Case Study 2. Who Plays data Games

Code

Setup

Hide

```
# install.packages('moments')
# install.packages('gmodels')
# install.packages('e1071')
# install.packages('dplyr')
library(moments)
library(gmodels)
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
```

Hide

```
library(plyr)
```

```
--------------------------------------------------------------------------------------
-------------------------------------------------------
You have loaded plyr after dplyr - this is likely to cause problems.
If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
library(plyr); library(dplyr)
--------------------------------------------------------------------------------------
-------------------------------------------------------

Attaching package: ‘plyr’

The following objects are masked from ‘package:dplyr’:

    arrange, count, desc, failwith, id, mutate, rename, summarise, summarize
```

Hide

```
library(e1071)
```

```


Attaching package: ‘e1071’

The following objects are masked from ‘package:moments’:

    kurtosis, moment, skewness
```

Hide

```
library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode
```

Hide

```
N = 314  # Population size of 314 students in the course
n = 91  # Sample size of 91 students in the course who completed the survey
quantile.95 = qnorm(.975)
data <- read.table("videodata.txt", header=TRUE)
data[data == 99] <- NA  # Set unanswered/improperly answered to NA
head(data)
```

	time <dbl>	like <int>	where <int>	freq <int>	busy <int>	educ <int>	sex <int>	age <int>	home <int>
1	2.0	3	3	2	0	1	0	19	1
2	0.0	3	3	3	0	0	0	18	1
3	0.0	3	1	3	0	0	1	19	1
4	0.5	3	3	3	0	1	0	19	1
5	0.0	3	3	4	0	1	0	19	1
6	0.0	3	2	4	0	0	1	19	0

6 rows | 1-10 of 15 columns

Scenario 1

Begin by providing an estimate for the fraction of students who played a data game in the week prior to the survey. Provide an interval estimate as well as a point estimate for this proportion.

Hide

```
# Point estimate
point.estimate <- length(which(data$time > 0)) / n
point.estimate
```

```
[1] 0.3736264
```

Hide

```
# Simple confidence interval via CLT
standard.error <- sqrt(point.estimate * (1-point.estimate) / n)
lower <- point.estimate - quantile.95 * standard.error
upper <- point.estimate + quantile.95 * standard.error
c(lower, upper)
```

```
[1] 0.2742318 0.4730210
```

Hide

```
# Confidence interval via CLT with finite sample population correction
standard.error = sqrt((point.estimate * (1-point.estimate)) / (n-1) * (N-n) / N)
lower <- point.estimate - quantile.95 * standard.error
upper <- point.estimate + quantile.95 * standard.error
c(lower, upper)
```

```
[1] 0.2893996 0.4578531
```

Hide

```
# Confidence interval via bootstrap
set.seed(0)
bootstrap.population <- rep(data$time > 0, length.out = N)
bootstrap.means <- NULL
for (i in 1:1000) {
  bootstrap.means <- c(bootstrap.means, sum(sample(bootstrap.population, size = n, re
place = FALSE)) / n)
}
point.estimate <- mean(bootstrap.means)
point.estimate
```

```
[1] 0.364978
```

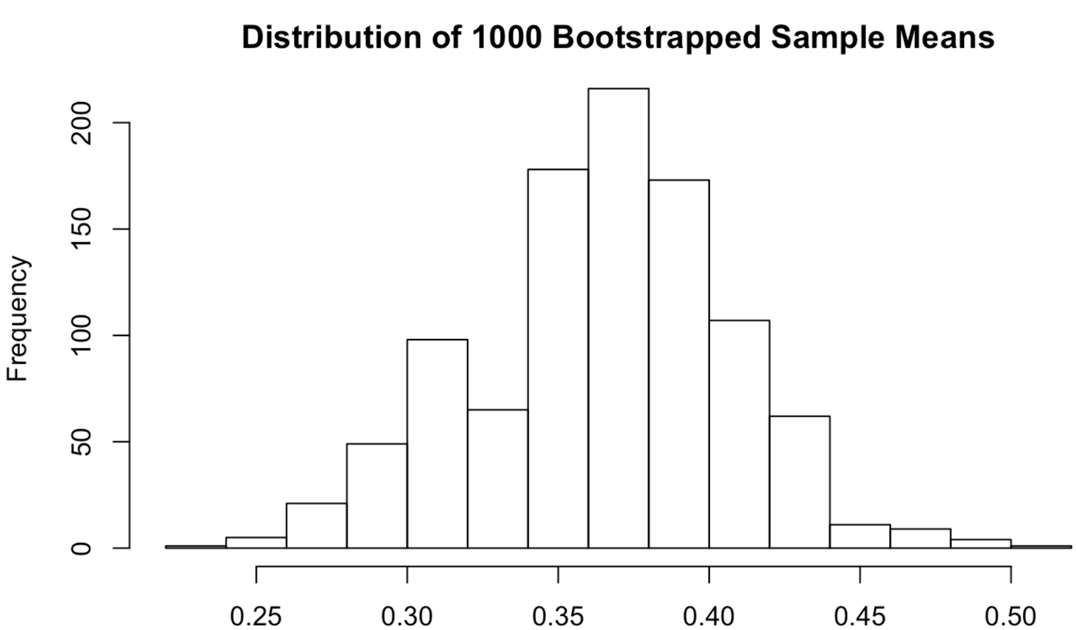
Hide

```
lower <- unname(quantile(bootstrap.means, .025))
upper <- unname(quantile(bootstrap.means, .975))
c(lower, upper)
```

```
[1] 0.2747253 0.4398352
```

Hide

```
hist(bootstrap.means, main='Distribution of 1000 Bootstrapped Sample Means', xlab='Pr
oportion of Students who Played')
```



