they contain no union members in that household.

We can also get the proportion of observations that are in each cell with the `prop.table` command, which you can create from a `table` object.

```r
t<-table(mydata$bush, mydata$unionhouse)

prop.table(t)
```

```
##
##             0         1
##   0 0.3717472 0.1201983
##   1 0.4411400 0.0669145
```

The idea of tabular analysis is to compare the actual data to what we would expect if there were no relationship between the variables. For example, say 35 percent of the sample support a certain policy and 65 percent do not support the policy. We want to know if being a Republican is statistically related to supporting the policy. If there were no relationship between being a Republican and supporting the policy, we would expect Republicans to reflect the same level of support for the policy that is seen in the entire population. In this example, if there were no relationship between partisanship and policy support, 35 percent of Republicans and 35 percent of Democrats would support this policy. The goal is to see how different our actual outcomes differ from the expected outcome. The more different the outcome from the expect, the more confidently we can say that there is a systematic relationship between the two variables. We can measure that difference with a *test statistic*

The test statistic used with Tabular Analysis is call χ^2, or chi squared. This test does exactly as described above: it compares the actual data to the expected results if there were no relationship. χ^2 is calculated using simple mathematical formula that only uses basic arithmetic. The basic formula is:

$\chi^2 = \sum \frac{(O-E)^2}{E}$

where O stands for the Observed quantity and E stands for the Expected quantity if the null hypothesis were true. This calculation ($\frac{(O-E)^2}{E}$) is done for all cells of the table and added together (as indicated by the summation symbol ($\sum$)).

An Example

In our union membership and votes for George W. Bush example above, we can use the frequency and proportion tables above to help us calculate χ^2 by hand. The first thing we need to do is to make a table of what we would expect if there were no relationship between union membership and voting for George W. Bush. With our data, we can use the `table` command to find out the percentage of our sample voted for Bush:

```r
#The table command tells you how many people voted for Bush (bush=1)
#and how many did not vote for Bush (bush=0),

table(mydata$bush)
```

```
##
##   0   1
## 399 412
```

```r
#But, there is some missing data, so we will need to use the table command
#for both variables again:

table(mydata$bush, mydata$unionhouse)
```

```
##
##       0   1
##   0 300  97
##   1 356  54
```

```
#Then we use basic math to calculate the percentage of the
#sample voted for Bush (the row where bush=1)

410/(397+410)
```

```
## [1] 0.5080545
```

This tells us that 50.8% of our total sample voted for George W. Bush. If there is no relationship between union membership and vote choice, that same percentage should be reflected among union households. Therefore, we can fill out the table to look like:

Vote Choice	Union House	Not Union House
Bush	50.8%	50.8%
Not Bush	49.2%	49.2%

The chi^2 calculation requires that we have the value of the expected outcome. This means we have to find the *number* of people in each category. We can do this with basic math again, by multiplying the number of people in a union by 50.8% to get the expected number of union households that voted for George W. Bush. We get the total number of union households in the sample by adding the two rows together for union membership in the from the results of the `table` command above. See below for the example:

Vote Choice	Union House	Not Union House
Bush	50.8% * 151	50.8% * 656
Not Bush	49.2% * 151	49.2% * 656

We can calculate this in R by:

```
#Expected Union members that voted for Bush
.508*151
```

```
## [1] 76.708
```

```
#Expected Union members that did not vote for Bush
.492*151
```

```
## [1] 74.292
```

```
#Expected Non-Union members that voted for Bush
.508*656
```

```
## [1] 333.248
```

```
#Expected Non-Union members that did not vote for Bush
.492*656
```

```
## [1] 322.752
```

Then you can plug in the numbers to get the expected outcome table:

Vote Choice	Union House	Not Union House
Bush	76.708	333.248
Not Bush	74.292	322.752

The actual outcome in the data is shown in the following table:

Vote Choice	Union House	Not Union House
Bush	54	356
Not Bush	97	300

We then plug these numbers into the chi^2 equation:

$\chi^2 = \sum \frac{(O-E)^2}{E}$

$\chi^2 = \frac{(54-76.708)^2}{76.708} + \frac{(97-74.292)^2}{74.292} + \frac{(356-333.248)^2}{333.248} + \frac{(300-322.752)^2}{322.752}$

