Bivariate Regression

Ben Gonzalez

2023-10-06

Contents

Chapter 1

About

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports; for example, a math equation $a^2 + b^2 = c^2$.

1.1 Usage

Each **bookdown** chapter is an .Rmd file, and each .Rmd file can contain one (and only one) chapter. A chapter *must* start with a first-level heading: `# A good chapter`, and can contain one (and only one) first-level heading.

Use second-level and higher headings within chapters like: `## A short section` or `### An even shorter section`.

The `index.Rmd` file is required, and is also your first book chapter. It will be the homepage when you render the book.

1.2 Render book

You can render the HTML version of this example book without changing anything:

1. Find the **Build** pane in the RStudio IDE, and
2. Click on **Build Book**, then select your output format, or select "All formats" if you'd like to use multiple formats from the same book source files.

Or build the book from the R console:

```
bookdown::render_book()
```

To render this example to PDF as a `bookdown::pdf_book`, you'll need to install XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.org/tinytex/.

1.3 Preview book

As you work, you may start a local server to live preview this HTML book. This preview will update as you edit the book when you save individual .Rmd files. You can start the server in a work session by using the RStudio add-in "Preview book", or from the R console:

```
bookdown::serve_book()
```

Statistical Significance - Confidence Intervals - Effect Size

```
##
## 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
```

1.3.0.1 Sample Standard Deviation Formula

$$S = \sqrt{\frac{\sum(X - \overline{X})^2}{N - 1}}$$

```
N <- length(mtcars$hp)
deviations <- mtcars$hp-mean(mtcars$hp)

s <- deviations^2

m_plus<- sum(s)/(N-1)

sd_plus <- sqrt(m_plus)
print(sd_plus)
```

```
## [1] 68.56287
```

```
sqrt((sum((mtcars$hp-mean(mtcars$hp))^2))/(length(mtcars$hp)-1))
```

```
## [1] 68.56287
```

```
sd(mtcars$hp)
```

```
## [1] 68.56287
```

1.3.0.2 Standard Error Formula

$$SE \ = \frac{\sigma}{\sqrt{n}}$$

$SE \ = Standard \ Error \sigma \ = sample \ standard \ deviation n \ = number \ of \ samples$

1.3.0.3 How to calculate the standard error of a sample

```
#standard deviation/squareroot(n)

# length(mtcars$hp)
# nrow(mtcars)

###Shortcut to calculate the standard error of a sample
###The length function is utilized to find the number of observations in a data set
sd(mtcars$hp)/sqrt(length(mtcars$hp))
```

```
## [1] 12.12032
```

```
####Another shortcut to calculate the standard error of a sample
###The nrow function is utilized to find the number of observations in a data set
sd(mtcars$hp)/sqrt(nrow(mtcars))
```

```
## [1] 12.12032
```

```
####Long way to calculate the standard error of a sample
print(sqrt(sum((mtcars$hp - mean(mtcars$hp)) ^ 2/(length(mtcars$hp) - 1)))
      /sqrt(length(mtcars$hp)))
```

```
## [1] 12.12032
```

1.3.1 Computing Confidence Intervals

1.3.1.1 Confidence Interval Formula

The confidence interval (C.I.) according to (Hatcher, 2013) gives us a range of values for the population parameter being estimated.

Computed for:

- mean
- difference between means
- correlation coefficients
- etc.

