# MAP531 : Statistics
# PC1– Descriptive statistics

We correct exercises 1,3,4,6,9,10,12,13,14. We use the following notations in the solutions : PR := 'Probability Refresher : Lecture Notes : Master X-HEC, September 2018'.[1] Ex := Exercise. CBS := Cauchy-Bunyakovsky-Schwarz inequality.

'Statistics is a branch of mathematics dealing with the collection, organization, analysis and **presentation of data**.'

In this training session, we are going to study tools to describe and summarize data. The goal is to know how to present the data once it has been collected. The next step will be to model and then analyze the data. This description step can help in the choice of the model.

Our data is a collection $(x_1, \ldots, x_n)$ of $n$ observed values which are realizations of the vector of observations $(X_1, \ldots, X_n)$. The vector of observations $(X_1, \ldots, X_n)$ is a random vector and $(x_1, \ldots, x_n)$ is a realization, meaning that there exits $\omega$ such that $x_i = X_i(\omega)$ for all $i$. Here, we will only consider the case where $X_1, \ldots, X_n$ are i.i.d. with c.d.f. $F$. $(x_1, \ldots, x_n)$ is called sample, data set, data or observed values.

You will face different kinds of data :

— **univariate** : $x_i$ has only one component, e.g. $x_i \in \mathbb{R}$, or $x_i \in \{\text{blue}, \text{red}, \text{green}\}$ ;

— **multivariate** : $x_i$ has several components $x_i = (x_i^1, \ldots, x_i^d)$, e.g. $x_i \in \mathbb{R}^d$ with $d \geqslant 2$ or $x_i \in \mathbb{R} \times \{\text{blue}, \text{red}, \text{green}\}$.

When the observed values are numerical, they are called **quantitative**. **Qualitative** measurements have 'values' that are either categories, characteristics or conditions. Quantitative univariate observations are said **continuous** when $F$ admits a density function with respect to Lebesgue measure. Univariate quantitative observations are called **discrete** if it can take a finite number of values or a countably infinite number. Be careful, some probability measures are neither discrete nor absolutely continuous. For instance, it's the case of censored data where $X_i = \min(Z_i, c)$ with a continuous random variable $Z_i$ and a deterministic constant $c \in \mathbb{R}$. You have such data when you use a thermometer with maximal temperature 25°C.

Next, we are going to focus on quantitative data, and first on univariate quantitative data.

## 1 Univariate quantitative data

We first consider univariate quantitative data where $x_1, \ldots, x_n$ are $n$ real observed values. We consider in the following examples variables obtained from the dataset 'diamonds' from the package 'ggplot2' in R.

**Example 1.** *The first following dataset gives the width of the twelve first diamonds.*

| *Width (mm) :* | *3.98* | *3.84* | *4.07* | *4.23* | *4.35* | *3.96* | *3.98* | *4.11* | *3.78* | *4.05* | *4.28* | *3.90* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

**Example 2.** *The following dataset gives the quality of the cut of the twelve first diamonds with the following correspondence :* 1 *meaning ideal,* 2 *premium,* 3 *very good,* 4 *good and* 5 *fair.*

| *Quality of the cut :* | *1* | *2* | *4* | *2* | *4* | *3* | *3* | *3* | *5* | *3* | *4* | *1* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

### 1.1 Numerical summaries

#### 1.1.1 Measure of location

A measure of location is a measure of a 'center' of a dataset. The first measure of location we can think of is the empirical mean.

[1] Link : `https://www.lpsm.paris/pageperso/rebafka/lecturenotes_all.pdf`

**Definition 1.** *The* **empirical mean***, also called arithmetic mean, of a sample* $(x_1, \dots, x_n)$ *is defined as*

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

*Similarly, the empirical mean of the observations* $X_1, \dots, X_n$ *is* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

While $\bar{x}_n$ is deterministic, $\bar{X}_n$ is random. The empirical mean $\bar{X}_n$ approaches the expectation of $F$ : $\bar{X}_n \to \mathsf{E}_F(X)$ almost surely when $\mathsf{E}_F[|X|] < \infty$. This measure of location (like the expectation) can be misleading due to outliers. We present here two other measures of location which are robust, i.e. insensitive to outliers.

One robust measure of the location is the median. To define it, let's first define the order statistics which are the re-ordered values of the sample.

**Definition 2.** *The* **order statistics** *of a sample* $(x_1, \dots, x_n)$ *are defined as* $x_{(1)}, \dots, x_{(n)}$ *where*

$$x_{(j)} \in \{x_1, \dots, x_n\}, \text{ for all } j \leqslant n \quad \text{and} \quad x_{(1)} \leqslant x_{(2)} \leqslant \dots \leqslant x_{(n)}.$$

*The* **order statistics** $X_{(1)}, \dots, X_{(n)}$ *of the observations* $(X_1, \dots, X_n)$ *are defined in the same way.*

The $\alpha$-empirical quantile is a value which cuts the ordered sample into two sets with proportion $\alpha 100\%$ and $(1-\alpha)100\%$.

**Definition 3.** *Let* $\alpha \in (0,1)$*, the* $\alpha$-**empirical quantile** *of a sample* $(x_1, \dots, x_n)$ *(of the observations* $(X_1, \dots, X_n)$*) is*

$$x_\alpha(n) = x_{(\lceil \alpha n \rceil)} \quad \left(X_\alpha(n) = X_{(\lceil \alpha n \rceil)}\right).$$

*Then the* **median** *is just the* $1/2$*-empirical quantile :*

$$x_{1/2}(n) = x_{(\lceil n/2 \rceil)} \quad \left(X_{1/2}(n) = X_{(\lceil n/2 \rceil)}\right).$$

*Similarly, the first, second and third quartiles are*

$$x_{1/4}(n) = x_{(\lceil n/4 \rceil)}, \quad x_{2/4}(n) = x_{1/2}(n) = x_{(\lceil 2n/4 \rceil)}, \quad \text{and} \quad x_{3/4}(n) = x_{(\lceil 3n/4 \rceil)},$$

$$\left(X_{1/4}(n) = X_{(\lceil n/4 \rceil)} \quad X_{2/4}(n) = X_{1/2}(n) = X_{(\lceil 2n/4 \rceil)}, \quad \text{and} \quad X_{3/4}(n) = X_{(\lceil 3n/4 \rceil)}\right).$$

The empirical quantiles approach the theoretical ones which are defined as follows (see Exercise 4).

**Definition 4.** *Let* $F$ *be a c.d.f. on* $\mathbb{R}$ *and* $\alpha \in (0,1)$*, its* $\alpha$-**quantile** *is defined as*

$$Q_F(\alpha) = \inf\{t \in \mathbb{R},\, F(t) \geqslant \alpha\}.$$

*The* $\alpha$*-quantile is also called* $100 * \alpha$*-percentile.*

Another robust measure of location is the $\alpha$-trimmed (or truncated) mean which is the empirical mean computed with only a central $(1-\alpha)$-proportion of the sample :

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

**Definition 5.** *Let* $\alpha \in [0,1)$*, the* $\alpha$-**trimmed mean** *of a sample* $(x_1, \dots, x_n)$ *(of the observations* $X_1, \dots, X_n$*) is*

$$\bar{x}_\alpha(n) = \frac{1}{n - 2\lceil \alpha n/2 \rceil} \sum_{i=\lceil \alpha n/2 \rceil + 1}^{n - \lceil \alpha n/2 \rceil} x_{(i)} \qquad \left(\bar{X}_\alpha(n) = \frac{1}{n - 2\lceil \alpha n/2 \rceil} \sum_{i=\lceil \alpha n/2 \rceil + 1}^{n - \lceil \alpha n/2 \rceil} X_{(i)}\right).$$

**Exercise 1. Empirical mean, median and trimmed-mean**

1. Compute the empirical mean, the median and the 0.2-trimmed mean of the data set of Example 1.

2. Give an example of a dataset where the difference between the median and the empirical mean is greater than $n$ for all $n \in \mathbb{N}$. What is the 0.2-trimmed mean in this case ?
3. Can you write the empirical mean and median as $\alpha$-trimmed mean for some $\alpha$ ?
4. How do you explain that the median GDP per capita in the US, around 17,600$ in 2014, is far from its mean, around 54,600$ in 2014 ?

