Topic 3 Exercise Answers

Steve Paterson and Jordan Jones

2024-08-19

> Exercise 1
>
> Read in the “pima.20.txt” data file. The data file contains a random sub-sample of 10 negative and 10 positive individuals taken from the larger file. Check the data has been read into R correctly.

Read the data into R.

```
pima.20 <- read.table("pima.20.txt", header = TRUE)
```

Check the data has been read in correctly.

```
head(pima.20)
```

```
##   pregnant glucose diastolic triceps insulin  bmi age     test
## 1        4     158        78      NA      NA 32.9  31 positive
## 2        4     136        70      NA      NA 31.2  22 positive
## 3       10     168        74      NA      NA 38.0  34 positive
## 4        1     128        98      41      58 32.0  33 positive
## 5        2     174        88      37     120 44.5  24 positive
## 6        4     123        62      NA      NA 32.0  35 positive
```

```
summary(pima.20)
```

```
##     pregnant        glucose        diastolic        triceps
##  Min.   : 0.00   Min.   : 80.0   Min.   :46.00   Min.   :11.00
##  1st Qu.: 1.00   1st Qu.:113.2   1st Qu.:64.75   1st Qu.:21.00
##  Median : 2.00   Median :127.0   Median :71.00   Median :28.00
##  Mean   : 3.65   Mean   :125.7   Mean   :71.75   Mean   :27.18
##  3rd Qu.: 4.25   3rd Qu.:138.2   3rd Qu.:78.00   3rd Qu.:34.50
##  Max.   :15.00   Max.   :174.0   Max.   :98.00   Max.   :41.00
##                                                  NA's   :9
##     insulin            bmi             age           test
##  Min.   : 50.00   Min.   :22.10   Min.   :21.00   Length:20
##  1st Qu.: 65.75   1st Qu.:28.38   1st Qu.:22.00   Class :character
##  Median : 98.50   Median :32.00   Median :31.00   Mode  :character
##  Mean   :114.00   Mean   :33.10   Mean   :30.40
##  3rd Qu.:120.00   3rd Qu.:37.33   3rd Qu.:35.25
##  Max.   :285.00   Max.   :45.00   Max.   :47.00
##  NA's   :10
```

We can see that `test` is incorrectly labelled as a *character*. Change `test` to a *factor* and double check to see if it has been changed correctly.

```
pima.20$test<- factor(pima.20$test)
summary(pima.20)
```

```
##     pregnant        glucose        diastolic        triceps
##  Min.   : 0.00   Min.   : 80.0   Min.   :46.00   Min.   :11.00
##  1st Qu.: 1.00   1st Qu.:113.2   1st Qu.:64.75   1st Qu.:21.00
##  Median : 2.00   Median :127.0   Median :71.00   Median :28.00
##  Mean   : 3.65   Mean   :125.7   Mean   :71.75   Mean   :27.18
##  3rd Qu.: 4.25   3rd Qu.:138.2   3rd Qu.:78.00   3rd Qu.:34.50
##  Max.   :15.00   Max.   :174.0   Max.   :98.00   Max.   :41.00
##                                                  NA's   :9
##     insulin            bmi             age             test
##  Min.   : 50.00   Min.   :22.10   Min.   :21.00   negative:10
##  1st Qu.: 65.75   1st Qu.:28.38   1st Qu.:22.00   positive:10
##  Median : 98.50   Median :32.00   Median :31.00
##  Mean   :114.00   Mean   :33.10   Mean   :30.40
##  3rd Qu.:120.00   3rd Qu.:37.33   3rd Qu.:35.25
##  Max.   :285.00   Max.   :45.00   Max.   :47.00
##  NA's   :10
```

We can see now that `test` is now correctly labelled as a factor with two levels (negative and positive).

1. Analyse this smaller dataset to get the mean, standard deviation and standard error for glucose in the positive and negative groups and conduct a t-test to test whether there's a difference in glucose level between the two groups.

