THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Analysis of Engineering and Scientific Data

Semester 1 – 2019

Sabrina Streipert s.streipert@uq.edu.au

Descriptive Statistics

- Visualisation of the data.
- Analysis and presentation of characteristics of the data.

Data types

Possible data types:

1. *Continuous quantitative* data $\longrightarrow$ values in continuous range (height, width, length, temperature, humidity, volume, area, and price)
2. *Discrete qualitative* data (*factor / categorical variable*) $\longrightarrow$ values in discrete range (number of family members, gender (male or female), count of objects).

   Discrete Sub-types:

- *Nominal* factors = variables without order, such as males and females.
- *Ordinal* factors = variable with a certain order, such as *age group*.

Data configurations

- Many possible data configurations
- Each configuration will consist of continuous and discrete (ordinal and nominal) variables.
- **Major configuration types:**
  - A single sample configuration consists of m scalars:
    $\mathcal{D} = \{x_1, x_2, \ldots, x_m\}$.
    Nr of fisherman per day; $m = 365$ and $x_i = 0, 1, \ldots$.
  - Two (or more) sets of samples:
    $$\mathcal{D} = \left\{ \left\{x_1^1, \ldots, x_{m_1}^1\right\}, \left\{x_1^2, \ldots, x_{m_2}^2\right\}, \ldots, \left\{x_1^k, \ldots, x_{m_k}^k\right\} \right\}.$$
    Nr of fisherman per day in k different regions.
  - Data tuples: $\mathcal{D} = \{(x_{1,1}, x_{1,2}), (x_{2,1}, x_{2,2}), \ldots, (x_{m,1}, x_{m,2})\}$.
    $x_{i,1}$ = Nr of fisherman at ith day, $x_{i,2}$ is the number of fishing nets used at day i.
  - Generalization of tuples to vectors:
    $$\mathcal{D} = \{(x_{1,1}, \ldots, x_{1,n}), \ldots, (x_{m,1}, \ldots, x_{m,n})\}$$
    $x_{i,1}$ = Nr of fisherman at ith day, $x_{i,2}$ is the number of fishing nets used at day i, $x_{i,3}$ = Sea-Surface temperature at day i, etc.

1. Data tables

The table **rows** represent observed measurements for *independent* variables (**columns**).

Observ.	variable 1	variable 2	$\cdots$	variable i	$\cdots$	variable n
1	.	.	.	.	.	.
2	.	.	.	.	.	.
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	.	.	.	.	.	.

```
library(carData)
D <- Arrests
tail(D)
#or alterantive:
library(data.table)
print(data.table(D))
```

```
     released colour year age    sex employed citizen checks
5221      Yes  White 2002  22   Male      Yes     Yes      0
5222      Yes  White 2000  17   Male      Yes     Yes      0
5223      Yes  White 2000  21 Female      Yes     Yes      0
5224      Yes  Black 1999  21 Female      Yes     Yes      1
5225       No  Black 1998  24   Male      Yes     Yes      4
5226      Yes  White 1999  16   Male      Yes     Yes      3
```

Figure: Data on police arrests in Toronto for possession of marijuana.

Data summarization

A *statistic* is a numerical quantity, such as a proportion, that is computed from a sample $x_1, \ldots, x_m$.

```
library(dplyr)
```

```r
D1 <- D %>% group_by(sex) %>% summarize(Count_Arrests = n(), Proportion = Count_Arrests/nrow(D))
D1
```

```
  sex    Count_Arrests Proportion
  <fct>          <int>      <dbl>
1 Female           443     0.0848
2 Male            4783     0.915
```

Study a **correlation** between the two factor variables using the so called **contingency table**:

```r
D2 <- D %>% mutate(sex = ifelse(sex=="Female",1,0), employed = ifelse(employed == "Yes",1,0)) %>%
  select(sex, employed,age, year)
round(cor(D2), digits = 3)
```

```
            sex employed    age   year
sex       1.000   -0.039 -0.011 -0.020
employed -0.039    1.000 -0.117  0.030
age      -0.011   -0.117  1.000 -0.005
year     -0.020    0.030 -0.005  1.000
```