**Solution**

1. mean : $\bar{x}_n = 4.0442$ ; median : $n = 12$, $\lceil 1/2 \times 12 \rceil = \lceil 6 \rceil = 6$ ; the ordered values are

$$\begin{array}{llllll} x_{(1)} = 3.78, & x_{(2)} = 3.84, & x_{(3)} = 3.90, & x_{(4)} = 3.96, & x_{(5)} = 3.98, & x_{(6)} = 3.98, \\ x_{(7)} = 4.05, & x_{(8)} = 4.07, & x_{(9)} = 4.11, & x_{(10)} = 4.23, & x_{(11)} = 4.28, & x_{(12)} = 4.35 \end{array}$$

   hence $x_{(6)} = 3.98$ ; 0.2-trimmed mean : $\lceil 0.2 \times 12/2 \rceil = \lceil 1.2 \rceil = 2$, $12 - \lceil 0.2 \times 12/2 \rceil + 1 = 12 - 2 + 1 = 11$. Hence

$$x_{0.2}(12) = \frac{x_{(3)} + x_{(4)} + x_{(5)} + x_{(6)} + x_{(7)} + x_{(8)} + x_{(9)} + x_{(10)}}{8} = 4.035.$$

2. $x_1 = 1$, $x_2 = 1$, $x_3 = n$. In this case $\lceil 1/2 \times 3 \rceil = 2$ hence $x_{1/2}(3) = 1$, but $\bar{x}_3 = \frac{n+2}{3}$. Since $\lceil 0.2 \times 3 \rceil = \lceil 0.6 \rceil = 1$, $3 - 1 + 1 = 3$, we have $x_{0.2}(3) = x_{(2)} = 1$.
3. No ; it would require $\lceil \alpha n/2 \rceil = 0$ which can not hold since $\alpha > 0$, $n > 0$ ($\Rightarrow \lceil \alpha n/2 \rceil \geq 1$).
4. There are a few millionaires (=outliers) ; hence the mean is much larger than the median.

In the next exercise, we show that in the case of symmetric density functions the expectation and 1/2-quantile (i.e. the theoretical median) are equal.

**Exercise 2. Symmetric density functions**

1. Assume that $F$ admits a density function $f$ (with respect to the Lebesgue measure) which is symmetric with respect to $\mu \in \mathbb{R}$, i.e. $f(\mu - x) = f(\mu + x)$ for almost all $x$. Prove that for all odd $k$, if $\mathsf{E}_F(|X|^k) < \infty$ then

$$\mathsf{E}_F[(X - \mu)^k] = 0.$$

   Deduce that if $f$ is positive in the neighborhood of $\mu$ then

$$\mathsf{E}_F[X] = Q_F(1/2).$$

2. Give an example of a density function where its 1/2-quantile and its expectation are different.

### 1.1.2 Measures of dispersion

A measure of dispersion measures the scatteredness of a sample around its 'center'. A measure of dispersion together with a measure of dispersion are a good summary of a sample.

**Definition 6.** *Here are some measures of dispersion of a sample $x = (x_1, \dots, x_n)$ :*

— *the* **range** *of the data : $x_{(n)} - x_{(1)}$,*

— *the* **empirical standard deviation** *: $s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$,*

— *the* **empirical variance** *: $s_x^2$*

— *the* **interquartile range** *: $x_{3/4}(n) - x_{1/4}(n)$,*

— *and the* **median absolute deviation from the median** *which is the median of $|x_1 - x_{1/2}(n)|$, $|x_2 - x_{1/2}(n)|$, ..., $|x_n - x_{1/2}(n)|$ .*

These measures of dispersions can be defined in the same way for the vector of observations $(X_1, \dots, X_n)$.

**Exercise 3. Measures of dispersion**

1. Compute theses measures of dispersion for the dataset of Example 1.
2. Which of these measures of dispersion are not sensitive to outliers ?

**Solution**

1. range : $x_{(12)} - x_{(1)} = 0.57$ ; empirical std : $s_x = 0.1748$ ; empirical variance : $s_x^2 = 0.0306$ ; interquantile range : $\lceil 1/4 \times 12 \rceil = 3$, $\lceil 3/4 \times 12 \rceil = 9 \Rightarrow x_{(9)} - x_{(3)} = 0.21$ ;median absolute deviation from the median : 0.09. The sorted values (where we need the 6th) :

   $$0 \quad 0 \quad 0.02 \quad 0.07 \quad 0.08 \quad 0.09 \quad 0.13 \quad 0.14 \quad 0.20 \quad 0.25 \quad 0.30 \quad 0.37.$$

2. The last two measures are not (that) sensitive to outliers.

**Exercise 4. Consistent estimators**

Let $X_i$ be i.i.d. observations with c.d.f. $F$ and $X_{1:n} = (X_1, \dots, X_n)$. Show that under some appropriate assumptions the following limits hold

$$\bar{X}_n \xrightarrow[n\to\infty]{F-\text{proba}} \mathsf{E}_F[X], \quad s^2_{X_{1:n}} \xrightarrow[n\to\infty]{F-\text{proba}} Var_F[X] \quad \text{and} \quad X_{1/2}(n) \xrightarrow[n\to\infty]{F-\text{proba}} Q_F(1/2).$$

We will see later that this property means that the empirical mean, the empirical median and the empirical variance are consistent estimators of the expectation, the 1/2-quantile and the theoretical variance respectively.

**Solution**

1. We saw in PR : Theorem 6.2.1 (weak law of large numbers) and its consequence ($L_2$ convergence $\Rightarrow$ convergence in probability). Condition : $\mathrm{var}(X_1) < \infty$ (weak law of large numbers). [2]
2. Assume that $X_{1:n}$ is coming from a continuous distribution and $n = 2m+1$. Let $M = Q_X\left(\frac{1}{2}\right)$ be the true median ; its empirical estimator is $M_n := X_{1/2}(n) = X_{\left(\frac{n+1}{2}\right)}$. The goal is to show that $\mathbb{P}(|M_n - M| > \epsilon) \xrightarrow{P} 0$ for any fixed $\epsilon > 0$. We will prove that $\mathbb{P}(M_n - M > \epsilon) \xrightarrow{P} 0$ ; $\mathbb{P}(M_n - M < -\epsilon) \xrightarrow{P} 0$ can be done similarly. The idea is to reformulate the event in terms of a binomial variable, have its mean appear and apply the Chebysev inequality.

