PC1: Descriptive Statistics

Wei Jiang, Geneviève Robin, Zoltán Szabó, Elodie Vernet

Monday, September 24. 2018

Statistics

A branch of mathematics dealing with data

- Collection
- Organization
- Analysis
- **Presentation** (summaries, graphical displays, etc.)

Population?
Sample known

Today: Exercises 1, 3, 4, 6, 9, 10, 12, 13, 14

Data set and random variables

- $x_1, \ldots, x_n$ are n observed values: $1, \ldots, n$ index individuals, censors, etc.
- If we repeat an experience, the collection of data $(x_1, \ldots, x_n)$ might take different values -$>$ randomness.
- We assume that the **data set** $(x_1, \ldots, x_n)$ is a realization of a random vector $(X_1, \ldots, X_n)$ (statistical modeling).
- $X_1, \ldots, X_n$ are n random variables, assumed i.i.d here with c.d.f. F

Different types of data

- Quantitative (numerical): price of diamonds
- Qualitative (categories, characteristics): color of diamonds
- Univariate data: x_i has only one component.

i	1	2	3	4
x_i (price)	0.23	0.21	0.23	0.29

- Multivariate data: x_i has several components.

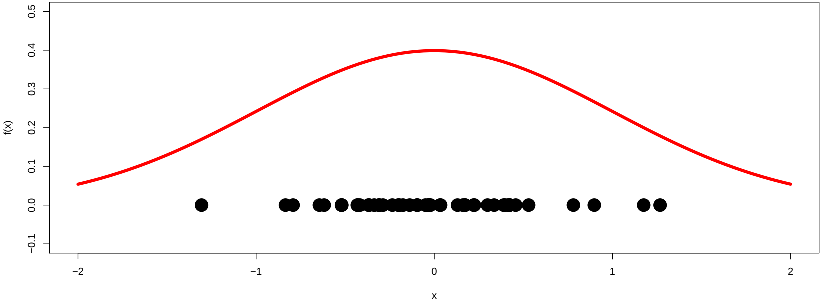
i	1	2	3	4
carat	0.23	0.21	0.23	0.29
color	E	E	E	I

The diamonds data

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
0.23	Very Good	H	VS1	59.4	61	338	4.00	4.05	2.39
0.30	Good	J	SI1	64.0	55	339	4.25	4.28	2.73
0.23	Ideal	J	VS1	62.8	56	340	3.93	3.90	2.46

Univariate quantitative data

- $x_1, \ldots, x_n$ are n observed real values
- We assume that the **data set** $(x_1, \ldots, x_n)$ is a realization of a random vector $(X_1, \ldots, X_n)$ (statistical modeling).
- $X_1, \ldots, X_n$ are n real random variables, assumed i.i.d here with c.d.f. F

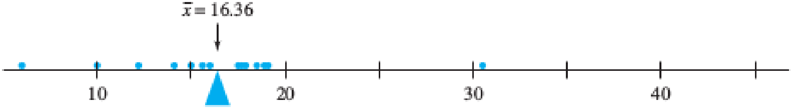


Measures of location

- Find **numerical summaries** about the location of the data points in the observation space ($\mathbb{R}$).

Empirical mean

- empirical mean of the sample: $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$



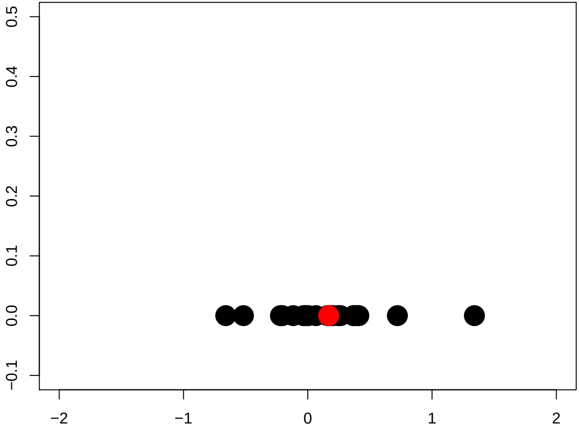
- empirical mean of the observations: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (also a random variable)
- LLN: $\bar{X}_n \to \mathbb{E}_F(X)$ almost surely when $\mathbb{E}_F[|X|] < \infty$
- $\bar{x}_n$ is not a robust summary of the location (exercise1 Q1)

Order statistics, quantiles, median

- **Order statistics**: $x_{(1)}, \dots, x_{(n)}$, where $x_{(i)}$ is the i-th smallest value in $(x_1, \dots, x_n)$.
- α-**empirical quantile**: $x_\alpha(n) = x_{(\lceil \alpha n \rceil)}$.
- The **median** is the $1/2$-empirical quantile $x_{1/2}(n) = x_{(\lceil n/2 \rceil)}$.

Median

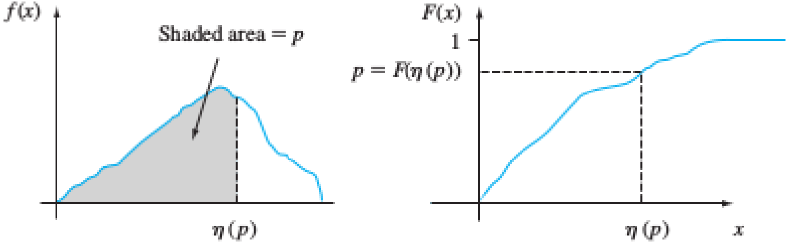
- Find **numerical summaries** about the location (= "center") of the data points in the observation space ($\mathbb{R}$).
- Median is the 1/2-empirical quantile $x_{1/2}(n) = x_{(\lceil n/2 \rceil)}$: as many values in $(x_1, \ldots, x_n)$ are below $x_{1/2}(n)$ than above $x_{1/2}(n)$.



Theoretical quantiles

- Let F be a probability distribution. The α-**theoretical quantile** is defined by

$$Q_F(\alpha) = \inf\{t \in \mathbb{R}; F(t) \geq \alpha\}$$



Trimmed mean

The α-**trimmed mean** is the mean computed using $x_{(\lceil \alpha n \rceil)}, \dots, x_{(\lceil (1-\alpha) n \rceil)}$, *i.e.* after removing the $\lceil \alpha n/2 \rceil$ smallest and the $\lceil \alpha n/2 \rceil$ largest values.

$$x_{(1)},\ x_{(2)},\ \dots\ ,\ x_{(\lceil \alpha n/2 \rceil)},\ \underbrace{x_{(\lceil \alpha n/2 \rceil + 1)}, \quad \cdots \quad, x_{(n - \lceil \alpha n/2 \rceil)}}_{\text{part used to compute the } \alpha\text{-trimmed mean}}\ ,\ x_{(n - \lceil \alpha n/2 \rceil + 1)},$$

$$\dots\ ,\ x_{(n)}.$$

Exercises

- Exercise 1: empirical mean, median and trimmed mean
- (Exercise 2: symmetric density functions)

Exercise 1, question 1

```r
library(ggplot2)
data_small <- diamonds[1:12,]
mean(data_small$y)
```

```
## [1] 4.044167
```

```r
sort(data_small$y)
```

```
##  [1] 3.78 3.84 3.90 3.96 3.98 3.98 4.05 4.07
##  [9] 4.11 4.23 4.28 4.35
```

```r
quantile(data_small$y, 1/2, type=1)
```

```
##  50%
## 3.98
```

```r
mean(data_small$y, trim=0.2/2)
```