$\chi^2 = 6.722288 + 6.940899 + 1.553358 + 1.603874$

$\chi^2 = 16.8204$

That is our χ^2 value. However, that does not tell us if these results are *significantly* different from what we would expect if there were not relationship between union membership and vote choice. For this, we need to find the **degrees of freedom** which are essentially the amount of variance that we have to explain. In the χ^2 hypothesis test, we calculate the degrees of freedom by:

$d.f. = (r-1)(c-1)$

where r is the number of rows and c is the number of columns in the table. For our example, we have 2 rows and 2 columns and calculate the degrees of freedom as:

$d.f. = (2-1)(2-1) = 1$

Therefore, we have 1 degree of freedom. We use this information to then find the *critical value*. The critical value in this context is the minimal value of χ^2 at which statistical significance is achieved. You use the degrees of freedom and your chosen alpha level (or the probability of making a type I error). In political science, we usually look for .05. The next step is to look up the critical value in a table (found here https://www.mathsisfun.com/data/chi-square-table.html). Looking at the table for 1 degree of freedom and a type I error rate of .05, we find that the critical value of χ^2 is 3.841. Because our χ^2 value is 16.8204, which is greater than 3.841, we can reject the null hypothesis and conclude that there is a relationship between union membership and vote choice.

However, this does take time. Especially if there are more rows and columns. The good news is that R can estimate a tabular analysis easily and quickly.

Estimation

Estimating a tabular analysis model is quick and simply in R. You will first need to install two packages: `gmodels` and `gtools`. You then load those two packages. Using the command `CrossTable`, you can get a frequency table and the χ^2 score for the two variables.

```
#Install Packages needed (commented out here because already installed on my computer)

#install.packages("gmodels")
#install.packages("gtools")


#Load packages

library(gmodels)
library(gtools)

#The Model:
```

```
CrossTable(y=mydata$bush , x=mydata$unionhouse, prop.c=FALSE,
           prop.t=FALSE, prop.chisq=FALSE, chisq=TRUE, format="SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
##
## Total Observations in Table:  807
##
##                    | mydata$bush
## mydata$unionhouse |         0  |         1  | Row Total |
## ------------------|------------|------------|-----------|
##                 0 |       300  |       356  |       656  |
##                   |   45.732% |   54.268% |   81.289% |
## ------------------|------------|------------|-----------|
##                 1 |        97  |        54  |       151  |
##                   |   64.238% |   35.762% |   18.711% |
## ------------------|------------|------------|-----------|
##      Column Total |       397  |       410  |       807  |
## ------------------|------------|------------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  16.82047     d.f. =  1     p =  4.108774e-05
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  16.08816     d.f. =  1     p =  6.046096e-05
##
##
##        Minimum expected frequency: 74.28377
```

Interpretation

As you can see, R gives not only the actual χ^2 score, but also the p-value associated with that χ^2 value. In this case, the p-value is .0000410877, which much less than our .05 threshold. The correct interpretation of this is that, if the null hypothesis (no relationship) were true, we would get these results or something more extreme 0.00410877 percent of the time. Therefore, we can conclude that there is a systematic relationship between union membership and vote choice.

Difference of Means

While the χ^2 test is good for testing associations between two categorical variables, the difference of means test can test for associations between a categorical independent variable and a continuous dependent variable. A difference of means test can also be called a t-test. A difference of means tests essentially tells you how

different the mean of the dependent variable is in one category than in another. For example, say the average age of those that support one policy is 37 and the average age that opposes that policy is 58. how do we know that that difference is statistically significant? The difference of means test can tell us this. In this section, we will show an example of how to do a difference of means test by hand and then how to do the same calculations in R.

