Class Example 1 – STAT2201 Student Weights and Heights

In lectures, the Height and Weight of the attending students was recorded. These have been anonymised and put into the file `Height_Weight_STAT2201.xlsx`. To read this file in the following commands are used:

```
> library(readxl)
> library(tidyverse)
> HWSTAT2201 <- read_excel("Height_Weight_STAT2201.xlsx",skip = 1)
```

The data set imported should have 33 rows and 2 columns. Now using this data, determine the following, where X is defined as the Height of a student and Y is defined as the Weight of a student:

(a) $P(X > 170\,\text{cm} \;\&\; Y < 80\,\text{kg})$

Solution: Here we simply need to count how many students have a height greater than 170 cm and a weight less than 80 kg. To do this in R we run the following command

```
> mean(HWSTAT2201$Height>170 & HWSTAT2201$Weight<80)
```

```
[1] 0.3939394
```

(b) $E(X)$

Solution: Here we simply need to take the mean of the heights of students which can be done in R using the following command:

```
> mean(HWSTAT2201$Height)
```

```
[1] 173.5758
```

(c) $E(X|Y \geq 70)$

Solution: Here we first need to filter out the cases of students with a Weight that is greater than 70 kg and then take the mean height of these students. To do this in R we first use the `filter` command to create a new data set, then use `mean` to get the expected value of X in the new data set.

```
> temp <- filter(HWSTAT2201,Weight>=70)
> mean(temp$Height)
```

```
[1] 180.5455
```

Now let us create two new variables, X which Height in categories of 10 cm between 155 and 185 cm, and Y which is Weight in categories of 10 kg. To do this in R we use the following commands:

```
> HeightF<-with(HWSTAT2201,cut(Height,breaks=c(0,155,165,175,185,300),
+ labels=c("<155","155-165","165-175","175-185","185+")))
> WeightF<-with(HWSTAT2201,cut(Weight,breaks=c(0,45,55,65,75,85,200),
+ labels=c("<45","45-55","55-65","65-75","75-85","85+")))
>
> HWSTAT2201<-bind_cols(HWSTAT2201,HeightF=HeightF,WeightF=WeightF)
```

We now can tabulate the counts of students who fall into the different categories.

```
> tab1<-with(HWSTAT2201,table(HeightF,WeightF))
> print(tab1)
```

```
         WeightF
HeightF   <45 45-55 55-65 65-75 75-85 85+
  <155      1     0     0     0     0   0
  155-165   1     3     2     2     0   0
  165-175   0     2     4     2     1   0
  175-185   0     0     4     3     2   3
  185+      0     0     0     0     0   3
```

(d) Are X' and Y' independent?

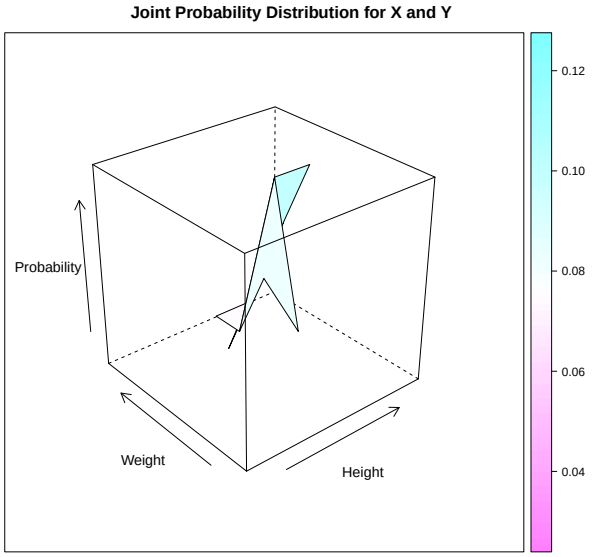
Solution: To work out if the new random variables are independent we need to calculate the joint probability mass function for them. To do this we do the following in R

```
> proptab1 <- prop.table(tab1)
> proptab1 <- as.data.frame(proptab1)
> proptab1 <- filter(proptab1,Freq>0)
```

