8
Distributions of Random Variables

8.1 Random variables

A *random variable* (rv) is a function that assumes its values with a given probability P. When one of the numbers that are observed (from instance from an experiment), those numbers are referred as *random variables*. In this notes the notation for a random variable is X. The set of all possible values of X is called *range* of X. Random variables can be *discrete* or *continuous*. A *discrete random variable* (*drv*) comes from a countable set of discrete values, while a *continuous random variable* (*crv*) can take any number in a continuous sample space.

8.1.1 Discrete random variables

In many application in data analysis, the outcome of the experiment is numerical. A **random variable** (r.v.) often denoted by capital letters like $X, YandZ$ is a numerical value obtained from the experiment. Since the experiment can produce a variety of outcomes usually we refer to single elements as $x_1, x_2, x_3, ..., x_n$ which are outcomes of the general random variable X.

X is a discrete random variable if the range of X is a countable set:

$$S_X = [x_1, x_2, ..., x_n] \tag{8.1}$$

8.1.2 Continuos random variables

A continuos random variable is a random variable defined by a continuous set of numbers, referred as an *interval*. The interval contains all of the real numbers between two limits. For instance

$$[x_1, x_2] = (x | x_1 \leq x \leq x_2) \tag{8.2}$$

is a closed interval defined by all the real numbers between x_1 and x_2 including both x_1 and x_2.

There are many variables measured through experiments that lead to a *crv*, such as the arrival time of a particle, the voltage across a resistor, the photon reaching a solar radiation sensor.

8.2 Distributions

As introduced in the previos chapter, a probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes of

a random phenomena. The sample space, often denoted by Ω is the set of all possible outcomes of a random phenomenon being observed; it may be any set of real numbers, a set of vectors, a set of arbitrary non-numerical values. Many distributions have been derived describing different phenomena. For instance for some processes or phenomena, experiments provided information to apply a specific distribution. There are many distributions, the most common ones are: *Uniform, Bernoulli, Binomial, Geometric, Hypergeometric, Exponential, Poisson, Weibull, Normal, Log-Normal, Chi−Squared, Student's t, Gamma* and *Beta*. Clearly, some of them are strictly related both in terms of mathematical form and applications such as the Normal, Log-Normal, Chi−Squared and Student's t. When studying a process and analyzing data it is important to observe the data and identify the most suitable distribution to be applied to derive probabilities.

8.3 Uniform distribution

The uniform distribution is a distribution where equal probability is assigned to the random variable, within a given interval. It is applied when there are no reason to think that an occurence, an event would have higher probability to occur with respect to another event for the same process. It is often used in random sampling, when numbers should have the same probability, it is employed in finance and economics. In physics the emission of radioactive particles is often described with uniform distributions. Another example is the presence of a winning raffle ticket in a number N of tickets sold to people. At the beginning of the raffle the probability is the same of each ticket to be the winning one.

The uniform distribution can be continuous and discrete. The continuous uniform distribution the PDF is constant over the possible values of x.

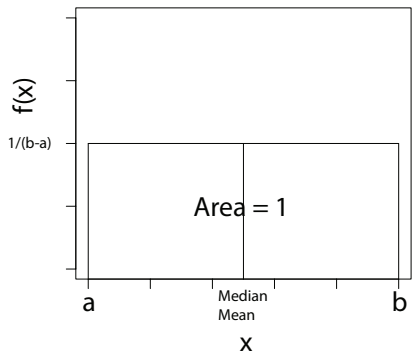


Fig. 8.1 Uniform distribution

The minimum value and maximum value that x can take are called a and b respectively (Fig.8.3). Any intervals in the interval $[a, b]$ or (a, b) are equally likely to occur.

Since the geometry is a rectangle, and the Area is equal to 1, it leads to:

$$\text{Area} = (b-a)f(x) = 1 \tag{8.3}$$

and

$$f(x) = \frac{1}{b-a} \tag{8.4}$$

Formally:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{elsewhere} \end{cases}$$

The median of the distribution is the value that splits the distribution in two equal parts and, since it is a symmetric distribution, the mean and median are equal and it is the midpoint between a and b. So the median is:

$$\text{median} = \frac{a+b}{2} \tag{8.5}$$

and so is the mean

$$\mu = \frac{a+b}{2} \tag{8.6}$$

The variance is

$$\sigma^2 = \frac{1}{12}(b-a)^2 \tag{8.7}$$

Usually for continuous distribution, finding the area under the curve requires integration. They can be analytical or numerical integrations, depending upon the mathematical form of the distribution, that may have an analytical integral or not.