   $$\begin{aligned}\mathbb{P}(M_n - M > \epsilon) &= \mathbb{P}(M_n > M+\epsilon) = \mathbb{P}(\text{at least } (n+1)/2 \text{ of the } X_i\text{-s exceed } M+\epsilon) = \mathbb{P}\left(B_n \geq \frac{n+1}{2}\right) \\ &= \mathbb{P}\Big(B_n - \underbrace{np}_{\mathsf{E}(B_n)} \geq \underbrace{\frac{n+1}{2} - np}_{n(1/2-p)+1/2}\Big) \leq \mathbb{P}\Big(B_n - \mathsf{E}(B_n) \geq \underbrace{n(1/2-p)}_{a}\Big) \overset{(*)}{\leq} \frac{\mathrm{var}(B_n)}{a^2} = \frac{np(1-p)}{n^2(1/2-p)^2} \\ &= \frac{p(1-p)}{n(1/2-p)^2} \xrightarrow{n\to\infty} 0,\end{aligned}$$

   where $B_n = B(n,p)$, $p = \mathbb{P}(X_i > M+\epsilon) > \frac{1}{2}$, $\mathsf{E}(B_n) = np$, $\mathrm{var}(B_n) = np(1-p)$, in (*) we used the Chebysev inequality.
3. Using the weak law of large numbers (PR : Theorem 6.2.1) and with PR : Prop. 6.1.3. $[g(a) = a^2]$ one gets

   $$\tilde{S}^2_{X_{1:n}} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i^2}_{\xrightarrow{P}\mathsf{E}(X^2)} - \underbrace{\bar{X}_n^2}_{\xrightarrow{P}\mathsf{E}^2(X)} \quad .$$

   Thus, by PR : 6.1.3 : 2nd bullet $[g(a,b) = a+b]$ $\tilde{S}^2_{X_{1:n}} \xrightarrow{P} \mathsf{E}(X^2) - \mathsf{E}^2(X) = S^2_X$. Since $S^2_{X_{1:n}} = \underbrace{\frac{n}{n-1}}_{\xrightarrow{P}1}\underbrace{\tilde{S}^2_{X_{1:n}}}_{\xrightarrow{P}S^2_X}$, the claimed result follows by invoking PR : 6.1.3 : 2nd bullet with $g(a,b) = ab$.

[2] Using the strong law of large numbers $\mathsf{E}(X_1) < \infty$ is enough.

### 1.1.3 Some other statistics

Here are two other statistics :

— the **empirical skewness** of a sample $x = (x_1, \dots, x_n)$ is $\hat{\alpha}_x = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s_x^3}$,

— the **empirical kurtosis** of a sample $x = (x_1, \dots, x_n)$ is $\widehat{\beta}_x = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s_x^4} - 3$.

**Exercise 5. Empirical skewness and kurtosis**

1. Compute these measures of dispersion for the dataset of Example 1.
2. Which deterministic quantity $\alpha$ and $\beta$ the statistics $\hat{\alpha}_X$ and $\hat{\beta}_X$, associated to the vector of observations $(X_1, \dots, X_n)$ approach ?
3. What is the value of $\alpha$ when $F$ admits a symmetric density ?
4. Show that when $\beta$ exists $\beta \geq -2$.
5. Here are the values of $\beta$ for different c.d.f. $F$ with density function $f$ symmetric at 0, with standard deviation equal to one. How would you interpret $\beta$ ?

| Distribution | density $\propto$ | $\beta$ |
|---:|:---:|:---:|
| Pearson VII ($m > 5/2$) | $(1 + x^2/(2m-3))^{-m}$ | $6/(2m-5)$ |
| Laplace | $\exp(\lvert x\rvert/\sqrt{2})$ | 3 |
| Normal | $\exp(x^2/2)$ | 0 |
| Wigner semicircle | $\sqrt{4-x^2}\mathbb{1}_{(-2,2)}(x)$ | $-1$ |
| Uniform | $\mathbb{1}_{(-\sqrt{3},\sqrt{3})}(x)$ | $-1.2$ |

## 1.2 Graphical representations

### 1.2.1 Box plot

Box plots are good summaries of data, representing the location, the dispersion, the symmetry and even the skewness of the data as can be seen in Figure 1.

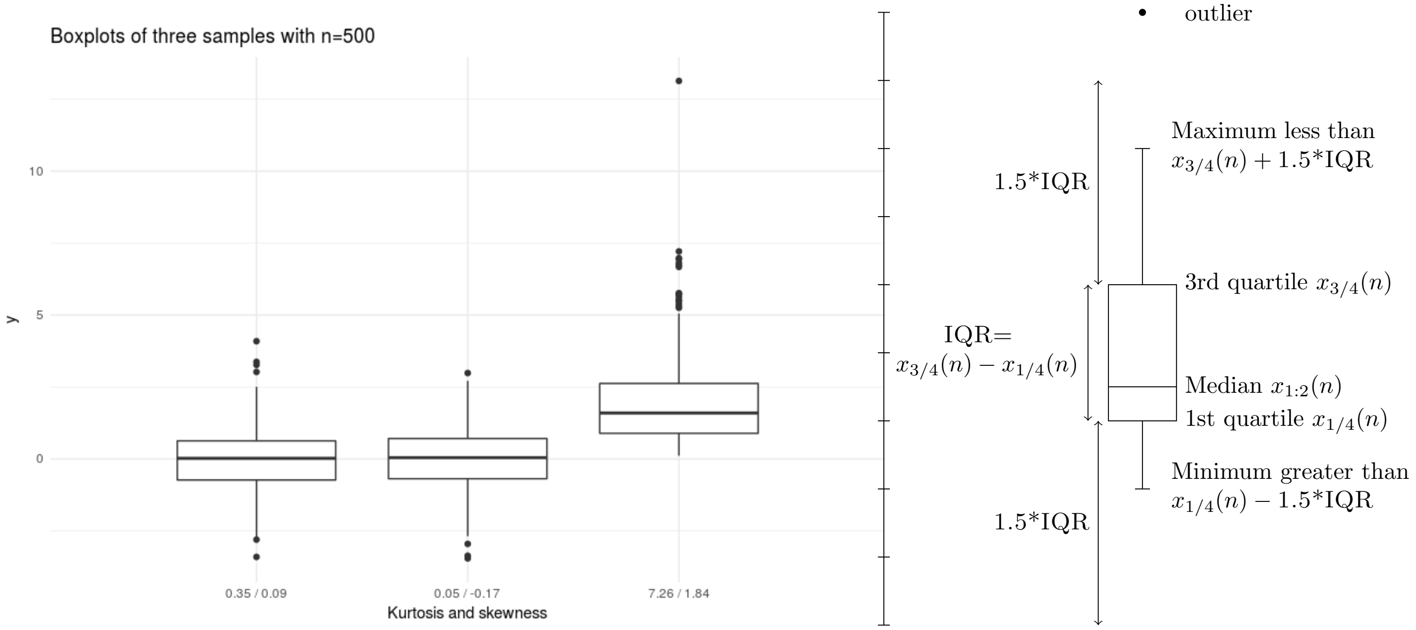
**Definition 7.** *Box plots show the empirical median, first and third quartiles thanks to a rectangle divided by a segment as shown in Figure 1b. The most extreme data point within a distance of* 1.5 *times the interquartile range below the first quartile and above the third are represented via a segment above and below the rectangle. Possible outliers, i.e. data points beyond a distance of* 1.5 *times the interquartile range below the first quartile and above the third quartiles are plotted with circles (or asterisks).*

**Exercise 6. Box plots**

1. Draw the box plots of the sample of Example 1.
2. Figure 1a represents the box plots of three samples. Determine characteristics of the distribution associated to these datasets. And propose a family of distributions which is likely to have generated them.

**Solution**

1. $x_{1/2}(n) = 3.98$ (Ex1), $\mathrm{IQR} = 0.21$ (Ex3), $x_{1/4}(n) = x_{(3)} = 3.90$ (Ex3), $x_{3/4}(n) = x_{(9)} = 4.11$ (Ex3) $\Rightarrow x_{1/4}(n) - 1.5 \times \mathrm{IQR} = 3.90 - 1.5 \times 0.21 = 3.585$, $x_{3/4}(n) + 1.5 \times \mathrm{IQR} = 4.11 + 1.5 \times 0.21 = 4.425$ $\Rightarrow$ upper whisker $= x_{(12)} = 4.35$ (Ex1), lower whisker $= x_{(1)} = 3.78$ (Ex1), no outlier.
2. — symmetric around 0, many outliers, large kurtosis $\Rightarrow$ heavy-tailed distribution.
   — symmetric around 0, e.g. $N\left(0, \sigma^2\right)$.
   — not symmetric, support $\subset \mathbb{R}_{>0}$, skewed, e.g. $exp(\lambda)$.