Given a data vector of numbers $\mathbf{x} = (x_1, \ldots, x_n)$, we have:

- **Sample mean**:

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

```r
D <- Arrests
mean(D$age)
> 23.84654
#or alternative:
```

```
sum(D$age)/nrow(D)
> 23.84654
```

Describing quantitative data

- **Range** of data:

$$\text{range} = \max_{1 \le i \le n} x_i - \min_{1 \le i \le n} x_i.$$

- The *order statistics*.

  First, sort the data to obtain $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, and observe the following.

  1. The minimum: $x_{(1)}$.
  2. The maximum: $x_{(n)}$.
  3. The median $\tilde{\mathbf{x}} =$ "middle" of data.

     (order the data: $x_1 \le x_2 \le \cdots \le x_n$):

$$\begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+1}{2}\right)}\right) & \text{if } n \text{ is even.} \end{cases}$$

```
D <- Arrests
R <- max(D$age) - min(D$age)
> 54
Min_age <- min(D$age)
> 12
Max_age <- max(D$age)
> 66
Med_age <- median(D$age)
> 21
```

- **Sample Variance** (data - spread):

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

where $\overline{\mathbf{x}}$ is the sample mean.

Sample Standard Deviation $= s = \sqrt{s^2}$

- **Sample Correlation Coefficient:**

$$r_{\mathbf{xy}} = \frac{\sum_{i=1}^{n} (x_i - \overline{\mathbf{x}}) (y_i - \overline{\mathbf{y}})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2 \ \sum_{i=1}^{n} (y_i - \overline{\mathbf{y}})^2}}$$

```
D <- Arrests
# Sample Variance
Sample_Var <-  var(D$age)
> 69.15807
#or alterantively
mean_age <- mean(D$age)
D3 <- D %>% mutate(Diff = age-mean_age, Diff_squ = Diff*Diff
    )
Sample_Var <- sum(D3$Diff_squ)/(nrow(D3)-1)
> 69.15807
# Sample Standard Deviation
sd(Sampel_Var)
> 8.316133
# or alternatively:
Sample_STD <- sqrt(Sample_Var)
> 8.316133
```

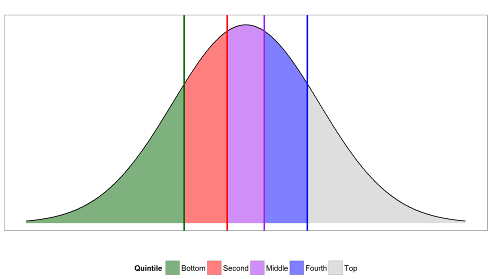
```
D <- Arrests
D4 <- D %>% mutate(sex = ifelse(sex=="Female",1,0))
# Sample Correlation Coefficient
Sample_cor <- cor(D4$age, D4$sex)
> -0.01148502
#or alterantively
mean_age <- mean(D4$age)
mean_sex <- mean(D4$sex)
D5 <- D4 %>% mutate( Num = (age-mean_age)*(sex-mean_sex),
    Denom1 = (age-mean_age)**2, Denom2 = (sex-mean_sex)**2)

Sample_cor <- sum(D5$Num)/sqrt(sum(D5$Denom1)*sum(D5$Denom2)
    )
> -0.01148502
```

- **p-quantile** $(0 < p < 1)$

  $= z$ such that $F(z) = P(X \leq z) = p$

  Common values: 0.25, 0.5, 0.75 quantiles (=25, 50, and 75 percentiles /first, second, and third quartiles)



```
D <- Arrests
quantile(D$age)
```

>	0%	25%	50%	75%	100%
	12	18	21	27	66

```
quantile(D$age, seq(0,1,by=.2))
```

>	0%	20%	40%	60%	80%	100%
	12	17	20	23	30	66

The quantile of a probability distribution

Let f be a prob. density function for a R.V. X.

