**Inferential Statistics** **First semester, 2021**

# Lecture 2: Descriptive statistics and statistical models

*Instructor: Erlis Ruli (ruli@stat.unipd.it), Department of Statistical Sciences*



In Probability Theory, we know $F_X$ and its parameter $\theta$ and we are interested in calculating quantities such as $P(X \in B)$, $E(X)$, $\text{var}(X)$, etc.

In Statistics, we observe data $x_1, x_2, \ldots, x_n$, which we assume they have been generated from $F_X$, at some parameter value $\theta_0$. Hence, we know the family to which $F_X$ belongs to (i.e. $F_X$ could be exponential or normal or gamma, etc.) but $\theta_0$ is unknown. The aim is to guess $\theta_0$.

For instance, Nature gives you the sample: 0.318, 1.765, 0.259, 0.450, 0.730, 0.235, 0.017, 1.010, 1.418, 0.480 and it reveals you that the sample was generated from $\text{Exp}(\lambda_0)$. Assuming Nature never lies, we will see how to produce a guess for $\lambda_0$ with *theoretical guarantees*. By the way, producing a guess for $\lambda_0$ is called *estimation*, since we are trying to estimate/guess the value of a parameter of a distribution from the observed data. We will see more on this in Lecture 4.

The list of values $x_1, \ldots, x_n$ is called the *observed sample*. The easiest situation is when each $x_i$ is generated from the same distribution and independently of each other. This type of sample is called *i.i.d. observed sample*. Hence, in practice, we observe $x_1, \ldots, x_n$ an i.i.d. sample, which is made of $n$ independent realisations of the r.v. $X \sim F_\theta$. Equivalently, we can say that $x_1, \ldots, x_n$ is an i.i.d. sample, where each $x_i$ is a realisation of $X_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$. We switched from $F_X$ to $F_\theta$ to highlight the fact that we know all about $F_X$ except $\theta$, our object of interest.

It's crucial to distinguish $X_1, \ldots, X_n$ from $x_1, \ldots, x_n$. The former is a vector of r.v.'s (or a r.ve.), thus it is a function, $x_i$ is a number. We may call $X_1, \ldots, X_n$ an i.i.d. random sample. The number of (random or observed) samples $n$ is called *sample size*. To further clarify the difference between $X_i$ and $x_i$, think about the difference between a voltmeter and a voltage,

you can use a voltmeter, i.e. $X_i$, to produce a voltage, i.e. a number $x_i$.

## 2.1 Descriptive statistics

Descriptive statistics are numerical and graphical tools used to extract information from a sample. The sample can be either random or observed. Section 2.2 deals with univariate samples, i.e. $X_i$ is a r.v. and $x_i$ is a scalar for all $i$. Section 2.3 deals with bivariate samples.

A descriptive statistic applied to a random sample is just an instance of a transformation of a r.ve. (Lecture 1, §1.1.5.), though such a transformation may not be smooth or bijective.

## 2.2 Univariate samples

*Location* and *spread* are important features of a distribution and they are important also for a sample.

For a r.v. $X \sim F$, typical measures of location are the expectation $\mu$, the median $\xi_{0.5}$, the $p$th quantile $\xi_p$, the mode, etc. For symmetric distributions, the expectation is equal to the median, which is equal to the mode. When the distribution has an elongated right tail, i.e. the distribution is skewed to the right, $\mu > \xi_{0.5}$. Typical measures of dispersion are the variance $\sigma^2$ or the standard deviation $\sigma = \sqrt{\sigma^2}$, the inter-quartile range, the median absolute deviation from the median, etc.. In this section we define some of them and we will study their properties latter.

Let $X_1, \ldots, X_n$ be r.v.'s with common distribution $F$.

### 2.2.1 Moment-based statistics

The *sample average*, defined by

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n},$$

and the *sample variance*

$$S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

are respectively measures of location and spread of a sample. Note that $\overline{X}$ and $S^2$ are r.v.s.

Given $x_1, \ldots, x_n$ an observed sample, we define the *observed sample average* and the *observed sample variance* by, respectively

$$\overline{x} = \tfrac{x_1+\cdots+x_n}{n}, \quad s^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

We define the *sample moment of order k* and the *observed sample moment of order k*, respectively by

$$\overline{X^k} = \tfrac{1}{n}\sum_{i=1}^{n} X_i^k \quad \text{and} \quad \overline{x^k} = \tfrac{1}{n}\sum_{i=1}^{n} x_i^k.$$

### 2.2.2 Order statistics

Let $F$ be continuous with p.d.f. $f(x)$. Let $X_{(1)}$ be the smallest of $X_i$ for all $i$, $X_{(2)}$ be the next smallest $X_i$,..., and $X_{(n)}$ be the largest of all $X_i$. The list $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is called *order statistics* of the random sample; $X_{(1)}$ is called the first order statistic, and so on, $X_{(n)}$ is called the last order statistic.

Thus $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ is the ordered arrangement of $X_i$, with order being increasing. Each $X_{(i)}$ is a r.v. since it is obtained from $X_1, \ldots, X_n$ by suitable ordering function, i.e. $X_{(1)} = \min(X_1 \ldots, X_n)$, ..., $X_n = \max(X_1, \ldots, X_n)$.