HeightF	WeightF	Freq
<155	<45	0.030
155-165	<45	0.030
155-165	45-55	0.091
165-175	45-55	0.061
155-165	55-65	0.061
165-175	55-65	0.121
175-185	55-65	0.121
155-165	65-75	0.061
165-175	65-75	0.061
175-185	65-75	0.091
165-175	75-85	0.030
175-185	75-85	0.061
175-185	85+	0.091
185+	85+	0.091

We can visualise this distribution using the following commands

```
> library(lattice)
> wireframe(Freq~HeightF*WeightF,data=proptab1,drape=TRUE,
+ main= "Joint Probability Distribution for X and Y", xlab="Height",ylab="Weight",
+ zlab="Probability")
```



Looking at the probability mass function we can predict the value of Weight from the value of height for less than 150 cm and greater than 185 cm. As such the variables are not independent.

To confirm this we need to calculate the marginal distributions of X and Y. In R we do this as follows

```
> xproptab1<-proptab1%>%group_by(HeightF)%>%summarise(Freq=sum(Freq))
> yproptab1<-proptab1%>%group_by(WeightF)%>%summarise(Freq=sum(Freq))
```

Here `xproptab1` is the marginal distribution of X and `yproptab1` is the marginal distribution of Y. For the two variables to be independent we need that

$$f_{X,Y} = f_X f_Y$$

Taking the $X < 155\,\text{cm}$ and $Y < 45\,\text{kg}$ case we find that

$$P(X < 155, Y < 45) = 0.030303$$

$$\begin{aligned} P(X < 155)P(Y < 45) &= 0.030303 \times 0.0606061 \\ &= 0.0018365 \end{aligned}$$

$$0.030303 \neq 0.0018365$$

As the two probabilities are not equal, the variables are not independent.

(e) Plot Histograms of the marginal distributions $f_{X'}$ and $f_{y'}$

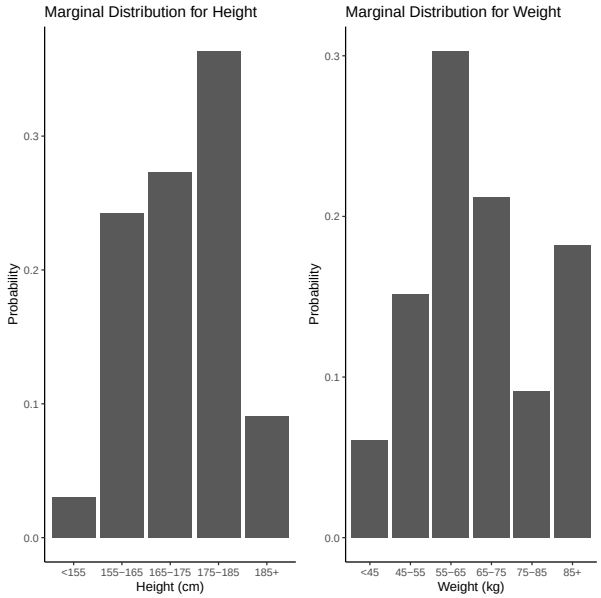
Solution: Now having calculated the marginal distributions in the previous part we just need to plot a histogram of each distribution. As we know the Probabilities for each level we just need plot a barchart for each distribution.

```
> p1<-ggplot(xproptab1,aes(x=HeightF,y=Freq))+geom_col()+
+ labs(title="Marginal Distribution for Height",x="Height (cm)",
+ y="Probability") + theme_classic()
```

```
> p2<-ggplot(yproptab1,aes(x=WeightF,y=Freq))+geom_col()+
+ labs(title="Marginal Distribution for Weight",x="Weight (kg)",
+ y="Probability") + theme_classic()
>
> library(gridExtra) #to arrange plots


Attaching package: 'gridExtra'
The following object is masked from 'package:dplyr':
    combine

> grid.arrange(p1,p2,nrow=1)
```



Looking at these histograms we see that Height is left skewed and Weight is slightly right skewed.