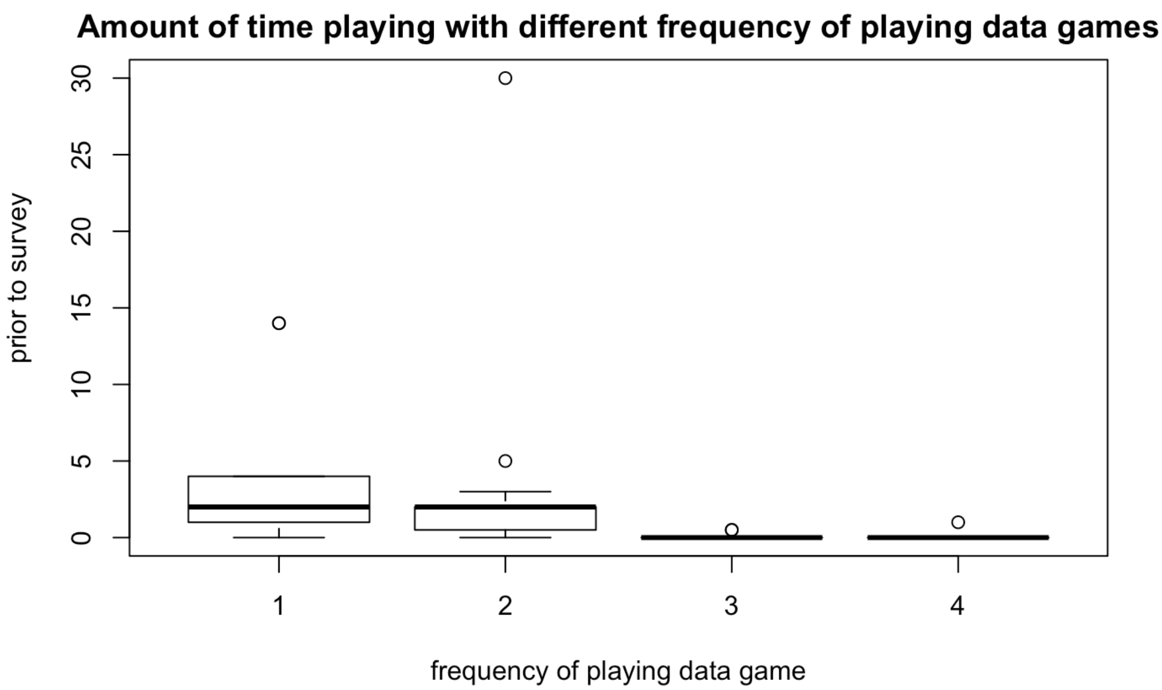
Proportion of Students who Played

Scenario 2

Check to see how the amount of time spent playing videgames in the week prior to the survey compares to the reported frequency of play (daily, weekly, etc). How might the fact that there was an exam in the week prior to the survey affect your previous estimates and this comparison?

Hide

```
#Generate box plot between freqency and times students play data game
time_play <- data$time
frequency_play <- data$freq
boxplot(time_play~frequency_play, data = data, main = "Amount of time playing with di
fferent frequency of playing data games", xlab ="frequency of playing data game", yla
b="number of hours played in the week
        prior to survey")
```



Hide

```
#Finding the average of the amount of time playing with different frequency
freq_1 <- data$time[which(data$freq==1)]
mean_time_1 <- mean(freq_1)
freq_2 <- data$time[which(data$freq==2)]
mean_time_2 <- mean(freq_2)
freq_3 <- data$time[which(data$freq==3)]
mean_time_3 <- mean(freq_3)
freq_4 <- data$time[which(data$freq==4)]
mean_time_4 <- mean(freq_4)
c(mean(data$time[which(data$busy==1 & data$freq==1)]), mean(data$time[which(data$busy
==0 & data$freq==1)]))
```

```
[1] 7.2 1.0
```

Hide

```
c(mean(data$time[which(data$busy==1 & data$freq==2)]), mean(data$time[which(data$busy
==0 & data$freq==2)]))
```

```
[1] 4.000000 1.594118
```

Hide

```
c(mean(data$time[which(data$busy==1 & data$freq==3)]), mean(data$time[which(data$busy
==0 & data$freq==3)]))
```

```
[1] 0.00000000 0.05882353
```

Hide

```
c(mean(data$time[which(data$busy==1 & data$freq==4)]), mean(data$time[which(data$busy
==0 & data$freq==4)]))
```

```
[1] NaN   0
```

Scenario 3

Consider making an internal estimate for the average amount of time spent playing data games in the week prior to the survey. Keep in mind the overall shape of the sample distribution. A simulation study may help determine the appropriateness of an interval estimate.

Hide

```
# Point estimate
point.estimate <- mean(data$time)
point.estimate
```

```
[1] 1.242857
```

Hide

```
# Simple confidence interval via CLT
standard.error <- sd(data$time) / sqrt(n)
lower <- point.estimate - quantile.95 * standard.error
upper <- point.estimate + quantile.95 * standard.error
c(lower, upper)
```

```
[1] 0.4668263 2.0188880
```

Hide

```
# Confidence interval via CLT with finite sample population correction
standard.error = sd(data$time) / sqrt(n) * sqrt((N-n) / N)
lower <- point.estimate - quantile.95 * standard.error
upper <- point.estimate + quantile.95 * standard.error
c(lower, upper)
```

```
[1] 0.5888739 1.8968403
```