(a)

(b) Description of a box plot

Figure 1 – Box plots

### 1.2.2 Empirical cumulative distribution function

**Definition 8.** *The* **empirical cumulative distribution function** *(ecdf) of a sample $x = (x_1, \dots, x_n)$ is the following function*

$$\widehat{F}_x : \begin{cases} \mathbb{R} & \to [0,1] \\ t & \mapsto \widehat{F}_x(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{x_i \leq t} = \frac{\#\{i : x_i \leq t\}}{n} \end{cases} .$$

The empirical cumulative distribution function (ecdf) of a vector of observations $X = (X_1, \dots, X_n)$ is $\widehat{F}_X$.
The empirical cumulative distribution function is a non-decreasing step function where

$$\widehat{F}_x(t) = 0 \text{ if } t < x_{(1)} = \min\{x_1, \dots x_n\}, \quad \widehat{F}_x(t) = 1 \text{ if } t \geqslant x_{(n)} = \max\{x_1, \dots x_n\}$$

and jumps at each $t$ in $\{x_1, \dots, x_n\}$ with a jump of height $1/n$ in the case of a unique observed value is equal to $t$ or a jump of height $k/n$ if $k$ observed value equals $t$.

**Exercise 7.** Draw the empirical cumulative distribution function of the datasets of Examples 1 and 2.

**Exercise 8. Empirical cumulative distribution function** Let $X_i$ be i.i.d. observations with c.d.f. $F$ and $X_{1:n} = (X_1, \dots, X_n)$.

1. Show that for all $\alpha \in (0,1)$

$$X_\alpha(n) = \inf\{t \in \mathbb{R}, \widehat{F}_{X_{1:n}}(t) \geqslant \alpha\} =: \widehat{F}^{-1}_{X_{1:n}}(\alpha),$$

where $\widehat{F}^{-1}_{X_{1:n}}$ is the generalized inverse of the empirical cumulative distribution function.

2. Fix $t \in \mathbb{R}$, what is the distribution of $n\widehat{F}_{X_{1:n}}(t)$ ? Can you complete the following limits :

$$\widehat{F}_{X_{1:n}}(t) \xrightarrow[n\to\infty]{F-\text{proba}} ?? \quad \text{and} \quad \sqrt{n}\left(\widehat{F}_{X_{1:n}}(t) - ??\right) \xrightarrow[n\to\infty]{F-\text{dist.}} \mathcal{N}(0, ??)\,?$$

**Exercise 9. Empirical and theoretical quantiles and cumulative distribution function**

1. Compute the theoretical quantiles of a uniform distribution on $\mathcal{U}(0,\theta)$, for $\theta \in \mathbb{R}_+$.

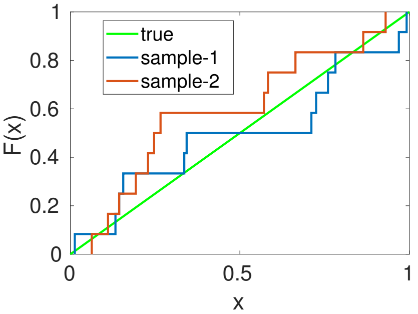


FIGURE 2 – Ex9 : True and empirical cdf-s.

2. Draw the theoretical cumulative distribution function and the density of a uniform $\mathcal{U}(0,1)$ distribution in two different graphs.
3. Here are two 12-samples from a $\mathcal{U}(0,1)$ distribution :

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| sample 1 : | 0.725 | 0.133 | 0.156 | 0.992 | 0.144 | 0.711 | 0.013 | 0.336 | 0.343 | 0.760 | 0.782 | 0.969 |
| sample 2 : | 0.864 | 0.144 | 0.229 | 0.111 | 0.193 | 0.664 | 0.930 | 0.063 | 0.583 | 0.571 | 0.247 | 0.266 |

Draw the empirical cumulative distribution functions of each sample in the same graph as the theoretical cumulative distribution function.

### Solution

1. The cdf of $U[0,\theta]$ is $F(z) = \frac{1}{\theta} z \mathbf{1}_{[0,\theta]}$. Hence for an $\alpha \in (0,1)$

$$Q_F(\alpha) = \inf\{z \,:\, F(z) \geq \alpha\} = \{z \,:\, F(z) = \alpha\} = \left\{z \,:\, \frac{z}{\theta} = \alpha\right\} \Rightarrow z = \alpha\theta.$$

2. The pdf is $f(z) = \frac{1}{\theta}\mathbf{1}_{[0,\theta]}$ ; the cdf is $F(z) = \frac{1}{\theta} z \mathbf{1}_{[0,\theta]}$.
3. The ordered samples for plotting (Fig 2) :

$$x_{(1)} = 0.0130, \quad x_{(2)} = 0.1330, \quad x_{(3)} = 0.1440, \quad x_{(4)} = 0.1560, \quad x_{(5)} = 0.3360, \quad x_{(6)} = 0.3430,$$
$$x_{(7)} = 0.7110, \quad x_{(8)} = 0.7250, \quad x_{(9)} = 0.7600, \quad x_{(10)} = 0.7820, \quad x_{(11)} = 0.9690, \quad x_{(12)} = 0.9920.$$

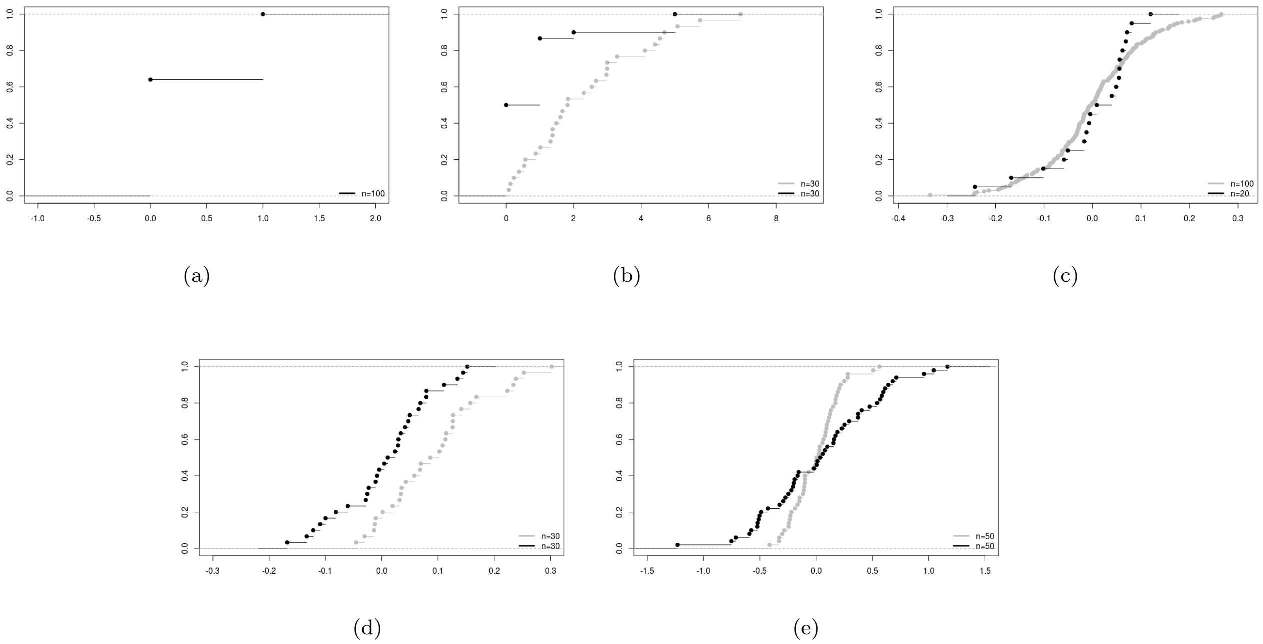
$$x_{(1)} = 0.0630, \quad x_{(2)} = 0.1110, \quad x_{(3)} = 0.1440, \quad x_{(4)} = 0.1930, \quad x_{(5)} = 0.2290, \quad x_{(6)} = 0.2470,$$
$$x_{(7)} = 0.2660, \quad x_{(8)} = 0.5710, \quad x_{(9)} = 0.5830, \quad x_{(10)} = 0.6640, \quad x_{(11)} = 0.8640, \quad x_{(12)} = 0.9300.$$