For the uniform distribution the areas are simply rectangles. Suppose that the distribution has a value of $a = 20$ and $b= 25$, $f(x)$ is therefore $f(x) = \frac{1}{25-20} = 0.2$ for values of $20 \le x \le 25$. The area is therefore:

$$\text{Area} = \frac{1}{5}(25-20) = 1 \tag{8.8}$$

What is the probability that $P(x > 23)$. Probabilities are areas under the curve, therefore the area to the right of 23 must be found:

$$\text{P}(x > 23) = \text{base} \times \text{height} = (25-23)\frac{1}{5} = 0.4 \tag{8.9}$$

Using R it is easy to generate and analyse the uniform distribution. In the example below, where x values are generated within the interval 15 to 30. A random generation of uniform numbers is performed. Then the PDF and CDF of the distribution are computed and plotted.

```
x <- seq(15, 30, by = 0.01)
xx <- runif(x)
plot(xx)
dx <- dunif(x, min = 20, max = 25, log = FALSE)
px <- punif(x, min = 20, max = 25, lower.tail = TRUE, log.p = FALSE)
plot(x,dx,type="l",ylab=("f(x)"))
plot(x,px,type="l",ylab=("CDF"))
hist(xx,prob=T)
```

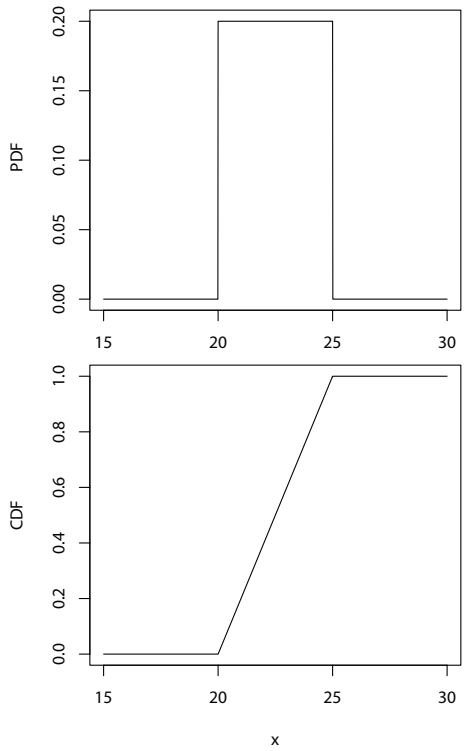


Fig. 8.2 PDF and CDF for the uniform distribution generated with R

8.4 Bernoulli distribution

Let us consider the flipping of a coin one time. What is the distribution of the number of heads in a single toss. This example has a Bernoulli distribution. We assume to have a single trail where each can have two possible mutually exclusive outcomes:

$$\begin{cases} a & \text{P(success)} = \text{p} \\ b & \text{P(failure)} = 1 - \text{p} \end{cases}$$

We define the random variable $X = 1$ if a success occurs and $X = 0$ if a failure occurs. Then X has a *Bernoulli* distribution:

$$P(X = x) = p^x(1-p)^{1-x} \tag{8.10}$$

for $x = 0$ or $x = 1$. It is also called probability mass function of the *Bernoulli* distribution. Now writing this distribution for the two individual values. For $x = 1$:

$$P(X = x) = p^1(1-p)^{1-1} = p \tag{8.11}$$

and for $x = 0$

$$P(X = x) = p^0(1-p)^{1-0} = 1-p \tag{8.12}$$

which is the same of what was written above, for success and failure, but with the formulation of eq. 8.10 it is written in one single equation. The mean of a Bernoulli random variable is

$$\mu = p \tag{8.13}$$

and the variance is

$$\sigma^2 = p(1-p) \tag{8.14}$$

In Italy there are 403,454 medical doctors over a population of 59,55 milion people. The ratio is 0.006, therefore there are about 6 medical doctors every 1000 Italians. There is therefore 1 medical doctor every 166.6. Let us approximate to 160. One Italian citizen is selected, what is the probability that he or she is a medical doctor ? and what is the distribution of the number of medical doctors ? We have one single trail. So the condition for the Bernoulli distributions are met so the number of medical doctor will have a Bernoulli distribution with parameter p and it is simply $p = \frac{1}{160}$. If the variable X is the number of medical doctor for a sample of size one, then the probability is:

$$P(X = x) = (\frac{1}{160})^x(1 - \frac{1}{160})^{1-x} \tag{8.15}$$

and this is for $x = 0$ or $x = 1$ since the person will a medical doctor or not. Substituting these values will lead to, for $x = 1$:

$$P(X = x) = (\frac{1}{160})^1(1 - \frac{1}{160})^{1-1} = \frac{1}{160} \tag{8.16}$$

and for $x = 0$

$$P(X = x) = (\frac{1}{160})^0(1 - \frac{1}{160})^{1-0} = \frac{159}{160} \tag{8.17}$$

So one may wonder why to make this calculation for something so simple, since we knew that there was one person every 160. First the *Bernoulli* distribution is a concise and mathematically precise description of the probability. Moreover, this distribution is the base for other common distributions that are built from the *Bernoulli* distribution and from the assumption of *independent Bernoulli trials*, such as the Binomial or the geometric distribution.

8.5 Binomial distribution

Let us consider n *independent Bernoulli trials*, where independent means that the outcome of one trial does not affect the outcome of the following trials. As an example we choose the outcome of getting heads from flipping a fair coin five times. Therefore the *r.v.* $X =$ number of heads from flipping a fair coin 5 times. The possible outcomes are many, we could have $THTHT, HHHTT, HHTTH$ and so forth. The probability of success is denoted p ant it is constant for each experiment. It is important to remember that the values n and p are known and constant and that each experiment has the identical probability of any other.