**Example 2.1** *Assume that $X_i$ are i.i.d.. We determine the distribution $X_{(1)}$ and $X_{(n)}$. For any $t \in \mathbb{R}$ we have*

$$\begin{aligned} F_{X_{(n)}}(t) &= P(X_{(n)} \le t) = P(\max(X_1 \ldots, X_n) \le t) \\ &= P(\{X_1 \le t\} \cap \{X_2 \le t\} \cap \cdots \cap \{X_n \le t\}) \\ &= \prod_i P(X_i \le t) = [F(t)]^n. \end{aligned}$$

*The density function is*

$$f_{X_{(n)}}(t) = \frac{d}{dt}F_{X_{(n)}}(t) = n[F(t)]^{n-1}f(t).$$

*For the first order statistic we have* $f_{X_{(1)}}(t) = n(1-F(t))^{n-1}f(t)$. *In general, for any* $k$, $1 \leq k \leq n$ *we have*

$$f_{X_{(k)}} = \frac{n!}{(k-1)!(n-k)!}[F(t)]^{k-1}[1-F(t)]^{n-k}f(t).$$

Order statistics are useful for defining measures of location and spread alternative to the those based on moments. For instance, the sample median is defined by

$$Q_2 = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & n \quad \text{odd} \\ \frac{1}{2}\left(X_{(n/2)} + X_{(n/2+1)}\right) & n \quad \text{even.} \end{cases}$$

In general for any $k = [p(n+1)]$ with $p \in (0,1)$, we define the sample quantile by $X_{(k)}$; $[x]$ denotes the greatest integer $\leq x$. As for the median, some sample quantiles have special names, $Q_1 = X_{([0.25(n+1)])}$ and $Q_3 = X_{([0.75(n+1)])}$ are called respectively, the first and third sample quartile. $Q_1, Q_2, Q_3$ are all measures of location.

Order statistics are useful because $X_{(k)}$ approximately targets $\xi_p$, the $p$th quantile of $X$.

Indeed, the probability of observing values of $X$ less than $X_{(k)}$ is $F(X_{(k)})$. On average, such a probability is equal to $k/(n+1)$, i.e.

$$E(F(X_{(k)})) = \int_{-\infty}^{\infty} F(t)f_{X_{(k)}}(t)dt = \frac{k}{n+1} \approx p.$$

In computing the above integral we have made the substitution $F(t) = s$, with $t = F^{-1}(s)$ so that $\frac{dt}{ds} = \frac{1}{f(t)}$ and we have used the properties of the beta function $\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1}dx$.

The same quantities can be defined on an observed sample $x_1, \ldots, x_n$. Here the observed ordered statistics are $x_{(1)} \leq x_{(2)} \leq \ldots x_{(n)}$, where $x_{(i)}$ is called the $i$th observed ordered statistics; in particular, $x_{(1)}$ is the observed minimum, $x_{(n)}$ is the observed maximum, and so on. For instance, if the observed sample is : 1.1, 0.5, 0.4, 3, 2.2, the observed ordered statistics are: 0.4, 0.5, 1.1, 2.2, 3.0. In this case, $x_1 = 1.1 = x_{(3)}$, $x_2 = 0.5 = x_{(2)}$ and so on.

The *observed lower quartile* is defined as $q_1 = x_{([0.25(n+1)])}$, the *observed upper quartile* is

defined by $q_3 = x_{([0.75(n+1)])}$. The observed sample median is defined by

$$\mathrm{q}_2 = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \quad \text{odd}, \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & n \quad \text{even}. \end{cases}$$

The *sample inter-quartile range* (IQR) is defined by $iqr = q_3 - q_1$.

The sample median is often preferred to $\overline{x}$ as a measure of location when there are outlying observations. In this case, also the $iqr$ is preferred to the sample variance (or sample standard deviation).

**Remark 2.1** *We learned in L0 that the variance $\sigma^2$ it's a parameter and measures the dispersion (or spread) of the distribution of $X$. $S^2$ also deals with dispersion, but of a random sample $X_1, \ldots, X_n$ of $X$. Thus, $\sigma^2$ is a feature of $X$ (sometimes called population) and $S^2$ is a feature of its sample. As we will see in the incoming Lectures, $\sigma^2$ and $S^2$ are linked, i.e. $S^2 \xrightarrow{P} \sigma^2$ as $n \to \infty$; similar remark applies to the other statistics.*

### 2.2.3 Histograms

For a continuous r.v. $X$, we can get an idea about the shape of its distribution by comparing measures of location and dispersion, but graphical representation of the data could be more useful. A simple graphical technique is the *histogram*. Here we consider only its observed sample version.

Consider a partition $a_0 < a_1 < a_2 < \cdots < a_m$ that covers the range of data $x_1, \ldots, x_n$. The histogram is the function that, on each interval $(a_{j-1}, a_j]$, takes on the value equal to the number of sample points $x_i$ belonging to that interval divided by $n$ times the length of the interval, $j = 1, \ldots, m$.