To do this, we need to first subset the data set by positive and negative test.

```
#for negative subset
pima.20.negative <- pima.20[pima.20$test=="negative",]

#for positive subset
pima.20.positive <- pima.20[pima.20$test=="positive",]
```

We have now created two subsets of data. Now we can go ahead and find the mean, standard deviation and standard error for each subset.

```
table(pima.20$test) #10 samples in each
```

```
##
## negative positive
##       10       10
```

```
#for negative
mean(pima.20.negative$glucose)
```

```
## [1] 113.7
```

```
sd(pima.20.negative$glucose)
```

```
## [1] 22.56866
```

```
sd(pima.20.negative$glucose)/sqrt(10)
```

```
## [1] 7.136837
```

```
#for positive
mean(pima.20.positive$glucose)
```

```
## [1] 137.7
```

```
sd(pima.20.positive$glucose)
```

```
## [1] 23.09425
```

```
sd(pima.20.positive$glucose)/sqrt(10)
```

```
## [1] 7.303044
```

We will now run a t-test to determine whether there is a difference in glucose levels between the two test groups. We use the t.test function and set glucose as the dependent (y) variable and test as the explanatory (x) variable.

```
t.test(glucose~test, data = pima.20)
```

```
##
##  Welch Two Sample t-test
##
## data:  glucose by test
## t = -2.3504, df = 17.99, p-value = 0.03036
## alternative hypothesis: true difference in means between group negative and group positive is not equal to 0
## 95 percent confidence interval:
##  -45.453779  -2.546221
## sample estimates:
## mean in group negative mean in group positive
##                  113.7                  137.7
```

From the model output we can see that $p < 0.05$. We can reject the null hypothesis and accept the alternative hypothesis.

2. Repeat for BMI in both the full dataset (pima_cleaned.txt) and the reduced dataset (pima.20.txt).

```
#for full dataset
#for negative
mean(pima.negative$bmi)
```

```
## [1] 30.97495
```

```
sd(pima.negative$bmi)
```

```
## [1] 6.571093
```

```
sd(pima.negative$bmi)/sqrt(475)
```

```
## [1] 0.3015024
```

```
#for positive
mean(pima.positive$bmi)
```

```
## [1] 35.21653
```

```
sd(pima.positive$bmi)
```

```
## [1] 6.422097
```

```
sd(pima.positive$bmi)/sqrt(248)
```

```
## [1] 0.4078036
```

```
#for reduced dataset
#for negative
mean(pima.20.negative$bmi)
```

```
## [1] 31.28
```

```
sd(pima.20.negative$bmi)
```

```
## [1] 7.649808
```

```
sd(pima.20.negative$bmi)/sqrt(10)
```

```
## [1] 2.419082
```

```
#for positive
mean(pima.20.positive$bmi)
```

```
## [1] 34.92
```

```
sd(pima.20.positive$bmi)
```

```
## [1] 4.928556
```

```
sd(pima.20.positive$bmi)/sqrt(10)
```

```
## [1] 1.558546
```

```
#t-test for full data set
t.test(bmi~test, data = pima)
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi by test
## t = -8.3635, df = 511.23, p-value = 5.826e-16
## alternative hypothesis: true difference in means between group negative and group positive is not equal to 0
## 95 percent confidence interval:
##  -5.237952 -3.245218
## sample estimates:
## mean in group negative mean in group positive
##               30.97495               35.21653
```

```
#t-test for reduced data set
t.test(bmi~test, data = pima.20)
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi by test
## t = -1.2649, df = 15.373, p-value = 0.2247
## alternative hypothesis: true difference in means between group negative and group positive is not equal to 0
## 95 percent confidence interval:
##  -9.760671  2.480671
## sample estimates:
## mean in group negative mean in group positive
##                  31.28                  34.92
```

3. What's the effect of reducing the sample size?

The most obvious effect of reducing the group size comes from the result from the t-tests we have conducted. In the full data set, our t-test is telling us our means are significantly different from each other, whereas in the reduced data set, our t-test is telling us they are not.