Hide

```
# Confidence interval via bootstrap
```

```
set.seed(0)
bootstrap.population <- rep(data$time, length.out = N)
bootstrap.means <- c()
for (i in 1:1000) {
  bootstrap.means <- c(bootstrap.means, mean(sample(bootstrap.population, size = n, r
eplace = FALSE)))
}
point.estimate <- mean(bootstrap.means)
point.estimate
```

```
[1] 1.123867
```

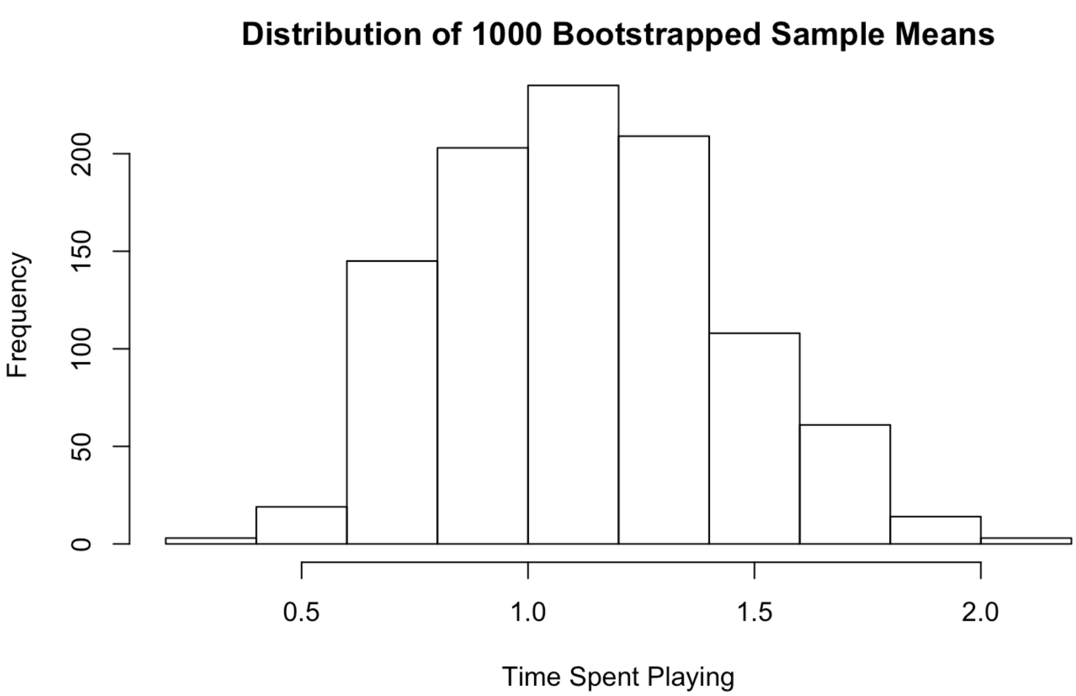
Hide

```
lower <- unname(quantile(bootstrap.means, .025))
upper <- unname(quantile(bootstrap.means, .975))
c(lower, upper)
```

```
[1] 0.6053846 1.7802747
```

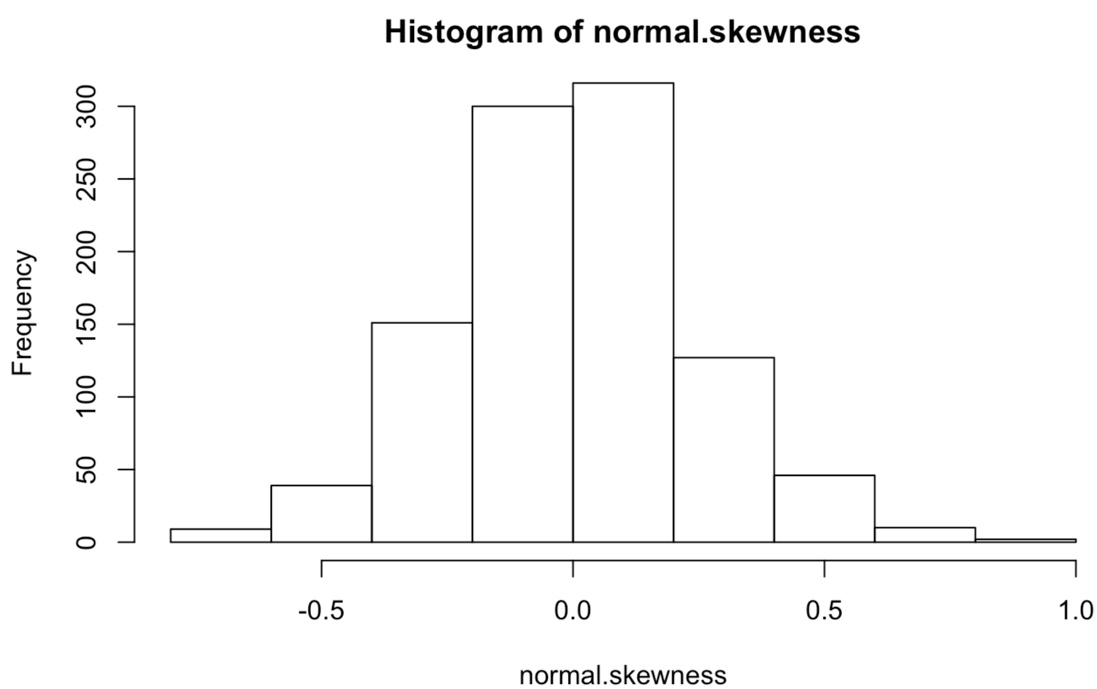
Hide

```
hist(bootstrap.means, main='Distribution of 1000 Bootstrapped Sample Means', xlab='Ti
me Spent Playing')
```