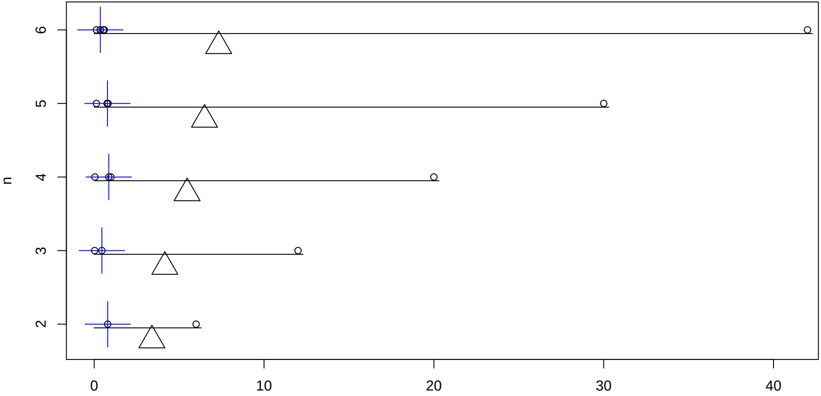
```
## [1] 4.04
```

```r
mean(sort(data_small$y)[2:11])
```

```
## [1] 4.04
```

Exercise 1, question 2

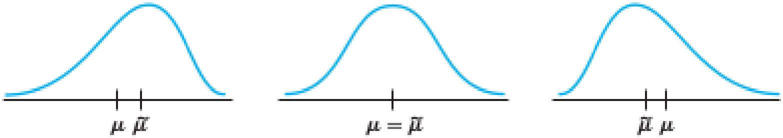
- $(x_1, \ldots, x_{n-1})$ all in $(0, 1)$
- and one outlier $x_n = n * (n + 1)$
- for all n, the median is in $(0, 1)$, but the mean is greater than $n + 1$.



Exercise 1, questions 3 and 4

- The empirical mean is the 0-trimmed mean.
- The median is the $1/2 + 1/n$-trimmed mean when n is uneven.
- The median GDP is much lower than the average GDP because of outliers, i.e. individuals with very large wage compared to the rest of the population. The larger the gap, the more inequality there is in a country.

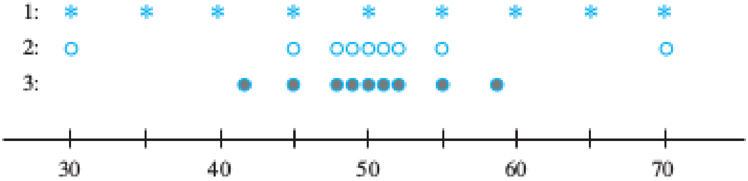
Exercise 2



(a) Negative skew (b) Symmetric (c) Positive skew

Measures of dispersion

- Numerical summaries measuring how the data points are dispersed around their centers (center = measure of location).



Some measures of dispersion

- Empirical **range**:

$$x_{(n)} - x_{(1)}$$

- Empirical **standard deviation**:

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

- Empirical **variance**: s_x^2
- **Interquartile range**: $x_{3/4}(n) - x_{1/4}(n)$
- **Median absolute deviation**: the median of

$$|x_1 - x_{1/2}(n)|, |x_2 - x_{1/2}(n)|, \ldots, |x_n - x_{1/2}(n)|$$

Exercises

- Exercise 3: measures of dispersion
- Exercise 4: consistent estimators

Exercise 3

```
library(ggplot2)
data_small <- diamonds[1:12,]
sy <- sort(data_small$y)
sy
```

```
##  [1] 3.78 3.84 3.90 3.96 3.98 3.98 4.05 4.07
##  [9] 4.11 4.23 4.28 4.35
```

```
sy[12]-sy[1]
```

```
## [1] 0.57
```

```
c(quantile(data_small$y, 3/4, type=1),quantile(data_small$y, 1/4, type=1))
```

```
##  75%  25%
## 4.11 3.90
```

```
quantile(data_small$y, 3/4, type=1)-quantile(data_small$y, 1/4, type=1)
```

```
##  75%
## 0.21
```

Exercise 3

```
sd(data_small$y)
```

```
## [1] 0.1748484
```

```
var(data_small$y)
```

```
## [1] 0.03057197
```

```
sort(abs(data_small$y-quantile(data_small$y, 1/2, type=1)))
```

```
##  [1] 0.00 0.00 0.02 0.07 0.08 0.09 0.13 0.14
##  [9] 0.20 0.25 0.30 0.37
```

```
quantile(abs(data_small$y-quantile(data_small$y, 1/2, type=1)), 1/2, type=1)
```

```
##  50%
## 0.09
```

Exercise 4

Other statistics

- Empirical **skewness**:

$$\hat{\alpha}_x = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^3}{s_x^3}$$

- Empirical **kurtosis**:

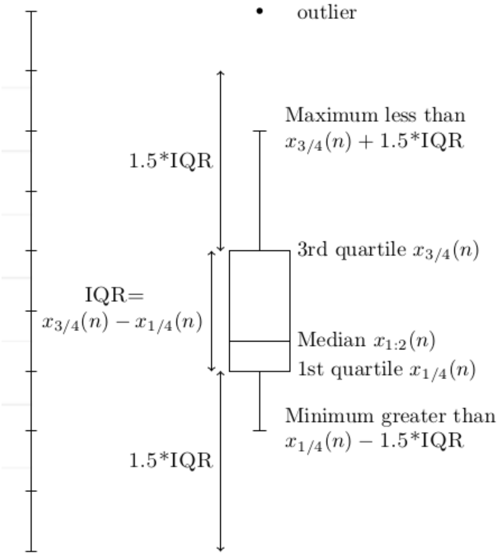
$$\hat{\beta}_x = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^4}{s_x^4} - 3$$

- (Exercise 5)

Graphical representations

- **Box plots** represent the location, dispersion, symmetry and skewness of the data
- The **empirical cumulative distribution function** represents the fraction of observations that are below a specific value.
- **Bar plots** and **histograms** display the shape of the distribution of the data.
- **Q-Q plots** compare the quantiles of a sample with he theoretical quantiles or another sample.

Box plots



outlier
Maximum less than
$x_{3/4}(n) + 1.5$*IQR
1.5*IQR
3rd quartile $x_{3/4}(n)$
IQR=
$x_{3/4}(n) - x_{1/4}(n)$
Median $x_{1:2}(n)$
1st quartile $x_{1/4}(n)$
Minimum greater than
$x_{1/4}(n) - 1.5$*IQR
1.5*IQR

Box plots

- Exercise 6

Exercise 6, question 1

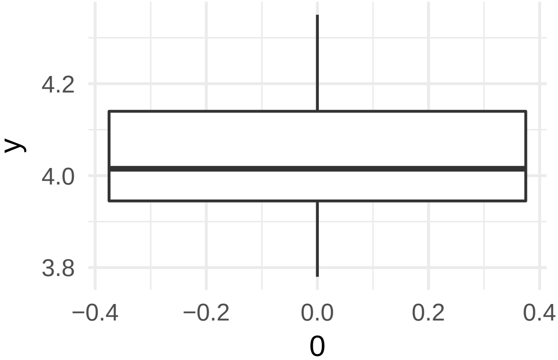
```
c(Q1-1.5*IQR,Q1,Med,Q3,Q3+1.5*IQR)
```

```
##   25%   25%   50%   75%   75%
## 3.585 3.900 3.980 4.110 4.425
```

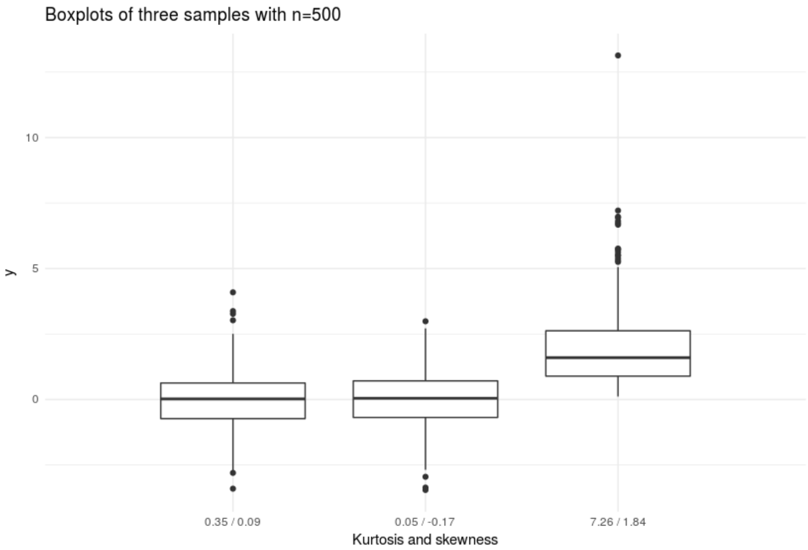
```
sort(data_small$y)
```

```
##  [1] 3.78 3.84 3.90 3.96 3.98 3.98 4.05 4.07
##  [9] 4.11 4.23 4.28 4.35
```