1.3.1.1.1 Lower Bound of the C.I. for a Sample Mean

$$CI = \overline{X} \pm (SE_m)(t_{crit})$$

$CI = Confidence\ Interval \overline{X} = observed\ sample\ mean SE_m = standard\ error\ of\ the\ mean t_{crit} = t$

```
###Calculate the mean of the hp
mean_hp <- mean(mtcars$hp)
print(paste0("Mean of horsepower: ",mean_hp))
```

```
## [1] "Mean of horsepower: 146.6875"
```

```
###Calculate the sd of the hp
sd_hp <- sd(mtcars$hp)
print(paste0("Standard deviation of horsepoweer: ",sd_hp))
```

```
## [1] "Standard deviation of horsepoweer: 68.5628684893206"
```

```
###Calculate the square root of the hp
n<- sqrt(length(mtcars$hp))
print(n)
```

```
## [1] 5.656854
```

```
###Confidence level of 0.95% e.g. two-tailed with 2.5%
t_value <- 1.96
###How to calculate the t-value properly
###Take the p-value: 0.05 and the degrees of freedom: 32-1
tval <- qt((1-0.95)/2, df=32-1)
print(paste0("t-critical value: ",tval))
```

```
## [1] "t-critical value: -2.03951344639641"
```

```
###Standard error of the sample mean
se_sample_mean <- (sd(mtcars$hp)/sqrt(length(mtcars$hp)))
print(paste0("Standard Error of the Sample Mean",se_sample_mean))
```

```
## [1] "Standard Error of the Sample Mean12.1203173116"
```

```
sd(mtcars$hp)/sqrt(length(mtcars$hp))
```

```
## [1] 12.12032
```

```
ci_lower_bound <-mean(mtcars$hp)-(se_sample_mean*tval)

ci_upper_bound<- mean(mtcars$hp)+(se_sample_mean*tval)
print(ci_lower_bound)
```

```
## [1] 171.4071
```

```
print(ci_upper_bound)
```

```
## [1] 121.9679
```

```
# head(HumanResourcesDataAA5221_all_zscores[,c(1,11,2,12)]) %>%
#   kbl(row.names = T) %>%
#   kable_styling(row_label_position = "l", full_width = F) %>%
#   footnote(general = "Head of data frame.") %>%
#   add_header_above(c(" " = 1, "Satisfaction Level" = 2, "Last Evaluation" = 2))
```

1.3.1.1.2 Utilizing the linear model formula in R Here we can also calculate the CI by utilizing the linear model funciton.

```
mean(mtcars$hp)+(1.96*12.1203)
```

```
## [1] 170.4433
```

```
# Calculate the mean and standard error
l.model <- lm(hp ~ 1, mtcars)
summary(l.model)
```

```
##
## Call:
## lm(formula = hp ~ 1, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -94.69 -50.19 -23.69  33.31 188.31
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   146.69      12.12    12.1 2.79e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.56 on 31 degrees of freedom
```

```
# Calculate the confidence interval
confint(l.model, level=0.95)
```

```
##                 2.5 %   97.5 %
## (Intercept) 121.9679 171.4071
```

References

Hatcher, L. (2013). *Advanced statistics in research: Reading, understanding, and writing up data analysis results. Shadow Finch Media.*

https://www.statology.org/t-distribution-table/

https://www.scribbr.com/frequently-asked-questions/critical-value-of-t-in-r/#:~:text=You%20can%20use%20the%20qt,the%20significance%20level%20by%20two.

1.3.2 Chi-square Assumptions

- Chi-Square test – statistical test used to compare observed results with expected results. This is a test of association.
- Used with Nominal or Categorical data.
- We can utilize Steven's four scales of measurements to check our data.

1.3.2.1 Steven's four scales of measurement

Characteristic of Scale

Nominal

Ordinal

Interval

Ratio

Applies names or numbers to categories?

Yes

Yes

Yes

Yes

Orders categories according to quantity?

Yes

Yes

Yes

Displays equal intervals between consecutive numbers?

Yes

Yes

Diplays a “true zero point?”

Yes

1.3.3 Student Peformance Data Set Information

1.3.3.0.1 Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

- 1 school - student’s school (binary: ‘GP’ - Gabriel Pereira or ‘MS’ - Mousinho da Silveira)
- 2 sex - student’s sex (binary: ‘F’ - female or ‘M’ - male)
- 3 age - student’s age (numeric: from 15 to 22)
- 4 address - student’s home address type (binary: ‘U’ - urban or ‘R’ - rural)
- 5 famsize - family size (binary: ‘LE3’ - less or equal to 3 or ‘GT3’ - greater than 3)
- 6 Pstatus - parent’s cohabitation status (binary: ‘T’ - (living) together or ‘A’ - apart)
- 7 Medu - mother’s education (numeric: 0 - none, 1 - primary education (4th grade), 2 (5th to 9th grade), 3 (secondary education) or 4 (higher education))