Class Example 2 – The Correlation Coefficient and Bivariate Normal Distribution

Consider a dataset consisting of the daily maximum temperature (in °C) recorded at Brisbane and Gold Coast each day from the start of 2015 to the middle of February 2017. The data is from the Australian Bureau of Meteorology, `http://www.bom.gov.au/`.
Denote the observations for Brisbane and Gold Coast (respectively) by:

$$x_1, \ldots, x_n \quad \text{and} \quad y_1, \ldots, y_n$$

Here n = 777. We will now perform the following:

(a) Load the data file and ensure you have the right dimensions.

(b) Calculate the following from the data:

(i) The sample means for temperatures at both locations.

(ii) The sample standard deviations for temperatures at both locations.

(iii) The correlation coefficient estimate.

(c) Use these estimates to describe the nature of the data in about 3 lines.

(d) Draw a scatter plot of the data.

(e) Assume the data comes from a bivariate Normal distribution.

(i) Draw a contour plot diagram of the estimated distribution.

(ii) Draw a 3D graph of the estimated distribution.

(iii) Write an expression for the probability of having a day in Brisbane with temperature less than 30 degrees and temperature in Gold Coast greater than 25 degrees.

Solution:

(a) Load the .csv file as follows:

```
> temps <- read.csv("BrisGCtemp.csv")
> head(temps)
```

```
  Year Month Day BrisMaxTemp.C. GoldCoastMaxTemp.C.
1 2017     2  15           30.4                 30.4
2 2017     2  14           30.3                 29.9
3 2017     2  13           35.6                 31.5
4 2017     2  12           37.6                 33.0
5 2017     2  11           37.0                 32.6
6 2017     2  10           33.0                 29.0
```

Then noticing that the data points begin on the second row and the respective temperatures are in the 4^{th} and 5^{th} columns, we extract arrays for Brisbane and Gold Coast as follows:

```
> Bris <- temps$BrisMaxTemp.C.
> GC <- temps$GoldCoastMaxTemp.C.
> c(length(Bris), length(GC))
```

```
[1] 777 777
```

(b) (i) The formulas for calculating the sample mean are as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{and} \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

```
> c(mean(Bris),mean(GC))
```

(ii) The **sample variance** is calculated using:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

and the sample standard deviation, s, is the positive square root of the sample variance. We calculate this here using both the formula and the built-in R `sd()` function.

```
> sqrt((sum(Bris^2)-length(Bris)*mean(Bris)^2)/(length(Bris)-1))
> sd(Bris)
> sqrt((sum(GC^2)-length(GC)*mean(GC)^2)/(length(GC)-1))
> sd(GC)
```

(iii) The correlation coefficient estimate is calculated using:

$$r_{xy} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\left[\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{\frac{1}{2}}}$$

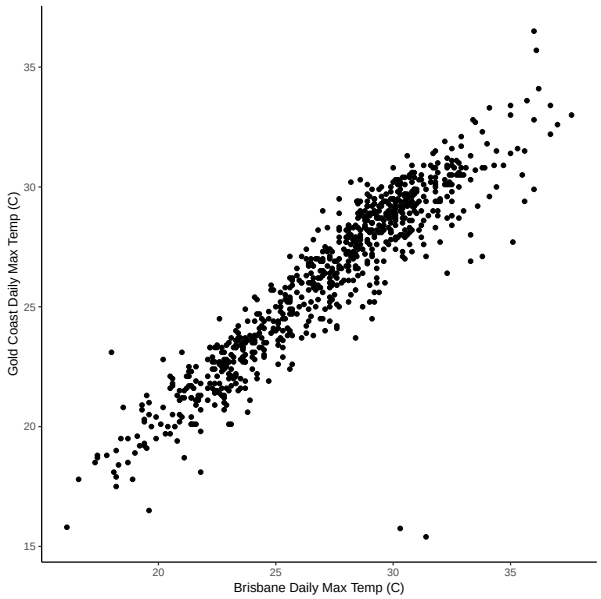
In R we use the following:

```
> cor(Bris,GC)
```

(c) We see that the estimated mean temperature in Brisbane is 27.16 and the estimated mean temperature in the Gold Coast is 26.16. These are with corresponding sample standard deviations 4.02 and 3.52. The correlation coefficient is 0.92 hinting that on days where the temperature is high in Brisbane, it is more likely to be high in the Gold Coast, and vice versa.
Note: Since the data is daily data over a long period, it is very plausible that it encompasses seasonal effects. Hence when considering the variation of the data (standard deviations of around 4) we need to be cognisant of the fact that much of the variation is perhaps due to seasonal effects. Separating the seasonal effects and random effects, is a matter of further study.