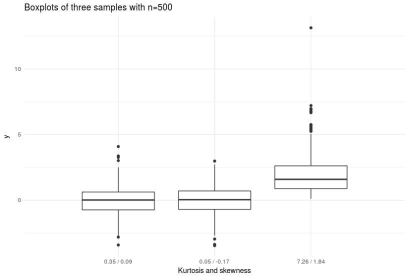
```
ggplot(data=data_small,aes(x=0,y=y))+geom_boxplot() +theme_minimal()
```



Exercise 6, question 1



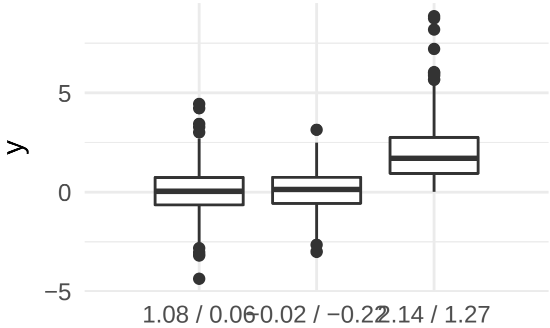
Exercise 6, question 1



- symmetric around 0, many outliers, large kurtosis, family of symmetric distribution with heavy tail, e.g. $\{\mathcal{T}(d), d > 0\}$,
- symmetric around 0, e.g. $\{\mathcal{N}(0, \sigma^2), \sigma^2 \in \mathbb{R}_+\}$,
- not symmetric, support in $\mathbb{R}_+$, skewed, e.g. $\{\text{Exp}(\lambda), \lambda \in \mathbb{R}_+\}$.

Exercise 6, question 1

```r
library(e1071)
nx <- 500
dfx <-10
y1 <-rt(nx,df=dfx)
k1 <-kurtosis(y1)
s1 <- skewness(y1)
y2<-rnorm((nx))
k2 <-kurtosis(y2)
s2 <- skewness(y2)
y3 <-rgamma(nx,shape=2,scale=2)
k3 <-kurtosis(y3)
s3 <- skewness(y3)
data=data.frame(x=c(rep(1,nx),rep(2,nx),rep(3,nx)), y=c(y1,y2,1/2*y3))
library(ggplot2)
ggplot(data = data, aes(x=x,y=y,group=x))+geom_boxplot() +# ggtitle(paste("Boxplots of three samples with
theme_minimal()+ scale_x_discrete(name ="Kurtosis and skewness",breaks=c("0","1","2"),labels=c(paste(as.ch
```



Empirical cumulative distribution function

Defined for a sample $(x_1, \dots, x_n)$ by

$$t \in \mathbb{R} \mapsto \hat{F}_x(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \leq t} = \frac{\mathrm{Card}(\{i : x_i \leq t\})}{n}$$

- $\hat{F}_x$ is a non-decreasing step function: it jumps of $1/n$ at each $t \in \{x_1, \dots, x_n\}$.

Empirical cumulative distribution function

- Exercises 9, 10

Exercise 9, question 1

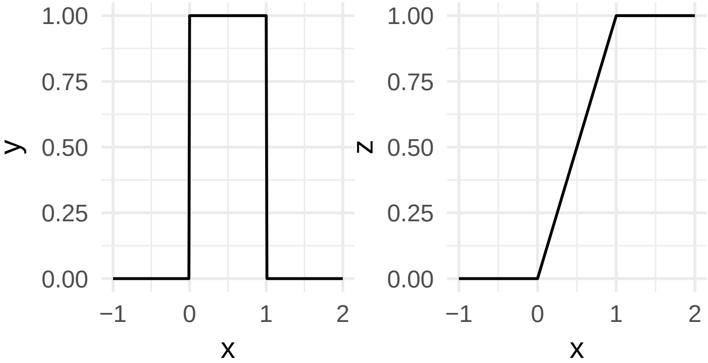
- cdf of $\mathcal{U}(0,\theta)$:

$$F_\theta(t) = \frac{1}{\theta}\int_0^\theta \mathbb{1}_{x \leqslant t} dx = \begin{cases} 0 & \text{if } t \leqslant 0 \\ \frac{t}{\theta} & \text{if } t \in (0,1) \\ 1 & \text{otherwise.} \end{cases}$$

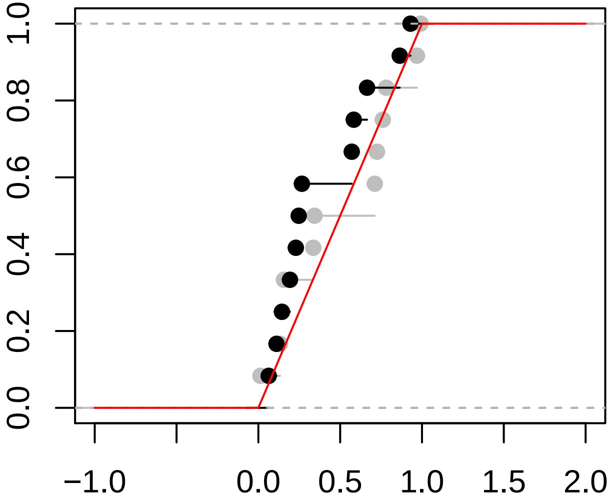
- Theoretical quantile, $\alpha \in (0,1)$,

$$Q_{\mathcal{U}(0,\theta)}(\alpha) = \inf_{t \in \mathbb{R}} \{F(t) \geqslant \alpha\} = \alpha\theta$$

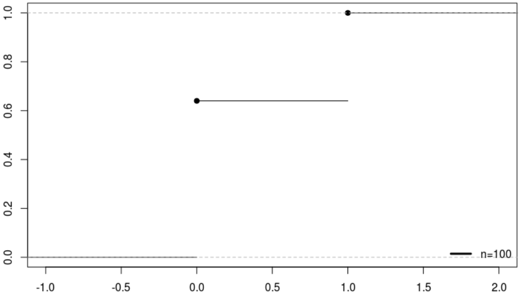
Exercise 9, question 2



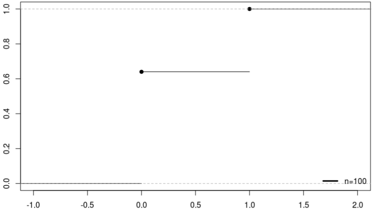
Exercise 9, question 3



Exercise 10



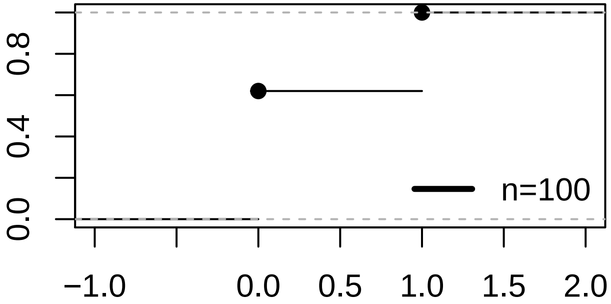
Exercise 10



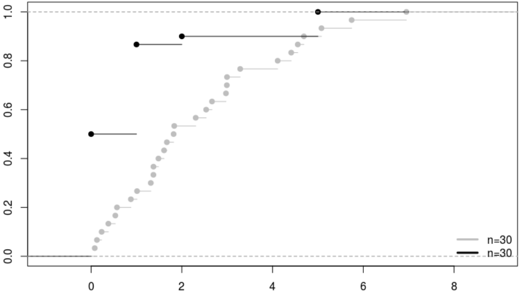
- Two possible values 0 and 1,
- Bernoulli distribution $\{\mathcal{B}(\theta), \theta \in (0,1)\}$,
- Looking at the graphic θ should be around 0.3.