- Given $\alpha \in [0,1]$, what is x such that $P(X \leq x) = \alpha$?

- By definition:

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(u)\mathrm{d}u = \alpha.$$

Example: $X \sim Exp(1)$ and $\alpha = 0.3$, find x.

$$0.3 = \int_0^x \lambda e^{-\lambda \hat{x}}\, \mathrm{d}\hat{x} = \int_0^x e^{-\hat{x}}\, \mathrm{d}\hat{x} = -e^{-x} - (-e^{-0}) = 1 - e^{-x}$$

Therefore

$$0.7 = e^{-x} \Rightarrow x = -\log(0.7)$$

Data Analysis

- 1st step: Data-Table (+ (statistic) summary of data)
- 2nd step: **Visualisation** with the aim of:

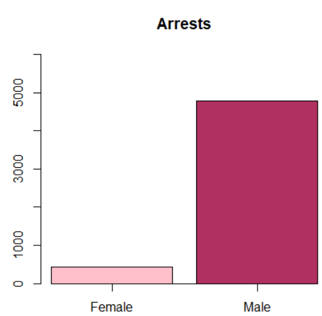
1. Identifying the most common values (for each variable)
2. Determining the amount of variability (for each variable)
3. Recognising unusual observations.
4. Exploring trends in the data.

Visualization of Discrete Data: Bar chart

Visualization for **factor variables**

(Nominal factor):

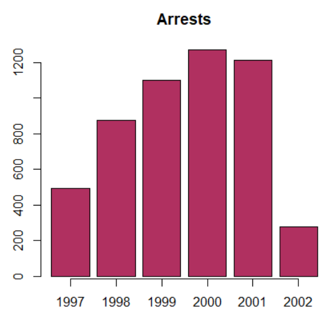
```
barplot(table(D$sex), main='Arrests', ylim=c(0,6000), axis.
    lty=1, col=c("Pink", "Maroon"))
```



Barplot for Ordinal factor:

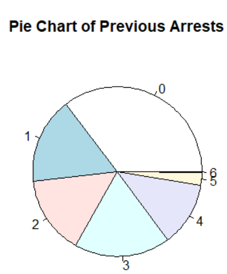
```r
barplot(table(D$year), main='Arrests', axis.lty=1, col= "
    Maroon")
```

What will this code produce?



Visualization of Discrete Data: Pie chart

```r
slices <- table(D$checks)
pie(slices, labels = rownames(slices), main = "Pie Chart of
    Previous Arrests")
```

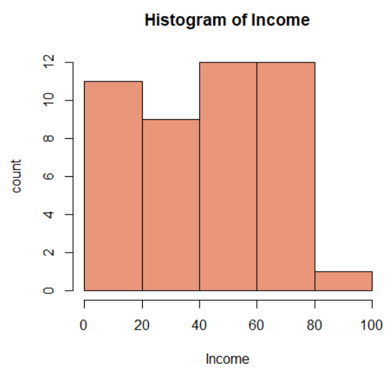


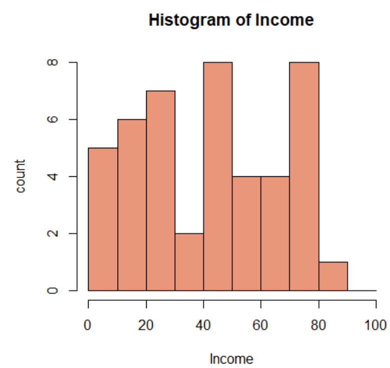
Visualization of "Continuous" Data: Histogram

Continuous analogue of bar plot

Idea:

- Divide the range of a continuous variable into interval-bins
- Plot the associated frequencies for each bin.






```
D_c <- Duncan # data set in carData - library
#left image:
hist(D_c$income, breaks = seq(0,100,20), col="DarkSalmon",
    main = "Histogram of Income", xlab = "Income", ylab = "
    count")
```