The histogram is a piecewise constant function, defined by

$$h_n(x) = \frac{1}{n(a_j - a_{j-1})} \sum_{i=1}^{n} 1_{(a_{j-1}, a_j]}(x_i), \quad \text{for all } x \in (a_{j-1}, a_j],$$

where the indicator function $1_{(a_{j-1}, a_j]}(x_i)$ equals 1 if $x_i \in (a_{j-1}, a_j]$ and 0 otherwise. It can

provide a good summary provided the partition $a_0 < a_1 < a_2 < \cdots < a_m$ has been chosen well and that the sample size $n$ is not too small.

### 2.2.4 Empirical distribution function

Let $X_1, \ldots, X_n$ be an i.i.d. random sample from $X \sim F$. Then the *empirical distribution function* (EDF) is defined by

$$F_n(x) = n^{-1} \sum_{i=1}^{n} \mathcal{I}_{X_i}(x), \quad x \in \mathbb{R},$$

where $\mathcal{I}_{X_i}(x)$ is a Bernoulli r.v. which assumes value 1 in the event $X_i \leq x$ and 0 otherwise. $F_n$ is a random step function, that is for each fixed $x$, $F_n$ is a r.v. since it is a function of Bernoulli r.v.s.

If the observed sample $x_1, \ldots, x_n$ is used, then the observed empirical distribution function is defined by

$$\widehat{F}_n(x) = n^{-1} \sum_{i=1}^{n} 1_{x_i}(x), \quad x \in \mathbb{R},$$

where $1_{x_i}(x)$ denotes the indicator function which assumes value 1 if $x_i \leq x$ and 0 otherwise.

While the histogram aims at approximating the p.d.f. of $X$, the empirical distribution function approximates $F$. In practice, the EDF should be preferred to the histogram because it:

(1) has better theoretical guarantees (Glivenko-Cantelli Theorem),

(2) does not require a partition of the sample.

### 2.2.5 Boxplots

A *boxplot* is a graphical summary of the observed sample that provides indications about the:

- location

- dispersion
- symmetry of the distribution
- presence of outliers.

The bottom of the "box" is drawn at $q_1$, and the top at $q_2$ the data. The lower (respectively, upper) quartile of the data is the value $x$ for which one fourth of the data points are less (respectively, greater) than $x$. The width of the box is arbitrary. The box has a horizontal line at $q_2$ of the data (The median is the middle value in a sorted row of data.). At the top and bottom of the box, whiskers are drawn. The whisker at the top links the box to the greatest observation that lies within $q_3 + 1.5 \times iqr$. The whisker at the bottom is at the lowest observation within $q_1 - 1.5 \times iqr$. Observations that lie beyond the whiskers are indicated separately, for example by a star, a small circle, or a dash; these are considered outlier observations.

Figure 2.1 shows three boxplots of data simulated from three distributions. The samples from the exponential and $t$-Student distributions have outliers, shown by the small circles beyond the whiskers. The boxplot in the middle shows that the data generated from the standard normal distribution are quite symmetric with respect to the median and do not contain outliers.

### 2.2.6 QQ-plots

The *QQ-plot* or *quantile-quantile plot* is used for assessing if a given sample $x_1, \ldots, x_n$ is compatible with a given distribution $F$. To motivate it, suppose the r.v. $X$ has d.f. $F\left(\frac{x-\mu}{\sigma}\right)$, where $F(x)$ is a known function but $\mu$ and $\sigma > 0$ may not be.[1] Let $Z = (X - \mu)/\sigma$, then

$$F_Z(z) = P(Z \leq z) = P((X - \mu)/\sigma \leq z) = P(X \leq \sigma z + \mu) = F(\sigma z + \mu).$$

Because $F(z)$ is known, we can compute its quantiles of any level $p$, call them $\xi_{Z,p}$. But then

$$p = P(X \leq \xi_{X,p}) = P\left(Z \leq \tfrac{\xi_{X,p}-\mu}{\sigma}\right) = P\left(Z \leq \xi_{Z,p}\right),$$

[1] $\mu$ and $\sigma$ are parameters, which in this particular case are called location and scale parameters, respectively.

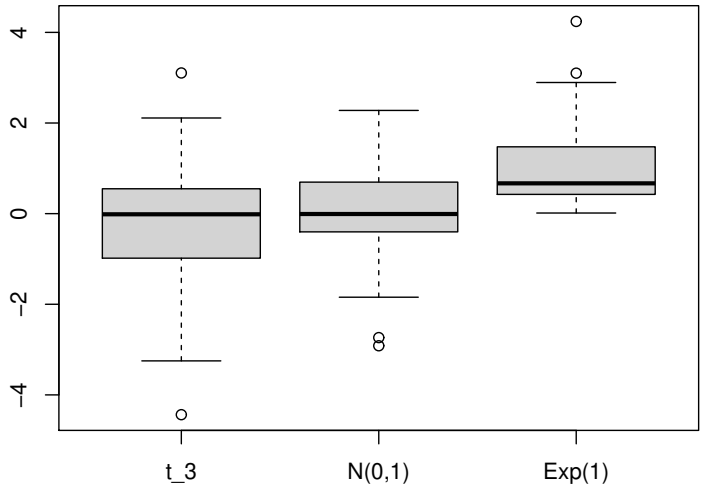


Figure 2.1: Boxplots of samples of size 50 from the $t$-Student distribution with 3 degrees of freedom (left), the standard normal distribution (middle), and the exponential distribution with unit rate (right).

and $\xi_{X,p} = \sigma\xi_{Z,p} + \mu$. Thus we can get the quantiles of $X$ by suitably shifting and scaling those of $Z$. However, in practice $\mu$ and $\sigma$ are unknown but we can still estimate $\xi_{X,p}$ through the order statistics $X_{(k)}$. Indeed, for $k = 1, \ldots, n$ let $p_k = k/(n+1)$, then $X_{(k)} \approx \xi_{X,p}$.