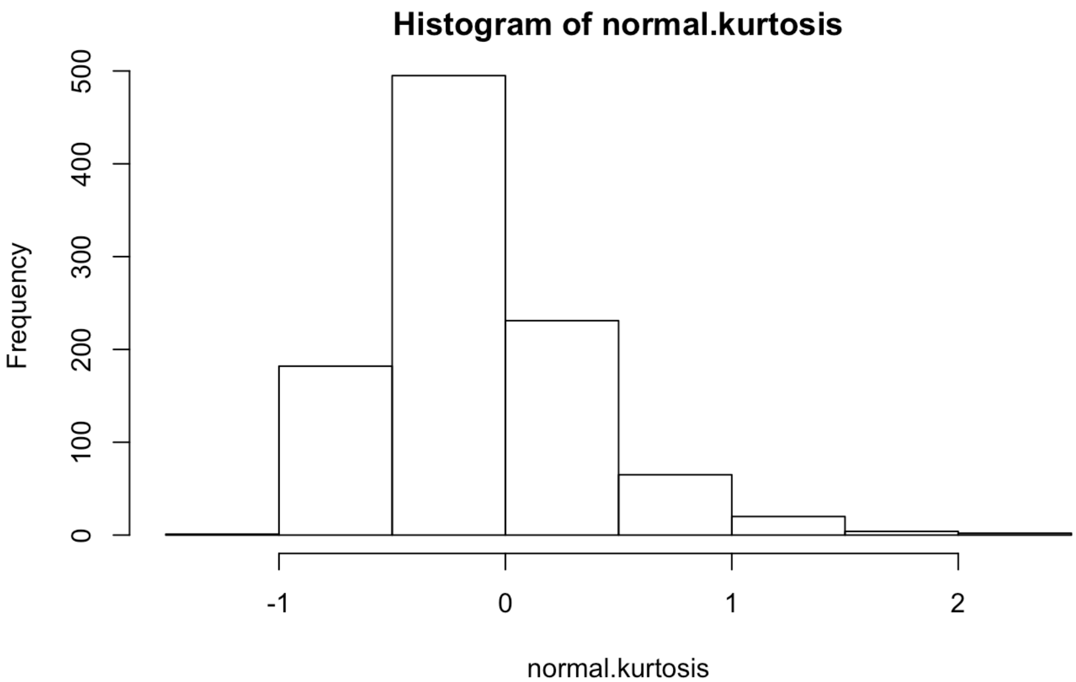
Hide

```
normal.skewness <- NULL
normal.kurtosis <- NULL
for (i in 1:1000) {
  normal.skewness <- c(normal.skewness, skewness(rnorm(n)))
  normal.kurtosis <- c(normal.kurtosis, kurtosis(rnorm(n)))
}
hist(normal.skewness)
```



Hide

```
hist(normal.kurtosis)
```



Hide

```
c(skewness(bootstrap.means), kurtosis(bootstrap.means))
```

```
[1]  0.3070444 -0.3732171
```

Scenario 5

Look for the differences between those who like to play data games and those who don't. To do this, use the questions in the last part of the survey, and make comparisons between male and female students, those who work for pay and those who don't, those who own a computer and those who don't. Graphical display and cross-tabulations are particularly helpful in making these kinds of comparisons. Also, you may want to collapse the range of responses to a question down to two or three possibilities before making these comparisons.

Hide

```
# Clean out the "Never played data"
data.clean <- data[which(data$like != 1),]
# Regroup the 'like' value
data.clean$like[data.clean$like == 2 | data.clean$like == 3] <- "Like"
data.clean$like[data.clean$like == 4 | data.clean$like == 5] <- "Dislike"
data.w <- data.clean[which(!is.na(data.clean$work)),]
# Regroup the 'sex' value
data.w$sex[data.w$sex == 0 ] <- "Female"
data.w$sex[data.w$sex == 1] <- "Male"
# Regroup the 'work' value
data.w$work[data.w$work > 0 ] <- "Work"
data.w$work[data.w$work == 0] <- "No Work"
# Regroup the 'own' value
data.w$own[data.w$own == 0 ] <- "No PC"
data.w$own[data.w$own == 1] <- "Own PC"
# Cross tabulations between like and sex
CrossTable(data.w$like, data.w$sex)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

```
Total Observations in Table:  86


             | data.w$sex 
 data.w$like |    Female |      Male | Row Total | 
-------------|-----------|-----------|-----------|
     Dislike |        12 |         8 |        20 | 
             |     1.132 |     0.896 |           | 
             |     0.600 |     0.400 |     0.233 | 
             |     0.316 |     0.167 |           | 
             |     0.140 |     0.093 |           | 
-------------|-----------|-----------|-----------|
        Like |        26 |        40 |        66 | 
             |     0.343 |     0.272 |           | 
             |     0.394 |     0.606 |     0.767 | 
             |     0.684 |     0.833 |           | 
             |     0.302 |     0.465 |           | 
-------------|-----------|-----------|-----------|
Column Total |        38 |        48 |        86 | 
             |     0.442 |     0.558 |           | 
-------------|-----------|-----------|-----------|
```

Hide

```
chisq.test(table(data.w$like, data.w$sex))
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  table(data.w$like, data.w$sex)
X-squared = 1.8731, df = 1, p-value = 0.1711
```

Hide

```
# Cross tabulations between like and work
CrossTable(data.w$like, data.w$work)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  86


             | data.w$work 
 data.w$like |   No Work |      Work | Row Total | 
-------------|-----------|-----------|-----------|
     Dislike |        14 |         6 |        20 | 
             |     1.387 |     1.453 |           | 
             |     0.700 |     0.300 |     0.233 | 
```

