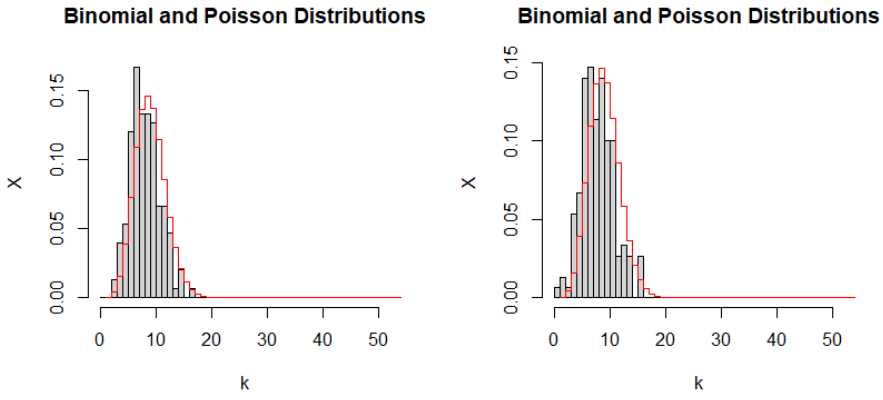
Statistical Data Analysis
Assignment 1

Matt Kuodis & James Zoryk.
2640428 : 2663347
Group: 36
Due Date: 15.02.2022.

Question 1.5.

For the first assignment, we plotted samples from the binomial and Poisson distributions, and compared the resulting histograms. We varied the input parameters for the binomial distribution functions, while keeping $\lambda = size \times probability$ as a constant, namely $\lambda = 7.5$. We looked at the histograms with n values at 150 and 500, and for each case we varied the *size* and *probability* inputs, under the restraint of their product being equal to 7.5. We plotted the graphs individually, as plotting multiple histograms would have required us to write a program with built in constants, or with too many inputs.

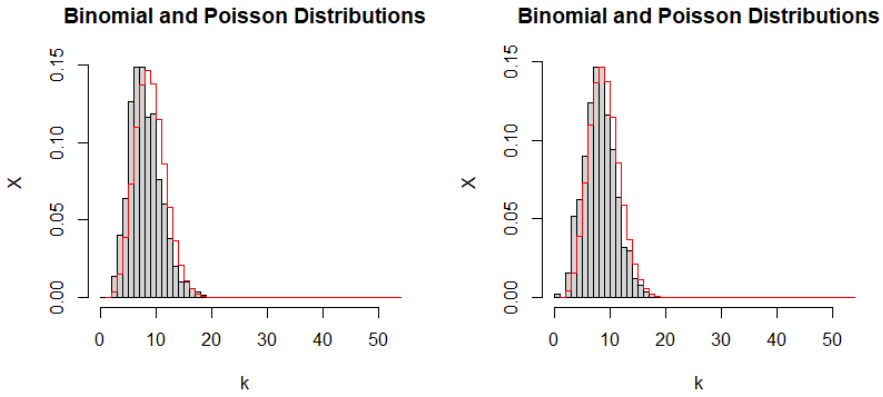


(a) $size = 1500, p = 0.005$ (b) $size = 3000, p = 0.0025$

Figure 1: Histograms for n = 150, Poisson shown in red

We note that the binomial samples vary quite noticeably under the changes in size and probability, but show no convergence to the Poisson distribution as the *size* variable is increased for small n, specifically $n = 150$. In fact, it seems that the binomial distribution maintains a structure much more similar to that of the Poisson for smaller

size in the case that n is relatively small, but for both histograms we note unexpected peaks and dips that remove any notion of a uniform structure.



(a) $size = 1500, p = 0.005$ (b) $size = 3000, p = 0.0025$

Figure 2: Histograms for n = 500, Poisson shown in red

The opposite, however, seems to be true for large n, as it seems that the distributions are much more similar overall, but in particular for larger *size*. The shape of the distribution for $n = 500$ is much more uniformly that of the Poisson, despite that there seems to be a slight deviation to the left in the case of the binomial. An increase in *size* seems to improve the overlap of the two histograms in the latter case, as the left-offset decreases as *size* increases.

Overall, it would seem that the binomial distribution is less unpredictable for large n, and seems to converge to the Poisson as n increases. The *size* hardly seems to alter the likeliness of the two distributions in either case, but appears to reduce the shift slightly to the left for the binomial distribution.