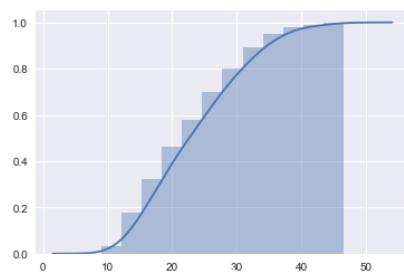
How do you have to change the R-code to get the image on the right?

```
#right image:
hist(D_c$income, breaks = seq(0,100,10), col="DarkSalmon",
    main = "Histogram of Income", xlab = "Income", ylab = "
    count")
```

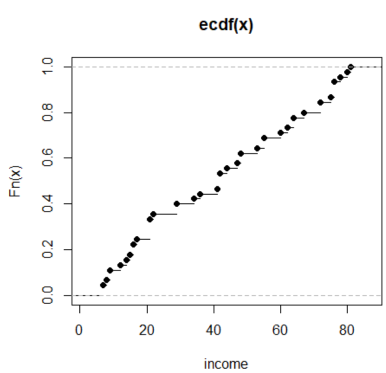
Empirical Cumulative Distribution Function (ECDF):

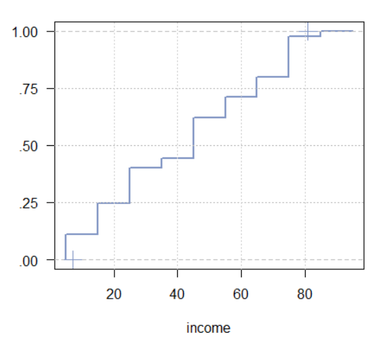
$$\hat{F}(x) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{x_i \leq x\}},$$

where $1_{\{\cdot\}}$ is the indicator function.



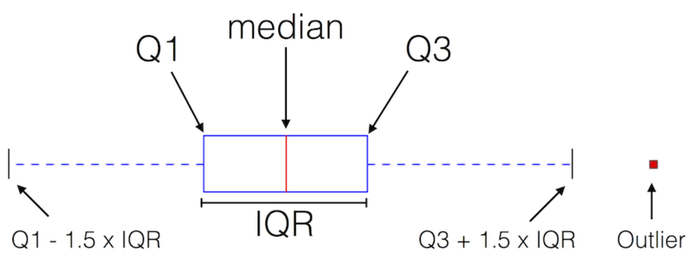

```
D_c <- Duncan
#left image:
plot.ecdf(D_c$income, xlab = 'income')
#right image:
install.packages("DescTools")
library(DescTools)
PlotECDF(D_c$income, seq(0,100,10), xlab = 'income')
```



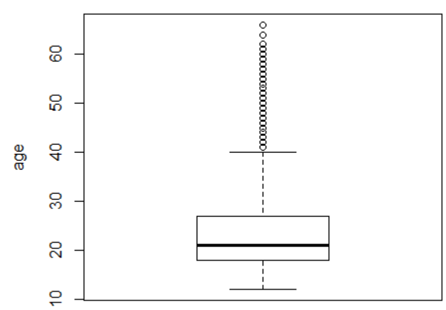


Box Plot

- describes centre of the data,
- describes spread of the data,
- describes departure from symmetry,
- describes identification of outliers of the data




```r
D <- Arrests
boxplot(D$age,  ylab = "age")
```



Scatter Plot - Visualization of relations between variables

Idea:

Plot the observations in the x and y diagram

$\longrightarrow$ Relation between x and y becomes apparent

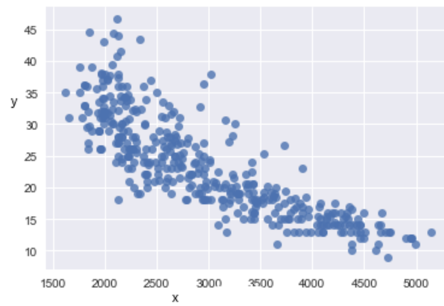
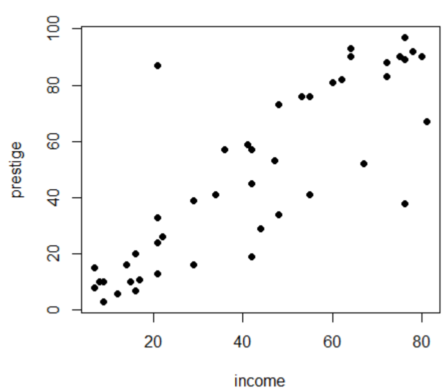