The question is: *what is the probability of success in n trials* ? For instance, to obtain 30 heads in 100 throws, $\mathsf{P}\{X = k\text{success}, n \text{ trials}\}$. If X is a discrete *r.v.* whose field is $\{0, 1, 2, \ldots, n\}$. We indicate the success with 1 and failure with 0. For instance $n = 10, k = 4$: $1, 1, 0, 0, 0, 1, 0, 0, 1, 0$. Defining with A the set of success and B the set of failure, the probability of the previous outcome is given by the law of combined probability for independent events:

$\mathsf{P}\{A \cap B\} = \mathsf{P}\{A\}\,\mathsf{P}\{A\}$. In general:
$\mathsf{P}\{\bigcap A_i\} = \prod_i \mathsf{P}\{A_i\}$ $\mathsf{P}\{1, 1, 0, 0, 0, 1, 0, 0, 1, 0\} =$

$\mathsf{P}\{1\} \times \mathsf{P}\{1\} \times \cdots \times \mathsf{P}\{1\}$

Therefore $\prod_i \mathsf{P}\{A_i\} = pp(1-p)(1-p)(1-p)p(1-p)(1-p)p(1-p) = p^4(1-p)^{10-4}$
The probability of the outcome of k success in n trials is:
$p^k(1-p)^{n-k}$.
How many of these sequences are possible, with n and k successes ? They are a number equal to the many different ways of partitioning the n in different positions in the sequence of k success and $n-k$ failures. In other words the number of n objects that are equal to eachother. The number of sequences is given by:

$$\frac{n!}{k!(n-k)!} \tag{8.18}$$

Now, the computation of the outcome of number of heads (H) from flipping a coin five times, is computed by using eqn.8.18, where from combinatory theory:

$$5^{C_0} = \frac{n!}{k!(n-k)!} = \frac{5!}{0!(5-0)!} = \frac{5\times4\times3\times2\times1}{(1)(5\times4\times3\times2\times1)} = \frac{120}{120} = 1$$

$$5^{C_1} = \frac{n!}{k!(n-k)!} = \frac{5!}{1!(5-1)!} = \frac{5\times4\times3\times2\times1}{(1)(4\times3\times2\times1)} = \frac{120}{24} = 5$$

$$5^{C_2} = \frac{n!}{k!(n-k)!} = \frac{5!}{2!(5-2)!} = \frac{5\times4\times3\times2\times1}{(2\times1)(3\times2\times1)} = \frac{120}{12} = 10$$

$$5^{C_3} = \frac{n!}{k!(n-k)!} = \frac{5!}{3!(5-3)!} = \frac{5\times4\times3\times2\times1}{(3\times2\times1)(2\times1)} = \frac{120}{12} = 10$$

$$5^{C_4} = \frac{n!}{k!(n-k)!} = \frac{5!}{4!(5-4)!} = \frac{5\times4\times3\times2\times1}{(4\times3\times2\times1)(1)} = \frac{120}{24} = 5$$

$$5^{C_5} = \frac{n!}{k!(n-k)!} = \frac{5!}{4!(5-5)!} = \frac{5\times4\times3\times2\times1}{(5\times4\times3\times2\times1)(1)} = \frac{120}{120} = 1$$

Therefore the probabilities are:

$$P(x=0) = \frac{5^{C_0}}{32} = \frac{1}{32} = 0.031$$

$$P(x=1) = \frac{5^{C_1}}{32} = \frac{5}{32} = 0.156$$

$$P(x=2) = \frac{5^{C_2}}{32} = \frac{10}{32} = 0.312$$

$$P(x=3) = \frac{5^{C_3}}{32} = \frac{10}{32} = 0.312$$

$$P(x=4) = \frac{5^{C_4}}{32} = \frac{5}{32} = 0.156$$

$$P(x=5) = \frac{5^{C_5}}{32} = \frac{1}{32} = 0.031$$

The binomial r.v. $\mathcal{B}(k;n,p)$ represents the probability of k success in n indipendent trials, each with a probability of success p. We can write:

$$\mathcal{B}(k;n,p) = X_1 + X_2 + \cdots + X_n$$

where $X_i = 1$, if the i-th trial is successful, otherwise $= 0$. Since $X_i = p$ and $X_i = p(1-p)$, it follows:

$$\frac{\mathcal{B}-np}{\sqrt{np(1-p)}} = \frac{\mathcal{B}-p}{\sqrt{\frac{p(1-p)}{n}}} \approx \mathcal{N}(0,1)$$

In **R** the binomial distribution is implemented into the function `rbinom(n, ns, p)`. This function has three arguments: the number of random draws (n), the number of coins being flipped on each draw (ns), and the probability of a heads (p) as the outcome. With the function `rbinom(n, ns, p)`, simulates 5 *random* flips of a single coin using a bimodal distribution:

```
n<- 5
ns<- 1
p<- .5
y <- rbinom(n, ns, p)
y
```

The outcomes could be:

```
0 1 0 0 1
1 0 0 0 0
1 0 1 0 1
...
```

where the head is 1 and the tail is 0. To estimate the exact probability density at a given point, the function `dbinom(n, ns, p)` is used. The arguments here are the density being estimated (1 head), the number of coins (5), and the probability of producing a head (0.5).

```
n<- 1
ns<- 5
p<- .5
y <- dbinom(n, ns, p)
y
```

The results of this code is 0.15625, as expected. Replacing the value of `n` in the function will produce the expected values of probabilities described above. To compute all the values the number of successes are computed from 0 to 5:

```
n<- 0:5
ns<- 5
p<- .5
y <- dbinom(n, ns, p)
y
plot(type="h",n,y,xlab="Number of heads",ylab="P(X)")
```

The results are the probabilities as listed above.

```
[1] 0.03125 0.15625 0.31250 0.31250 0.15625 0.03125
```

Figure 8.5 shows the probabilities of heads for trials of 5 throws.