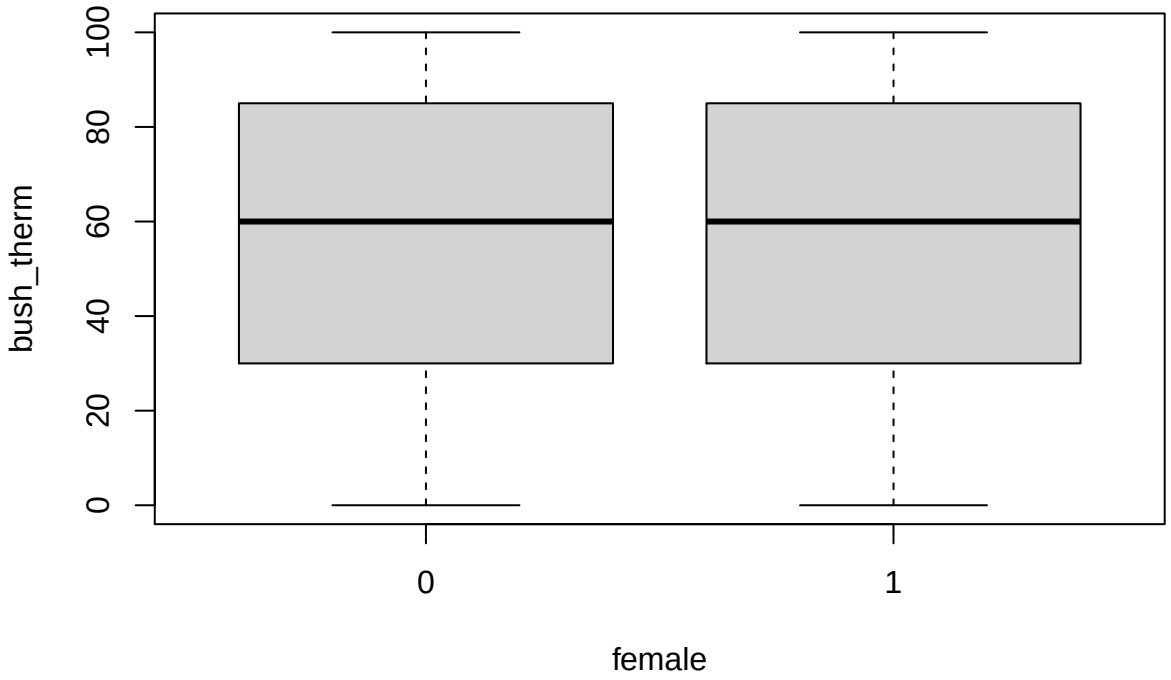
Example

The best way to start understanding how a difference of means test work is to visually look at the data. Our running example is that we believe that those that identify as female give a different score to President George W. Bush on the feeling thermometer, which is a 0-100 scale where 0 means the respondent has nothing nice to say about the subject and 100 means they have nothing negative to say. Respondents in this survey were all asked to give President Bush a score on the feeling thermometer.

To visually inspect the data to see if there is something there, we can use what is called a Box and Whisker plot. This is a plot that shows the distribution of a continuous variable. In R, this is a fairly easy plot:

```
#To make a basic boxplot

boxplot(bush_therm~female, data=mydata)
```



We can also use the `dplyr` package to get the means by group:

```
#install.packages("dplyr")

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
```

```
m<-mydata %>% group_by(female) %>%
  summarize_at(vars(bush_therm), funs(mean(., na.rm=TRUE), sd, length))
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
m
```

```
## # A tibble: 2 x 4
##   female  mean    sd length
##    <int> <dbl> <dbl>  <int>
## 1      0  56.6  32.8    565
## 2      1  53.5  34.2    642
```

So, we can see that the mean Bush Thermometer score for those respondents that identify as female is 53.52, while the mean for others is 56.56. From there, we need to calculate the *t-score*, which is the test statistic for the difference of means hypothesis test. The formula for the t-score is:

$t = \frac{\bar{Y}_1 - \bar{Y}_2}{se(\bar{Y}_1 - \bar{Y}_2)}$

where $\bar{Y}_1$ is the mean of Y in group 1, $\bar{Y}_2$ is the mean of Y in group 2, and $se(\bar{Y}_1 - \bar{Y}_2)$ is the *joint* standard error. The standard error is calculated as:

$se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

where n_1 is the number of observations in group 1, n_2 is the number of observations in group 2, s_1 is the standard deviation in group 1, and s_2 is the standard deviation in group 2.