Figure 1: Scatter plot of two variables x and y.

```
D_c <- Duncan
plot(D_c$income, D_c$prestige, pch=16,xlab='income', ylab =
    'prestige')
```



Mixing variable types

To get the relation between two variables (*"conditioned"*) one the value of one variable, we can use boxplots.

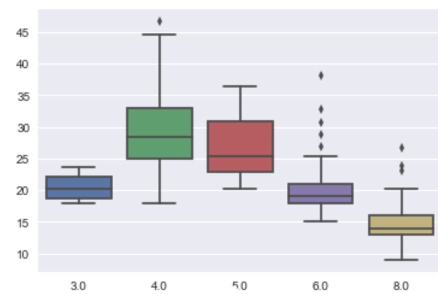
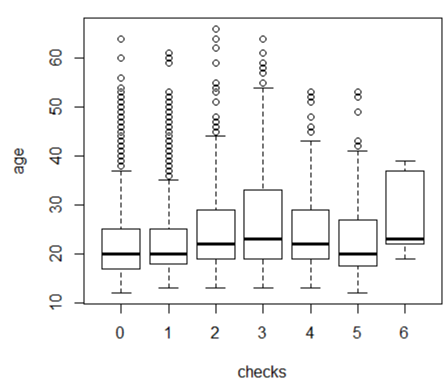
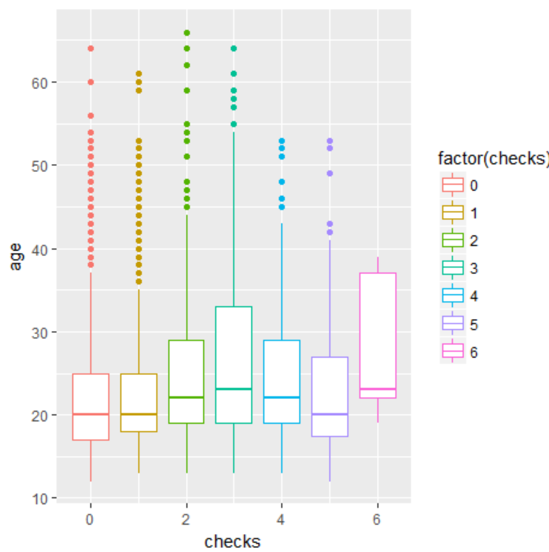


Figure 2: Box plot by category.

```
D <- Arrests
boxplot(age~checks, data = D, xlab='checks', ylab='age')
```



```
install.packages("ggplot2")
library(ggplot2)
ggplot(D, aes(x=checks, y=age, color=factor(checks))) + geom
    _boxplot()
```



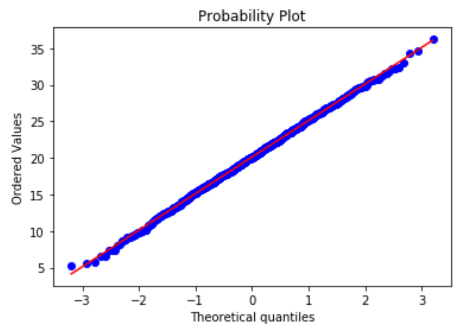
QQ plots

Plots the quantiles of the first data set against the quantiles of the second data set.

