ST663 - Semester 1 - Assignment 3 - Solutions

Sean O'Riogain (18145426)

12 November 2018

```
knitr::opts_chunk$set(echo = TRUE)
getwd()
```

```
## [1] "C:/Users/oriogain/Dropbox/Maynooth/Statistical Methods/Semester 1 - Assignment 3"
```

Question 1

The data in file census1.csv is obtained from www.censusatschool.ie. It contains observations from the Phase 14 census at school Ireland project. The data contains a sample of 97 students.

The variables are Height FootR (right foot size) FootL (left foot size) in cm and Gender. Read the data into R and form a female only subset using the code below.

```
census <- read.csv("census1.csv")
censusF <- subset(census, Gender=="F")
str(censusF)
```

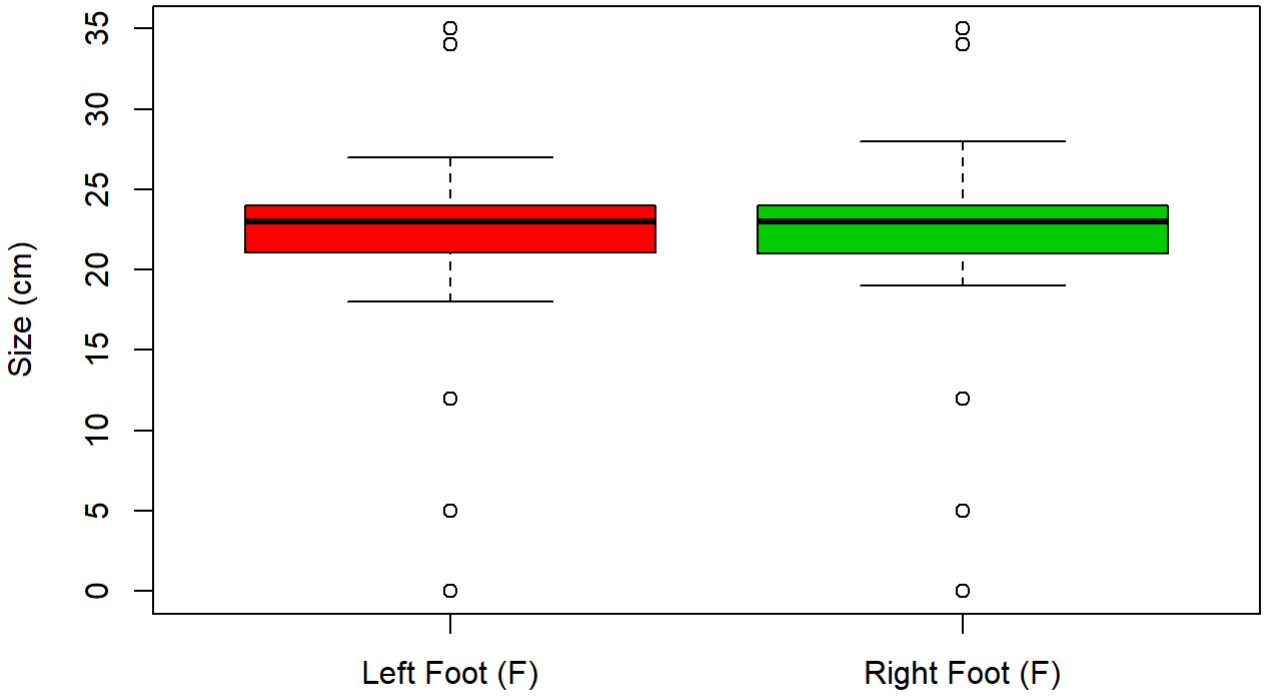
```
## 'data.frame':    54 obs. of  4 variables:
##  $ Height: num  168 156 162 179 165 126 0 176 154 178 ...
##  $ FootR : num  24 25 25 24 24 12 0 26 35 28 ...
##  $ FootL : num  24 25 25 24 23 12 0 26 35 27 ...
##  $ Gender: Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(censusF)
```

```
##    Height FootR FootL Gender
## 1     168    24    24      F
## 3     156    25    25      F
## 5     162    25    25      F
## 9     179    24    24      F
## 10    165    24    23      F
## 13    126    12    12      F
```

a. Construct a graph comparing FootR and FootL for females.

```
boxplot(censusF$FootL,censusF$FootR,col=c(2:3),
        names=c("Left Foot (F)","Right Foot (F)"),
        ylab="Size (cm)")
```



Also use a suitable hypothesis test.

State the appropriate null and alternative hypotheses.