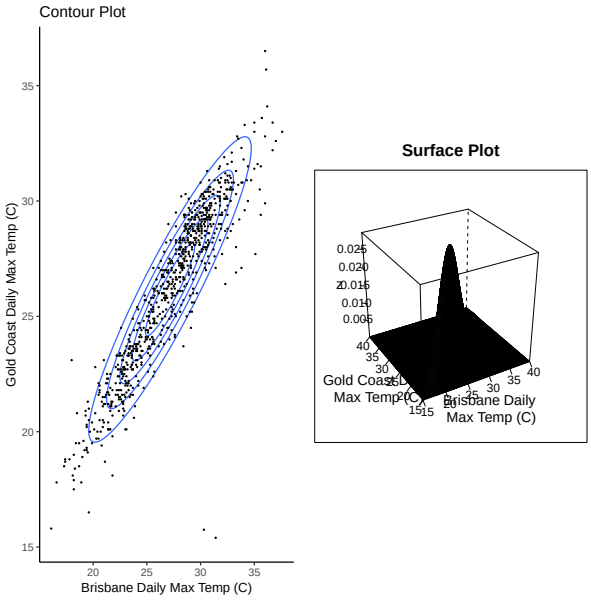
(d) We can visualise the data using a scatter plot:

```
> library(ggplot2)
> temp2 <- data.frame(Bris,GC)
> ggplot(temp2,aes(x=Bris,y=GC)) + geom_point() +
+ labs(x="Brisbane Daily Max Temp (C)",y="Gold Coast Daily Max Temp (C)") +
+ theme_classic()
```



(e) Understanding the code below is beyond the scope of the course. Nevertheless, it is good to see as it provides a neat representation of the data together with the fitted bivariate Normal Distribution, answering (i)-(ii).

```
> library(mvtnorm) #contains dmvnrom/rmvnorm
> library(ggplot2)
> library(gridExtra)
> library(lattice) #for wireframe
>
> n<-200
> mu <- c(mean(Bris),mean(GC))
> sigma <- matrix(c(sd(Bris)^2,cor(Bris,GC)*sd(GC)*sd(Bris),
+ cor(Bris,GC)*sd(GC)*sd(Bris),sd(GC)^2),ncol=2,byrow=TRUE)
>
> originaldat <- data.frame(Bris,GC)
> domain <- seq(15,40,length=n)
> xygrid <- expand.grid(x=domain,y=domain)
> z <-dmvnorm(xygrid,mean=mu,sigma=sigma)
> plotmat <- cbind(xygrid,z)
>
> #contour plot
> p1<-ggplot()+geom_contour(data=plotmat,aes(x=x,y=y,z=z))+
+ geom_point(data=originaldat,aes(Bris,GC),size=.2) +
+ theme_classic() + labs(title ="Contour Plot",x="Brisbane Daily Max Temp (C)",
+ y="Gold Coast Daily Max Temp (C)")
> # surface plot
> p2<-wireframe(z~x*y,data=plotmat,pretty=TRUE,
+ scale=list(arrows=FALSE),screen = list(z = 30, x = -60),
+ xlab="Brisbane Daily\n Max Temp (C)",ylab="Gold Coast Daily\n Max Temp (C)",
+ main="Surface Plot")
>
> grid.arrange(p1,p2,nrow=1)
```



(iii) Write an expression for the probability of having a day in Brisbane with temperature less than 30 degrees and temperature in Gold Coast greater than 25 degrees.

$$P(X < 30, Y > 25) = \int_{x=-\infty}^{30} \int_{y=25}^{\infty} f_{X,Y}(x, y; s_X, s_Y, \bar{x}, \bar{y}, \hat{\rho})\ dx\ dy$$

where $f_{X,Y}(\cdot)$ is the bivariate normal density:

$$f_{X,Y}(x, y; s_X, s_Y, \bar{x}, \bar{y}, \hat{\rho}) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}$$