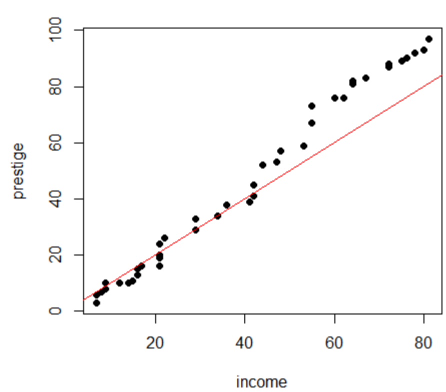
Idea:

- Calculate quantiles of the dataset for x.
- Calculate quantiles of the dataset for y.
- Plot quantiles of x against quantiles of y.

⟹ If the line is on the 45-degree reference line, the two sets come from a population with the same distribution.



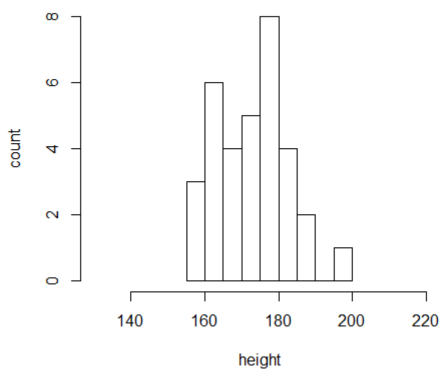

```
D_c <- Duncan
qqplot(D_c$income, D_c$prestige, xlab='income', ylab='
    prestige', pch=16)
abline(0,1,col='red')
```



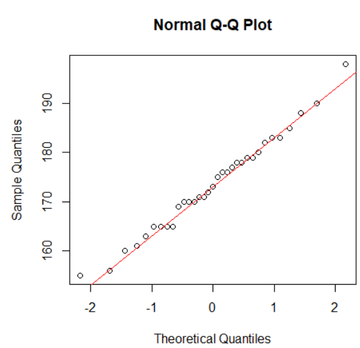
We see that income and prestige do not seem to come from the same distribution for all values, but for an income that is below 50, they do seem to come from the same distribution.

Often QQ-Plots are used to compare sample data to the Normal Distribution.

Stat2201 height distribution:

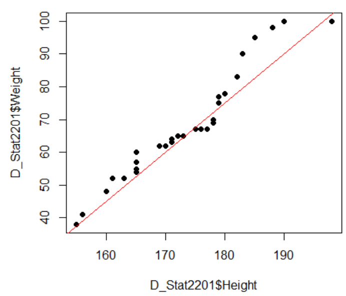



```r
library(xlsx)
D_Stat2201 <- read.xlsx("Height_Weight_STAT2201.xlsx", 1)
qqnorm(D_Stat2201$Height)
abline(173,10,col='red')
```

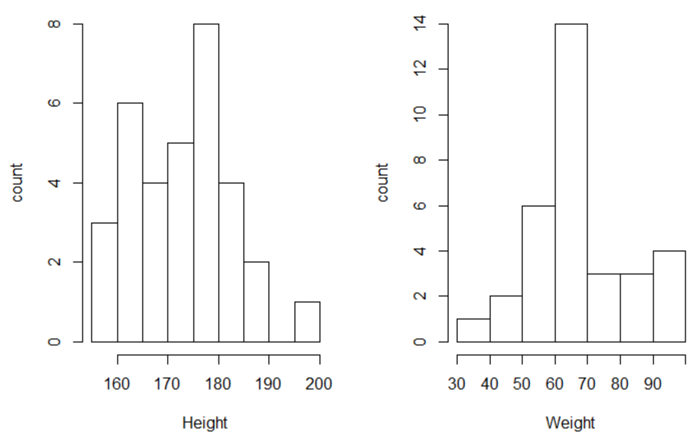


We see in the Normal QQ-plot, that the Height seems to be normal distributed, as the QQ plot is reveals a linear relation of the quantiles.

```
qqplot(D_Stat2201$Height, D_Stat2201$Weight, pch=16)
abline(-195,1.5,col='red')
```



However, the Height and the Weight do not seem to come from the same distribution. For a height that is below 180cm, they do seem to come from the same distribution.