So, why are we getting different results?

This is where the importance of sample size in hypothesis testing comes in. As the sample size increases, your standard error decreases. Remember we calculate the standard error by dividing the standard deviation by the square root of the sample size. The square root of a larger sample size number means you are dividing the standard deviation by a larger number and so, the standard error will be smaller.

From comparing the standard errors that we calculated from the reduced and full data set we can see indeed, they are larger in the reduced data set compared to the full.

	Negative group SE	**Positive group SE**
Reduced data set	2.419	1.559
Full data set	0.302	0.408

When we have a smaller standard error, our test has more statistical power. In turn, increased statistical power means we are less likely to make a Type II error (failing to reject the null hypothesis). For example, if we only analysed the reduced data set here, we would have accepted the null hypothesis and failed to correctly reject it - leading to the wrong conclusion which could have major implications in some types of studies.

So in general, a larger sample size will give us a more precise and accurate p-value.

Please refer to the answer video if you want more explanation on this.

Exercise 2

Read in the "chickwts_edited.csv" file. The data set contains the weight (g) of chickens from two groups which have been fed different types of diet. Check the data has been read into R correctly.

Read in the "chickwts_edited.csv" file.

```
chickwts_edited <- read.csv("chickwts_edited.csv")
```

Check the data has been read into R correctly.

```
head(chickwts_edited)
```

```
##   weight      feed
## 1    408 sunflower
## 2    292 sunflower
## 3    283 sunflower
## 4    416 sunflower
## 5    378 sunflower
## 6    312 sunflower
```

```
summary(chickwts_edited)
```

```
##      weight          feed
##  Min.   : 80.0   Length:60
##  1st Qu.:225.2   Class :character
##  Median :292.0   Mode  :character
##  Mean   :285.3
##  3rd Qu.:336.0
##  Max.   :418.0
```

We can see that `feed` is incorrectly labelled as a *character*. Change `feed` to a *factor* and double check to see if it has been changed correctly.

```
chickwts_edited$feed<- factor(chickwts_edited$feed)

summary(chickwts_edited)
```

```
##      weight             feed
##  Min.   : 80.0   soybean  :30
##  1st Qu.:225.2   sunflower:30
##  Median :292.0
##  Mean   :285.3
##  3rd Qu.:336.0
##  Max.   :418.0
```

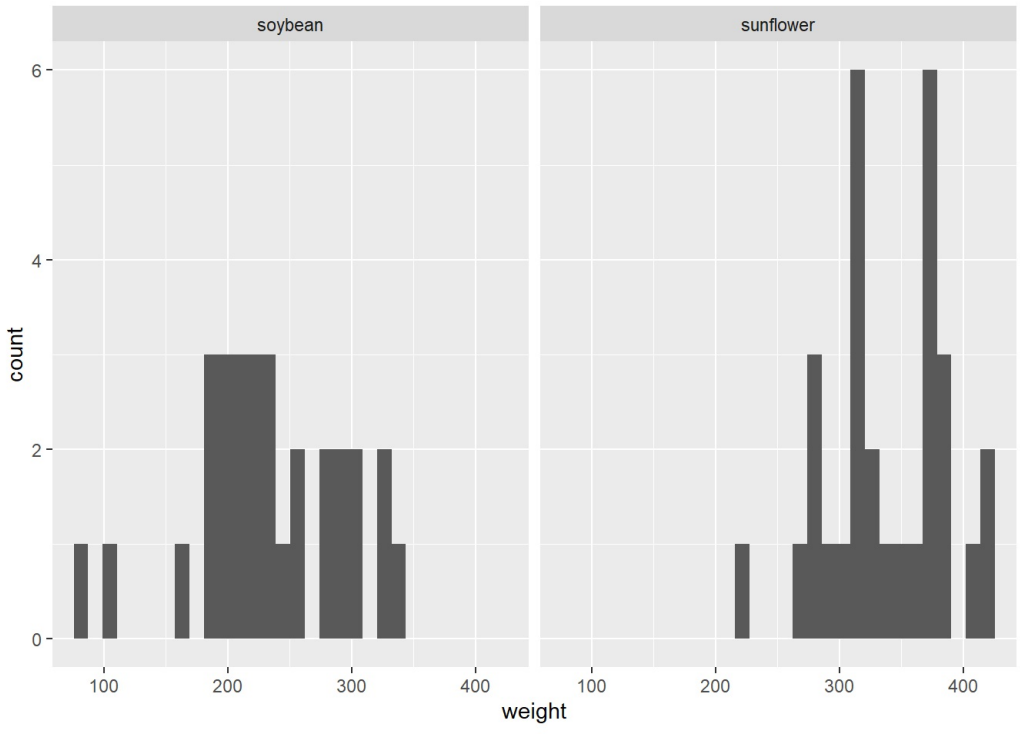
We can see now that `feed` is now correctly labelled as a factor with two levels (soybean and sunflower).