With an observed sample $x_1, \ldots, x_n$, the $k$th observed order statistic $x_{(k)}$ is a realisation of $X_{(k)}$. The QQ-plot then consists in drawing the pairs

$$\left\{\left(x_{(j)}, \mu + \sigma F^{-1}\left(\tfrac{j}{n+1}\right)\right), j = 1, 2, \ldots, n\right\}.$$

If the pairs lie along the $y = x$ line, then this indicates that observations $x_1, \ldots, x_n$ are compatible with the distribution $F$. Typically $F$ is chosen to be the normal distribution.

Deviations from the half-plane line could be of different kinds. The points could be rotated, shifted, or could have a curved shape. Shifts and rotations with respect to the $y = x$ line indicate differences in terms of location and scale, respectively. Whereas $U$-shaped or $S$-shaped QQ-plots indicate differences in terms of asymmetry or in terms of the length of the tails, respectively.

Figure 2.2 shows examples of QQ-plots for six different observed samples $x_1, \ldots, x_n$, all of size $n = 100$, versus the standard normal distribution, i.e. vs $X \sim \mathrm{N}(0,1)$, $\mu = 0, \sigma = 1$. In panel $(a)$ we may conclude that the observed sample is compatible with the standard normal distribution. Thus we can consider this observed sample as if it was generated from the $N(0,1)$ distribution. In $(b)$ we note that the tails of the observed sample are much longer than those of the theoretical distribution, we can see this by comparing the range of the values in the two axes. In panel $(c)$ we have differences in terms of symmetry (besides other issues, see below), in the sense that the sample quantiles are from some asymmetric distribution with the right tail being much longer than the left one. In $(d)$ we have a difference in location, i.e. the location of the data is higher than that of the $\mathrm{N}(0,1)$ distribution. In $(e)$ we have a difference in terms of scale, i.e. the observed sample is more dispersed than the $\mathrm{N}(0,1)$ distribution and in $(f)$ we have $(d)$ and $(e)$.

## 2.3 Bivariate samples

Observed data $x_i$ may be a vector. For instance, a meteorological station located in a city, can provide real-time measurements of: temperature, pressure, humidity, rain fall, wind speed, solar irradiation, $PM_{2.5}$, $PM_{10}$ CO, $CO_2$, etc.. Thus, on a given day $i$ we observe $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, where $x_{i1}$ may be the temperature at day $i$, $x_{2i}$ the pressure and so on.

In the case with $p = 2$, let $(x_1, y_1), \ldots, (x_n, y_n)$ be the observed sample of size $n$ obtained as realisations of the r.ve.'s $(X_i, Y_i)$, $i = 1, \ldots, n$. For the random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, we define the *sample covariance* by

$$S_{XY} = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}),$$

and the *sample correlation coefficient* is defined by

$$R_{XY} = \frac{S_{XY}}{\sqrt{S_X^2}\sqrt{S_Y^2}}.$$

With the observed pairs $(x_1, y_1), \ldots, (x_n, y_n)$ we get the *observed sample covariance* and the *observed sample correlation coefficient*, given respectively by