```
              |     0.318 |     0.143 |           |
              |     0.163 |     0.070 |           |
-------------|-----------|-----------|-----------|
        Like |        30 |        36 |        66 |
              |     0.420 |     0.440 |           |
              |     0.455 |     0.545 |     0.767 |
              |     0.682 |     0.857 |           |
              |     0.349 |     0.419 |           |
-------------|-----------|-----------|-----------|
Column Total |        44 |        42 |        86 |
              |     0.512 |     0.488 |           |
-------------|-----------|-----------|-----------|
```

Hide

```
chisq.test(table(data.w$like, data.w$work))
```

```

	Pearson's Chi-squared test with Yates' continuity correction

data:  table(data.w$like, data.w$work)
X-squared = 2.7838, df = 1, p-value = 0.09522
```

Hide

```
# Cross tabulations between like and own
CrossTable(data.w$like, data.w$own)
```

```


   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  86


             | data.w$own
 data.w$like |     No PC |    Own PC | Row Total |
-------------|-----------|-----------|-----------|
     Dislike |         3 |        17 |        20 |
             |     1.031 |     0.377 |           |
             |     0.150 |     0.850 |     0.233 |
             |     0.130 |     0.270 |           |
             |     0.035 |     0.198 |           |
-------------|-----------|-----------|-----------|
        Like |        20 |        46 |        66 |
             |     0.313 |     0.114 |           |
             |     0.303 |     0.697 |     0.767 |
             |     0.870 |     0.730 |           |
             |     0.233 |     0.535 |           |
-------------|-----------|-----------|-----------|
Column Total |        23 |        63 |        86 |
             |     0.267 |     0.733 |           |
```

```
-------------|-----------|-----------|-----------|
```

Hide

```
chisq.test(table(data.w$like, data.w$own))
```

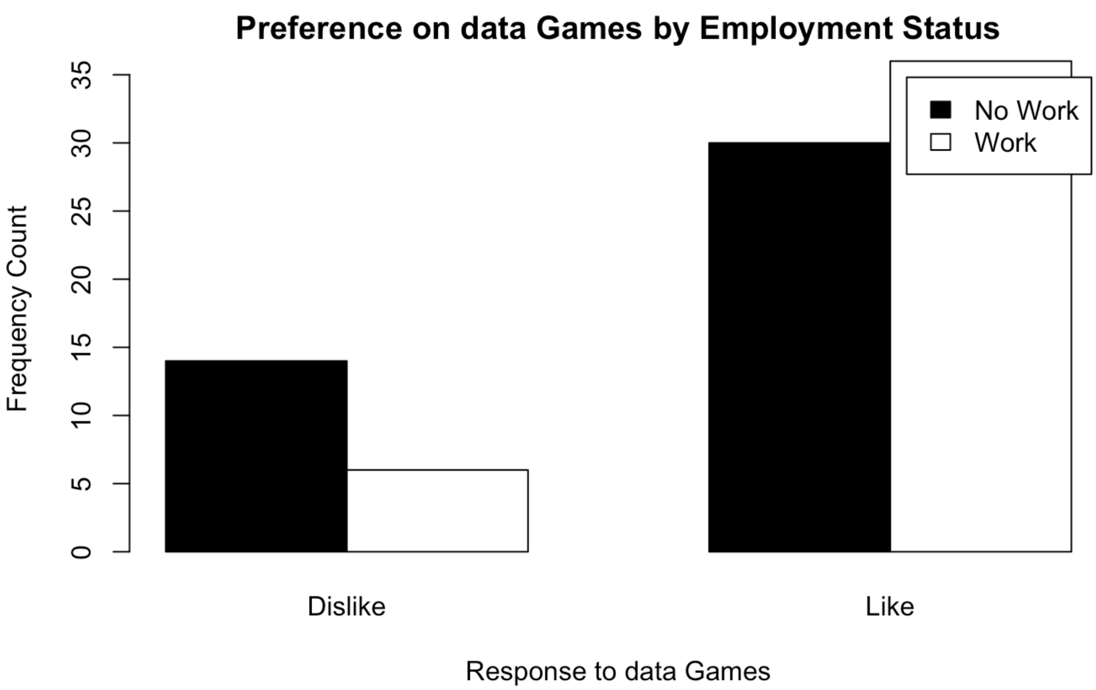
```

    Pearson's Chi-squared test with Yates' continuity correction

data:  table(data.w$like, data.w$own)
X-squared = 1.1367, df = 1, p-value = 0.2863
```

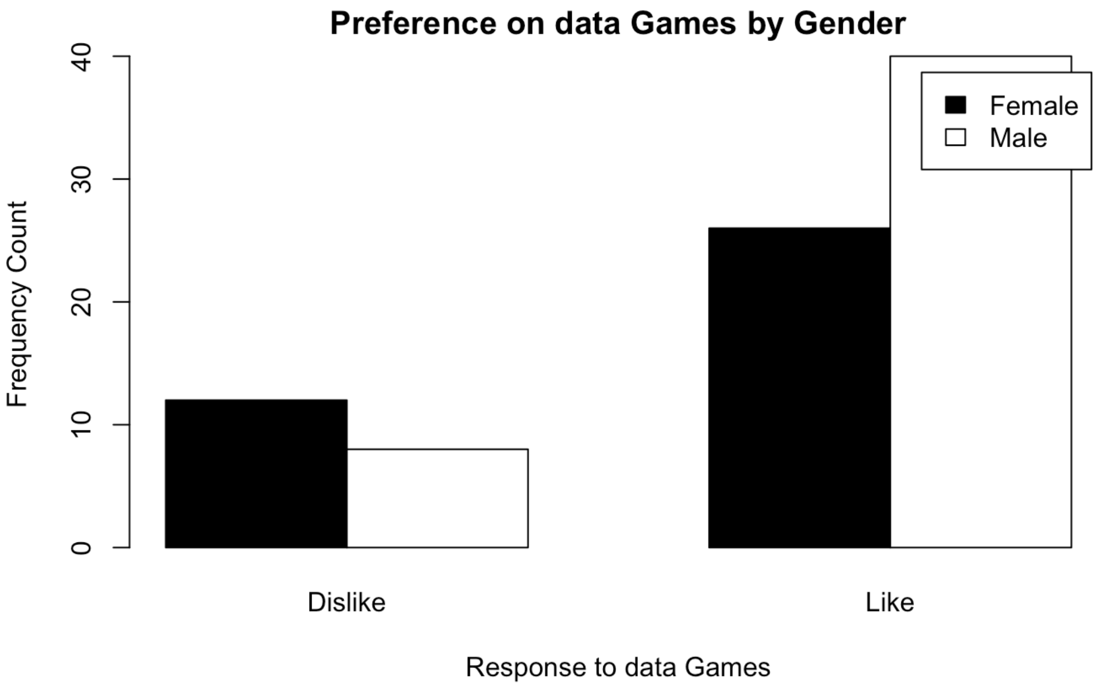
Hide