### Exercise 10. Description of data and ecdf

1. Figure 3a represents the ecdf of some sample of size 100. Deduce the characteristics of the distribution of the sample and propose a distribution that is likely to have generated the data.
2. Each sub-figure 3(b-d) represents the ecdfs of two samples. For each sub-figure, compare the characteristics of the distributions of each sample.

### Solution

1. $X \sim$Bernoulli($p$), $1-p = \mathbb{P}(X=0) \approx 0.65$, $p = \mathbb{P}(X=1) \approx 0.35$. Range = 1.
2. — (b) : Both are discrete, black is concentrated on fewer points $X \in \{0,1,2,5\}$ ; alternatively gray is continuous on $\mathbb{R}_{>0}$. Range = 5.
   — (c) : Both are coming from normal distribution, with zero mean, gray finer 'resolution' due to the larger number of samples. Range = 0.6.
   — (d) : Both are uniform with the same support size, black is concentrated on smaller numbers (=pdf shifted to the left).
   — (e) : Both are coming from the same distribution, but the black has bigger variance.



(a) (b) (c)

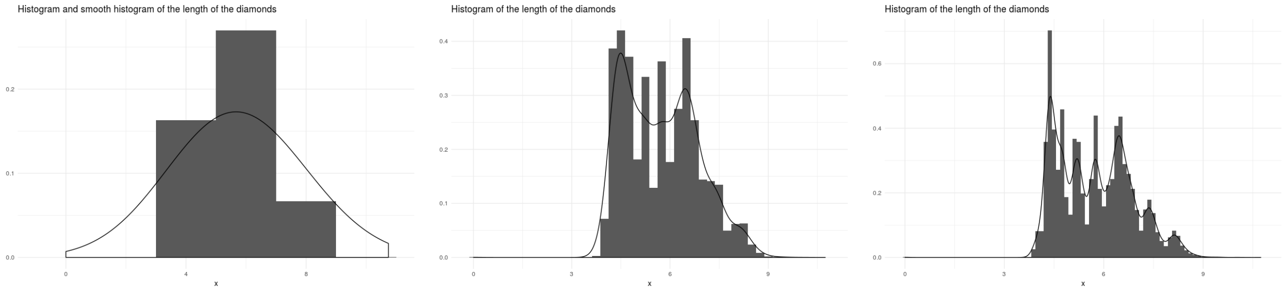
(d) (e)

FIGURE 3 – ECDFs for some samples

### 1.2.3 Bar plots and histograms

Bar plots and histograms display the shape of the distribution of data values. The first one is used for discrete data and the second one for non-discrete data.

**Definition 9.** *To represent discrete data, one can use* **bar plots** *where each bar has a null width (it is a stick). And the height of a bar at a taken value $k$ is the proportion $\hat{p}_{x_{1:n}}(k) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i=k} = \frac{\#\{i:x_i=k\}}{n}$ of values in the sample $x_{1:n} = (x_1, \ldots, x_n)$ with value $k$.*



(a) Bandwidth : 2

(b) Bandwidth : 1/4

(c) Bandwidth : 1/8

Figure 4 – Histograms and smooth histograms for the length of the diamonds

**Exercise 11. Bar plot**

1. Let $K = \{k_1, k_2, k_3, \ldots\}$ be a finite or countably infinite set and $X_i$, $i \leqslant n$ i.i.d. discrete random variables with values in $K$ and cumulative distribution function $F$. Fix $k \in K$, what is the distribution of $n\hat{p}_{X_{1:n}}(k)$ ? Can you complete the following limits :
$$\hat{p}_{X_{1:n}}(k) \xrightarrow[n\to\infty]{F-\text{proba}} ?? \quad \text{and} \quad \sqrt{n}\,(\hat{p}_{X_{1:n}}(k) - ??) \xrightarrow[n\to\infty]{F-\text{dist.}} \mathcal{N}(0, ??)\,?$$
2. Draw the bar plot of the sample of Example 2.

In the case of continuous data, we can represent the shape of the distribution of data values with a histogram.

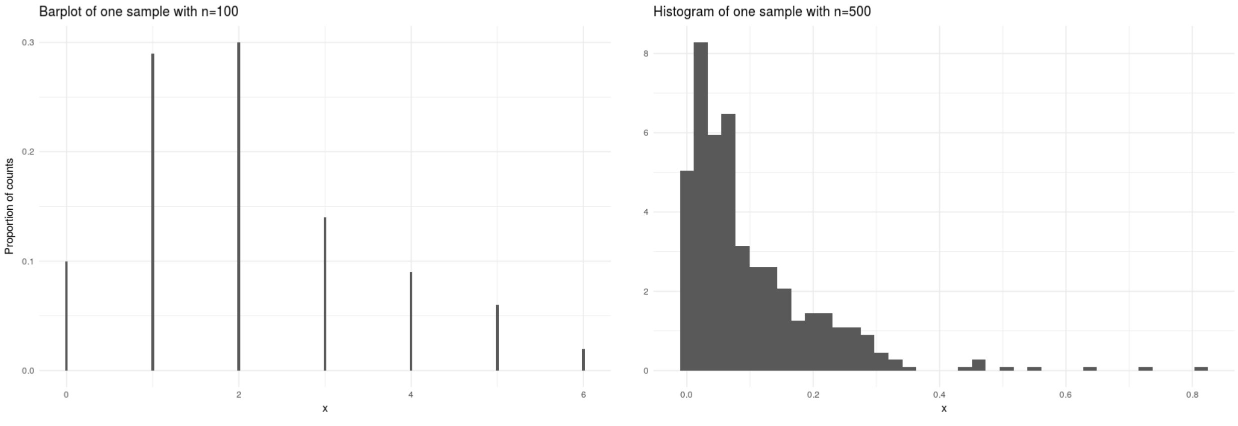
**Definition 10.** *Let $[a, b]$ be a fixed interval and $m$ an integer. Cut $[a, b)$ into $m$ bins of equal size $h = (b - a)/m$ such as $A_j = [a + \frac{j-1}{m}h, a + \frac{j}{m}h)$, for all $1 \leqslant j \leqslant m$. The* **histogram** *of the vector of observations $X = (X_1, \ldots, X_n)$ associated with the partition $\mathcal{H} = \{A_1, \ldots, A_m\}$ of $[a, b)$ is the following step function :*
$$\widehat{f}_X^{\mathcal{H}} : \left\{ \begin{array}{ll} \mathbb{R} & \to \mathbb{R} \\ t & \mapsto \widehat{f}_X^{\mathcal{H}}(t) = \sum_{j=1}^{m} \frac{\#\{i : X_i \in A_j\}}{nh} \mathbb{1}_{A_j}(t) \end{array} \right. .$$
*The nh factor in the denominator guarantees that $\widehat{f}_X^{\mathcal{H}}$ integrates to one.*

A histogram depends on the size of the window $h$, as illustrated in Figure 4. When the window is too small, then the histogram becomes too ragged. Contrariwise, if the window is too large, the shape is too over-smoothed. The choice of this window is an important one and can be done automatically, but it's not the topic of this course.