Question 1.6.

In this section we shall make a data summary that will compare the recent partly vaccinated percentage against COVID-19 to that of the Gross Domestic Product (GDP) per capita of countries that are in Asia. We are supplied with up-to-date data from the online source ourworldindata from which we carry out our analysis with the assistance of the software package RStudio. The RStudio code that we used to generate the data in this report can be found in the appendix.

Our first step to analysis the data is to process the data. We observe that the data is formatted into a CSV (Comma Separated Values) file, which once imported, the data is organised into named columns. Here we note that the column labeled 'Continent' has the data type of nominal scale, hence it can be grouped. This data type can be use to reduce the data down to countries that lie in Asia, since the raw data contains information about countries world wide, which are not in the interest for this report. Moreover, we are then interested in analysing the data in the two columns labeled 'gdp per capita' and 'partly vacc'. Observing that both these variables are continuous in nature, with 'partly vacc' being a percentage so it is scaled in the range of $[0, 100]$; and 'gdp per capita' being a qualitative ordinal variable. There are a couple of cases where there is no variable give for that specified field and instead it is marked 'NA', we can remove these from our summary while making note about how many we have encounter.

The data set of the two variables 'partly vacc' and 'gdp per capita' can now be numerical analysis with the results located in Table 1 and 2 respectfully.

Sample Size:	61
Mean:	54.67
SD:	24.33
Var:	592.32
Min:	1.13
1st qu.:	36.94
Median:	62.70
3rd qu.:	73.14
Max:	93.45
IQR:	36.20
NA's:	17

Table 1: Numerical analysis for partly vaccinated people in Asia

Sample Size:	61
Mean:	24719
SD:	28064.34
Var:	787606970
Min:	1479
1st qu.:	6289
Median:	12604
3rd qu.:	35237
Max:	116936
IQR:	28947.69
NA's:	15

Table 2: Numerical analysis for GDP per capita in Asia

The tables give us a lot of information about both the data sets, however this format is not that intuitive to analysis, with this in mind we can further our work by representing this data graphically. In order to do this we choose to plot both sets of data in a boxplot and also histogram . These both give us a better 'feel' for the distribution of the data. The plots for partly vaccinated people can be found in Figure 3 and the plots for GDP per capita can be found in Figure 5.

In Figure 3, we observe that the boxplot has a large spread between the 'minimum'

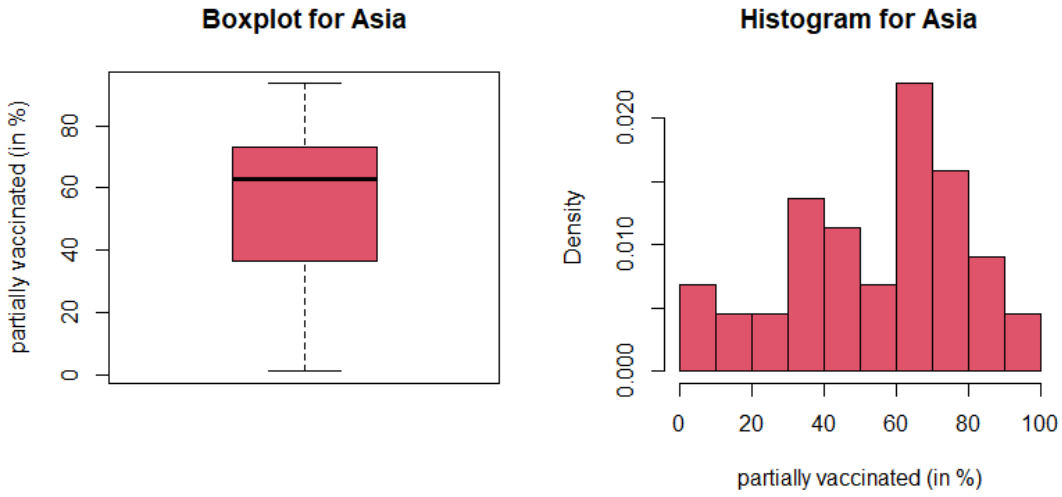


Figure 3: Graphical representation of partly vaccinated people for Asia. Left: Boxplot and Right: histogram.

and maximumn', which nearly covers the whole range possible for this variable, thus there are no outliers. From the boxplot and the histogram we can see that the data is skewed to the upper side, indicated by the black median line in the boxplot and the larger bars in the histogram.