```
# Bar Graph
counts <- table(data.w$work, data.w$like)
barplot(counts, main = "Preference on data Games by Employment Status",
        xlab='Response to data Games', ylab = 'Frequency Count',
        col=c('black','white'), legend = rownames(counts), beside=TRUE)
```



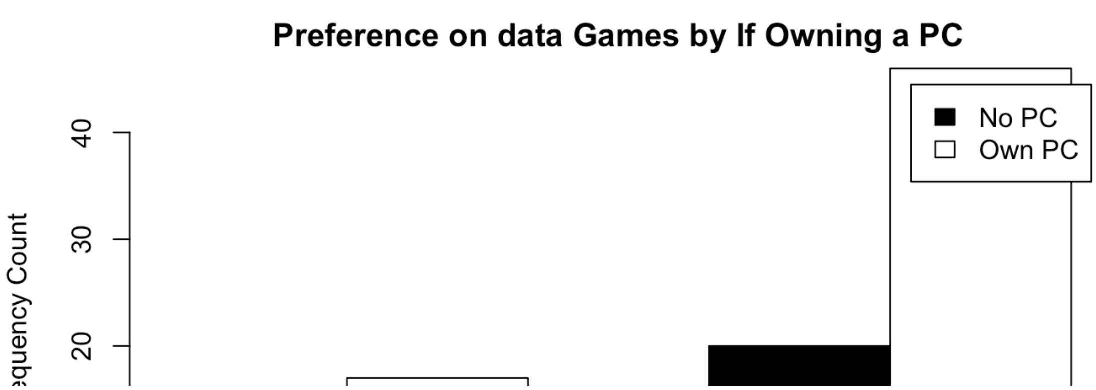
Hide

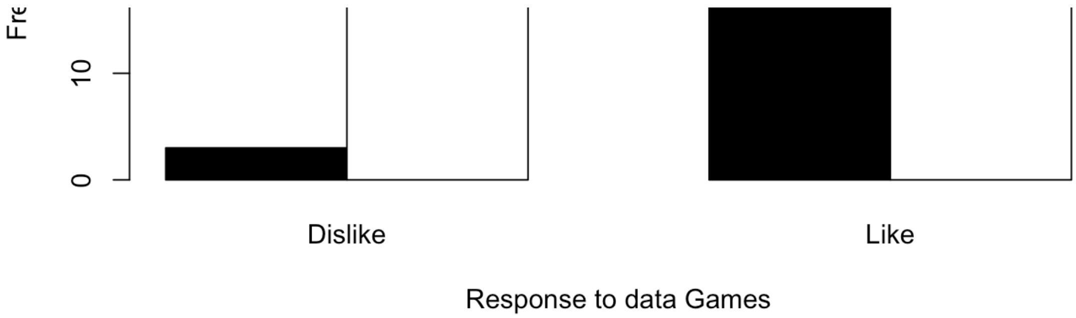
```
counts <- table(data.w$sex, data.w$like)
barplot(counts, main = "Preference on data Games by Gender",
        xlab='Response to data Games', ylab = 'Frequency Count',
        col=c('black','white'), legend = rownames(counts), beside=TRUE)
```



Hide