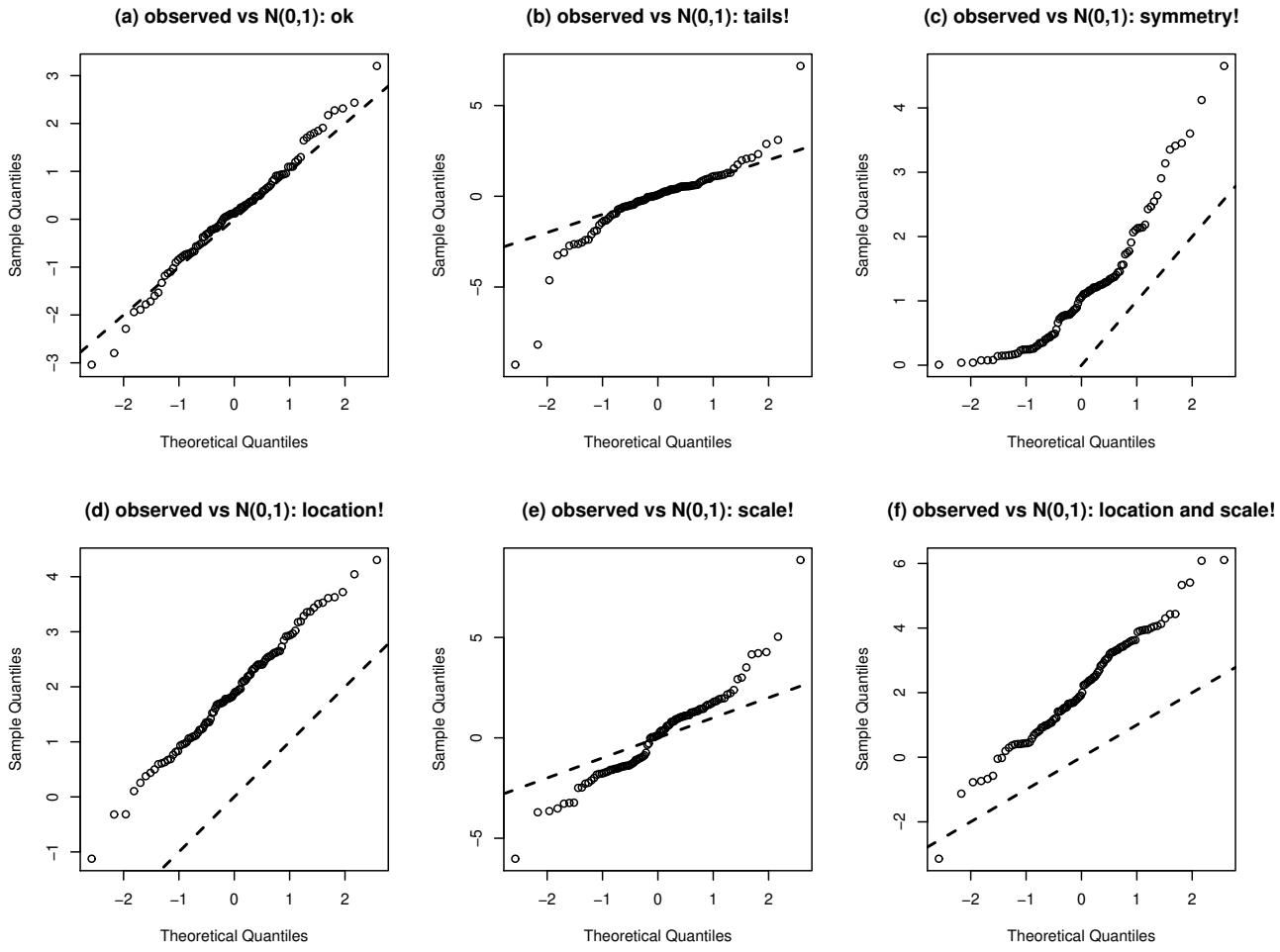


Figure 2.2: QQ-plots of six different observed samples each of size 100 vs the standard normal distribution.

$$s_{xy} = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}), \quad \text{and} \quad r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2}\sqrt{s_y^2}}.$$

Again note that $S_{XY}, R_{XY}$ are random variables wheres $s_{xy}, r_{xy}$ are numbers. Both $s_{xy}$ and $r_{xy}$ as well as their random versions are measures of linear association between the two variables involved. $s_{xy}$ ranges over the set of real numbers whereas $r_{xy} \in [-1, 1]$. The higher $r_{xy}$ the higher is the degree of linear association between two variables. A correlation coefficient equal to zero implies that there is no linear association between two variables, though the variables could be related in some non linear fashion as in Figure 2.3. Here the pairs of data are shown by means of a *scatter plot*. Only the plot on the left shows a strong

(linear) correlation. The plot on the right shows a clear quadratic relation between $x_i$ and $z_i$ (here $z_i = y_i^2$) but the correlation coefficient is essentially zero.

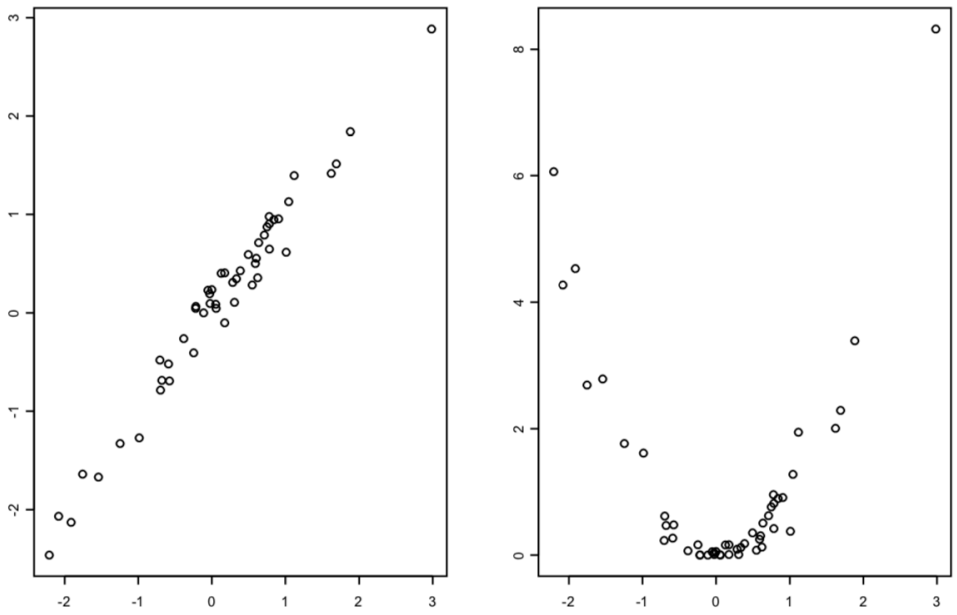


Figure 2.3: Examples of scatter plots. On the left it is shown $(x_i, y_i)$ with $r_{xy} = 0.95$. On the right it is shown $(x_i, z_i)$ for which $r_{xz} = -0.05$.