Exercise 10

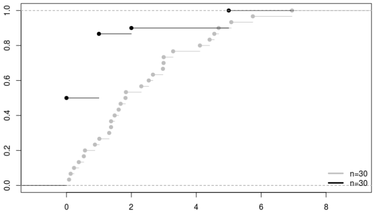
```
nx <- 100
x <-rbinom(nx,size=1,prob=0.35)
cdfx <-ecdf(x)
plot(cdfx,,xlab="",ylab="",main=NULL)
legend("bottomright","n=100",col = "black", pch=c(NA,NA), 
        lwd=c(3,3),bty="n", pt.cex=2)
```



Exercise 10



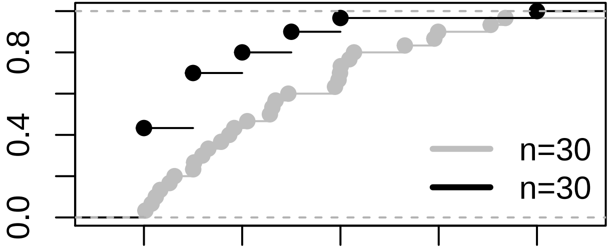
Exercise 10



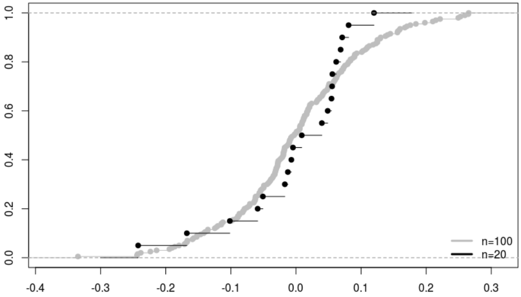
- One discrete variable (black) (small number of heights compared to *n*, different height of jumps, values in $\mathbb{N}$)
- The other (grey) looks continuous (small amount of jumps than *n*, same height of jumps), with values in $\mathbb{R}_+$

Exercise 10

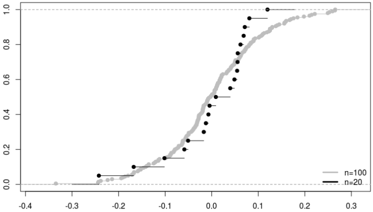
```r
nx <- 30
ny <- 30
x <-rexp(ny,0.4)
y <-rgeom(nx,0.4)
cdfx <-ecdf(x)
cdfy <- ecdf(y)
plot(cdfx,,xlab="",ylab="",main=NULL,col="grey",xlim=c(-1,
lines(cdfy)
legend("bottomright",c("n=30","n=30"),col = c("grey","black
        lwd=c(3,3),bty="n", pt.cex=2)
```



Exercise 10



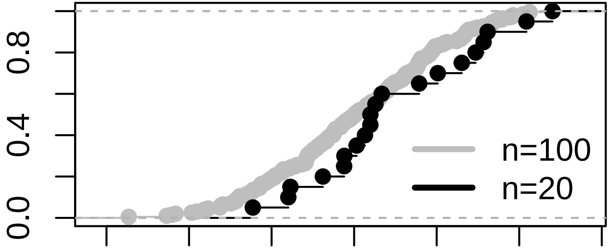
Exercise 10



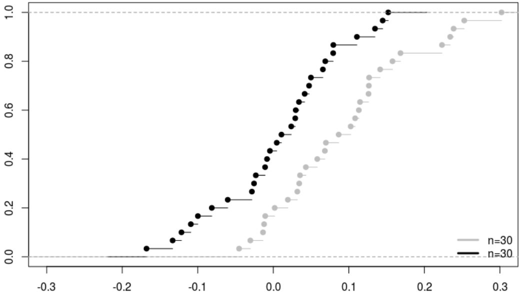
- These ecdfs look alike appart from the fact that one has less jumps and its sample size is smaller. They may be the ecdf of variables with the same distribution.

Exercise 10

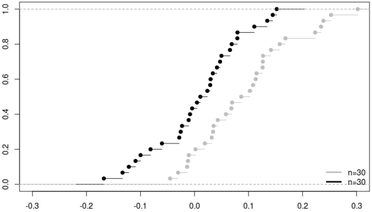
```
nx <- 200
ny <- 20
x <-rnorm(nx,mean=0,sd=0.1)
y <-rnorm(ny,mean=0,sd=0.1)
cdfx <-ecdf(x)
cdfy <- ecdf(y)
plot(cdfx,,xlab="",ylab="",main=NULL,col="grey")
lines(cdfy)
legend("bottomright",c("n=100","n=20"),col = c("grey","blac
        lwd=c(3,3),bty="n", pt.cex=2)
```



Exercise 10



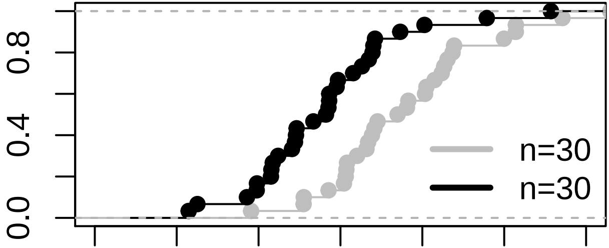
Exercise 10



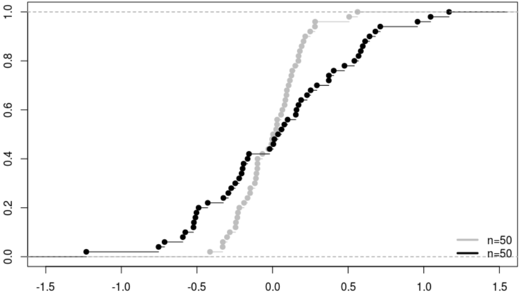
- The ecdfs $\hat{F}$ and $\hat{G}$ associated to the two samples look alike apart from the fact that they are translated,
- One could think that the distributions F and G of the associated random variables (describing the associated population) are translated versions, i.e. $F(\cdot) = G(\cdot - \mu)$ for some μ.

Exercise 10

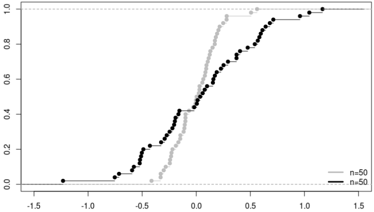
```
nx <- 30
ny <- 30
x <-rnorm(nx,mean=0.1,sd=0.1)
y <-rnorm(ny,mean=0,sd=0.1)
cdfx <-ecdf(x)
cdfy <- ecdf(y)
plot(cdfx,,xlab="",ylab="",main=NULL,col="grey",xlim=c(-0.3
lines(cdfy)
legend("bottomright",c("n=30","n=30"),col = c("grey","black
        lwd=c(3,3),bty="n", pt.cex=2)
```



Exercise 10



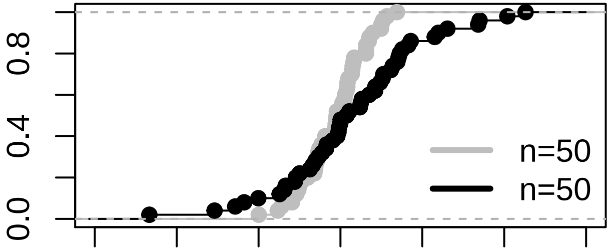
Exercise 10



- The ecdfs $\hat{F}$ and $\hat{G}$ associated to the two samples look alike apart from that the black one represents a more dispersed sample,
- One could think that the distributions F and G of the associated random variables (describing the associated population) are rescaled versions, i.e. $F(\cdot) = G(\cdot/\sigma)$ for some σ (scale parameter).