```
counts <- table(data.w$own, data.w$like)
barplot(counts, main = "Preference on data Games by If Owning a PC",
        xlab='Response to data Games', ylab = 'Frequency Count',
        col=c('black','white'), legend = rownames(counts), beside=TRUE)
```



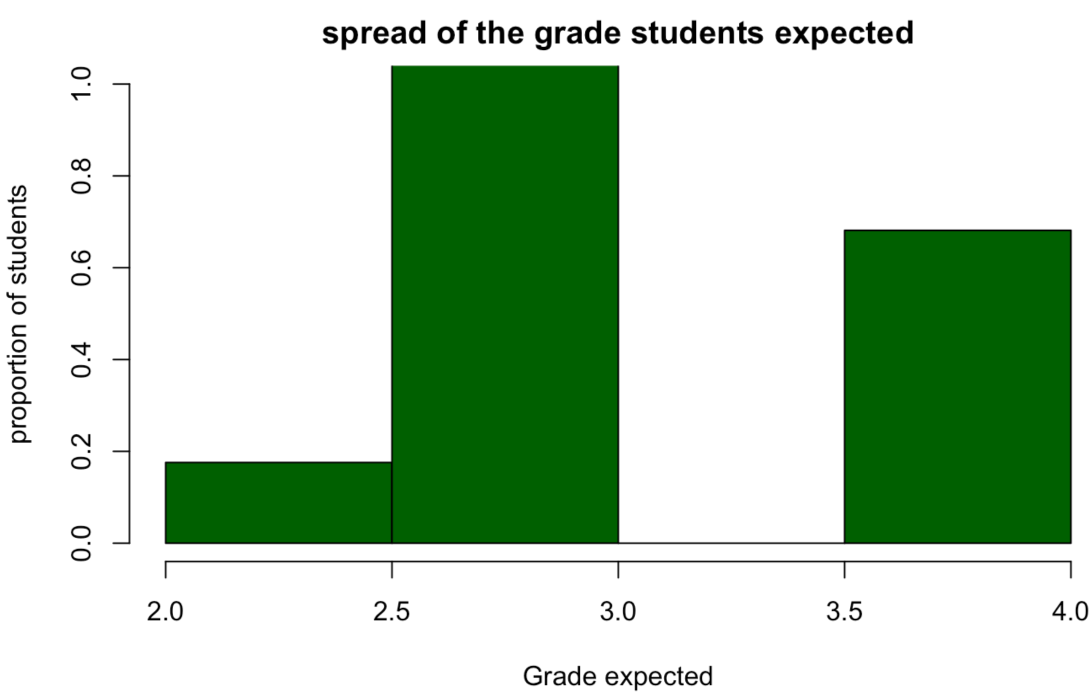


Scenario 6

Just for fun, further investigate the grade that students expect in the course. How will does it match the target distribution used in grade assignment of 20% A's, 30%B's, 40% c's and 10% D's or lower? If the nonrespondents were failing students who no longer bothered to come to the discussion section, would this change the picture ?

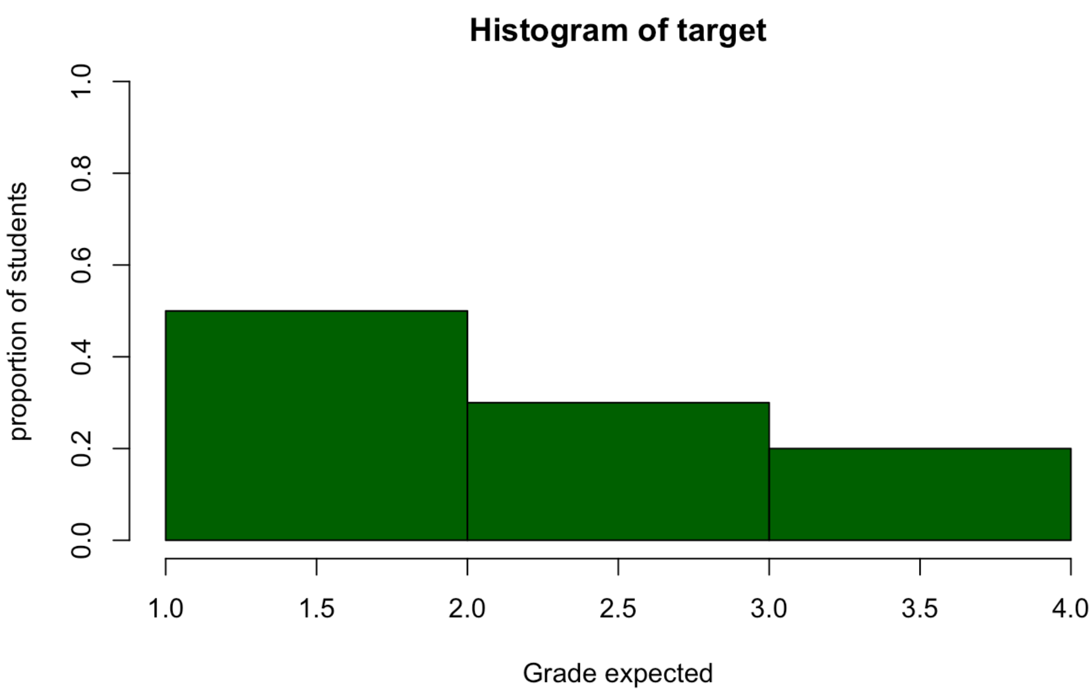
Hide

```
grade <- data$grade
p1<-hist(data$grade, main = "spread of the grade students expected", col="darkgreen",
ylab ="proportion of students", xlab
     ="Grade expected",freq=FALSE,breaks=5,ylim=c(0,1))
```



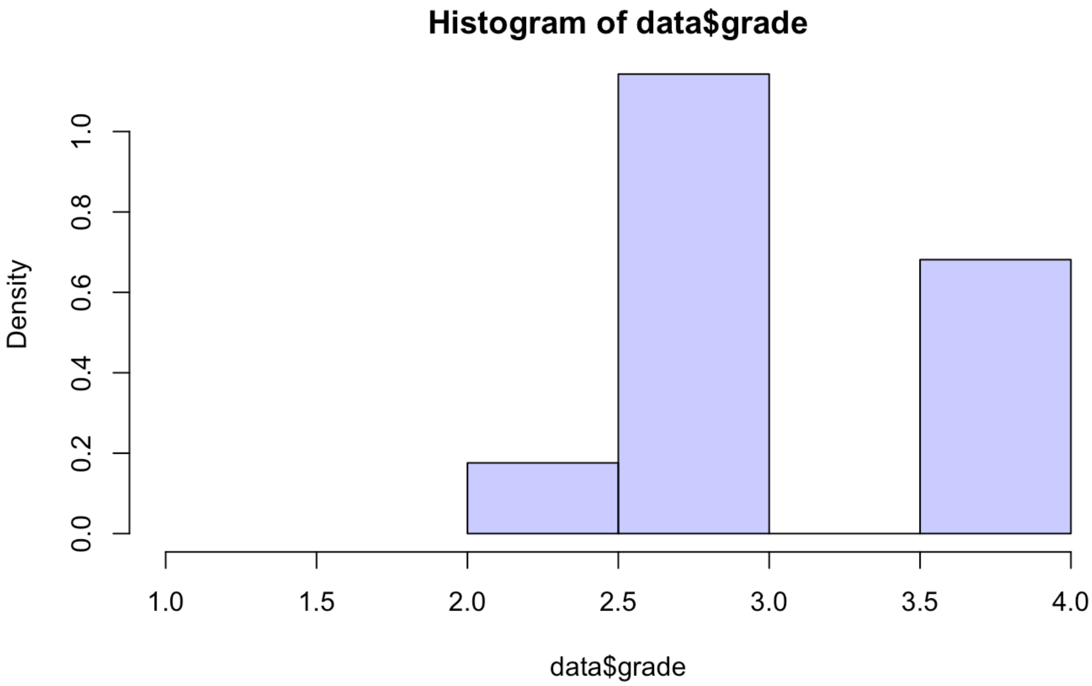
Hide