In order to have a smoother version of the distribution of the sample, a kernel probability density estimate is sometimes preferred. In the particular case of the Gaussian kernel $w(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$, for a bandwidth $h > 0$, the kernel probability density estimator is defined as follows :
$$\widehat{f}_{w,h,X}(t) = \frac{1}{nh} \sum_{i=1}^{n} w\left(\frac{t - X_i}{h}\right).$$
This estimate is a smooth version of the histogram, describing the datasets, not to be confused with the unknown density of $F$. Figure 4 represents 3 kernel probability density estimates of the same data for different bandwidth.

**Exercise 12. Bar plots and histograms**

1. Draw a histogram of the sample of Example 1.
2. Draw in the same graph the density of a uniform $\mathcal{U}(0, 1)$ distribution and the histograms of the two sample of Exercise 9 with a partition of the interval $[0, 1]$ with 2, 5 and 20 bins.
3. Figure 5 represents a bar plot of a sample of discrete data and a histogram of continuous data. Determine characteristics of the distribution associated to these datasets. And propose a family of distributions which is likely to have generated them.



(a)

(b)

FIGURE 5 – Bar plot and histogram

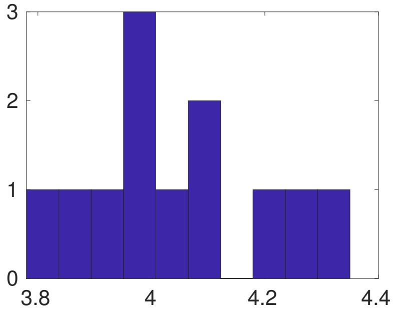


FIGURE 6 – Ex12 : Histogram of Example 1.

**Solution**

1. Fig. 6.
2. Fig. 7 and Fig. 8.
3. (a) : discrete concentrated on $\{0, 1, \ldots, 6\}$, not symmetric, light tailed, e.g. Poisson. (b) : not symmetric, not light tailed, e.g. exponential, $\gamma$ or $\chi^2$ distributions.

### 1.2.4 Q-Q plot

Quantile-quantile (Q-Q) plots are used to compare the distribution of a sample with a probability distribution or with the distribution of another sample. In a Q-Q plot, the empirical quantiles of some data are plotted against the quantile of some distribution or the empirical quantiles of another sample.

**Definition 11.** *In the case of the comparison of a sample $(x_1, \ldots, x_n)$ with some distribution $G$, the* **Q-Q plot** *contains the points $(Q_G(i/(n+1)), x_{(i)})$ for $i = 1, \ldots, n$. If the points are aligned on the line $y = x$, it is likely that the data points come from the distribution $G$. If the points are aligned on another line $y = ax + b$, it is likely that the data points come from a translated and rescaled version of the distribution $G$.*
*To compare two batches of $n$ numbers $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, a Q-Q plot is simply constructed by plotting the points $(x_{(i)}, y_{(i)})$ for $i = 1, \ldots, n$.*

**Exercise 13. Q-Q plots**

1. Compare sample 1 of Exercise 9 with the Uniform distribution $\mathcal{U}(0, 1)$ drawing a Q-Q plot.
2. Same question with the Uniform distribution $\mathcal{U}(2, 3)$.
3. Same question with the Exponential distribution $\mathcal{E}(1)$ .

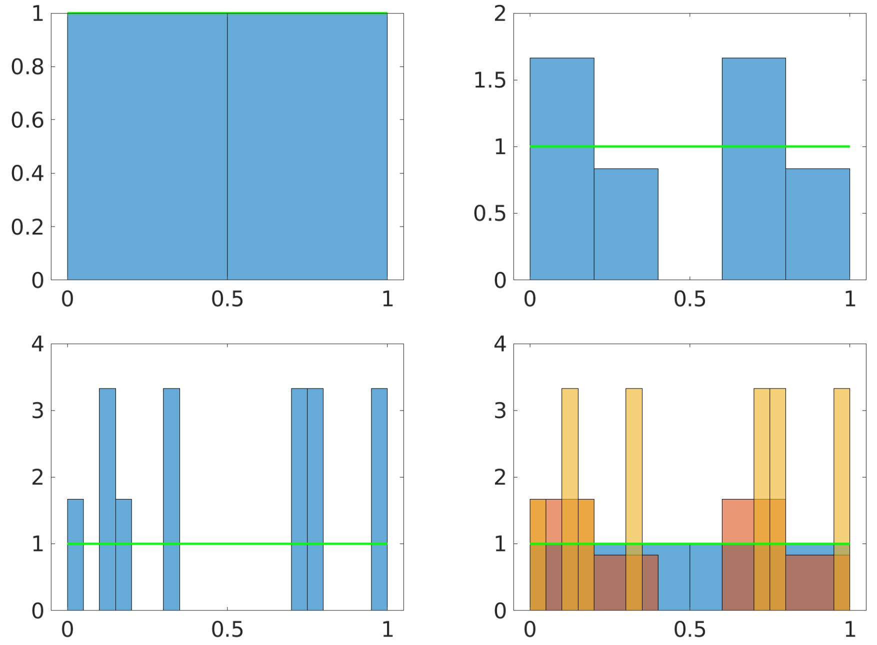


Figure 7 – Ex12 : Histogram of Ex9 :1, with 2, 5, 20 bins and all together (row-continuously).

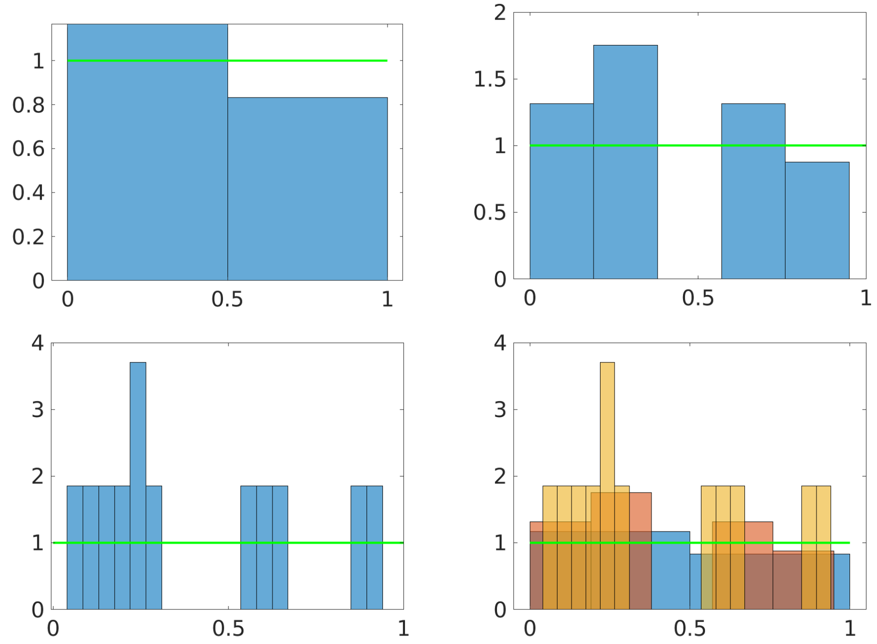


Figure 8 – Ex12 : Histogram of Ex9 :2, with 2, 5, 20 bins and all together (row-continuously).