Answer:

$$H_0 : \mu_l - \mu_r = 0$$

$$H_A : \mu_l - \mu_r \neq 0$$

Perform the test……

```
tails<-2                          # Two-tailed test

n<-length(censusF[,1]);n          # Sample size
```

```
## [1] 54
```

```
df<-n-1;df                        # Degrees of Freedom
```

```
## [1] 53
```

```
diff<-censusF$FootL-censusF$FootR   # Compute difference

xbar_diff<-mean(diff);xbar_diff     # Sample mean of diffs
```

```
## [1] -0.07777778
```

```
s_diff<-sd(diff);s_diff             # Sample Std Dev of diffs
```

```
## [1] 0.682301
```

```
c<-0.95                             # Level of Conf. (assumed)
alpha<-1-c;alpha                    # Level of Significance
```

```
## [1] 0.05
```

```
se_diff<-s_diff/sqrt(n);se_diff     # Standard Error of diffs
```

```
## [1] 0.0928494
```

```
z<-(xbar_diff - 0)/se_diff;z        # z-value
```

```
## [1] -0.8376767
```

```
pval<-tails*pnorm(-abs(z));pval     # p-value
```

```
## [1] 0.4022123
```

```
# Use t.test to sanity-check the calculated p-value
t.test(diff,df=df,alternative="two.sided")
```

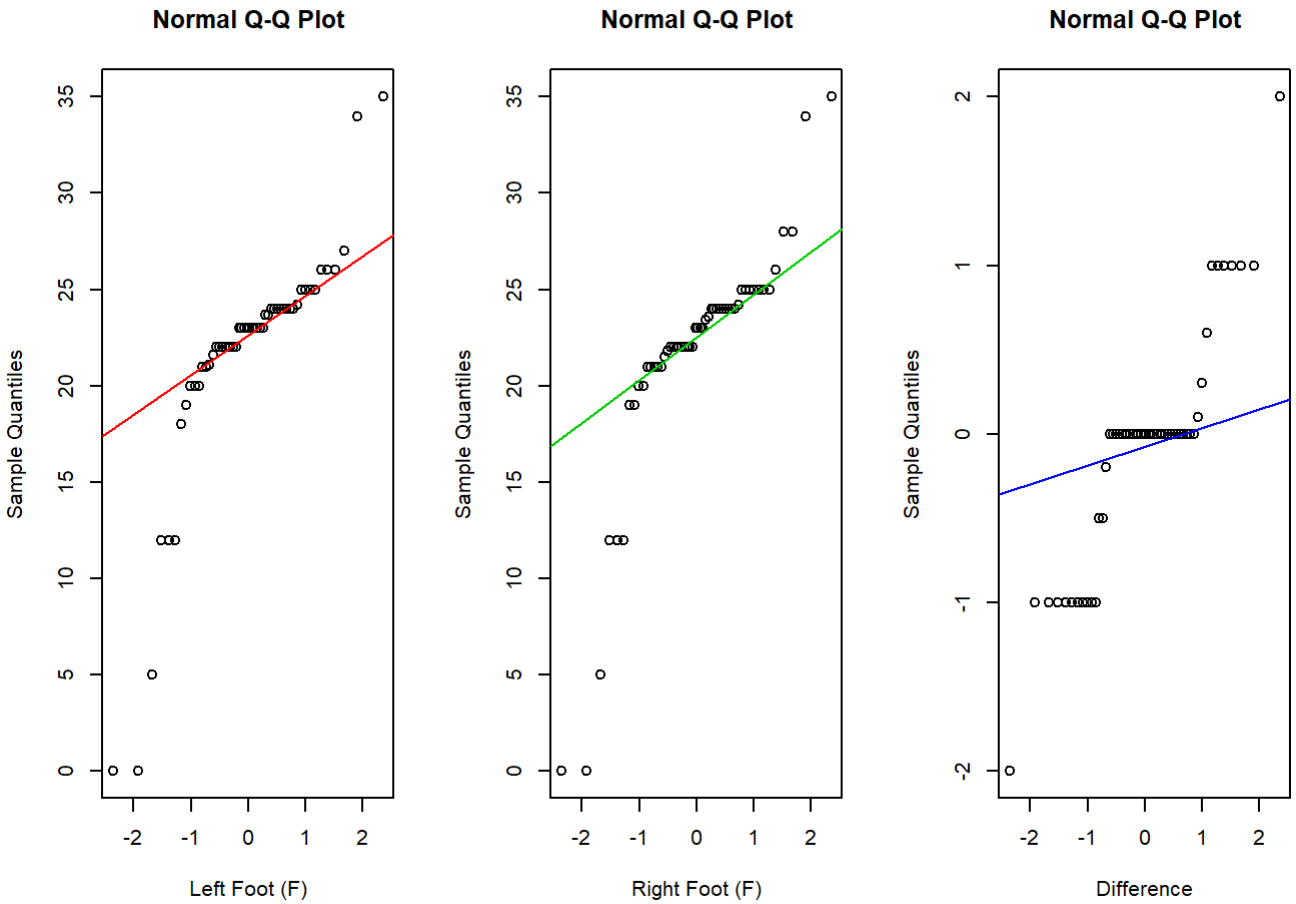
```
##
##  One Sample t-test
##
## data:  diff
## t = -0.83768, df = 53, p-value = 0.406
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2640101  0.1084545
## sample estimates:
##   mean of x
## -0.07777778
```

….and give your conclusions.