The histogram reveals the structure and is an indicator why for large heights, the height and weight do not seem to come from the same distribution. While the Height is still following a normal distribution for large values, the Weight seems to be nearly uniform distribution for large values.

Your First Data Analysis

```r
library(carData)
D_Q <- Depredations   #Wolf depredation in 1973
head(Depredations)
```

	longitude	latitude	number	early	late
1	-94.5	46.1	1	0	1
2	-93.0	46.6	2	0	2
3	-94.6	48.5	1	1	0
4	-92.9	46.6	2	0	2
5	-95.9	48.8	1	0	1
6	-92.7	47.1	1	0	1

a) What would be the very first step if someone gives you a dataset?

b) How do you determine the number of observations?

c) Which of the variables are continuous which ones are factors?

d) If you want to investigate the distribution of the latitude with respect to number of depredations, what type of plot (and what R-Code) would you use?

e) What variables do you suspect to be related and how would you test this?

f) Can you think of some other questions you would like to answer with that data set?

See Answers in World document created by RMarkdown.

Review Chapter 6: Data Description

- Summary Statistics

  a) Sample-Mean,

  b) Sample-Variance,

  c) Sample-Covariance & Sample-Correlation,

  d) Range of Data, Minimum, Maximum,

  e) Median,

  f) P-quantiles.

- Visualization:

  a) Bar-Plot (factor variable),

  b) Pie-Plot (factor variable),

  c) Histogram (continuous variable),

  d) ECDF-Plot,

  e) Box-Plot,

  f) Scatter-Plot (relation of two variables),

  g) QQ-Plot.

Chapter 7–9

- Statistical Inference
- Central Limit Theorem
- Confidence Intervals
- Hypothesis Testing

Statistical inference

Statistical Inference is the process of forming judgements about the parameters.

<u>Assumptions:</u>

- Assume that data $X_1, \ldots, X_n$ is drawn randomly from some ***unknown*** distribution (identically distributed).
- Assume that the data is independent

  $\longrightarrow$ X_i are i.i.d. (independent and identically distributed), i.e.,

  1. $X_i \sim G$ for all $1 \leq i \leq n$
  2. X_is are independent

A statistic

A **statistic** is any function of the observations in a random sample.

$\longrightarrow$ A statistic is itself a R.V.

Examples:

- $g(X_1, X_2, \ldots, X_n) = \overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ = Sample mean
- $g(X_1, X_2, \ldots, X_n) = \max\{X_1, X_2, \ldots, X_n\}$
- Sample variance and sample standard deviation
- Sample quantiles besides the median, (quartiles and percentiles)

Some notations:

- The probability distribution of a statistic is called the **sampling distribution**.

- A **point estimate** of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$.
- The statistic $\hat{\Theta}$ is called the point estimator.

Example:

Sample Mean = $\overline{X}$ = estimator of the population mean, μ.

Normal Distribution - Recap

$X \sim N(\mu, \sigma^2)$ then pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}.$$

- $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$
- If $\mu = 0$ and $\sigma = 1$ then

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}, \quad x \in \mathbb{R},$$

  = standard normal distribution

- $\frac{X-\mu}{\sigma} \sim \mathsf{N}(0,1)$ = standardization
- $X = \mu + \sigma Z, \quad Z \sim \mathsf{N}(0,1)$

Central Limit Theorem (for sample means)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population with mean μ and finite variance σ^2,then

$$\lim_{n\to\infty} \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = Z \sim N(0,1)$$

where $\bar{X}$ is the sample mean. Equivalently,

$$P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) = \Phi(x)$$

Regardless of X_i's distribution, the sum behaves (approximately) as the Gaussian random variable!

$$\bar{X} \quad \overset{n\to\infty}{\approx} \quad N\left(\mu, \frac{\sigma^2}{n}\right)$$