4. Compare sample 1 and sample 2 of Exercise 9 drawing a Q-Q plot.
5. Figure 9 represents the QQ-plot of four samples with respect to the Normal distribution $\mathcal{N}(0,1)$. Determine characteristics of the distribution associated to these datasets. And propose a family of distributions which is likely to have generated them.

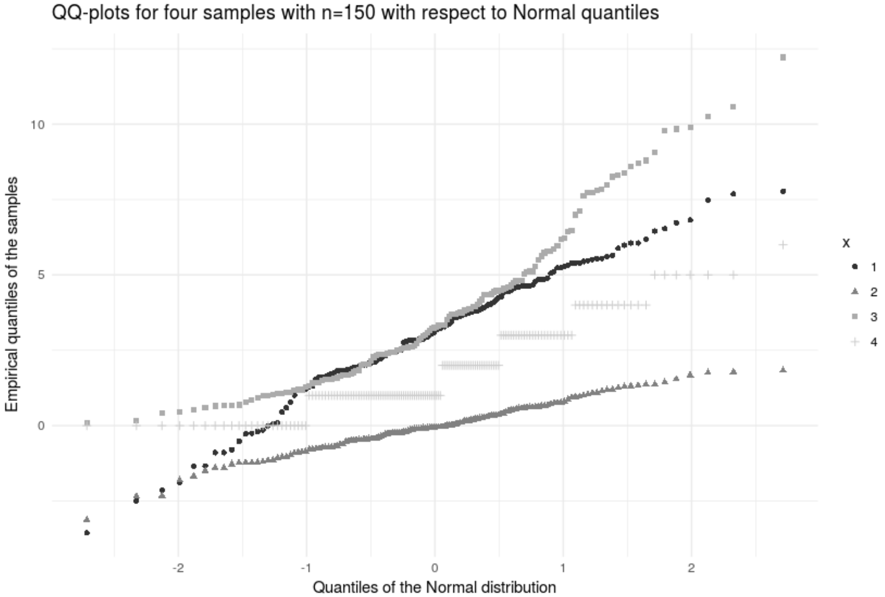


Figure 9 – QQ-plots

## 2 Bivariate quantitative data

In this part, we consider that we have a bivariate quantitative sample $\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}\right) \in (\mathbb{R}^2)^n$. The sample $\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}\right)$ is the realization of the vector $\left(\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}\right)$, where $\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$, for $i \leqslant n$, are i.i.d. from $F$ on $\mathbb{R}^2$. A typical question for this type of data is the relationship among the variables $x$ and $y$.

### 2.1 Graphical representations

It is natural to represent bivariate quantitative sample with **scatter plots**.

**Definition 12.** *A* **scatter plot** *is obtained by plotting the points $(x_i, y_i)$ in the xy plan.*

This representation can give an idea of the relationship between the variables $X_1$ and $Y_1$, see Exercise 15.

### 2.2 Measure of correlations

The empirical covariance and empirical correlation of a bivariate sample are descriptions of the relationship between the variables $x$ and $y$. They are defined as follows.

**Definition 13.** *The* **empirical covariance** *and* **empirical correlation** *of a sample $\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}\right)$ are defined as*

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}_n \bar{y}_n \quad \textit{and} \quad \rho_{xy} = \frac{s_{xy}}{s_x s_y}.$$

*Similarly, the empirical covariance and empirical correlation of the observations* $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ *are*

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}_n \bar{Y}_n \quad \textit{and} \quad \rho_{XY} = \frac{s_{XY}}{s_X s_Y}.$$

They are approximations of the theoretical covariance and correlation between two random variables $X_1$ and $Y_1$, admitting two moments (i.e. $\mathsf{E}(Y_1^2) < \infty$ and $\mathsf{E}(X_1^2) < \infty$), defined as

$$\mathrm{cov}(X_1, Y_1) = \mathsf{E}(X_1 Y_1) - \mathsf{E}(X_1)\mathsf{E}(Y_1) \quad \text{and} \quad \mathrm{cor}(X_1, Y_1) = \frac{\mathrm{cov}(X_1, Y_1)}{\sqrt{Var(X_1) Var(Y_1)}}.$$

**Exercise 14. Covariance and independence** We consider that the two real random variables $X_1$ and $Y_1$ have two moments.

1. Show that if the random variables $X_1$ and $Y_1$ are independent with non zero variances, then $\mathrm{cov}(X_1, Y_1) = \mathrm{cor}(X_1, Y_1) = 0$.
2. Show that $\mathrm{cor}(X_1, Y_1) \in [-1, 1]$.
3. Let $a \neq 0$ and $b$ two real constants. Compute the correlation between $X_1$ and $aX_1 + b$.
4. Assume now that $X_1 \sim \mathcal{U}(-1, 1)$ and for $a \in [0, 1]$

$$Y_a = \begin{cases} X_1 & \text{if } |X_1| > a \\ -X_1 & \text{if } |X_1| \leqslant a. \end{cases}$$

   Are $X_1$ and $Y_a$ independent ? What is the distribution of $Y_a$ ? Can you find a positive $a$ such that $X_1$ and $Y_a$ are uncorrelated ?

**Solution**

1. By PR : Prop. 3.4.2 with $\phi_1 = \phi_2 =$ identity, $\mathrm{cov}(X_1, Y_1) = \mathsf{E}(X_1 Y_1) - \mathsf{E}(X_1)\mathsf{E}(Y_1) = 0$. Hence $\mathrm{cor}(X_1, Y_1) = \frac{\mathrm{cov}(X_1, Y_1)}{\cdots} = 0$ also holds.
2. By noticing that $\mathrm{cov}(X_1, Y_1) = \langle X_1 - \mathsf{E}(X_1), Y_1 - \mathsf{E}(Y_1)\rangle$, $\mathrm{var}(X_1) = \mathsf{E}\left[X_1 - \mathsf{E}(X_1)\right]^2$, $\mathrm{var}(Y_1) = \mathsf{E}\left[Y_1 - \mathsf{E}(Y_1)\right]^2$, the result follow from the

$$\langle c, d\rangle \leq |\langle c, d\rangle| \overset{(*)}{\leq} \|c\|\,\|d\|$$

   CBS inequality holding for the $c := X_1 - \mathsf{E}(X_1), d := Y_1 - \mathsf{E}(Y_1) \in L^2$ variables ; see PR : Theorem 2.5.4.
3. By using PR : Prop. 2.2.5 :(iv) and $\mathrm{cov}(X_1, aX_1 + b) = \mathrm{cov}(X_1, aX_1) = a\,\underbrace{\mathrm{cov}(X_1, X_1)}_{\mathrm{var}(X_1)}$ we get

$$\mathrm{cor}(X_1, aX_1 + b) = \frac{\mathrm{cov}(X_1, aX_1 + b)}{\sqrt{\mathrm{var}(X_1)}\,\underbrace{\sqrt{\mathrm{var}(aX_1 + b)}}_{|a|\sqrt{\mathrm{var}(X_1)}}} = \frac{a}{|a|} = sign(a).$$

4. — (i) Not necessarily, for example in case of
   — $a = 1 : Y_1 = -X_1$, so by the previous bullet $\mathrm{cor}(Y_1, X_1) = -1 \neq 0$.
   — $a = 0 : Y_1 = X_1$ ; hence similarly $\mathrm{cor}(Y_1, X_1) = 1 \neq 0$.
   — More generally, one can assume w.l.o.g. that $a \neq 1$. By the graph of $(Y_a, X_1)$ e.g. taking a $b \in (-1, -a)$ value

$$F_{(Y_a, X_1)}(b, b) = \mathbb{P}\left(\begin{bmatrix} Y_a \\ X_1 \end{bmatrix} \leq \begin{bmatrix} b \\ b \end{bmatrix}\right) = \mathbb{P}(X_1 \leq b) = (b + 1)\frac{1}{2},$$

$$F_{Y_a}(b) = \mathbb{P}(Y_a < b) = \mathbb{P}(X_1 < b) = (b + 1)\frac{1}{2},$$

$$F_{X_1}(b) = \mathbb{P}(X_1 < b) = (b + 1)\frac{1}{2}.$$

Hence

$$(b+1)\frac{1}{2} = F_{(Y_a,X_1)}(b,b) \overset{?}{=} F_{Y_a}(b)F_{X_1}(b) = \left[(b+1)\frac{1}{2}\right]^2 \overset{(*)}{\Leftrightarrow} 1 \overset{?}{=} (b+1)\frac{1}{2} \Leftrightarrow 1 \overset{?}{=} b,$$

by using in (*) that $b \neq -1$. The last equality [and hence $F_{(Y_a,X_1)}(b,b) = F_{Y_a}(b)F_{X_1}(b)$] does not hold since $b < 0$.