Currently, we have all the pieces we need to calculate a t-score. The information we need is thus:

Group	Mean	Standard Deviation	Number of obs
female=0	56.55575	32.77019	565
female=1	53.52025	34.17844	642

And now we can plug those numbers into the equation for the standard error:

$se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{(565-1)32.77016^2+(642-1)34.17844^2}{565+642-2}} * \sqrt{\frac{1}{565} + \frac{1}{642}}$

$se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{605670.2+748794.3}{1205}} * \sqrt{\frac{1}{565} + \frac{1}{642}}$

$se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{1124.037} * 0.05768487$

$se(\bar{Y}_1 - \bar{Y}_2) = 1.933981$

With that standard error, we can now calculate the t-score:

$t = \frac{\bar{Y}_1 - \bar{Y}_2}{se(\bar{Y}_1 - \bar{Y}_2)}$

$t = \frac{56.55575 - 53.52025}{1.93398}$

$t = 1.569561$

Therefore, the value of our test statistic is 1.569561. However, we need to know the *critical value* for the t-score, the same process as the χ^2 test. In this case, you go find the t-score table (http://www.ttable.org). First, we need the degrees of freedom for the t-score. This is different than the χ^2 degrees of freedom calculation. In the case of the t-score, the degrees of freedom are the total sample size minus 2 ($d.f. = n - 2$). Therefore, the degrees of freedom here is 1205. We will also use the significance level of .05. One note on t-tests is that you must use the "two-tailed" .05 column. (We will discuss in class what this means). The table tells us that the critical value for the .05 significance level for 1000 degrees of freedom is 1.962. Our t-score is 1.569561, which is less than 1.962. As a result, we cannot reject the null hypothesis that the difference of means between the two groups is 0. We then cannot conclude that there is a relationship between Bush thermometer ratings and gender identity.

Estimation

R can do the above calculation easily and efficiently using the `t.test` command. Below shows the code. One thing to remember is the order of the variables. In the `t.test` command, you should always put the dependent variable first (before the ~ (called a tilde)).

```
t.test(mydata$bush_therm~mydata$female, alternative="two.sided", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mydata$bush_therm by mydata$female
## t = 1.5738, df = 1196.2, p-value = 0.1158
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7486974  6.8197034
## sample estimates:
## mean in group 0 mean in group 1
##        56.55575        53.52025
```

Interpretation

The interpretation of the difference of means test in R is really quite simple. Look for the p-value. In this case the p-value in this test is 0.1158, which is much larger than .05. Therefore we cannot reject the null hypothesis and cannot draw any conclusions.

Correlation Coefficient

When we have two continuous variables, we can use yet a different bivariate hypothesis test: the correlation coefficient. The correlation coefficient, also called *Pearson's r*, can quantify the general pattern of association between two continuous variables. Correlation begins with the concept of *covariance* which is a difference quantity that essentially measures how two continuous variables move together. It is calculated as:

$cov_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$

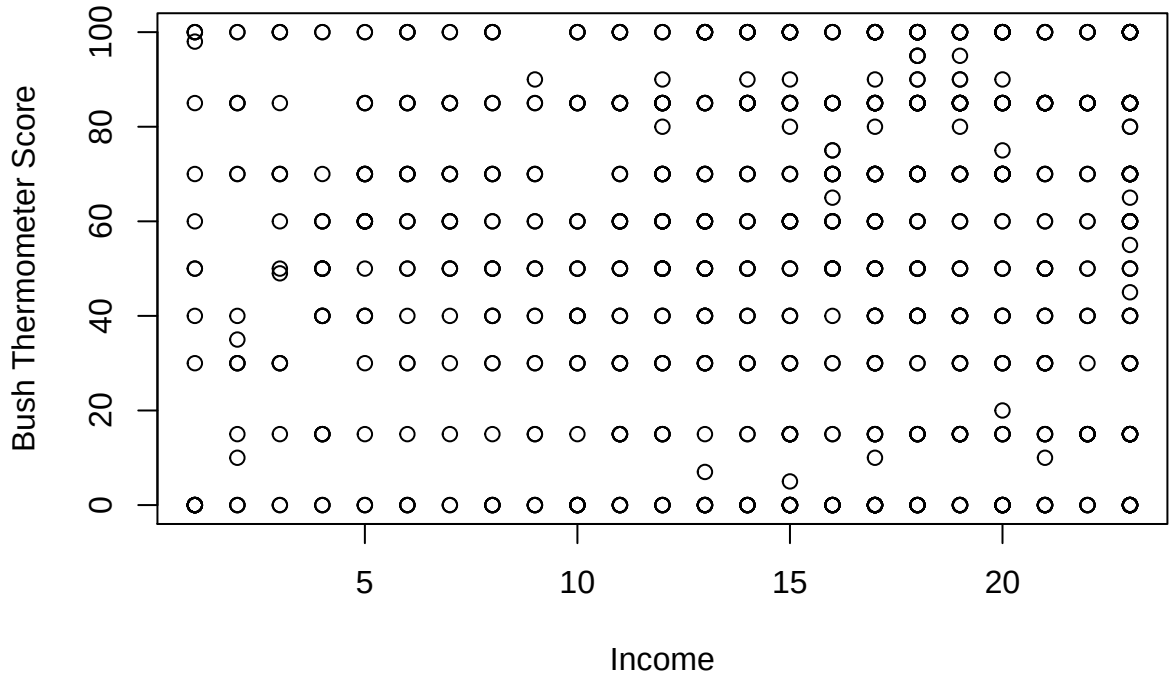
You turn the covariance into the correlation coefficient (Pearson's r) by scaling the covariance by the individual variances of the variables:

$r = \frac{cov_{XY}}{\sqrt{var_X var_Y}}$

Pearson's r is a score that ranges between -1 and +1. If there were a perfectly positive linear relationship between two variables, Pearson's r would be 1. On the other hand, if there were a perfectly negative linear relationship between two variables, Pearson's r would be -1. The weaker the relationship between the two variables, the closer Pearson's r is to zero. Below, I demonstrate this visually.

First, here is a scatterplot of two continuous variables: income and Bush's feeling thermometer scores:

```
plot(mydata$income, mydata$bush_therm, xlab="Income", ylab="Bush Thermometer Score")
```



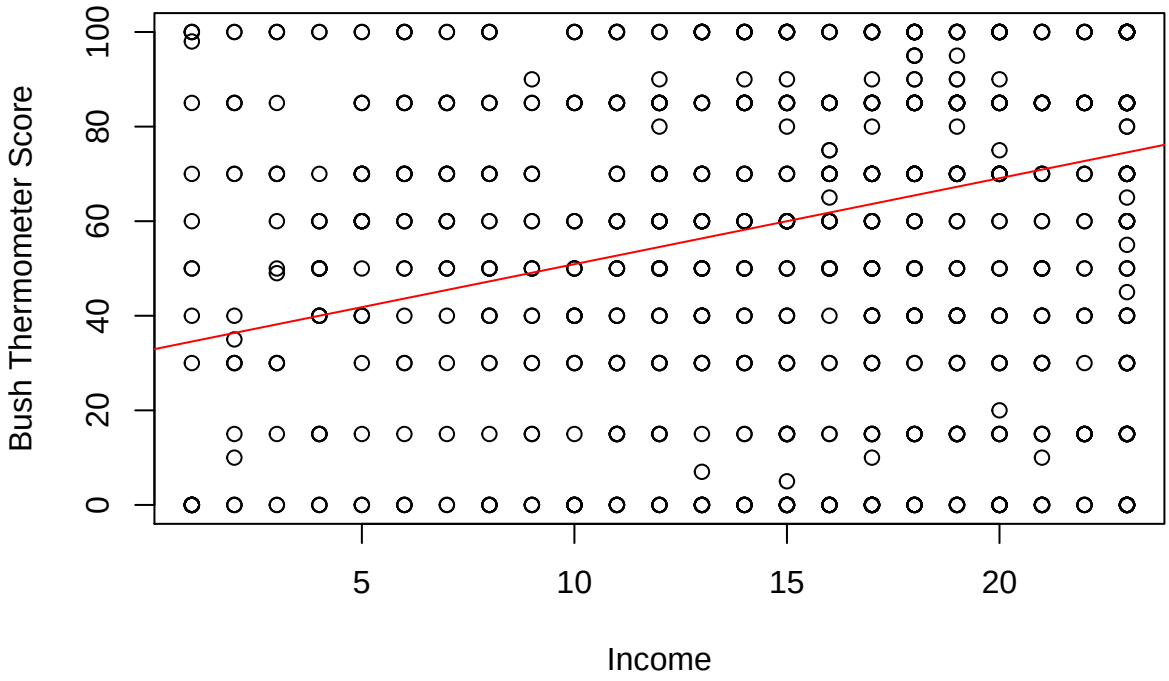
To get the line equation for line of best fit for the above scatter plot, we can use the `line` command:

```
line(mydata$income, mydata$bush_therm)
```

```
##
## Call:
## line(mydata$income, mydata$bush_therm)
##
## Coefficients:
## [1]  32.727   1.818
```