Exercise 10

```
nx <- 50
ny <- 50
x <-rnorm(ny,mean=0,sd=0.2)
y <-rnorm(nx,mean=0,sd=0.5)
cdfx <-ecdf(x)
cdfy <- ecdf(y)
plot(cdfx,,xlab="",ylab="",main=NULL,col="grey",xlim=c(-1.5
lines(cdfy)
legend("bottomright",c("n=50","n=50"),col = c("grey","blac
        lwd=c(3,3),bty="n", pt.cex=2)
```

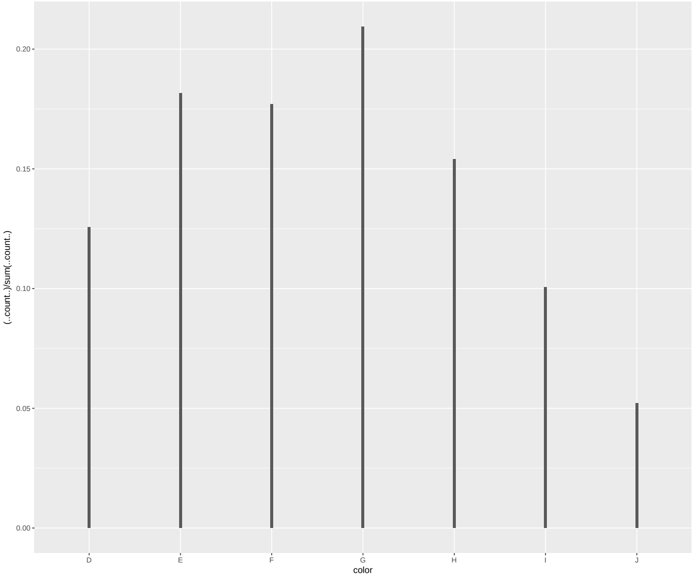


Bar plots

- The equivalent of histograms for **discrete** or **qualitative** data.
- One bar for every value the observations can take (example: Male, Female; Red, Blue, ... or 1, 2, 3, ...)
- The height of each bar is the proportion (or the number) of observations taking the corresponding value.

$$\hat{p}_x(k) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i=k} = \frac{\mathrm{Card}(\{i : x_i = k\})}{n}.$$

Bar plots



Histograms

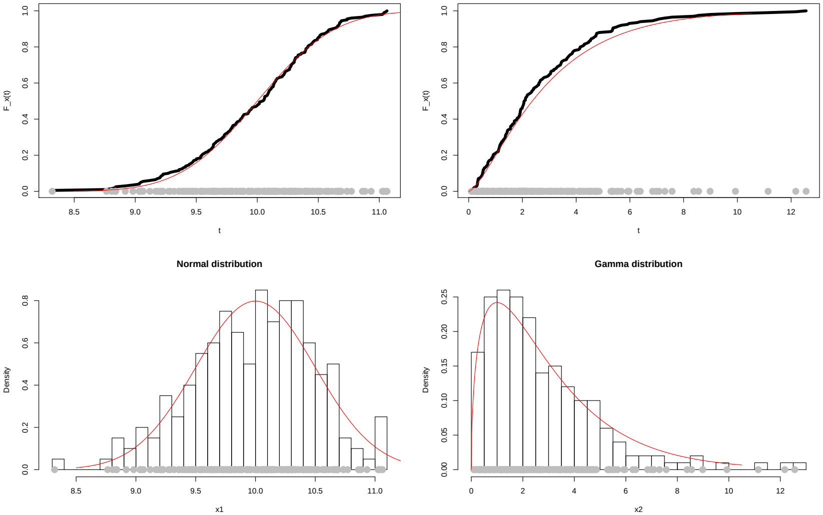
- Assume the sample $(x_1, \dots, x_n)$ is in $[a, b]$, *i.e.* $a \leq x_i \leq b$ for all i.
- Cut the interval $[a, b]$ into m **bins**; for example

$$A_i = \left[a + \frac{j-1}{m} h, a + \frac{j}{m} h\right),$$

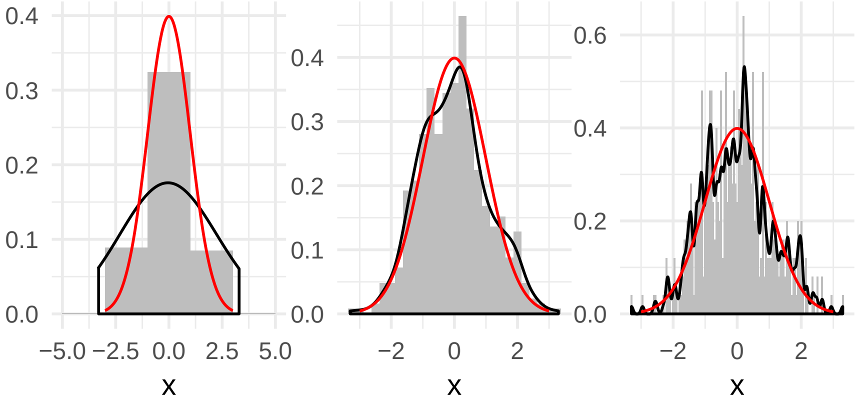
  $h = (b - a)/m$ and for $1 \leq j \leq m$.
- The **histogram** of $(x_1, \dots, x_n)$ represents the proportion of observations that fall in every bin. It is the following step function:

$$\hat{f}_x^H(t) = \sum_{j=1}^{m} \frac{\operatorname{Card}(\{i : x_i \in b_j\})}{nh} \mathbb{1}_{A_j}(t).$$

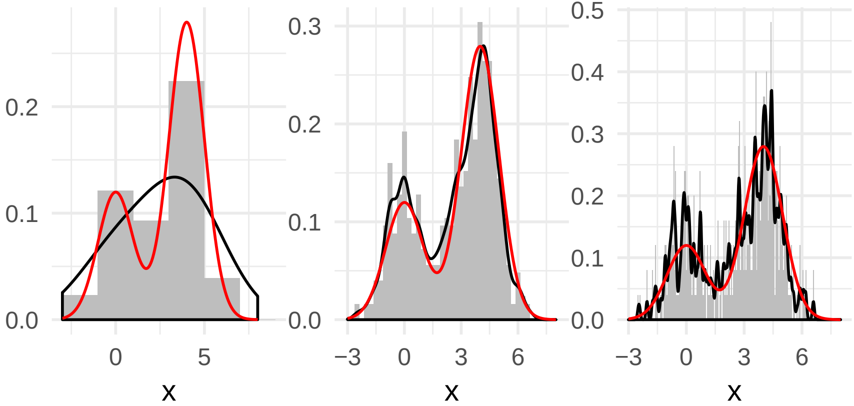
Histograms


F_x(t)
0.0
0.2
0.4
0.6
0.8
1.0
8.5
9.0
9.5
10.0
10.5
11.0
t
F_x(t)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
t
Normal distribution
Density
0.0
0.2
0.4
0.6
0.8
8.5
9.0
9.5
10.0
10.5
11.0
x1
Gamma distribution
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
12
x2

Histograms for Gaussian realizations



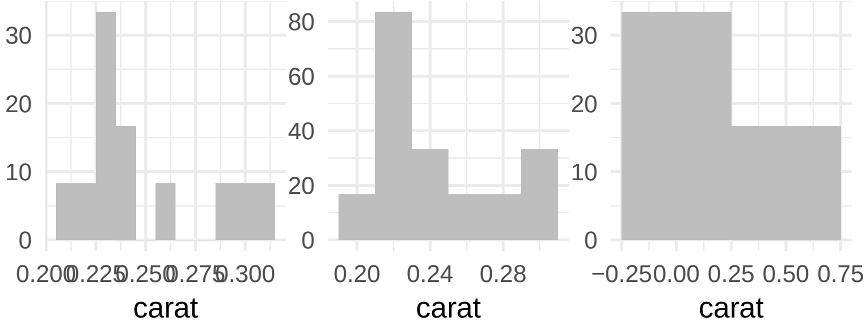
Histograms for mixture of Gaussians realizations