Answer: As the p-value (both manual & t.test) is greater than alpha (for a Level of Confidence assumed to be 0.95), we cannot reject H_0 and, therefore, we conclude (with the stated Level of Confidence) that, on average, there is no difference between left and right foot size in the general female population.

Check the normality of the relevant data using a QQ plot.

```
par(mfrow=c(1,3))

qqnorm(censusF$FootL,xlab="Left Foot (F)")  # QQPlot for Left
qqline(censusF$FootL,col=2)

qqnorm(censusF$FootR,xlab="Right Foot (F)") # QQPlot for Right
qqline(censusF$FootR,col=3)

qqnorm(diff,xlab="Difference")              # QQPlot for Diffs
qqline(diff,col=4)
```



Interpret your results.

Answer: None of the above plots exhibit any approximation to a normal distribution, all having significant tails on both the left and right. In particular, the difference data shows a stepped, linear distribution with significant clustering of data points at the zero (especially), -1 and +1 (cm) plateaux.

b. Construct a graph comparing FootR for males and females.

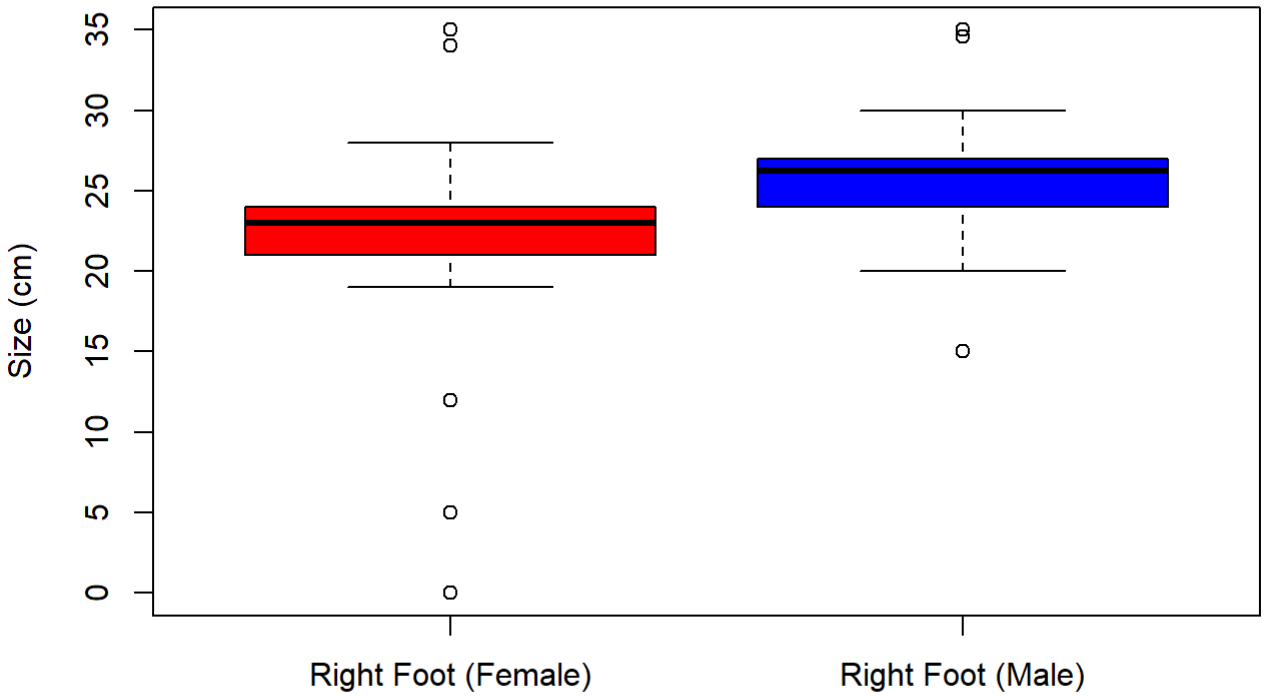
```
x<-subset(census, Gender=="F")$FootR # Extract the Female data
y<-subset(census, Gender=="M")$FootR # Extract the Male data

x<-na.omit(x)                        # Drop any NaNs
y<-na.omit(y)

head(x)                              # Take a peek at the data
```

```
## [1] 24 25 25 24 24 12
```

```
head(y)
```

```
## [1] 24.0 24.0 24.5 27.0 26.8 27.0
```

```
boxplot(x,y,col=c("red","blue"),     # Draw the box plots
        names=c("Right Foot (Female)","Right Foot (Male)"),
        ylab="Size (cm)")
```



