ECO 274 LAB: Descriptive Statistics and Graphics

Learning Objectives

By the end of this lecture students will be able to:

- Create summary statistics for a single variable, data frame, group and by different groups
- Interpret the values in a summary table
- Generate graphical display of data: histograms, empirical cumulative distribution, QQ-plots, box plots, bar plots, dot charts and pie charts

Summary statistics by data types in a data frame

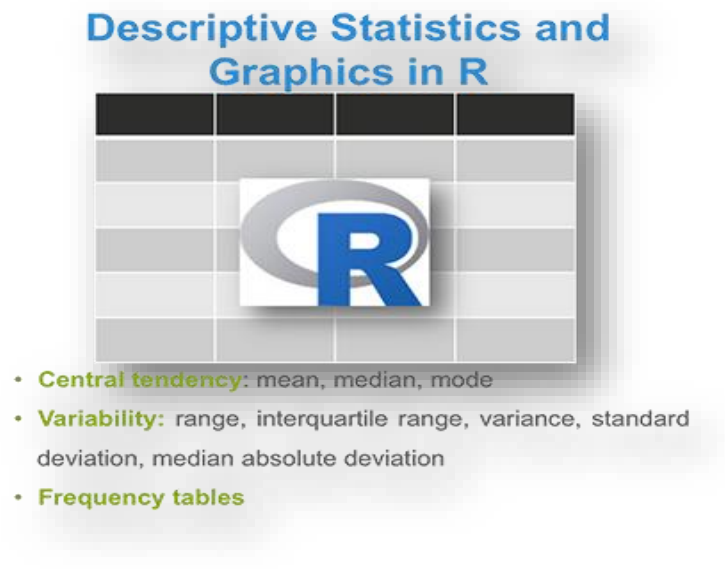
Descriptive statistics consist of describing simply the data using some summary statistics and graphics. We will explain each of the following data types and give examples of each:

Categorical (Binary as a special case)

Ordinal

Continuous

Time-to-Event.



Measure of central tendency: mean, median, mode

Roughly speaking, the central tendency measures the "average" or the "middle" of your data. The most commonly used measures include:

- the mean: the average value. It's sensitive to outliers.
- the median: the middle value. It's a robust alternative to mean.
- and the mode: the most frequent value

```
#we will be working with airquality data datasets and computing some
descriptive statistics.
# Recall that we can compute the mean Temp by "extracting" the variable Temp
from the dataset using the $ function as follows:

data("airquality")
View(airquality)
mean(airquality$Temp)
median(airquality$Temp)
var(airquality$Wind)
```

If we don't want to keep using the "$" sign to point to the data set, we a can use the attach command to keep the data set as the current or working one in R, and then just call the variables by name. For example, the above can then be accomplished by:

```
attach(airquality)

var(Wind)

# Once we are finished working with this data set, we can use the detach()
command to remove this data set from the working memory.
```

By default, we get the minimum, the maximum, and the three quartiles — the 0.25, 0.50, and 0.75 quantiles. The difference between the first and third quartiles is called the interquartile range (IQR) and is sometimes used as an alternative to the standard deviation.

Measures of variability gives how "spread out" the data are.

Range: minimum & maximum

Range corresponds to biggest value minus the smallest value. It gives you the full spread of the data. Quartiles divide the data into 4 parts. Note that, the interquartile range (IQR) - corresponding to the difference between the first and third quartiles - is sometimes used as a robust alternative to the standard deviation.

```
summary(Wind)
quantile(airquality$Temp)

To compute deciles (0.1, 0.2, 0.3, …., 0.9), use this:

quantile(airquality$Temp, seq(0, 1, 0.1))

To compute the interquartile range, type this:
```

```
IQR(airquality$Temp)

# It is also possible to obtain other quantiles; this is done by adding an
argument containing the desired percentage cut points. To get the deciles,
use the sequence function:

 pvec <- seq(0,1,0.1) #sequence of digits from 0 to 1, by 0.1

 pvec

quantile(airquality$Temp, pvec)

#We can also get summary statistics for multiple columns at once, using the
apply() command.

apply(airquality, 2, mean, na.rm=T)

# for a matrix 1 indicates rows, 2 indicates columns
```

Summary Function

There is also a summary function that gives a number of summaries on a numeric variable (or even the whole data frame!) in a nice vector format:

```
summary(airquality)

# Notice that "Month" and "Day" are coded as numeric variables even though
they are clearly categorical. This can be mended as follows, e.g.:

airquality$Month = factor(airquality$Month)

summary(airquality)
```

Descriptive statistics in R with pastecs package does bit more than simple describe () function. It also Calculates

- number of missing values and null of each column in R
- number of non missing values of each column
- sum , range ,variance and standard deviation etc for each column

```
library(pastecs)
stat.desc(airquality)
```

```
# Descriptive statistics in R with Hmisc package calculates the  distinct
value of each column, frequency of each value and proportion of that value in
that column.

library(Hmisc)
describe(airquality)

## Using stargazer package

library(stargazer)
stargazer(airquality, type = "text",
  summary.stat = c("min", "p25", "median", "p75", "max", "median", "sd"))

stargazer(airquality, type = "text", title = "Airquality:Summary Statistics",
out = "table.txt")
```

Which measure to use?

- Range. It's not often used because it's very sensitive to outliers.
- Interquartile range. It's pretty robust to outliers. It's used a lot in combination with the median.
- Variance. It's completely uninterpretable because it doesn't use the same units as the data. It's almost never used except as a mathematical tool
- Standard deviation. This is the square root of the variance. It's expressed in the same units as the data. The standard deviation is often used in the situation where the mean is the measure of central tendency.
- Median absolute deviation. It's a robust way to estimate the standard deviation, for data with outliers. It's not used very often.
- In summary, the IQR and the standard deviation are the two most common measures used to report the variability of the data.

```
##An example using actual financial assets data

library(zoo)
library(MASS)
library(vars)
library(psych)
df_ret <- read.csv(file = "ret_final.csv", header =
TRUE,stringsAsFactors=FALSE)

start_date = "1-03-2019"
end_date = "12-23-2021"

x = ts(df_ret[,1],start= c(2019,1),frequency = 214)
DATE = index(x)

df_ret1 <- df_ret[,-1]

k = ncol(df_ret1)
```

```
### Full sample Summary statistics_Return

SS = matrix(NA,ncol=5,nrow=k)
rownames(SS) = colnames(df_ret1)
colnames(SS) = c("Obs","Minimum","Mean","Maximum","SD")
for (i in 1:k) {
  SS[i,] =
c(nrow(df_ret1),min(df_ret1[,i]),mean(df_ret1[,i]),max(df_ret1[,i]),sd(df_ret
1[,i]))
}
round(SS,4)
library(stargazer)

sum_ret_full <- psych::describe(df_ret1)
sum_ret_full
```

```
## Homework#02: Replication (due on 27th Oct, 4 pm)
 ## Replicate the summary output Table 1 from the paper "Dynamic
Connectedness of UK Regional Property Returns." Dataset is available on my
github page.
Ref: Antonakakis, N., Chatziantoniou, I., Floros, C., & Gabauer, D. (2018).
The dynamic connectedness of UK regional property returns. Urban Studies,
55(14), 3110-3134.
```


Acknowledgement

1. Simon Ejdemyr (2015), https://sejdemyr.github.io/r-tutorials

2. https://statisticsglobe.com

3. Harvard Chan Bioinformatics Core, R for data science lab