8.6 Normal Distribution

Among the many distributions, the most common one is the Normal Distribution. It is a symmetric, unimodal, bell curve. It is very common simply because many variables tend to follow a normal distribution. For instance the heights of human adults follow a normal distribution. The general symbolic form is:

$$N = (\mu, \sigma) \tag{8.19}$$

where μ is the population mean and σ is the population standard deviation.

The density curve is:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad -\infty < x < \infty \tag{8.20}$$

where μ and σ are the mean and the standard deviation of the random variable x with density $p\,(x)$ (Fig.8.6).

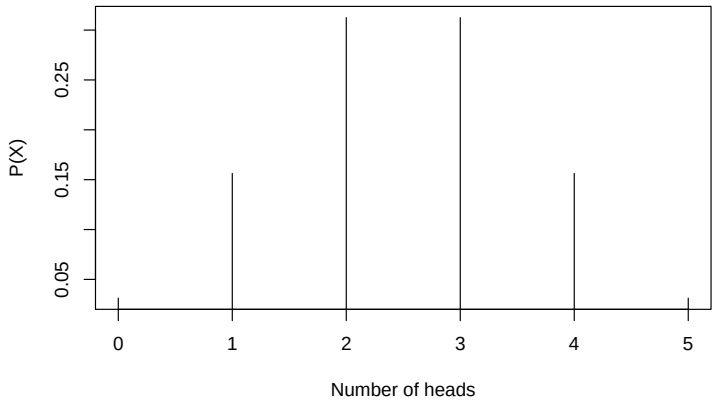


Fig. 8.3 Probabilities of heads for trials of 5 throws.

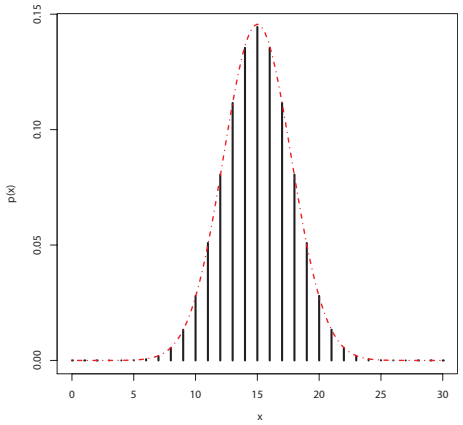


Fig. 8.4 Normal distribution

The bell curve becomes larger at increasing values of standard deviation as depicted in Fig. 8.6. Since it is a density distribution, the area under the curve is always

Usually, the cumulative form of the Gaussian function, is obtained by integration of eqn 8.20 with respect to x:

$$F(x) = \frac{1}{2}\left(1 + \text{erf}\left[\frac{(x-\mu)}{\sigma\sqrt{2}}\right]\right) \text{ for } (x > \mu) \tag{8.21}$$

$$F(x) = \frac{1}{2}\left(1 - \text{erf}\left[\frac{(x-\mu)}{\sigma\sqrt{2}}\right]\right) \text{ for } (x \leq \mu) \tag{8.22}$$

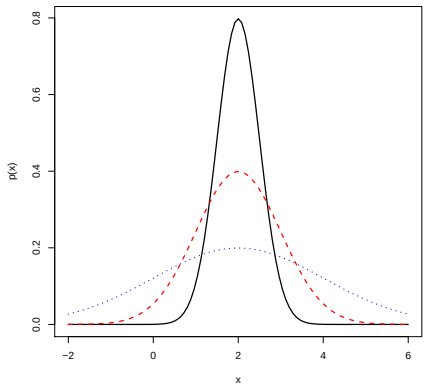


Fig. 8.5 Normal distributions with $\mu = 2$ and $\sigma = 0.5, 1, 2$ (continuos, interrupted lines and points.)

where erf[ ] is the error function. For the computation of the error function a numerical approximation from **?**) can be used:

$$\text{erf}(x) = 1 - (0.3480242T - 0.0958T^2 + 0.7478556T^3)\ \exp\ (-x^2) \tag{8.23}$$

where $T = 1/(1 + 0.47047x)$.
It is possible to generate random numbers having a normal distribution using the instruction `rnorm()`. A normalized distribution having mean equal to zero (`mean=0`) and standard deviation equal to one (`sd=1`) is obtained by specifying the values of mean and standard deviation, as shown below:

```
random_numb_normal <- rnorm(1000,mean=0,sd=1)
plot(random_numb_normal)
```

Figure 8.6 depicts the random numbers, from index 1 to 1000 is:
The numbers can be visualized by using frequency classes.

```
values <- rnorm(1000, mean= 0, sd = 1)
hist(values, col = 'green', freq=F)
```

The graph simple displays the generated numbers from 1 to 1000. Although the distribution is normal, it is not clear from a simple list of numbers. The distribution appears normal when the numbers are plotted by interval of frequency classes. The concept of frequency and classes is described in details in the next chapter. The plot (Fig.8.6) of the distribution looks now more familiar:

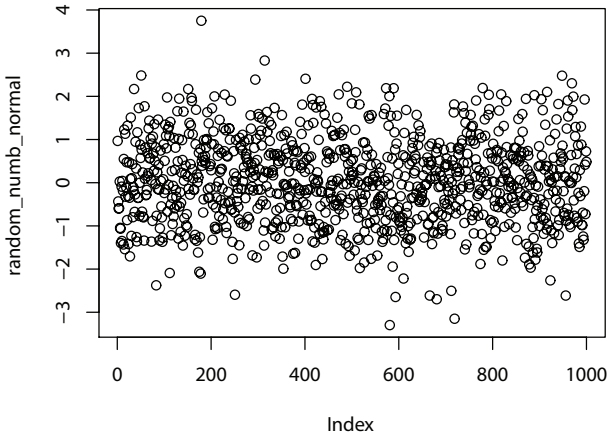


Fig. 8.6 Random numbers

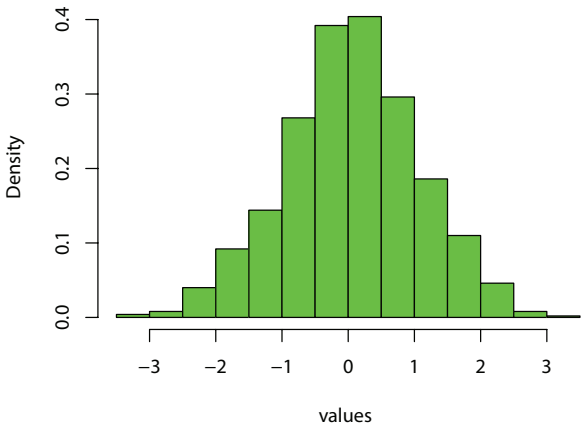


Fig. 8.7 Normal distribution of random numbers by classes

Note that the distribution is normalized with mean = 0 and standard deviation = 1. The code below (also shown in Chapter 7) allows for plotting the PDF and the CDF for the normal distribution.

```
# Create a sequence of numbers between -10 and 10 incrementing by 0.1.
x <- seq(-10, 10, by = .1)

# Choose the mean as 0 and standard deviation as 1.
y <- dnorm(x, mean = 0, sd = 1)
y_cum<-pnorm(x, mean = 0, sd = 1)

plot(x,y,type="l",col="blue",xlim = c(-5,5),ylab=("PDF"))
```

```
plot(y_cum,type="l",col="green",ylab=("CDF"))
```

Figure 8.6 show the PDF and CDF for the normal distribution.

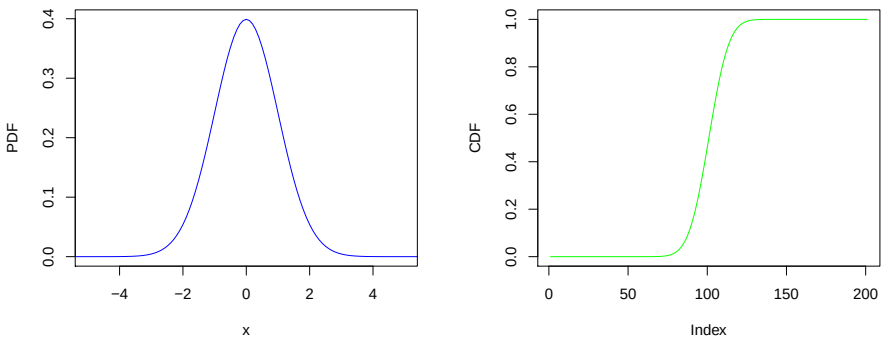


Fig. 8.8 PDF and CDF for a the normal distribution.

8.6.1 Transformations

If the frequency distribution for a soil property follows a normal probability density function, its position and dispersion are easily and conveniently described by the arithmetic mean and the variance. Probabilities of occurrence for various values of the property are easily found. The most useful of the transformations for soil physical properties is the log transform. The frequency distribution of the transformed data is symmetric with a mean of:

$$\langle \ln x \rangle = \frac{1}{n} \sum \ln x \tag{8.24}$$

and variance

$$\sigma^2 = \frac{1}{n-1} \sum (\ln x_i - \langle \ln x \rangle)^2 \tag{8.25}$$

8.6.2 Central limit theorem

The "central limit theorem" says that, under general conditions, the distribution of a *r.v.*, given by a summation of random variables, tends to be distributed as a normal distribution by increasing the number of the additive elements. Be n independent random variables $\{X_1, X_2, \ldots, X_n\}$, all having the same distribution, with finite mean μ and finite variance σ^2. The central limit theorem says that, for $n \to \infty$ (n large):

$$X_1 + X_2 + \cdots + X_n \longrightarrow \mathcal{N}(n\mu, n\sigma^2)$$

Note that this convergence is a *convergence in distribution*. Even if it is said that the sequence of r.v. converges in a distribution, are the partition functions that are converging, not the random variables.

In the standardized form:

$$\frac{X_1 + X_2 + \cdots + X_n - \mathrm{n}\mu}{\sigma\sqrt{\mathrm{n}}} \longrightarrow \mathcal{N}(0,1)$$

8.7 t-student distribution

One of the most important test within the branch of inferential statistics is the Students *t−test*. The Student's t−test for two samples is used to test whether two groups (two populations) are different in terms of a quantitative variable, based on the comparison of two samples drawn from these two groups. A Student's t−test for two samples allows for determining whether the two populations from which two samples are drawn are different. Here the t−student distribution is described, while in the section on inferential statistics an example of application of a t−student test is presented.