Also use a suitable confidence interval.

```
mx<-mean(x);mx                    # Mean (Female)
```

```
## [1] 21.67593
```

```
sx<-sd(x);sx                      # Standard Dev. (Female)
```

```
## [1] 6.288622
```

```
my<-mean(y);my                    # Mean (Male)
```

```
## [1] 26.12619
```

```
sy<-sd(y);sy                      # Standard Dev. (Male)
```

```
## [1] 3.660139
```

```
n<-length(x);n                    # Sample size (Female)
```

```
## [1] 54
```

```
m<-length(y);m                    # Sample size (Male)
```

```
## [1] 42
```

```
s<-sqrt(((n-1)*sx^2 + (m-1)*sy^2)/(n+m-2));s # Sample Std. Dev
```

```
## [1] 5.304795
```

```
se<-s*sqrt(1/n+1/m);se            # Standard Error of diffs
```

```
## [1] 1.091397
```

```
t<-qt(.975,df=n+m-2);t            # t-value
```

```
## [1] 1.985523
```

```
e<-t*se;e                         # Margin of error
```

```
## [1] 2.166994
```

```
cil<-mx - my - e;cil              # Confidence Interval(low)
```

```
## [1] -6.617259
```

```
ciu<-mx - my + e;ciu              # Confidence Interval(high)
```

```
## [1] -2.283271
```

```
t.test(x,y)
```

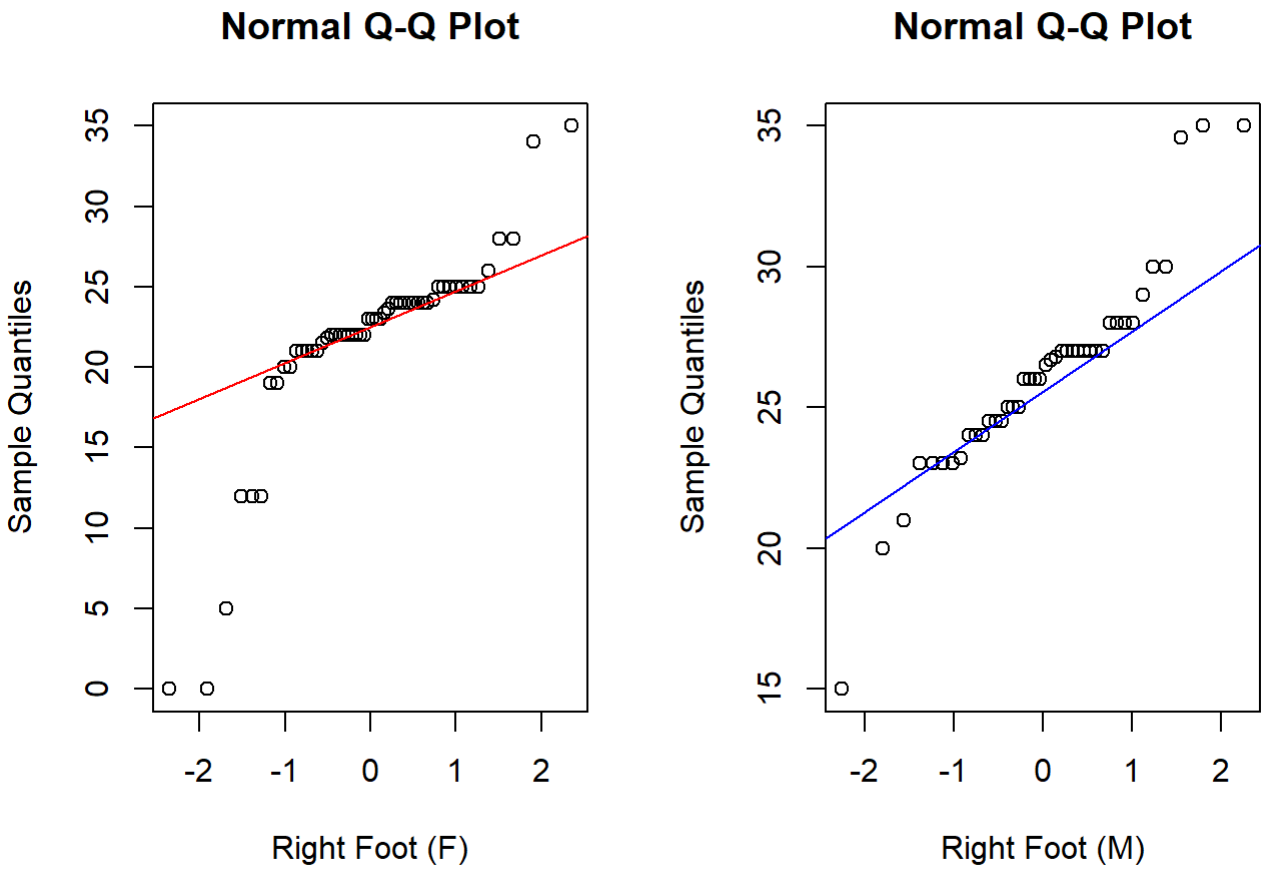
```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -4.3403, df = 87.713, p-value = 3.798e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.487998 -2.412531
## sample estimates:
## mean of x mean of y
##  21.67593  26.12619
```