```
target <- c(1,2,2,2,2,3,3,3,4,4)
p2<-hist(target, col="darkgreen", ylab ="proportion of students", xlab
     ="Grade expected",freq=FALSE,breaks=c(1,2,3,4),ylim=c(0,1))
```



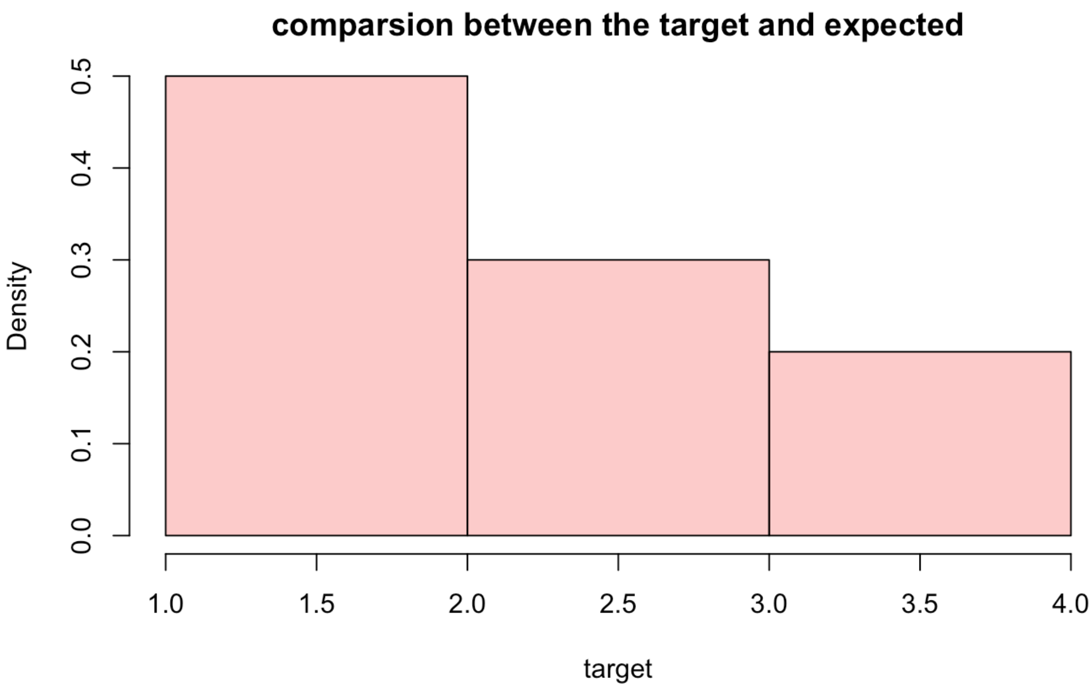
Hide

```
plot(p1,col=rgb(0,0,1,1/4),xlim=c(1,4),freq=FALSE,breaks=c(1,2,3,4,5))
```



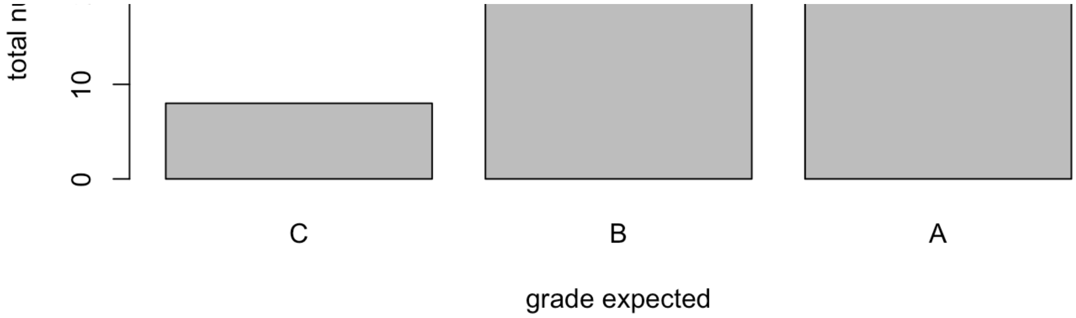
Hide

```
plot( main="comparsion between the target and expected", p2, col=rgb(1,0,0,1/4), xlim
=c(1,4), freq=FALSE)
```



Hide

```
table<-table(data$grade)
barplot(table,main = "number of students with expected grade", xlab = "grade expected
",ylab = "total number of students",name=c("C","B","A"))
```



Hide

```
proportion_C = length(which(data$grade==2))/sample
```

```
Error in length(which(data$grade == 2))/sample :
  non-numeric argument to binary operator
```