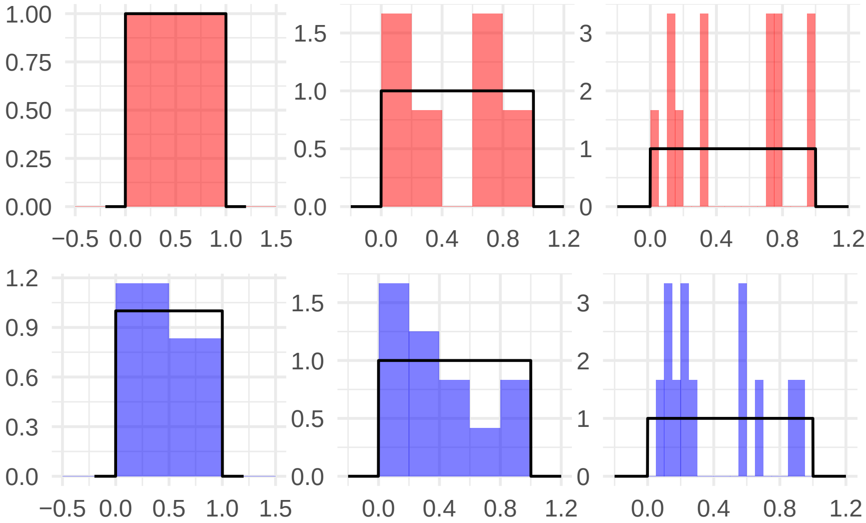
Bar plots

- (Exercise 11)
- Exercise 12

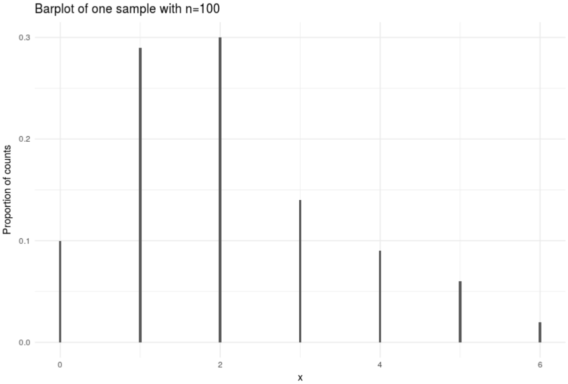
Exercise 12, question 1



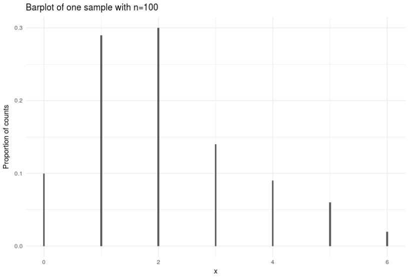
Exericse 12, question 2



Exercise 12, question 3



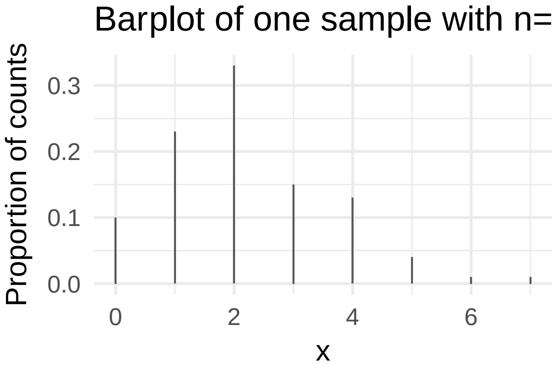
Exercise 12, question 3



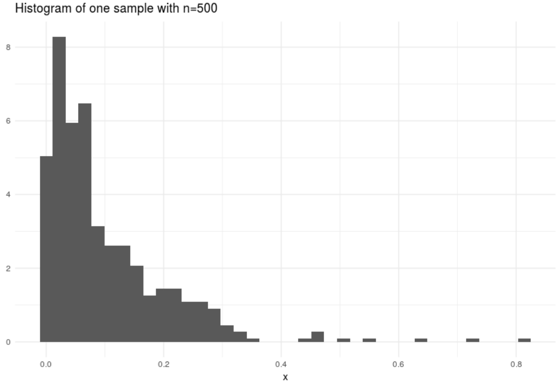
- The observed values are in $\{0, 1, 2, 3, 4, 5, 6\} \subset \mathbb{N}$
- The bar plot is not symmetrical
- The tail is light
- The observed values could be the realizations of Poisson random vairables $\{\mathcal{P}(\lambda), \lambda \in \mathbb{R}_+\}$.

Exercise 12, question 3

```
nx <- 100
lambdax <-2
x <-rpois(nx,lambda=lambdax)
data=data.frame(x=x)
library(ggplot2)
ggplot(data = data, aes(x=x))+geom_bar(width=0.03, aes(y =
```

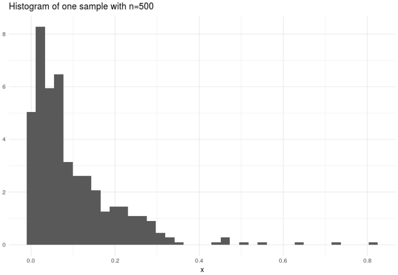


Exercise 12, question 3



- The observed values are in $\mathbb{R}_+$
- The bar histogram is not symmetrical
- The tail is not light.
- The observed values could be the realizations of Exponential random vairables $\{\mathcal{E}(\lambda), \lambda \in \mathbb{R}_+\}$.

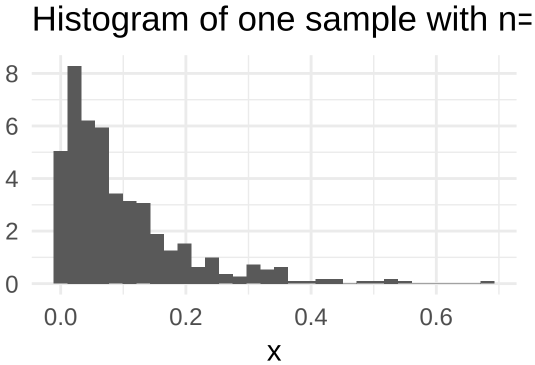
Exercise 12, question 3



- The ecdfs $\hat{F}$ and $\hat{G}$ associated to the two samples look alike apart from that the black one represents a more dispersed sample,
- One could think that the distributions F and G of the associated random variables (describing the associated population) are rescaled versions, i.e. $F(\cdot) = G(\cdot/\sigma)$ for some σ (scale parameter).