Interpret your interval.

Answer: We are 95% confident that the average difference in the size of the right foot between males and females in the general population falls within the range of 2.28 to 6.62 cm (in favour of the males).

Check the normality of the relevant data using a QQ plot.

```
par(mfrow=c(1,2))

qqnorm(x,xlab="Right Foot (F)")             # QQPlot for women
qqline(x,col="red")

qqnorm(y,xlab="Right Foot (M)")             # QQPlot for men
qqline(y,col="blue")
```



Interpret your findings.

Answer: Neither the female or male sample data sets follow a normal distribution, having significant tails to the right and left.

Question 2

Of a random sample of 750 local residents, 400 were strongly opposed to the location of a hospital in the area. Does the sample provide sufficient evidence to conclude that a majority of residents oppose the new hospital?

a. What is the null hypothesis? What is the alternative hypothesis?

Answer:

$$H_0 : p_{opposed} <= 0.50$$

$$H_A : p_{opposed} > 0.50$$

b. Perform the hypothesis test.

```r
p<-0.50                                 # Population proportion (as per H0)
n<-750                                  # Sample size
m<-400                                  # Sample 'successes'
phat<-m/n;phat                          # Sample proportion
```

```
## [1] 0.5333333
```

```r
c<-0.95                                 # Level of Confidence (assumed)
alpha<-1-c;alpha                        # Level of Significance
```

```
## [1] 0.05
```

```r
tails<-1;                               # This is a single (right)-tailed test

if(n*phat > 10 & n*(1-phat) > 10){
  print("The Normal distribution approximation condition is satisfied.")
} else {
  print("The Normal distribution approximation condition is not satisfied.")
}
```

```
## [1] "The Normal distribution approximation condition is satisfied."
```

```r
se<-sqrt((p*(1-p))/n);se                # Standard error
```

```
## [1] 0.01825742
```

```r
z<-(phat-p)/se;z                        # Test Statistic (z-value)
```

```
## [1] 1.825742
```

```r
pvalue<-tails*pnorm(-abs(z));pvalue     # p-value
```

```
## [1] 0.03394458
```

```r
## Let's check the calculated value of the p-value using prop.test (without correction)
prop.test(x=m,n=n,p=p,alternative="greater",conf.level=c,correct=F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  m out of n, null probability p
## X-squared = 3.3333, df = 1, p-value = 0.03394
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.5033032 1.0000000
## sample estimates:
##         p
## 0.5333333
```

Find the p-value.

Answer: 0.03394 is the calculated p-value.

State your conclusions.

Answer: As the p-value is less that alpha (0.05), we can reject the Null Hypothesis (H_0) and conclude that the majority of residents are opposed to the proposed location of the hospital (at the 95% Level of Confidence).

Question 3

Suppose you have been hired to estimate the percentage of adults who are illiterate. You take a random sample of 600 adults, and find (using a standard literacy test) that 65 of them are illiterate.