The normalized Normal distribution is:

$$\mathcal{Z} = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \tag{8.26}$$

Examples in R were presented where a sample of the random variable was drawn from the population. Commonly, the population standard deviation σ is not known, so it is not possible to use the value σ in the formula, but instead the sample standard deviation s is used:

$$\mathcal{Z} = \frac{(\bar{x} - \mu)}{s/\sqrt{n}} \tag{8.27}$$

the sample standard deviation is not a fixed number but it has a statistical distribution that varies from sample to sample. Therefore it is defined a t-student distribution:

$$t = \frac{(\bar{x} - \mu)}{s/\sqrt{n}} \tag{8.28}$$

This distribution has a similar shape of the normal distribution but with greater variance. It has $(n - 1)$ degree of freedom. As the degree of freedom increases the t-student will tend toward a normal distribution. The code below shows how the t-student distribution approaches the normal distribution at increasing values of df.

Density, distribution function, quantile function and random generation for the t distribution with df degrees of freedom is called with these instructions:

```
dt(x, df, ncp, log = FALSE)
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
rt(n, df, ncp)
```

```
x <- seq(-4, 4, length=100)
mean<-mean(x)
sd<-sd(x)
sd
```

```
#Apply a normal distribution
y_norm<-dnorm(x,mean = 0, sd =1 )
mean
# Applying the dt() function
y_dt <- dt(x, df = 2)
y_dt_5df <- dt(x, df = 5)
y_dt_10df <- dt(x, df = 10)
y_dt_30df <- dt(x, df = 30)
# Plotting
plot(x,y_dt, type = "l", ylab =("PDF"),las=1,ylim=c(0,0.5),col="yellow")
lines(x,y_dt_5df,col="red")
lines(x,y_dt_10df,col="blue")
lines(x,y_dt_30df,col="green")
lines(x,y_norm,col="black",lty=2, lwd=3)
```

Figure 8.7 shows the t−student distribution for different degrees of freedom (2,5,10 and 30). The normalized normal distribution is also depicted in the graph (dotted line). Note that above 30 degrees of freedom the t−student distribution approaches the normal distribution.

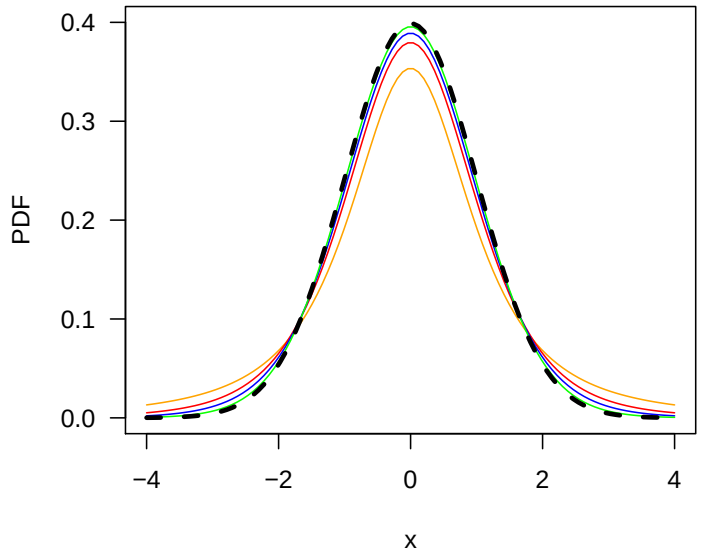


Fig. 8.9 t−student distribution for 2 (orange line),5 (red line), 10 (blue line) and 30 (green line) degrees of freedom. The normal distribution is plotted with black line.

8.8 Poisson distribution

The Poisson distribution is used when a **counting process** is analyzed.

A stocastic process $\{X(t), t \geqslant 0\}$ is considered a counting process if the $X(t)$ represents the total number of counted events. With the notion "up to time t" it means "before and *at* time t". If the event comes exactly at time t, it is counted. This a process with coutinuos time and with discrete states $s = N$.

The number of people who are entering into a store from the opening time until 10 AM is described as a counting process, but the number of people that are in the store at 10 AM is not a counting process. A counting process may be the number of births of a given animal until a given day, the number of songs written by a musician, the number of car accidents.

A counting process $\{X(t), t \geqslant 0\}$ is defined by listing some properties:

1. $\{X(t)\} \geqslant 0$.
2. $\{X(t\})$ it takes only integer values.
3. if $s < t$ then $X(s) \leqslant X(t)$.
4. if $s < t$ then the difference $X(t) - X(s)$ counts the number of events in the time interval $(s, t]$.

Therefore the differences $X(t) - X(s)$ are only non−negative, integer values. In particular if $X(0) = 0$, $\{X(t)\}$ represents the number of counted events in the time interval $[0, t]$. These processes are called *purely discontinuous*. For the counted events, other terms can be used such as "arrivals" or "emissions" from a source. The positive random variables $X(t) - X(s)\ \forall t, s$ are called *increments* of the process and they are particularly important when they are *indipendent* and *stationary*.