`line` gives you the y intercept and the slope for the line of best fit. You can add this line to the scatterplot using the `abline` command:

```
plot(mydata$income, mydata$bush_therm, xlab="Income", ylab="Bush Thermometer Score")
abline(a=32.727, b=1.818, col="red")
```



The slope of the line is a good indicator of a relationship, but the correlation coefficient is a standardized quantity, meaning that correlations can be directly compared. Slopes, on the other hand, depend on the scale of the variables in questions. Obtaining Pearson's r from R is quite simple. You use the `cor` command and enter your two variables inside the parentheses. However, `cor` only works when you have no missing observations. In order to use `cor` on a data set with missing values, you have to add `use="complete.obs"` at the end of the line.

Estimation

```
#Getting Pearson's r

r<-cor(mydata$income, mydata$bush_therm, use="complete.obs")

r
```

```
## [1] 0.08800578
```

You can also obtain a p-value for Pearson's r, which can tell you if the correlation is statistically significant, or significantly difference from zero. You do this by essentially calculating a difference of means test. In this t-test, the first group mean is the Pearson's r score. The second group mean is zero, essentially taking the place of a hypothetical second group that has no relationship between the two continuous variables (r=o). Below is the procedure to get the p-value:

```
#First, square pearson's r

r2<-cor(mydata$income, mydata$bush_therm, use="complete.obs")^2

#To Get a p-value for the Pearson's r

#First, we need sample size:
length(mydata$income)#Gives us the number of observations for this variable (1066)
```

```
## [1] 1207
```

```r
#To get the t score
t.cor<-(r*sqrt(1066-2))/(sqrt(1-r2))

t.cor
```

```
## [1] 2.881843
```

```r
#For a p-value, you can look it up or:

p<-2*pt(-abs(t.cor), df=1064)#pt gives you the area under the curve at that t-score and above

p
```

```
## [1] 0.004033012
```

You can also use one line of code to get both the p-value and the correlation coefficient:

```r
r.test<-cor.test(mydata$income, mydata$bush_therm, use="complete.obs")

r.test
```

```
##
##  Pearson's product-moment correlation
##
## data:  mydata$income and mydata$bush_therm
## t = 2.8818, df = 1064, p-value = 0.004033
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02811183 0.14727009
## sample estimates:
##       cor
## 0.08800578
```

Interpretation

The interpretation of Pearson's r is very similar to that of the other hypothesis tests in this document. Look at the p-value. A p-value less than .05 indicates statistical significance. Therefore, the p-value of .004033 is less than .05 and we can then reject the null hypothesis (that there is no relationship between a respondent's income and their rating of Bush on the feeling thermometer scale). We can then conclude that income is positively correlated with higher ratings of President Bush on the feeling thermometer scale.

Conclusion

Overall, bivariate hypothesis tests are useful starting points. However, remember that bivariate hypothesis testing does not allow you to control for confounding factors. This means that finding a *spurious correlation* is not only possible, but probable. You cannot make causal arguments with a bivariate hypothesis test. But, they are still useful to know and practice.