a. Compute a 95% confidence interval for the percentage of the population who are illiterate.

```
n<-600                                  # Sample size
m<-65                                   # Sample 'successes'
phat<-m/n;phat                          # Sample proportion
```

```
## [1] 0.1083333
```

```
c<-0.95                                 # Level of Confidence (assumed)
alpha<-1-c;alpha                        # Level of Significance
```

```
## [1] 0.05
```

```
# Check for Normal approximation

if(n*phat>10 & n*(1-phat) > 10){
  print("Normal distribution approximation condition is satisfied")} else{
  print("Normal distribution approximation condition is NOT satisfied")
}
```

```
## [1] "Normal distribution approximation condition is satisfied"
```

```
sep<-sqrt((phat*(1-phat))/n); sep     # Standard Error
```

```
## [1] 0.0126884
```

```
z<-round(abs(qnorm((alpha/2))),2);z   # z-value
```

```
## [1] 1.96
```

```
e<-z*sep;e                              # Margin of error
```

```
## [1] 0.02486925
```

```
cil<-phat-e;cil                         # Lower CI value
```

```
## [1] 0.08346408
```

```
ciu<-phat+e;ciu                         # Upper CI value
```

```
## [1] 0.1332026
```

```
## Let's check the calculated values of the Confidence Interval using prop.test (without correction)
prop.test(x=m,n=n,conf.level=c,correct=F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  m out of n, null probability 0.5
## X-squared = 368.17, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.08591052 0.13573947
## sample estimates:
##         p
## 0.1083333
```

Answer: (Calculated) 95% CI = (0.083, 0.133), when rounded to 3 decimal places

Interpret your results.

Answer: We can be 95% confident that the population illiteracy rate lies somewhere in the 8.3% to 13.3% range.

b. How many people would need to be surveyed if the percentage of the population who are illiterate is required to be estimated within + or - 1% with 95% confidence?

$$se_p = \sqrt{\hat{p}(1-\hat{p})/n}$$

$$se_p^2 = \hat{p}(1-\hat{p})/n$$

$$n = \hat{p}(1-\hat{p})/se_p^2$$

$$e = z * se_p$$

$$se_p = e/z$$

$$n = \hat{p}(1-\hat{p})/(e/z)^2$$

Answer: In this case, we must find the value of n where the Margin of Error (e)=0.01 - i.e. n = 3711 (as shown below)

```
n<-round((phat*(1-phat))/(0.01/z)^2);n
```

```
## [1] 3711
```

Question 4

In 1889, Geissler collected data on 6,115 families of size 12. For each family the number of females was recorded.

```
female <- 0:12
nfamilies <- c(7,45,181,478,829,1112,1343,1033,670,286,104,24,3)
d <- data.frame(female=female, nfamilies=nfamilies)
d
```

```
##    female nfamilies
## 1       0         7
## 2       1        45
## 3       2       181
## 4       3       478
## 5       4       829
## 6       5      1112
## 7       6      1343
## 8       7      1033
## 9       8       670
## 10      9       286
## 11     10       104
## 12     11        24
## 13     12         3
```

Are the data consistent with the assumption of a binomial distribution?

Answer this with a chi-squared test.

Note: you will need to adjust the degrees of freedom, similar to the soccer data. You may need to pool the last two categories.

```
dc<-d                                   # Copy the data
po<-dc$nfamilies/sum(dc$nfamilies)      # Calculate data probabilities
sum(po)                                 # Check observed probabilities
```

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

```
dcp<-cbind(dc,po)                       # Append the observed probs.
prob<-sum(dcp$female*dcp$nfamilies)/(12*sum(dcp$nfamilies));prob  # Female probability
```

```
## [1] 0.480785
```