1. Determine whether the data are normally distributed (*Hint:* You may need to play around with the "bins=" function as discussed in topic 2 as the data set is small (n = 60).

We can do this using what we learnt from Topic 2. Remember we need to check the distribution of both the soybean and sunflower data sets individually.

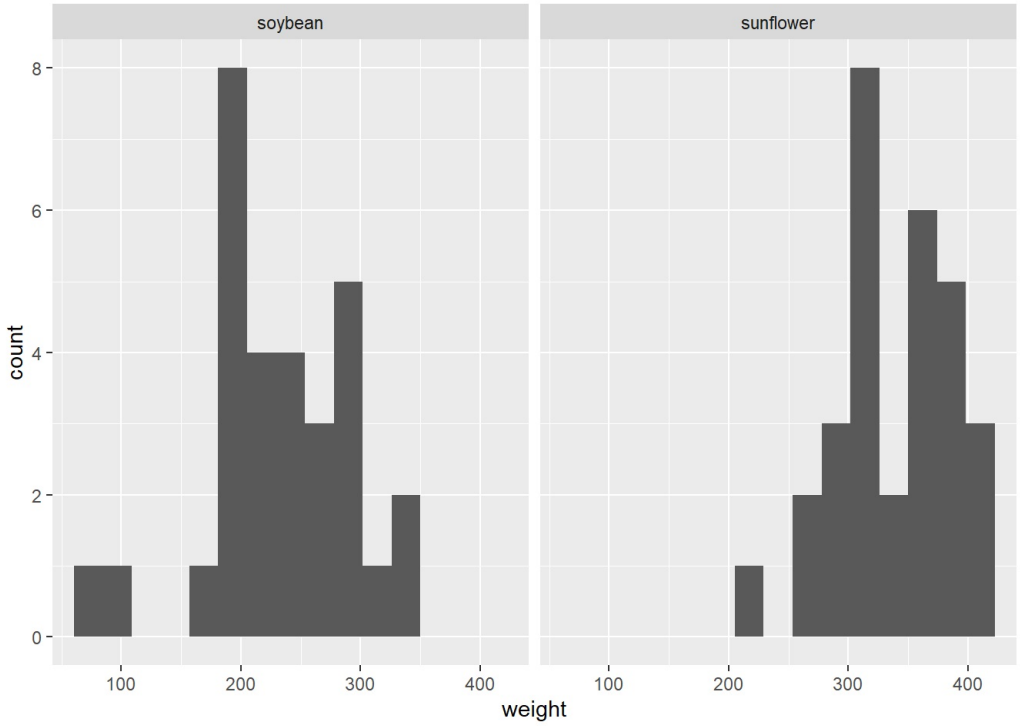
```
chick.hist <- ggplot(data=chickwts_edited, mapping = aes(x=weight))

chick.hist + geom_histogram() + facet_wrap(~feed)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The plot isn't very pretty. This is because we don't have a large data set (n = 60). We can change the plot to make it look a bit better by using the `bins=` function and playing around with the value. Since we have 30 data points for each feed group, it is best to use a value less than this. Let's go for 15.