## 2.4 Statistical models

In this section we introduce the concept of parametric statistical model. The word "parametric" is due to the fact that the statistical model is based on a distribution (either d.f. or p.d.f.) which depends on a finite number of parameters. The qualification parametric is essential in general because as we will see near the end of this course, there are other statistical models outside this framework.

A statistical model can be defined as a probabilistic representation of a physical system with the aim of modelling its current behaviour and predicting its future state. Typically, we have $x_1, \ldots, x_n$ measurements of some characteristics of the physical system and based on these measurements we might be interested in computing a measure of location, dispersion, etc. One approach for doing this is to assume that each observation is a realisation of a r.v. belonging to a given family with parameters to be "learned" from the observations. The aim

is then to learn or estimate the parameter of this model. Here is an example.

**Example 2.2** *Every washing machine (WM) sold in the UE market must be accompanied by technical documentation which describes, among other things, the energy consumed during a typical washing cycle. In order to measure the consumed energy, the WM is tested in a laboratory, say, n times under the same conditions. The resulting values of energy consumptions are* $x_1, \ldots, x_n$*. It is of interest to know both a measure of location and spread of the population* $X$*. A reasonable approach to this problem is to assume that each observation* $x_i$ *is a realisation of a r.v. with distribution* $\mathrm{N}(\mu, \sigma^2)$*, where* $\mu$ *and* $\sigma^2$ *are to be found. The normal assumption for measurements such as energy is very reasonable because the consumed energy of a WM is the sum of the energy consumed by its components (i.e. electric circuits, motor, heater, etc.). Thus by the CLT, the sum will be approximately normally distributed.*

We focus on statistical models with random i.i.d random samples or on statistical models based only on the assumption of independence of the samples. It is worth stressing that the assumption of data being realisations of identically and independently distributed r.v. or of independent but not identically distributed r.v. is not mathematically or statistically verifiable. This assumption is instead is founded on the way the data are collected. For instance, in Example 2.2 it is reasonable to assume that the data are i.i.d. since a washing cycle has no consequences on the next. On the other hand, in a particularly cold winter day, the water temperature in the laboratory could be lower than the temperature in summer days. Thus it seems reasonable to assume that the amount of energy consumed by the WM depends on the environment temperature. In the latter case the measurements may be independent but not identically distributed.

### 2.4.1 Identically and independently distributed r.v.

Let $X_1, \ldots, X_n$ be i.i.d. random variables with $X_i \sim F_\theta$, or more compactly $X_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$, where the distribution $F_\theta$, is indexed by the unknown parameter $\theta$. Then, since $Y_i$ are i.i.d., their joint distribution function is

teta could be a vector too

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} F_\theta(x_i)$$

and the joint probability density function is

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

A *statistical model* is a family of joint d.f. or joint p.d.f. for the r.v. $X_1,\ldots,X_n$ indexed by the parameter $\theta$, that is the set

$$\left\{F_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\right\}$$

or

$$\left\{f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\right\}.$$

The two formulations are essentially equivalent, but we will work with the latter.

**Example 2.3** *Let $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu,\sigma^2)$, $i = 1,\ldots,n$ where $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_{>0}$ are unknown parameters. The joint p.d.f. of these r.v.'s is*

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}, \end{aligned}$$

*and the statistical model*

$$\left\{\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} : \mu \in \mathbb{R}, \quad \sigma \in \mathbb{R}_{>0}\right\}$$

*is called the normal model, where the unknown parameter is $\theta = (\mu,\sigma^2)$.*

**Example 2.4** *Let $X_i \overset{\text{iid}}{\sim} \mathrm{Poi}(\lambda)$, $i = 1,\ldots,n$ where $\lambda \in \mathbb{R}_{>0}$ is the unknown parameter. The joint p.d.f. of these r.v.'s is*

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

*and the statistical model*

$$\left\{ \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} : \lambda \in \mathbb{R}_{>0} \right\}$$

*is called the Poisson model, where the unknown parameter is $\lambda$.*

**Example 2.5** *Let $X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, $i = 1,\ldots,n$ where $\theta \in (0,1)$ is the unknown parameter. The joint p.d.f. of these r.v.'s is*

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\lambda) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

*and the statistical model*

$$\left\{ \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} : \theta \in (0,1) \right\}$$

*is called the Bernoulli model, where the unknown parameter is $\theta$.*

### 2.4.2 Independently (but not Identically) distributed r.v.

Now suppose that the r.v.'s $X_1,\ldots,X_n$ are independently distributed, but we leave to each r.v. the freedom of having its own parameter, i.e. we assume that $X_i \sim F_{\theta_i}$, where $\theta_i$ is an unknown parameter, $i = 1,\ldots,n$. Unfortunately this formulation is of no practical use because the number of parameters grows with $n$ and we have not enough information for learning the parameters. Some constraints must be placed. We show how this is done by means of four examples.