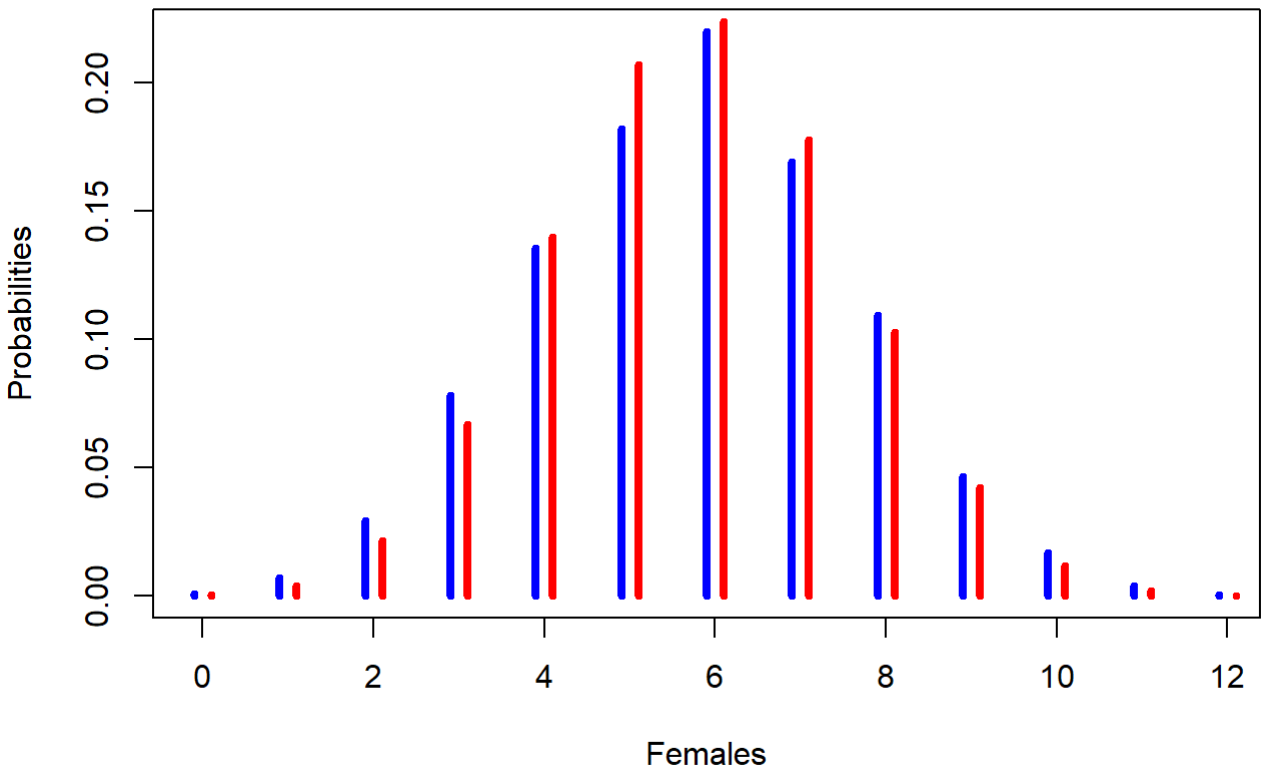
```
pt<-dbinom(x=c(0:max(dcp$female)),      # Generate the theoretical probs.
           size=length(dcp$female)-1,   # (as per Binomial distribution)
           prob=prob)
sum(pt)                                 # Check the theoretical probs.
```

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

```
dcp<-cbind(dcp,pt)                      # Append the theoretical probs.
dcp                                     # Display the data with po and pt
```

```
##    female nfamilies           po           pt
## 1       0         7 0.0011447261 0.0003838550
## 2       1        45 0.0073589534 0.0042653244
## 3       2       181 0.0295993459 0.0217229277
## 4       3       478 0.0781684383 0.0670502964
## 5       4       829 0.1355682747 0.1396969168
## 6       5      1112 0.1818479150 0.2069714320
## 7       6      1343 0.2196238757 0.2235943345
## 8       7      1033 0.1689288635 0.1774669993
## 9       8       670 0.1095666394 0.1027072791
## 10      9       286 0.0467702371 0.0422690325
## 11     10       104 0.0170073590 0.0117421375
## 12     11        24 0.0039247751 0.0019769154
## 13     12         3 0.0004905969 0.0001525494
```

```
plot(dcp$female-.1, dcp$po, lwd=4,         # Plot to compare po and pt
     type="h",col="blue",
     xlab="Females", ylab="Probabilities")
lines(dcp$female+.1, dcp$pt, lwd=4,
      type="h",col="red")
```



The above plot shows a reasonable match between the observed and theoretical (Binomial) probabilities

Let's perform the following hypothesis test:

H_0: $P_{observed}$ = $P_{theoretical}$

H_A: $P_{observed}$ <> $P_{theoretical}$

```
pop<-sum(po[1:2])                          # Pool the first 2 and last 2
pop<-c(pop,po[3:11])                       #   elements of po to avoid
pop[11]<-1-sum(pop)                        #   warning in chisq.test call

ptp<-sum(pt[1:2])                          # Pool the first 2 and last 2
ptp<-c(ptp,pt[3:11])                       #   elements of pt to avoid
ptp[11]<-1-sum(ptp)                        #   warning in chisq.test call

pop<-pop*sum(d$nfamilies)                  # Scale pop back up (values too
#   small for chisq.test)

ctest<-chisq.test(pop,p=ptp);ctest         # Perform the Chi Squared test
```

```
##
##  Chi-squared test for given probabilities
##
## data:  pop
## X-squared = 105.79, df = 10, p-value < 2.2e-16
```

```
pval<-1-pchisq(ctest$statistic, 9);pval
```

```
## X-squared
##         0
```

Answer: As the p-value achieved above is less than that of alpha (assumed to be 0.05, in this case), we must reject H_0 and conclude that the sample data is not sufficiently close (for a 95% Level of Confidence) to a Binomial distribution.

Question 5

Market researchers know that background music can influence the mood and purchasing behaviour of customers.