— (ii) $Y_a \sim U(-1,1)$. Indeed, by the graph of $(Y_a, X_1)$ for a $z \in (-1,1)$

$$\mathbb{P}(Y_a \leq z) = \begin{cases} \mathbb{P}(X_1 \leq z) = (z+1)\frac{1}{2}, & \text{if } z < -a, \\ \mathbb{P}(X_1 \leq z) = (z+1)\frac{1}{2}, & \text{if } z > a, \\ \mathbb{P}(X_1 \in [-1,-a] \cup [-z,a]) = \underbrace{[(-a+1)+(a+z)]}_{z+1}\frac{1}{2}, & \text{if } z \in [-a,a], \end{cases}$$

which is precisely the cdf of a random variable distributed uniformly on $(-1,1)$. Notice that in the last branch we have $[-z,a]$.

— (iii) Uncorrelatedness means that

$$\mathsf{E}(Y_a X_1) = \mathsf{E}(Y_a)\mathsf{E}(X_1). \tag{1}$$

— r.h.s. of (1) : Since $Y_a$ and $X_1$ are uniformly distributed on $(-1,1)$, we have $\mathsf{E}(Y_a) = \mathsf{E}(X_1) = 0$.

— l.h.s. of (1) : Using that

$$Y_a X_1 = \begin{cases} X_1^2, & \text{if } |X_1| > a, \\ -X_1^2, & \text{if } |X_1| \leq a. \end{cases} =: \phi_a(X_1)$$

by PR : Theorem 2.2.3 one gets

$$G(a) := \mathsf{E}(Y_a X_1) = \mathsf{E}\phi_a(X_1) = \int_{-1}^{-a} z^2(z+1)\frac{1}{2}\mathrm{d}z + \int_a^1 z^2(z+1)\frac{1}{2}\mathrm{d}z + \int_{-a}^{a} -z^2(z+1)\frac{1}{2}\mathrm{d}z.$$

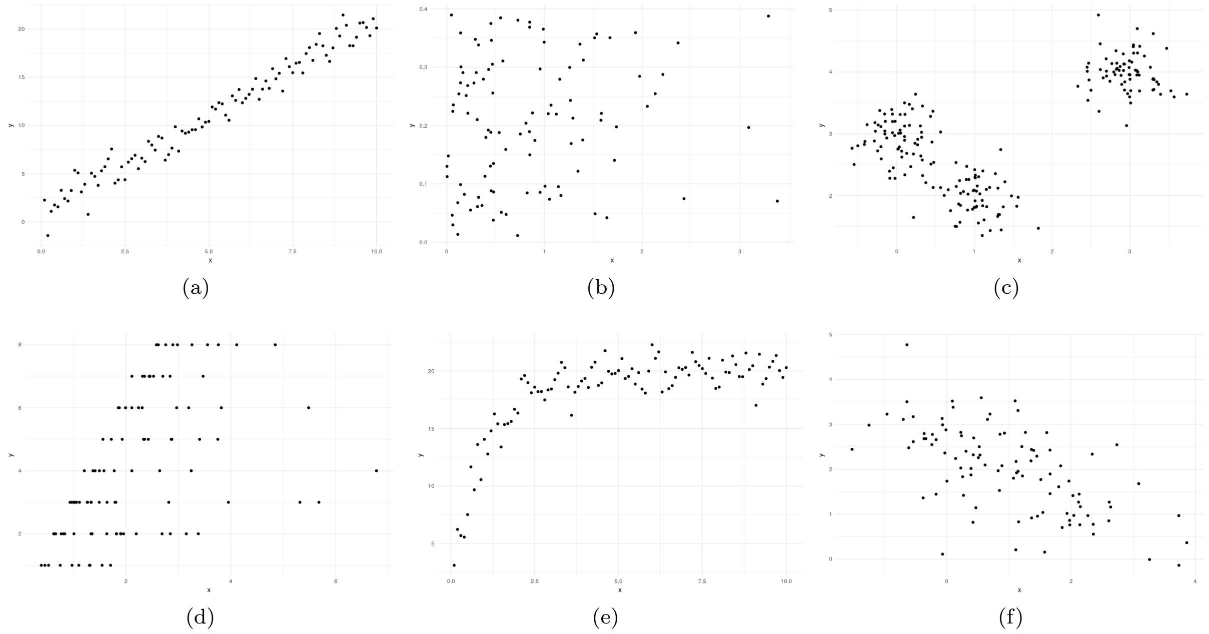
(a) Lazy solution (not constructive) : We saw in (i) that $G(0) = 1$ and $G(1) = -1$. By the Newton-Leibniz formula $G(a)$ is a polynomial function, hence continuous. Thus by the Weierstrass theorem there exists $a \in (0,1)$ such that $G(a) = 0$, which is equivalent to the asked uncorrelatedness.

(b) Not-lazy solution (constructive) :

$$\begin{aligned} G(a) &= \frac{1}{2}\left(\left[\frac{z^4}{4} + \frac{z^3}{3}\right]_{z=-1}^{z=-a} + \left[\frac{z^4}{4} + \frac{z^3}{3}\right]_{z=a}^{z=1} - \left[\frac{z^4}{4} + \frac{z^3}{3}\right]_{z=-a}^{z=a}\right) \\ &= \frac{1}{2}\left[\left(\frac{a^4}{4} - \frac{a^3}{3}\right) - \left(\frac{1}{4} - \frac{1}{3}\right) + \left(\frac{1}{4} + \frac{1}{3}\right) - \left(\frac{a^4}{4} + \frac{a^3}{3}\right) - \left(\frac{a^4}{4} + \frac{a^3}{3}\right) + \left(\frac{a^4}{4} - \frac{a^3}{3}\right)\right] \\ &= \frac{1}{2}\left(\frac{4a^3}{3} - \frac{2}{3}\right) = \frac{2a^3-1}{3} \end{aligned}$$

The asked uncorrelatedness is equivalent to $G(a) = \frac{2a^3-1}{3} = 0 \Leftrightarrow a = \sqrt[3]{\frac{1}{2}}$.

**Exercise 15. Scatter plots** Figures 10 represents the scatter plot of samples $\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_{100} \\ y_{100} \end{pmatrix}\right) \in (\mathbb{R}^2)^n$ which are realizations of the vectors $\left(\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_{100} \\ Y_{100} \end{pmatrix}\right) \in (\mathbb{R}^2)^n$, where $\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$ are i.i.d. from some distributions $F$ on $\mathbb{R}^2$. For each sub-figure, determine if $X_1$ and $Y_1$ are independent and their correlations.



(a) (b) (c)

(d) (e) (f)

FIGURE 10 – Scatter plot of some samples

# 3 Summary

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