```
chick.hist + geom_histogram(bins = 15) + facet_wrap(~feed)
```

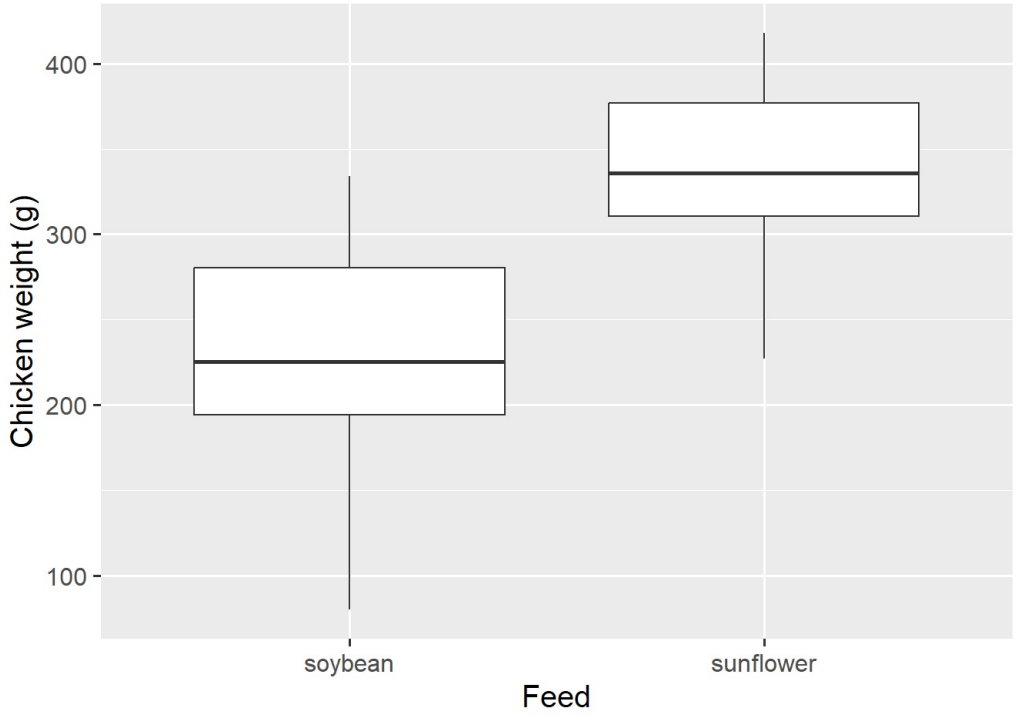


The plots look better but still slightly strange due to our sample size. However, you can see they follow a normal distribution.

2. Produce a plot of the data to publication standard using ggplot2.

As we are looking at the weight of chickens between two different types of feed, a boxplot would be best to visualise this data set.

```
p1 <- ggplot(chickwts_edited, aes(feed, weight))
p1 + geom_boxplot() +
  xlab("Feed") +
  ylab("Chicken weight (g)") +
  theme_grey(base_size = 15)
```



We give the plot new axis labels and increase the size to make it accessible to readers.

3. State the null and alternative hypotheses and conduct a t-test to determine whether we can reject or accept the null hypothesis.

Null hypothesis: There is no difference in the mean chicken weight between the soybean and sunflower feed
Alternative hypothesis: There is a difference in the mean chicken weight between the soybean and sunflower feed.
We use the t.test function and set weight as the dependent (y) variable and feed as the explanatory (x) variable.