Exercise 12, question 3

```r
nx <- 500
lambdax <-10
x <-rexp(nx,rate=lambdax)
data=data.frame(x=x)
library(ggplot2)
ggplot(data = data, aes(x=x))+geom_histogram( aes(y = 45*(
```



Q-Q plots

- Used to compare the distribution of a sample with a reference distribution (or another sample). For example, to assess whether a distribution “looks” Gaussian.
- Let $(x_{(1)}, \dots, x_{(n)})$ be the quantiles of the sample and $(y_{(1)}, \dots, y_{(n)})$ the quantiles of the other sample. A Q-Q plot represents the points $(x_{(i)}, y_{(i)})$ for all $1 \leq i \leq n$.
- In the case of a comarison with a distribution G, the empirical quantiles of the sample are plotted against the quantiles of the reference distribution

$$(Q_G(i/(n+1)), x_{(i)}), \quad \text{for } i = 1, \dots, n$$

Q-Q plots

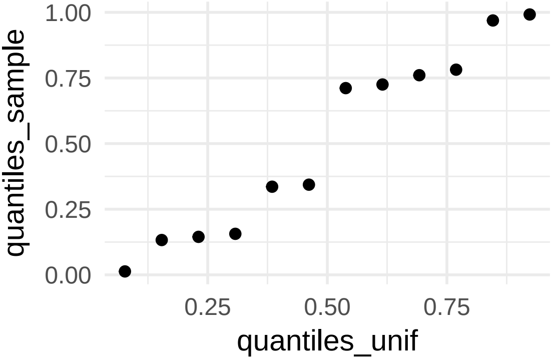
Q-Q plots

- Exercise 13

Exercise 13, question 1

- $Q_{\mathcal{U}(0,1)}\left(\frac{i}{n+1}\right) = \frac{i}{n+1}$

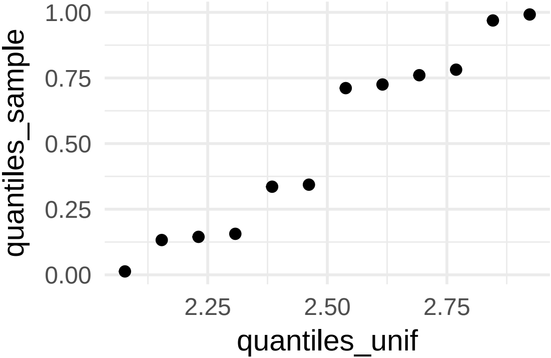
```
library(ggplot2)
n <-12
I <- 1/(n+1)*(1:n)
data_QQplot <- data.frame(quantiles_sample=sort(xunif),quan
ggplot(data=data_QQplot,aes(x=quantiles_unif,y=quantiles_sa
```



Exercise 13, question 2

- $Q_{\mathcal{U}(2,3)}\left(\frac{i}{n+1}\right) = \frac{i}{n+1} + 2$

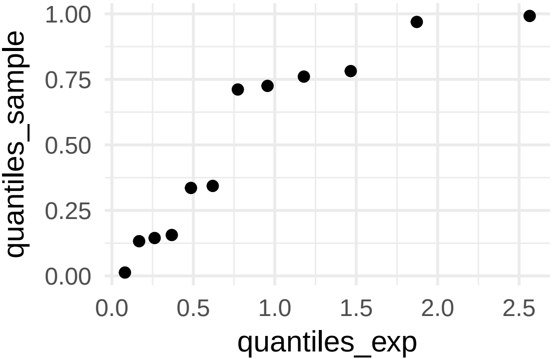
```
library(ggplot2)
n <-12
I <- 1/(n+1)*(1:n)
data_QQplot <- data.frame(quantiles_sample=sort(xunif),quan
ggplot(data=data_QQplot,aes(x=quantiles_unif,y=quantiles_sa
```



Exercise 13, question 3

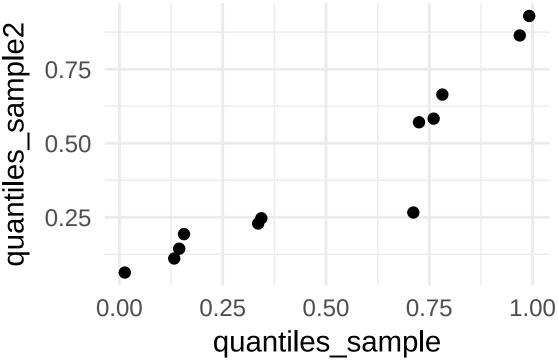
- $Q_{\mathcal{E}(1)}\left(\frac{i}{n+1}\right) = -\log(1-i/(n+1))$

```
library(ggplot2)
n <-12
I <- 1/(n+1)*(1:n)
data_QQplot <- data.frame(quantiles_sample=sort(xunif),qua
ggplot(data=data_QQplot,aes(x=quantiles_exp,y=quantiles_sa
```



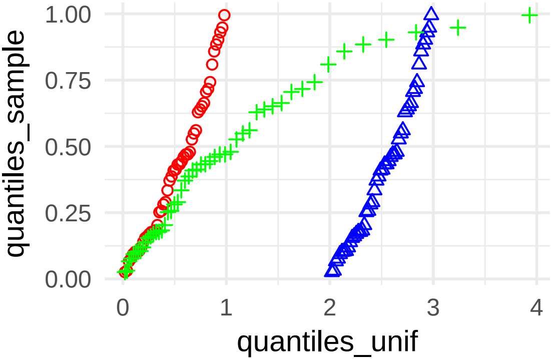
Exercise 13, question 4

```
library(ggplot2)
n <-12
I <- 1/(n+1)*(1:n)
data_QQplot <- data.frame(quantiles_sample=sort(xunif),qua
ggplot(data=data_QQplot,aes(x=quantiles_sample,y=quantiles_
```

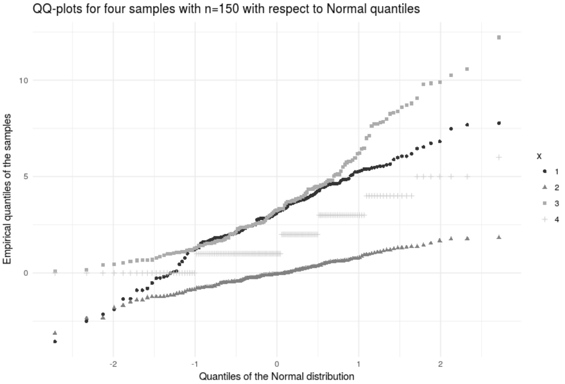


Exercise 13, $n = 500$

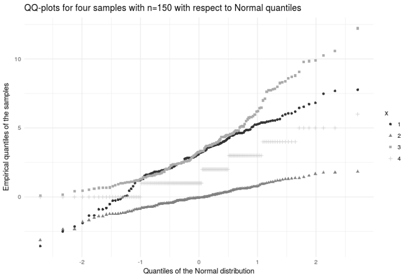
```
library(ggplot2)
n <-50
x <- runif(n)
I <- 1/(n+1)*(1:n)
data_QQplot <- data.frame(quantiles_sample=sort(x),quantile
ggplot(data=data_QQplot)+geom_point(aes(x=quantiles_unif,y=
```



Exercise 13, question 5

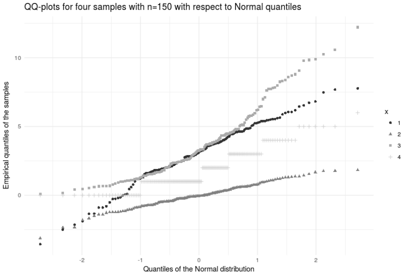


Exercise 13, question 5



- The Q-Q plots associated to samples 1 and 2 show aligned points so that the observed values might be the realizations of Normal random variables $\mathcal{N}(\mu, \sigma^2)$. For sample 2, $\mu \sim 0$ and $\sigma \sim 1$.

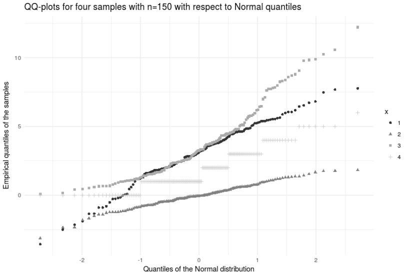
Exercise 13, question 5



- The points of the Q-Q plot of sample 3 are not aligned (not Gaussian at all). The smallest value is 0, all the values are positive. The left tail of the associated distribution is lighter than a Gaussian. While the right tail seems to have a heavier tail than the Gaussian distribution.

The sample could be the realizations of i.i.d. r.v. with Exponential (or Gamma) distributions $\{Exp(\lambda), \lambda > 0\}$.