**Example 2.6** *A manufacturer (P1) of motors for washing machines (WM) claims that his next generation motors (called NGM1) are energetically more efficient, while achieving the same speed as the previous sold, i.e. old, version motors (OM). Which one should we buy?*

*In order to verify this claim, two WM are taken, one is equipped with NGM1 and the other is equipped with OM. Suppose that the WM with OM is tested n times, whereas the WM with NGM1 is tested m times. Let $X_1, \ldots, X_n$ be the r.v.'s denoting the energy consumption of the WM with OM and let $Y_1, \ldots, Y_m$ be the r.v.'s denoting the energy consumption of the WM with NGM1. Measurements are reasonably independent from one other. Furthermore, we expect the WM with OM to possibly behave differently from the WM with NGM1. Thus the distribution of $X_i$ could be "different" from that of $Y_j$. By different we mean that they could have different parameter values but both d.f.'s must belong to the same family of distributions. A reasonable assumption is*

$$X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu_x, \sigma_x^2), i = 1, \ldots, n$$

*and*

$$Y_j \overset{\text{iid}}{\sim} \mathrm{N}(\mu_y, \sigma_y^2), j = 1, \ldots, m.$$

*Under these assumptions, the statistical model is*

$$\left\{ \frac{1}{(2\pi\sigma_x^2)^{n/2}} e^{-\frac{1}{2\sigma_x^2}\sum_{i=1}^n (x_i-\mu_x)^2} \frac{1}{(2\pi\sigma_y^2)^{m/2}} e^{-\frac{1}{2\sigma_y^2}\sum_{j=1}^m (y_j-\mu_y)^2} : (\mu_x, \mu_y) \in \mathbb{R}^2, \quad (\sigma_x^2, \sigma_y^2) \in \mathbb{R}^2_{>0} \right\},$$

*where the unknown parameters which are to be learnt from the data are $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$. In this model we impose the r.v.'s within the same type of motor to behave identically, but we leave the freedom for the d.f. of $X_i$ to have different parameters from the d.f. of $Y_j$, for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$.*

*In these type of problems we might be interested to know if $\mu_x = \mu_y$ or if $\sigma_x^2 = \sigma_y^2$. These are hypothesis testing (or confidence interval) problems for two samples; we will see them in detail in the incoming lectures.*

**Example 2.7** *Another manufacturer (P2) of motors for washing machines (WM) also claims that his next generation motors (called NGM2) are energetically more efficient, while achieving the same speed as previous sold, i.e. old, version motors (OM). P2 also claims that NGM2 are more efficient than NGM1. Which one should we buy?*

analisis of variance are the problems callde when we had different

*In order to verify this claim, three WM are taken, one is equipped with the OM, one with NGM1 and one with NGM2. Suppose that the WM with OM is tested $n$ times, the WM with NGM1 is tested $m$ times and the WM with NGM2 is tested $q$ times. Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ be as in Example 2.6 and let $Z_1, \dots, Z_q$ be the r.v.'s denoting the energy consumption of the WM with NGM2. Again, measurements are reasonably independent from one other. Furthermore, we expect the WMs with OM, NGM1 and NGM2 to behave differently from each other. In addition to those of Example 2.6, it is also reasonable to assume that $Z_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu_z, \sigma_z^2), i = 1, \dots, n$. Again, the assumed distribution depends on the type of motor the WM is equipped with.*

*Under these assumptions, the statistical model is*

$$\left\{ \frac{\exp\left[-\frac{1}{2}\left(\frac{\sum_{i=1}^{n}(x_i-\mu_x)^2}{\sigma_x^2}+\frac{\sum_{j=1}^{m}(y_j-\mu_y)^2}{\sigma_y^2}+\frac{\sum_{k=1}^{q}(z_k-\mu_k)^2}{\sigma_z^2}\right)\right]}{(2\pi)^{(n+m+q)/2}(\sigma_x^2)^{n/2}(\sigma_y^2)^{m/2}(\sigma_z^2)^{q/2}} : (\mu_x, \mu_y, \mu_z) \in \mathbb{R}^3, \quad (\sigma_x^2, \sigma_y^2, \sigma_z^2) \in \mathbb{R}^3_{>0} \right\}$$

*where the unknown parameters which are to be learned from the data are $(\mu_x, \mu_y, \mu_z, \sigma_x^2, \sigma_y^2, \sigma_z^2)$. In these type of problems we might be interested to know if $\mu_x = \mu_y = \mu_z$ while assuming that $\sigma_x^2 = \sigma_y^2 = \sigma_z^2$ (called homoscedasticity assumption). This is a hypothesis testing problem called analysis of variance (ANOVA).*

The next example extends Example 2.7 to a situation in which the r.v.'s vary continuously with some other fixed quantity.