```
t.test(weight ~ feed, data = chickwts_edited)
```

```
##
##  Welch Two Sample t-test
##
## data:  weight by feed
## t = -7.6084, df = 55.559, p-value = 3.582e-10
## alternative hypothesis: true difference in means between group soybean and group sunflower is not equal to 0
## 95 percent confidence interval:
##  -135.30373  -78.89627
## sample estimates:
##   mean in group soybean mean in group sunflower
##                231.7667                338.8667
```

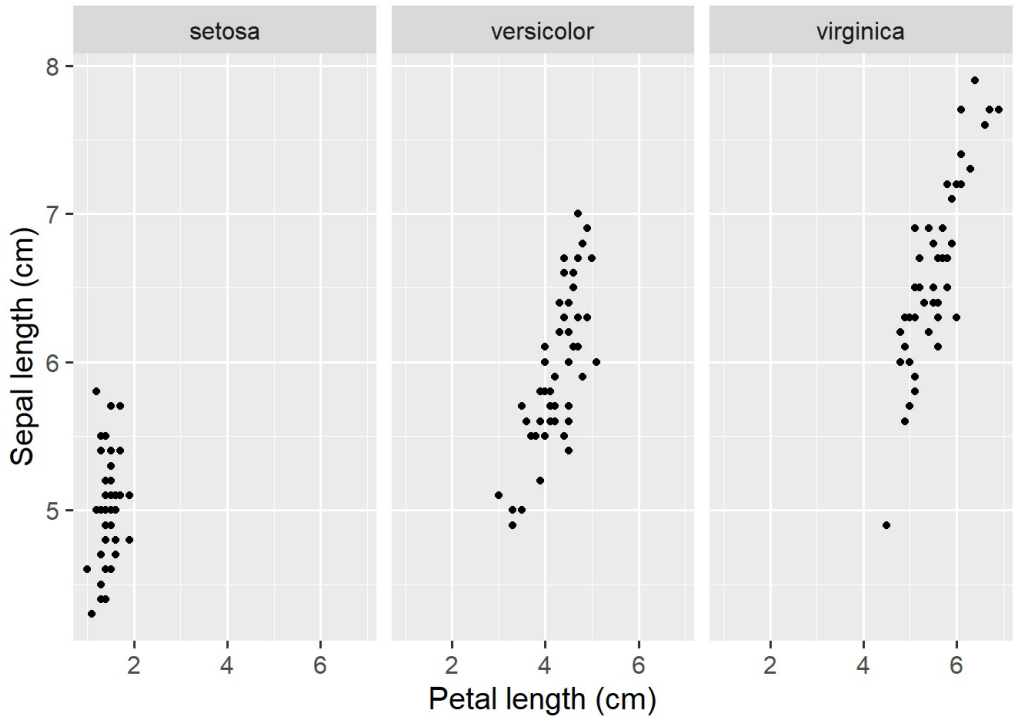
From the model output we can see that the p-value < 0.001. We can reject the null hypothesis and accept the alternative hypothesis.

Exercise 3

The “iris” data set (already loaded on *R*) contains sepal lengths (cm) and petal lengths (cm) of three different iris species. We would like to determine whether iris petal length is related to iris sepal length in three different species of iris.

1. Produce a plot of the data to publication standard using ggplot2.

```
p2 <- ggplot(iris, aes(Petal.Length, Sepal.Length))
p2 + geom_point() +
facet_wrap(~Species) +
theme_gray(base_size = 15) +
xlab("Petal length (cm)") +
ylab("Sepal length (cm)")
```



We give the plot new axis labels and increase the size to make it accessible to readers.

2. Subset the data for the three different iris species. Use the same approach as used at the beginning of this topic.

```
iris.setosa <- iris[iris$Species=="setosa",]

iris.versicolor <- iris[iris$Species=="versicolor",]

iris.virginica <- iris[iris$Species=="virginica",]
```

3. State the null and alternative hypotheses and conduct a correlation test to determine whether iris petal length is related to iris sepal length for each species. Is the result consistent across each species of iris?

Null hypothesis: Petal length and sepal length are not related in the setosa iris species.
Alternative hypothesis: Petal length and sepal length are related in the setosa iris species.