One study in a supermarket in Northern Ireland compared three treatments: no music, French accordion music and Italian string music. Under each condition, the researchers recorded the number of bottles of French, Italian and other wine purchased.

```
d<-matrix(c(30,39,30,11,1,19,43,35,35),nrow=3)
rownames(d)<-c("French wine","Italian wine","Other wine")
colnames(d)<-c("None","French music","Italian music")
d
```

```
##              None French music Italian music
## French wine    30           11            43
## Italian wine   39            1            35
## Other wine     30           19            35
```

a. Is there a relationship between the type of wine purchased and the type of music that is playing?

Answer:

H_0 : There is no relationship between the type of wine purchased and the type of music that is playing.

H_A : There is a relationship between the type of wine purchased and the type of music that is playing.

```
tab<-d
chisq.test(tab)
```

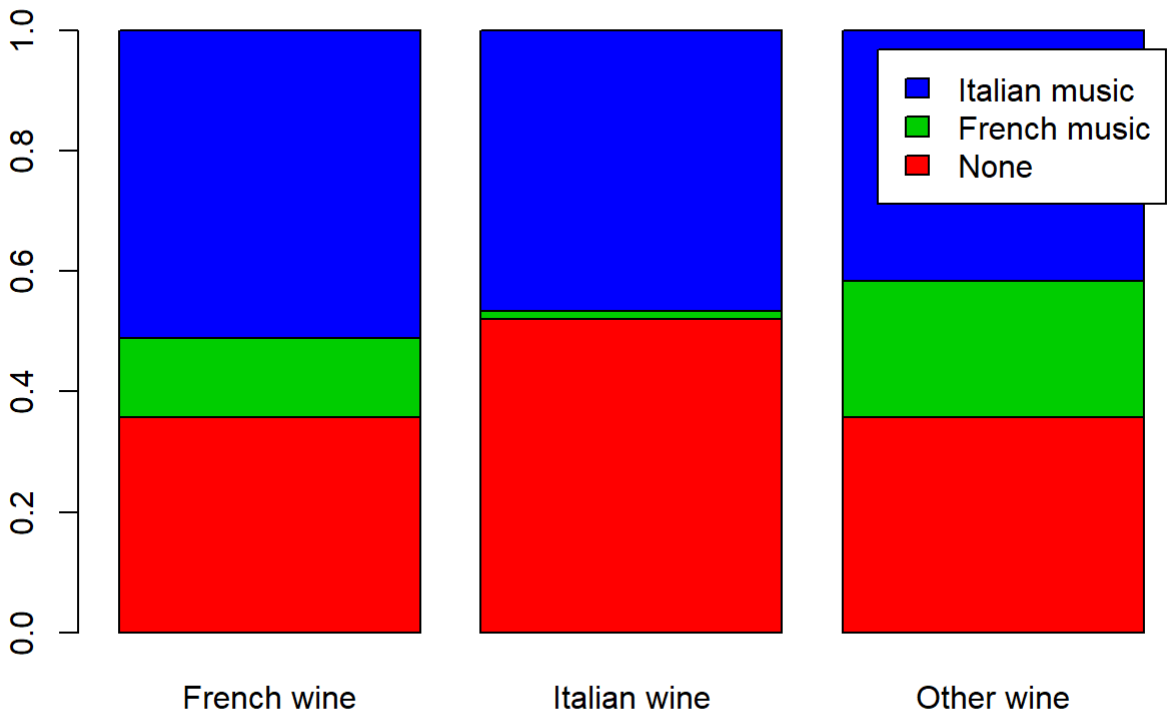
```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 18.279, df = 4, p-value = 0.001088
```

As the p-value achieved above is less than alpha (assumed to be 0.05, in this case), we reject H_0 and conclude that there is a relationship between the type of wine purchased and the type of music that is playing.

b. If the store manager wishes to improve the proportion of French wine sold, what music should she play?

Answer using a suitable barchart.

```
tab1 <- prop.table(tab,1)
barplot(t(tab1), legend=T, col=2:4)
```



Answer: The bar plots above indicate that, proportionately more bottles of French wine was sold while Italian music was playing than for any of the other two wine categories.Therefore, playing more Italian music (perversely) could help to sell more French wine.

c. Write a short summary of your conclusions.

Answer: In addition to the previous conclusion, the following points are worthy of note:

1. For this piece of analysis, we are assuming that, in a given period, the playing of music or not will not increase the average number of customers and that the same average number of bottles will be sold in total.
2. As the playing of French music appears to suppress the sales of Italian wine, it might also be argued that playing more French music could benefit the sales of the other wine categories (including French wine).

3. As the playing of no music appears to have more positive impact on the sales of Italian wine relative to the other 2 wine categories, playing less music might help to sell proportionally more Italian wine.
4. The sales of wine in the ‘Other’ category appears to be relatively less influenced by whether or not music is played, so modifying the choice of music should have relatively less impact on the proportion of sales made in that category.