- 8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 (5th to 9th grade), 3 (secondary education) or 4 (higher education))
- 9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- 10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- 11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- 12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
- 13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- 14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- 15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- 16 schoolsup - extra educational support (binary: yes or no)

- 17 famsup - family educational support (binary: yes or no)
- 18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- 19 activities - extra-curricular activities (binary: yes or no)
- 20 nursery - attended nursery school (binary: yes or no)
- 21 higher - wants to take higher education (binary: yes or no)
- 22 internet - Internet access at home (binary: yes or no)
- 23 romantic - with a romantic relationship (binary: yes or no)
- 24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- 25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- 26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- 27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- 28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- 29 health - current health status (numeric: from 1 - very bad to 5 - very good)
- 30 absences - number of school absences (numeric: from 0 to 93)

1.3.3.0.1.1 These grades are related with the course subject, Math or Portuguese:

- 31 G1 - first period grade (numeric: from 0 to 20)
- 31 G2 - second period grade (numeric: from 0 to 20)
- 32 G3 - final grade (numeric: from 0 to 20, output target)

1.3.4 Look at the structure of our data

```
str(student)
```

```
## 'data.frame':    395 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

```
## [1] "There are  395  observations in our student data set."
```

1.3.5 Chi-square

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$x^2 = chi - squared O_i = \ observed\ value E_i = \ expected\ value$$

1.3.6 Chi-square Assumptions

- Chi-Square test – statistical test used to compare observed results with expected results. This is a test of association.
- Used with Nominal or Categorical data.

1.3.7 Chi-Square Test of Association

Here we want to investigate if there is an association between the type of job a mother has and access to the internet.

```
student_chi <- table(student$Mjob, student$internet)
print(student_chi)
```

```
##
##             no yes
##   at_home   22  37
##   health     2  32
##   other     27 114
##   services  12  91
##   teacher    3  55
```

```
chisq.test(student$Mjob, student$internet, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  student$Mjob and student$internet
## X-squared = 28.861, df = 4, p-value = 8.341e-06
```

```
mjob_internet<- xtabs(formula = ~Mjob+internet,data = student)

chisq_result<- chisq.test(mjob_internet ,correct = FALSE)
chisq_result$observed
```

```
##           internet
## Mjob        no yes
##   at_home   22  37
##   health     2  32
##   other     27 114
##   services  12  91
##   teacher    3  55
```

```
chisq_result$expected
```

```
##           internet
## Mjob              no       yes
##   at_home   9.858228  49.14177
##   health    5.681013  28.31899
##   other    23.559494 117.44051
##   services 17.210127  85.78987
##   teacher   9.691139  48.30886
```

```
total_observed<- chisq_result$observed[,1]+chisq_result$observed[,2]

observed_expected<- data.frame(chisq_result$observed[,1],chisq_result$expected[,1],chisq_result$o


colnames(observed_expected) <- c("Observed","Expected","Observed","Expected","")
```

```
total_observed<- chisq_result$observed[,1]+chisq_result$observed[,2]
###Create data frame from chi-square results
observed_expected<- data.frame(chisq_result$observed[,1],round(chisq_result$expected[,1],digits =


overall_total<- rbind(observed_expected, c(colSums(round(observed_expected[,1:5]))))


colnames(overall_total) <- c("Observed","Expected","Observed","Expected","")
rownames(overall_total) <- c("at home","health","other","services","teacher","Total")
###Create chi-square table
kable(x = overall_total) %>%
  kable_styling(row_label_position = "l", full_width = F) %>%
  footnote(general = paste0("",sprintf(r'($X^2$)')," = (",chisq_result$parameter,", n = 395) = ",
  add_header_above(c("Guardian", "Internet No" = 2, "Internet Yes" = 2,"Total"=1))
```

We can see that our X^2 value is 28.861 and the p-value is $p > .05$. We check our X^2 value against the critical table value of 9.488. Since our X^2 value 28.861 is above our critical table value of 9.488 we **cannot** reject our null hypothesis. Therefore there is no significant association between Mjob type and internet.

Guardian	Internet No		Internet Yes		Total
	Observed	Expected	Observed	Expected	
at home	22	9.858	37	49.142	59
health	2	5.681	32	28.319	34
other	27	23.559	114	117.441	141
services	12	17.210	91	85.790	103
teacher	3	9.691	55	48.309	58
Total	66	67.000	329	328.000	395

Note:
X^2 = (4, n = 395) = 28.861 p = 8.341138e-06

	no	yes
father	12	78
mother	47	226
other	7	25

- In APA style, the proper reporting of chi-square test results is 2 = (degrees of freedom, n = number of scores) = chi-square score, p value.
- X^2 (4, n = 395) = 28.86, p = 8.341e-06. Chi-Square Critical Table

---

Next we can investigate if there is an association between the type of guardian a student has and access to the internet.