```
cor.test(iris.setosa$Sepal.Length, iris.setosa$Petal.Length, data = iris.setosa)
```

```
##
##  Pearson's product-moment correlation
##
## data:  iris.setosa$Sepal.Length and iris.setosa$Petal.Length
## t = 1.9209, df = 48, p-value = 0.0607
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01206954  0.50776233
## sample estimates:
##       cor
## 0.2671758
```

Null hypothesis: Petal length and sepal length are not related in the virginica iris species.
Alternative hypothesis: Petal length and sepal length are related in the virginica iris species.

```
cor.test(iris.virginica$Sepal.Length, iris.virginica$Petal.Length, data = iris.virginica)
```

```
##
##  Pearson's product-moment correlation
##
## data:  iris.virginica$Sepal.Length and iris.virginica$Petal.Length
## t = 11.901, df = 48, p-value = 6.298e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7714542 0.9210172
## sample estimates:
##       cor
## 0.8642247
```

Null hypothesis: Petal length and sepal length are not related in the versicolor iris species.
Alternative hypothesis: Petal length and sepal length are related in the versicolor iris species.

```
cor.test(iris.versicolor$Sepal.Length, iris.versicolor$Petal.Length, data = iris.versicolor)
```

```
##
##  Pearson's product-moment correlation
##
## data:  iris.versicolor$Sepal.Length and iris.versicolor$Petal.Length
## t = 7.9538, df = 48, p-value = 2.586e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6020680 0.8532995
## sample estimates:
##      cor
## 0.754049
```

No, the result is not consistent across species. Petal length and sepal length are significantly positively correlated in virginica and versicolor species but not in the setosa species.

Exercise 4

A biologist is interested in a bacterium which can protect insects from parasite infection. Specifically, they want to understand whether the bacterium, *Spiroplasma*, can protect fruit flies against trypanosomatid infection. Evidence from the lab suggests that they do. However, they would like to determine whether there is evidence of this from wild populations. To do this, a sample of flies were collected from the wild and screened for *Spiroplasma* infection and trypanosomatid infection using PCR methods.

Read in the "pcrscreen.csv" file. The data set contains the results from the PCR screen for *Spiroplasma* infection and trypanosomatid infection of individual fruit flies. Check over the data to ensure it has been read into R correctly.

Read in the "pcrscreen.csv" file.

```
pcrscreen <- read.csv("pcrscreen.csv")
```

Make sure you get into a habit of checking your data to ensure it has been read into R correctly.

```
head(pcrscreen)
```

```
##          spiro       tryp
## 1 S_uninfected T_infected
## 2 S_uninfected T_infected
## 3 S_uninfected T_infected
## 4 S_uninfected T_infected
## 5   S_infected T_infected
## 6 S_uninfected T_infected
```

```
summary(pcrscreen)
```

```
##     spiro              tryp
##  Length:556         Length:556
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

Again, we can see that R has read in `spiro` and `tryp` as *characters.* Change these to *factors* and check the data over again.

```
pcrscreen$spiro<- factor(pcrscreen$spiro)
pcrscreen$tryp<- factor(pcrscreen$tryp)

summary(pcrscreen)
```

```
##           spiro              tryp
##  S_infected  : 83   T_infected  :397
##  S_uninfected:473   T_uninfected:159
```

We can now see R is reading `spiro` and `tryp` as *factors* each with two levels, uninfected and infected.