A counting process has independent increments if the number of arrivals, realizations of the r.v. $X(t) - X(s)$, that are happening in any time interval $(s, t]$ are independent from the number of arrivals, realizations of the r.v. $X(v) - X(u)$, occured in any other time interval *disjointed* $(u, v]$, for instance in $[0, s]$. This definition means that the random variables representing the number of arrivals in disjointed intervals are indipendent. It is like the probability of n arrivals in $(s, t]$ does not change even though it is conditioned to the information about the arrivals in different time intervals from $(s, t]$. This independence of increments is the analogue of the independence of trials for the Bernoulli process. Process at independent increments could be the number of customers entering a store or the songs written by a musician, but it is not reasonable to apply it to the number of births. In this case if the number of birth is large enough it is not plausible that the future birth will not depend upon the births occured before.

A counting process has *stationary increments* or *homogeneous increments*, if the number of arrivals in any time interval $[t, t + \Delta t]$ depends only upon the length of the time interval Δt, but not on t, therefore on the position of the interval on the time axis. In other words, a r.v. "number of arrivals" in the interval Δt has the same distribution of the variable "number of arrivals" in $t + \Delta t$.

Hence, the stationarity property of the increments is written:

$$X(z + t + \Delta t) - X(z + t) \sim X(t + \Delta t) - X(t)$$

Or, given $\Delta t = t - s$, with $t > s$:

$$X(t+z) - X(s+z) \sim X(t) - X(s), \qquad \forall z > 0, \quad \forall t > s \tag{8.29}$$

This means that arrivals are " equally likely " for every t. This property is the continuous analogue of the constancy of the probability of success p in a Bernoulli process. In the shop example, that's not reasonable assume stationarity, as there are almost certainly intervals with greater influx of customers. Also in the other examples, the processes that describe them are not stationary. For births, stationarity would be plausible if the population of individuals remained constant; for the musician, his inspiration is likely to vary over time. Finally, note that the fact that the process has stationary increments does not mean that it is stationary, as exemplified by the Poisson process.

The counting process $\{X(t)\}$ is a Poisson process if:

a) $X(0) = 0$. In other words, the origin of the times arises at the moment that the count is started .

b) The process has independent increments.

c) The probability of n of arrivals in a range of finite length Δt is given by the random variable $X(\Delta t)$ with so-called Poisson distribution:

$$\mathsf{P}\{X(\Delta t) = n\} = \frac{(\lambda \Delta t)^n}{n!} e^{-\lambda \Delta t}, \qquad n = 0, 1, \ldots \tag{8.30}$$

The Poisson distribution defines the probability that the random variable X takes the values n. $\mathsf{P}\{X = n\}$ can take any non−negative whole number value. The mean of the distribution is $\mu = \lambda$ and the variance is also equal to lambda $\sigma^2 = \lambda$.

For example: one nanogram of Plutonium will have an average of 2.3 radioactive decays / 1 sec. What is the probability that in 2 seconds there are exactly 3 decays ? n represents the number of decays in 2 seconds:

$$\lambda \Delta t = 2.3 \times 2 = 4.6 \tag{8.31}$$

therefore X has a Poisson distribution with $\lambda = 4.6$ and with probability:

$$\mathsf{P}\{X = 3\} = \frac{4.6^3 e^{-4.6}}{3!} = 0.163 \tag{8.32}$$

The probability of having exactly 3 radioactive decays in 2 seconds is 16 %.

```
#POISSON
## lambda is known
set.seed(2)
deltat<- 3600 # 1 hr=3600 sec
lambda<- 0.0028
ev<- 0 # initial time t=0
t<- 0
while (t<deltat) {
```

```
       t<- t + rexp(1,lambda)
       ev<- c(ev,t) # adding the t values to zero; as well as ev<-c(0,t)
       print(ev) # print to see who events develop over time
} # end while

ev # arrival time are written
length(ev) # (count ev=0
plot(ev,0:(length(ev)-1),type="s",xlab="tempo (s)",ylab="n. arrival(t)")
# s is for "stair step" it makes steps
abline(v=deltat,lty=3)
# when the system changes state (new arrival) it takes a step on the graph;
# the time between one finish and the next is given by the exponential
```

```
##### The average number of arrivals (snow) in deltat is known
# practically like the first program, here snow lambda is calculated
set.seed(2)
deltat<- 3600
neventi<- 10
lambda<- neventi/deltat
lambda
1/lambda
ev<- 0
t<- 0
while(t<deltat){
       t<- t + rexp(1,lambda)
       ev <- c(ev,t)
}
ev
length(ev)

plot(ev,0:(length(ev)-1),type="s",xlab="tempo (s)",ylab="no. arrivi(t)")
abline(v=deltat,lty=3)
```

9
Descriptive statistics

9.1 Frequencies

An important representation of data is through the use of frequency histograms or relative frequency histograms. Both of these graphical techniques are applicable only to quantitative data. Before plotting the resultsm, data must be organized. There are different possibilities:

9.1.1 Absolute Frequency

The absolute frequency is the number of times a value appears. It is represented as

$$0 \leq n_i \leq n \qquad \sum_{i=1}^{k} n_i = n$$

where the subscript represents each of the values. So if a value or precipitation (50 mm) appears 4 times, its absolute frequency is 4.

9.1.2 Relative frequency

The relative frequency is the number of times a value appears, divided by the total number of data. It is obtained by dividing the absolute frequency of a certain value by the total number of data.

$f_i = n_i / n$

$$0 \leq f_i \leq 1 \qquad \sum_{i=1}^{k} f_i = 1$$

9.1.3 Percentual frequency

The percentual frequency is the number of times a value appears, divided by the total number of data and multiplied by 100.