**Example 2.8** *Suppose we are measuring some quantity which result is due in part to a variable under our control and in part due to randomness. For instances, suppose that you are studying a device for removing arsenic (which has negative health effects on human beings) from drinkable water but you know that the effectiveness of the removal depends on the pH of water; so it is of interest to assess how arsenic removal changes with water pH.*

*Let $Y_1 \dots, Y_n$ be independent r.v.s with $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$. All parameter are unknown. Furthermore, let $\mu_i = \alpha + \beta t_i$ where $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ are unknown parameters and $t_i$ is a non-random variable. For instance, in the problem of arsenic removal from water, $t_i$ could be the pH of the water at the $i$th water sample. Using the properties of the normal distribution (assuming again that $t_i$ are non stochastic) we could also write*

linear function taken from the experiment in the lab. When the arsenic grow PH did the same

$$Y_i - \alpha - \beta t_i \sim \mathrm{N}(0, \sigma^2)$$

*or in a more commonly used form*

mu-i

$$Y_i = \alpha + \beta t_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2), \quad i = 1, \ldots, n.$$

*This model thus tells that the measurements $Y_i$ are determined by a non-random part $\alpha + \beta t_i$ due to some environmental conditions or other factors over which we have control and by a random part, some times called error or noise; sometimes the model is also called a signal plus noise model.*

*The statistical model is*

$$\left\{ \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta t_i)^2 \right] : (\alpha, \beta) \in \mathbb{R}^2, \sigma^2 \in \mathbb{R}_{>0} \right\}$$

*Once we have an observed sample $y_1, \ldots, y_n$, also called called response variable, and the associated predictor values $t_1, \ldots, t_n$, the aim is to estimate the unknown parameters $(\alpha, \beta, \sigma^2)$.*

*This problem is known as (simple) linear regression and word "linear" is due to the linear equation $\alpha + \beta t_i$. The word "simple" it is because it deals with a single predictor variable $t_i$. A regression model with more than one predictor variables is called* multiple regression model.

In a regression problem the response variable need not be continuous. Here is an example of a regression problem involving Binomial r.v.'s.

**Example 2.9** *A polar station located in a remote area of the northern Polar region is powered by electricity generated by an outdoor generator. To deal with freezing outdoor temperatures, the generator has a special electric system made of 7 switches which run in parallel. The generator stops working if 5 or more switches breakdown. It is predicted that the next night will be exceptionally freezing, with predicted temperature equal to $-40°C$. Given this temperature, researchers living the in the polar station would like to know what is the probability that the generator will stop working.*

*Let $t_1, \ldots, t_n$ be temperature levels registered in $n$ occasions in the past in the same location of the polar station and let $Y_1 \ldots, Y_n$ be r.v.s with $Y_i \sim \mathrm{Bin}(7, \theta_i)$ which represent the number of failed switches when there are 7 switches overall. Thus we assume that the $n$ r.v.s have*

*different success probability $\theta_i$ and index equal to 7. Furthermore, it is reasonable to assume that the behaviour of the switches will depend on the temperature, thus we assume that the success probability $\theta_i$ is a function of temperature $t_i$, by $\theta_i = \frac{e^{\alpha+\beta t_i}}{1+e^{\alpha+\beta t_i}}$ where $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ are unknown parameters. Thus in a more compact form the model we built is*

$$\begin{aligned} Y_i &\sim \text{Bin}(7, \theta_i), \quad \text{with } Y_i \text{ independent from } Y_j, \text{ for all } i \neq j = 1, \ldots, n, \\ \text{logit}(\theta_i) &= \alpha + \beta t_i, \quad i = 1, \ldots, n, \end{aligned}$$

*where $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$ is the so-called logit function. Since $Y_i$'s are independent, the joint distribution of $Y_1, \ldots, Y_n$ is given by the product of their p.d.f.'s.*

*The statistical model is then*

$$\left\{ \prod_{i=1}^{n} \binom{7}{y_i} \left(\frac{e^{\alpha+bt_i}}{1+e^{\alpha+\beta t_i}}\right)^{y_i} \left(\frac{1}{1+e^{\alpha+\beta t_i}}\right)^{7-y_i} : (\alpha, \beta) \in \mathbb{R}^2 \right\}$$

*Given an observed sample $y_1, \ldots, y_n$, i.e. the response variable, and the associated temperature values $t_1, \ldots, t_n$, we can estimate the unknown parameters $\alpha, \beta$. This gives us also an estimate of the probability distribution for the behaviour of the generator under $-42^\circ C$, which is given by*

$$\text{Bin}\left(7, \frac{e^{\alpha+\beta\cdot 42}}{1+e^{\alpha+\beta\cdot 42}}\right).$$

*This model is known as the logistic regression model.*

In general, building or designing a good statistical model for a problem at hand may be a difficult task and it requires some patience and experience. Most importantly, it requires deep knowledge about statistical models. The latter is typically acquired through statistical modelling courses and is outside the scope of this course. With the the tools of statistical inference we will see how to estimate these unknown parameters from observed data and how to judge their quality.

# References

[HMC20] Hogg, R. V., McKeen, J. W. and Craig, A. T. (2018) *Introduction to Mathematical Statistics* (8th edition, global ed.), Pearson Education, Chapp. 3-4.