$S_n = \sum_{i=1}^{n} X_n$ is then distribution

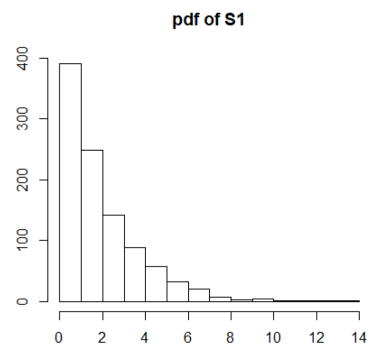
$$S_n \quad \overset{n\to\infty}{\approx} \quad N(n\mu, n\sigma^2)$$

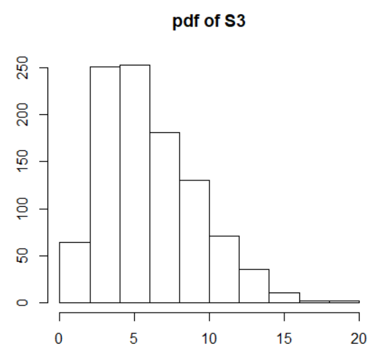
Example:

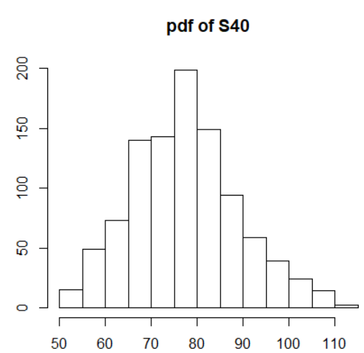
$X_i \sim Exp(0.5)$ (i.i.d.) $\to S_k = \sum_{i=1}^{k} X_i$

```
M <- matrix(0,50,1000)
M[1,] <- rexp(1000,lambda)
for (i in 2:50){
  M[i,] <- M[i-1,] + rexp(1000, 0.5)
}
```

```
hist(M[3,], main = 'pdf of S3', xlab='', ylab = '')
hist(M[40,], main = 'pdf of S40', xlab='', ylab = '')
```







We see that as we increase the number of X considered, the random variable $S_k = \sum_{i=1}^{k} X_i$ (=the sum of the X) behaves like a normal distribution, although each X is in fact an exponential distribution.

Note that the Central Limit Theorem also tells us something about the standard error of the sample mean $\bar{X}$:

- The standard error of $\overline{X}$ is given by $\frac{\sigma}{\sqrt{n}}$.
- In most practical situations σ is not known but rather estimated.
- The estimated standard error (SE) is:

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

Example:

For a temperature of 100°F and 550 watts, the following measurements of thermal conductivity were obtained:

41.60	41.48	42.34	41.95	41.86
42.18	41.72	42.26	41.81	42.04

→ sample mean is 41.924

→ estimated standard error is sample standard deviation s divided by $\sqrt{10}$, here $\frac{0.284}{\sqrt{10}} = 0.0898$

Confidence Interval

confidence interval for μ (the real mean):

$$l \leq \mu \leq u,$$

- Let $X_1, \ldots, X_n$ be collected data

- Endpoints are values of random variables $L = g_1(X_1, \ldots, X_n)$ and $U = g_2(X_1, \ldots, X_n)$ such that

$$P(L(\mathbf{X}) \leq \mu \leq U(\mathbf{X})) = 1 - \alpha, \quad \alpha \in (0, 1).$$

$\longrightarrow 1 - \alpha$ is called the **confidence level**.

((l, u) is the $100 \cdot (1 - \alpha)$ % confidence interval.)

Confidence Interval for Mean

Let X_i be i.i.d., then:

- Recall
$$\overline{X} \sim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

- That is, for some positive scalar value $z_{1-\alpha/2}$, we have

$$P\left(\overline{X} \leq \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = P\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha/2}\right) = \Phi(z_{1-\alpha/2})$$

$$P\left(\overline{X} \leq \mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = P\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq -z_{1-\alpha/2}\right) = \Phi(-z_{1-\alpha/2}) = 1 - \Phi(z_{1-\alpha/2})$$