Analysing Figure 4, we see that in this case we have three outliers, namely Qatar, Macao and Singapore. We also notice that the data in this case is positively skewed, which is seen by in the boxplot with the low median and in the histogram with the higher bars closer to the origin of the x-axis.

The next step in our analysis is to consider both these univariant data sets as a bivariant data set. A very effective method of showing the relationship between two variables is graphically through a scatter plot. In Figure 5 we plot GDP against the partly vaccinated people. This figure indicates that countries with higher GDP have more people partly vaccinated. The opposite, that countries with a lower GDP have less people vaccinated is not directly true. As we can see there are some countries with a relatively low GDP having higher partly vaccinated parentage than most other countries in Asia. One may speculate this maybe due to foreign aid given to this country.

We can also do a numerical summary of the bivariant data set, this can be seen in Tables 3, 4 and 5. These are matrix which numerical repentant the correlation between the GDP and the the partly vaccinated people.

Figure 4: Graphical representation of GPA per capita for Asia. Left: boxplot and Right: histogram.

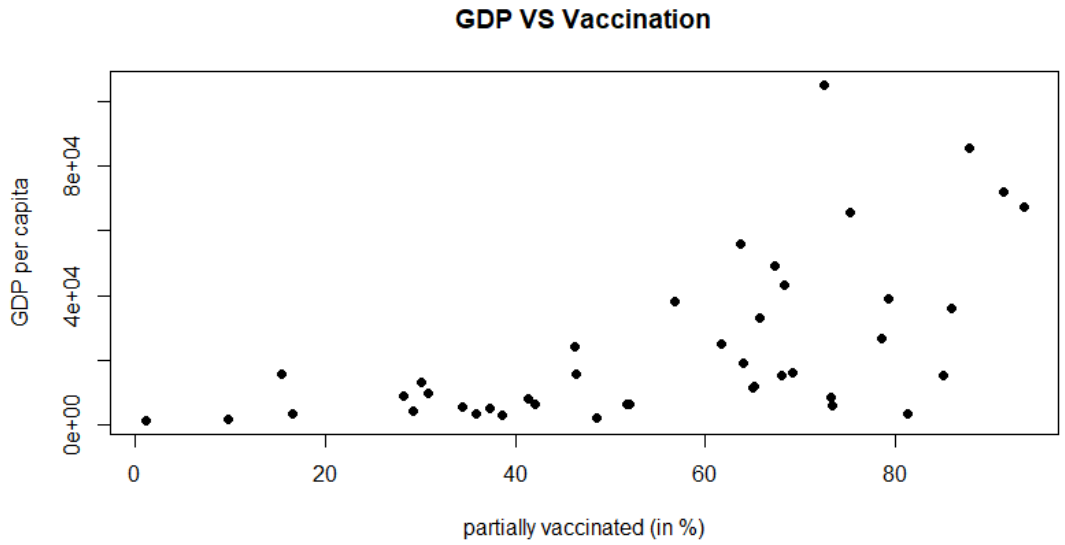


Figure 5: Scatter plot of GDP against partly vaccinated people.

Table 3: **Correlation matrix**

1.00	0.60
0.60	1.00

Table 4: **Spearman's rank correlation coefficients**

1.00	0.65
0.65	1.00

Table 5: **Kendall's rank correation coefficients**

1.00	0.47
0.47	1.00

1 Appendix

R code for exercise 1.4

```r
# Exercise 1.4

#Create function
lognorm <- function(n,mu,sigma){
  # set the seed 20220211
  # Our group number is 36
  input_seed = 20220211 + 36
  set.seed(input_seed)

  #create vector of random
  rnadom_data = rlnorm(n, mu , sigma )

  # Compute the Quantiles at %5 %10 %25 %50 %75 %90 %95.
  quants = quantile(rnadom_data, probs=c(0.05, 0.1,0.25,0.5,0.75,0.9,0.95),
       type=7)

  # Arithmetic Mean
  loc = mean(rnadom_data)

  # Standard Deviation
  spread = sd(rnadom_data)

  # Student numbers
  student_nu= c(2663347, 2640428)

  # put allcomputed variables intoa list
  mylist=list( quants, loc, spread, student_nu)

  # Save to disk the return of the list
  save(mylist, file="/home/james/SDA_files/myfile1_36.RData")
  }
```