```
###Chi-square calculation
guardian_internet<- xtabs(formula = ~guardian+internet,data = student)
guard_int_chisq_result<- chisq.test(guardian_internet ,correct = FALSE)

guard_int_chisq_result
```

```
##
##  Pearson's Chi-squared test
##
## data:  guardian_internet
## X-squared = 1.401, df = 2, p-value = 0.4963
```

1.3.7.0.1 Observed

	no	yes
father	15.037975	74.96203
mother	45.615190	227.38481
other	5.346835	26.65316

Guardian	Internet No		Internet Yes		Total
	Observed	Expected	Observed	Expected	
father	12	15.038	78	74.962	90
other	47	45.615	226	227.385	273
mother	7	5.347	25	26.653	32
Total	66	66.000	329	329.000	395

Note:
X^2 = (2, n = 395) = 1.401p =0.4963

```
total_observed<- guard_int_chisq_result$observed[,1]+guard_int_chisq_result$observed[,2]
###Create data frame from chi-square results
observed_expected<- data.frame(guard_int_chisq_result$observed[,1],round(guard_int_chisq_result$e


overall_total<- rbind(observed_expected, c(colSums(round(observed_expected[,1:5]))))


colnames(overall_total) <- c("Observed","Expected","Observed","Expected","")
rownames(overall_total) <- c("father","other","mother","Total")
###Create chi-square table
kable(x = overall_total) %>%
  kable_styling(row_label_position = "l", full_width = F) %>%
  footnote(general = paste0("",sprintf(r'($X^2$)')," = (",guard_int_chisq_result$parameter,", n =
  add_header_above(c("Guardian", "Internet No" = 2, "Internet Yes" = 2,"Total"=1))
```

1.3.7.0.2 Expected We can see that our X^2 value is 1.401 and the p-value is $p > .05$. We check our X^2 value against the critical table value of 5.991. Since our X^2 value 1.401 is below our critical table value of 5.991, therefore we **cannot** reject our null hypothesis. Therefore there is no significant relationship between guardian type and internet.

- In APA style, the proper reporting of chi-square test results is 2 = (degrees of freedom, n = number of scores) = chi-square score, p value.
- X^2 (2, n = 395) = 1.401, p = 0.4963. Chi-Square Critical Table

1.3.8 Correlation

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

1.3.9 Correlation Assumptions

The following assumptions are in place when utilizing the Pearson's r when calculating a correlation coefficient.

- **Interval or ratio scale** – both variables should be quantitative variables. Both should be assessed on an interval or ratio scale.
- **Normally-distributed sampling distribution** – sampling distribution of the statistic should be normally distributed. Likely to be met if the sample data are approximately normal or the sample is large.
- **Random Sample from bivariate normal distribution** – the data points used to compute the Pearson r are the pairs of scores on the predictor and criterion variables. The data points should be a random sample drawn from a bivariate normal distribution.

1.3.9.0.1 Index of Effect Size Table

- Index of effect size – a statistic that conveys the strength of the association between a predictor variable and criterion variable.
- r – the larger the absolute value of the correlation coefficient the larger the "effect".

Correlation Coefficient

Effect Size

Small Effect

r +/- .10

Medium Effect

r +/- .30

Large Effect

r +/- .50

1.3.9.1 Correlation between student absences and first period grade