Exercise 13, question 5

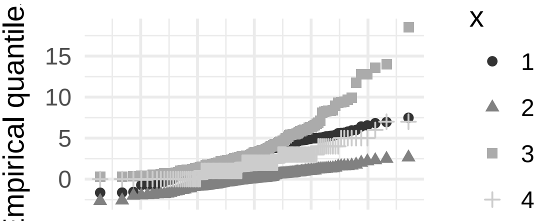


- The points of the Q-Q plot of sample 4 are not aligned (not Gaussian at all). There are only 6 possible values, the observations are discrete.

The sample could be the realizations of i.i.d. r.v. with Poisson distributions $\{\mathcal{P}(\lambda), \lambda > 0\}$

Exercise 13, question 4

```
library(e1071)
nx <- 150
dfx <-10
y1 <-rnorm(nx,mean=3,sd=2)
y2<-rnorm(nx)
y3 <-rgamma(nx,shape=2,scale=2)
y4<- rpois(nx,lambda=2)
data=data.frame(x=as.factor(c(rep(1,nx),rep(2,nx),rep(3,nx
library(ggplot2)
ggplot(data=data, aes(sample=y,group=x,colour=x,shape=x))+
  xlab("Quantiles of the Normal distribution")+ ylab("Empi
```



Bivariate quantitative data

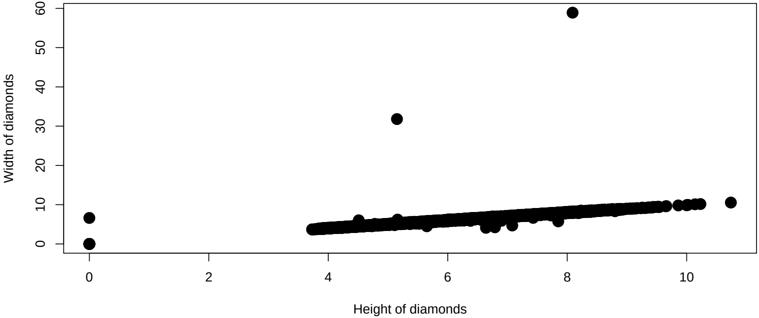
- Now for all $1 \leq i \leq n$ (individuals) we observe two quantitative values (e.g. height and weight): the data set is

$$\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix} \right),$$

- For all i, $(x_i, y_i) \in \mathbb{R}^2$
- Graphical representations & measures of correlation

Graphical representations

- Bivariate quantitative data are naturally displayed in **scatter plots**: plotting the points (x_i, y_i) in the xy plane.



Measures of correlation

- The **empirical covariance** and **empirical correlation** describe the relationships between the two variables x and y.
- Empirical covariance:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}_n \bar{y}_n$$

- Empirical correlation:

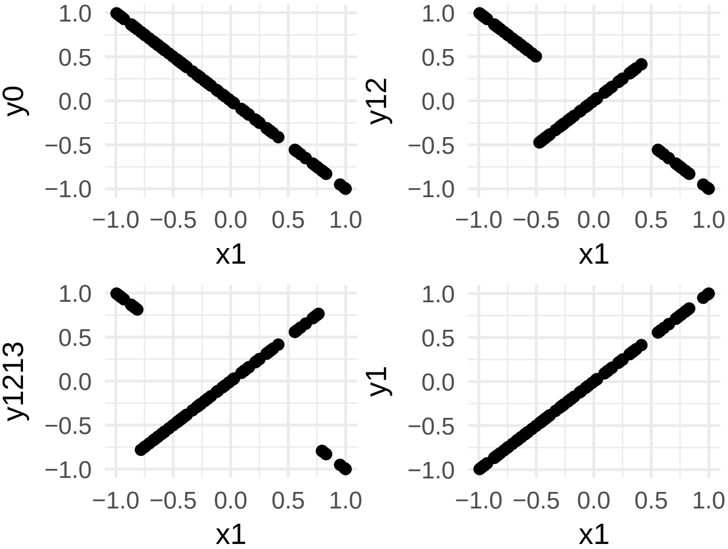
$$\rho_{xy} = \frac{s_{xy}}{s_x s_y}$$

Bivariate data

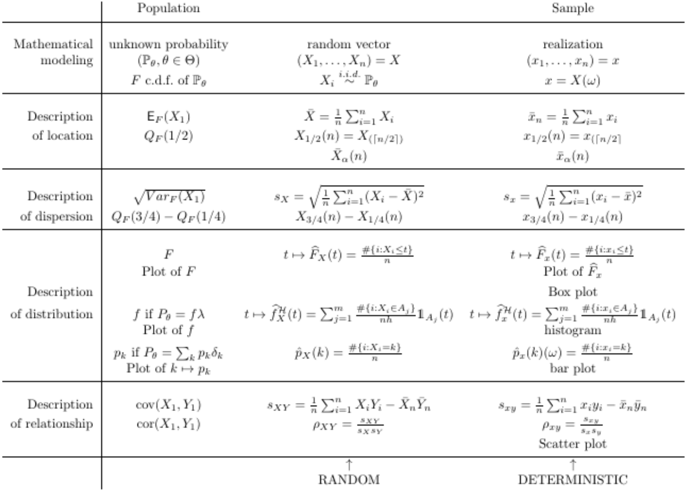
- Exercise 14
- (Exercise 15)

Exercise 14, question 4

Scatter plots of samples for $a = 0, 1/2, (1/2)^{(}1/3), 1$



Summary



	Population		Sample
Mathematical modeling	unknown probability $(\mathbb{P}_\theta, \theta \in \Theta)$ F c.d.f. of $\mathbb{P}_\theta$	random vector $(X_1, \ldots, X_n) = X$ $X_i \overset{i.i.d.}{\sim} \mathbb{P}_\theta$	realization $(x_1, \ldots, x_n) = x$ $x = X(\omega)$
Description of location	$\mathsf{E}_F(X_1)$ $Q_F(1/2)$	$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ $X_{1/2}(n) = X_{(\lceil n/2 \rceil)}$ $\bar{X}_\alpha(n)$	$\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ $x_{1/2}(n) = x_{(\lceil n/2 \rceil}$ $\bar{x}_\alpha(n)$
Description of dispersion	$\sqrt{Var_F(X_1)}$ $Q_F(3/4) - Q_F(1/4)$	$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}$ $X_{3/4}(n) - X_{1/4}(n)$	$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}$ $x_{3/4}(n) - x_{1/4}(n)$
Description of distribution	F Plot of F f if $P_\theta = f\lambda$ Plot of f p_k if $P_\theta = \sum_k p_k \delta_k$ Plot of $k \mapsto p_k$	$t \mapsto \widehat{F}_X(t) = \frac{\#\{i: X_i \leq t\}}{n}$ $t \mapsto \widehat{f}_X^{\mathcal{H}}(t) = \sum_{j=1}^m \frac{\#\{i: X_i \in A_j\}}{nh} \mathbb{1}_{A_j}(t)$ $\hat{p}_X(k) = \frac{\#\{i: X_i = k\}}{n}$	$t \mapsto \widehat{F}_x(t) = \frac{\#\{i: x_i \leq t\}}{n}$ Plot of $\widehat{F}_x$ Box plot $t \mapsto \widehat{f}_x^{\mathcal{H}}(t) = \sum_{j=1}^m \frac{\#\{i: x_i \in A_j\}}{nh} \mathbb{1}_{A_j}(t)$ histogram $\hat{p}_x(k)(\omega) = \frac{\#\{i: x_i = k\}}{n}$ bar plot
Description of relationship	$\mathrm{cov}(X_1, Y_1)$ $\mathrm{cor}(X_1, Y_1)$	$s_{XY} = \frac{1}{n}\sum_{i=1}^n X_i Y_i - \bar{X}_n \bar{Y}_n$ $\rho_{XY} = \frac{s_{XY}}{s_X s_Y}$	$s_{xy} = \frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}_n \bar{y}_n$ $\rho_{xy} = \frac{s_{xy}}{s_x s_y}$ Scatter plot
		↑ RANDOM	↑ DETERMINISTIC