1. Create and name a contingency table of *Spiroplasma* infection status x trypanosomatid infection status.

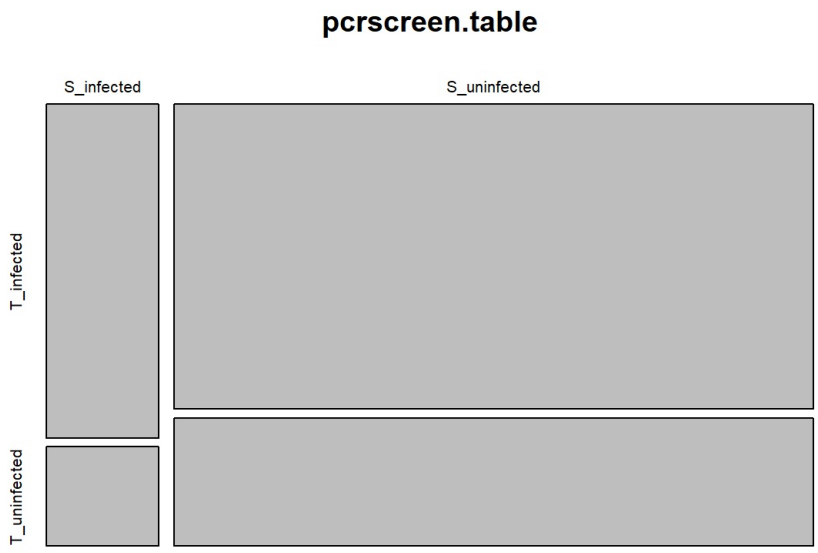
```
pcrscreen.table <- table(pcrscreen$spiro,pcrscreen$tryp)
pcrscreen.table
```

```
##
##                T_infected T_uninfected
##   S_infected           64           19
##   S_uninfected        333          140
```

2. Using the code `mosaicplot(Insert_your_table_name_here)` , create a plot of the contingency table. Would you predict there to be a relationship between *Spiroplasma* infection and trypanosomatid infection in fruit flies from the plot (does *Spiroplasma* protect against trypanosomatid infection)? Why do you predict this?

You would not typically present a mosaic plot in a publication/dissertation. However, if you prefer to visualise numbers - mosaic plots are a really useful way of visualising the data.

```
mosaicplot(pcrscreen.table)
```



The mosaic plot visually represents the contingency table. Each box represents the counts for each category. The boxes are wider for *Spiroplasma* uninfected flies as a greater number of flies were caught which were uninfected with *Spiroplasma* (n = 473) compared to infected with *Spiroplasma* (n=83).

From the mosaic plot, I wouldn't predict there to be a relationship between *Spiroplasma* infection and trypansomatid infection. I can see that in both the *Spiroplasma* infected fly group and the *Spiroplasma* uninfected fly group, we have roughly the same proportion of flies which are infected with trypansomatids (both boxes on the bottom row are roughly the same height).

If we were to see a smaller proportion of flies uninfected with trypansomatids in the *Spiroplasma* infected fly group compared to the *Spiroplasma* uninfected fly group (both boxes on the bottom row at different heights) then I might predict that there is a relationship and that *Spiroplasma* is protecting flies from trypansomatid infection.

Did you predict differently?

Arguably, from just eyeballing the plot, the proportion is slightly less but not by a lot. This is where statistical testing comes in - to formally tell us whether what we are observing could have arisen by random chance or not.

3. State the null and alternative hypothesis and conduct a Chi-square test. Was your prediction correct?

Null hypothesis: There is no relationship between *Spiroplasma* infection and trypanosomatid infection in fruit flies.
Alternative hypothesis: There is a relationship between *Spiroplasma* infection and trypansomatid infection in fruit flies.

```
chisq.test(pcrscreen$spiro,pcrscreen$tryp)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  pcrscreen$spiro and pcrscreen$tryp
## X-squared = 1.2443, df = 1, p-value = 0.2646
```

Our prediction was correct as the p-value > 0.05. We can accept the null hypothesis - there is no relationship between *Spiroplasma* infection and trypansomatid infection in fruit flies.

So it seems that at least from the wild, we don't see evidence of *Spiroplasma* protecting fruit flies against trypansomatid infection.