```
cor.test(student$absences,student$G1)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student$absences and student$G1
## t = -0.6149, df = 393, p-value = 0.539
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.12927845  0.06787576
## sample estimates:
##        cor
## -0.0310029
```

Here we can see that our correlation r = -0.031 and our $p > .05$. Since our $p - value$ is greater than 0.05 we can say there is no relationship between *absences* and $G1$.

1.3.9.2 Correlation between student absences and final grade

```
cor.test(student$absences,student$G3)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student$absences and student$G3
## t = 0.67933, df = 393, p-value = 0.4973
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.06464215  0.13247070
## sample estimates:
##        cor
## 0.03424732
```

Here we can see that our correlation $r = 0.034$ and our $p > .05$. Since our $p - value$ is greater than 0.05 we can say there is no relationship between *absences* and $G3$.

1.3.10 Point-biserial Correlation Assumptions

- **Point-biserial correlation coefficient** – appropriate when one variable is a roughly continuous , multi-value variable assessed on an interval or ratio scale and the other variable is a dichotomous(2 values) variable.

- – must have a true dichotomy e.g. sex (male vs female)

1.3.10.1 Point-biserial correlation: Student gender and final grade

Here we are looking at a true dichotomy sex (male vs female) and the association with final $G3$ grade. We need to recode the variable into a binary variable of 1 and 0 with 1 being $Male$ and 0 being $Female$.

```
student$sexrecoded<- ifelse(student$sex=="M",1,0)
cor.test(student$G3,student$sexrecoded)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student$G3 and student$sexrecoded
## t = 2.062, df = 393, p-value = 0.03987
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.004833966 0.200084200
## sample estimates:
##       cor
## 0.1034556
```

1.3.11 Biserial Correlation

- **Biserial correlation coefficient** – appropriate in exactly the same situation where the point-biserial correlation would be used; one continuous variable and one dichotomous variable.
- – appropriate when the dichotomous variable is not a true dichotomy e.g. pass/fail

1.3.11.1 Biserial Correaltion: Family educational support and final grade

Here we can look at whether there is an association between a students family support and their final grade. Again we need to re-code the variable so it is binary e.g. 0 and 1

```
###Recode the variable into a binary variable
student$familysuprecoded <- ifelse(student$famsup=="yes",1,0)
cor.test(student$G3,student$familysuprecoded)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student$G3 and student$familysuprecoded
## t = -0.77686, df = 393, p-value = 0.4377
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.13729770  0.05974472
## sample estimates:
##         cor
## -0.03915715
```

1.3.11.2 Biserial Correaltion: Extra educational support and final grade

Here we can look at if there is an association between *extra educational support schoolsup* and final grade $G3$ + Again we need to re-code the variable so it is binary e.g. 0 and 1

```
student$schoolsuprecoded <- ifelse(student$schoolsup=="yes",1,0)

cor.test(student$G3,student$schoolsuprecoded)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student$G3 and student$schoolsuprecoded
## t = -1.6469, df = 393, p-value = 0.1004
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17998895  0.01601362
## sample estimates:
##         cor
## -0.08278821
```

1.3.11.3 Biserial Correaltion: Attended nursery school and final grade

Here we can look at if there is an association between those who attended nursery school *nursery* and final grade $G3$ + We need to re-code the variable so it is binary e.g. 0 and 1

```
student$nurseryrecoded <- ifelse(student$nursery=="yes",1,0)

cor.test(student$G3,student$nurseryrecoded)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student$G3 and student$nurseryrecoded
## t = 1.0237, df = 393, p-value = 0.3066
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04734402  0.14947834
## sample estimates:
##       cor
## 0.0515679
```

1.3.12 Spearman Correlation

Spearman correlation coefficient for ranked data (Spearman's Rho or r_s) – displays the correlation between two variables whose values have been ranked.

In the student data set we see that studytime and traveltime have been ranked. Therefore we can utilize the Spearman's Rho r_s to see the association between these variables and final grade $G3$.

```
cor.test(student$studytime,student$traveltime,method="spearman")
```

1.3.12.0.1 Spearman Correlation: Study time and final grade

```
## Warning in cor.test.default(student$studytime, student$traveltime, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  student$studytime and student$traveltime
## S = 11360053, p-value = 0.03526
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.1059694
```