code/exercise1.4.R

R code for exercise 1.5

```r
# Exercise 1.5

bin <- function(n,size,prob){
  #binomial data random create
  binomial_data = rbinom(n, size, prob)

  #poisson data create
  lambda_user= size * prob
  poisson_data = dpois( seq(0,size,1), lambda_user)

  #breaks
  breaks = 0:max((7*size*prob), max(binomial_data)+1)-0.001

  #plot historgram and line function
  { hist(binomial_data, prob=T, breaks )
    lines(poisson_data, type="s", col="red") }
}
```

code/exercise1.5.R

R code for exercise 1.6

```r
#Exercise 1.6 Covid Asia Data

# Import data from file and store as variable
attach(covid_data)
covid_data <- read.csv("owid_covid_data_selection.csv", header=TRUE)
covid_data = owid_covid_data_selection

# Create data with only Asain
covid_data_asia <- covid_data[covid_data$continent == "Asia", ]


#######################################
#             Vaccines
#######################################

# Set up a display two pllot
par(mfrow=c(1,2))

# Create first plot - Box plot
boxplot(covid_data_asia$partly_vacc,main="Boxplot of part vac for Asia",
        ylab="partially vaccinated (in %)", names="Asia", col=2)

# Create second plot -histgram
par(new=F)
hist(covid_data_asia$partly_vacc, main="Histogram of part vac  for Asia",
     xlab="partially vaccinated (in %)",
     prob=T,xlim=c(0,100), col=2)


# Partly vaccinated asia data
covid_data_asia_vac = covid_data_asia[,'partly_vacc']
covid_data_asia_vacNA <- covid_data_asia_vac[!is.na(covid_data_asia_vac)]
vac_sample_size  = nrow(covid_data_asia_vac)
vac_NAS = sample_size - length(covid_data_asia_vacNA)
vac_sum = summary(covid_data_asia_vacNA)
vac_iqr=IQR


#######################################
#             GDP
#######################################
# Set up a display two plot
par(mfrow=c(1,2))

# Create first plot - Box plot
boxplot(covid_data_asia$gdp_per_capita,main="Boxplot of GDP for Asia",
        ylab="GDP per Capita", names="Asia", col=2)

# Create second plot -histgram
par(new=F)
hist(covid_data_asia$gdp_per_capita, main="Histogram of GDP for Asia",
     xlab="GDP per Capita ($)",
     prob=T,xlim=c(0,116936), col=2)
```

```r


# gdp asia data
covid_data_asia_gdp = covid_data_asia['gdp_per_capita']
covid_data_asia_gdpNA <- covid_data_asia_gdp[!is.na(covid_data_asia_gdp)]
gdp_sample_size   = nrow(covid_data_asia_gdp)
gdp_NAS = sample_size - length(covid_data_asia_gdp)
gdp_sum = summary(covid_data_asia_gdpNA)
gdp_iqr=IQR(covid_data_asia_gdpNA)
sd(covid_data_asia_gdpNA)



########################################
#             Bivariate
########################################

# Bivariate data for Asia: percent vaccinated and gdp
bivariate_asia <- cbind(covid_data_asia$partly_vacc, covid_data_asia$gdp_
    per_capita)

# Column means
colMeans(bivariate_asia, na.rm=T)

# Remove data fields tha contain NA
bivariate_withoutNA <- bivariate_asia[-which(is.na(covid_data_asia$gdp_per_
    capita)
                                             | is.na(covid_data_asia$partly
                                                 _vacc)), ]

#Correlation matrix
cor(bivariate_withoutNA)
#Spearman's rank correlation coefficients
cor(bivariate_withoutNA, method = "spearman")
# Kendall's rank correation coefficients
cor(bivariate_withoutNA, method = "kendall")

# Plot bivariate scatter plot
par(mfrow=c(1,1))
plot(bivariate_withoutNA[, 1], bivariate_withoutNA[, 2],
     main="Scatterplot for Asia GDP VS Vaccination",
     ylab="GDP per capita", xlab="partially vaccinated (in %)", pch=19)
```

code/exercise1.6.R