$p_i = n_i / n \times 100$

$$0 \leq p_i \leq 100 \qquad \sum_{i=1}^{k} p_i = 100$$

9.1.4 Example

We have $n = 25$ statistical units (people), with the character $X =$ drink (what they drink with pizza). There are 4 modalities, modality = 4 (beer, water, Coke and wine)

The qualitative data are assigned a code, and identifier 1=beer, 2 = water, 3 = Coke and 4 = wine. The data are for 25 customers: 3, 4, 1, 1, 3, 4, 3, 3, 1, 3, 2, 1, 2, 1, 2, 3, 2, 3, 1, 1, 1, 1, 4, 3, 1.

Table 9.1 Distribution of the character X

Modality	Absolute f.	Relative f.	Percentual f.
beer	10	0.40	40
water	4	0.16	16
coke	8	0.32	32
wine	3	0.12	12
Sum	25	1	100

The data organized in the table can be used to construct a *frequency histogram* or a *relative frequency histogram*. In this case on the x-axis, the modality is represented since it is a qualitative value (beer, wine, etc.). In the next example, on the x-axis will be represented classes or intervals, in case a quantitative variable is used.

It is possible to plot these frequencies with, on the left, absolute frequencies and on the right, relative frequencies.

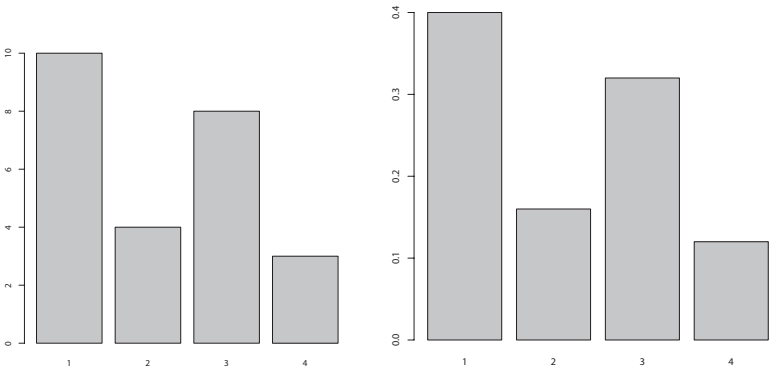


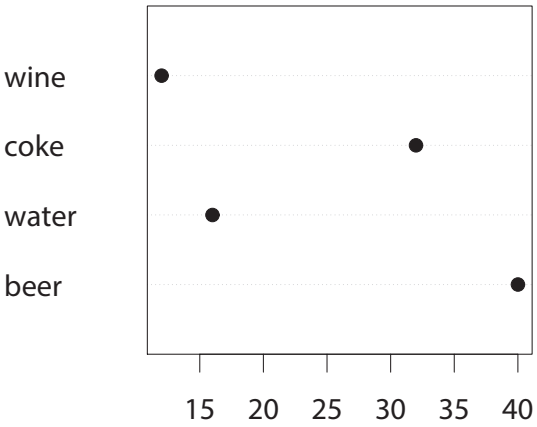
Fig. 9.1 Left: Absolute frequencies; Right: Relative frequencies

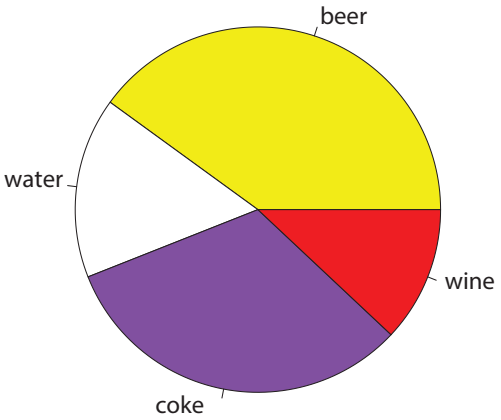
As discussed above the first thing to do is to plot the data. Some data visualizations are better than others. There are many different options. It is possible to use dot graphs, such the one below.

Another possibility are pie charts. Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.

The size of acute angles tends to be underestimated, and the size of obtuse angles overestimated. This is one reason pie charts are usually a bad idea. We also misjudge areas poorly. We have known for a long time that area-based comparisons of quantities are easily misinterpreted or exaggerated.

Cleveland (1985): "Data that can be shown by pie charts always can be shown by





a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements." This statement is based on the empirical investigations of Cleveland and McGill as well as investigations by perceptual psychologists.

The computation of these frequencies can be performed with R. First a vector `Vtr1` is created to create an index of the 25 measurements.

```
> Vtr1 <- c((1:25))
```

Then, a second vector `Vtr2` stores the recorded values:

```
> Vtr2<-c(3, 4, 1, 1, 3, 4, 3, 3, 1, 3, 2, 1, 2,
       1, 2, 3, 2, 3, 1, 1, 1, 1, 4, 3, 1)
```

Next, a `data.frame` named `drink` is created from the two vectors

```
> drink=data.frame(Vtr1,Vtr2)
```

If the `drink` data.frame is called it returns:

```
drink
   Vtr1 Vtr2
1 1 3
2 2 4
3 3 1
4 4 1
5 5 3
6 6 4
7 7 3
8 8 3
9 9 1
10 10 3
11 11 2
12 12 1
13 13 2
14 14 1
15 15 2
16 16 3
17 17 2
18 18 3
19 19 1
20 20 1
21 21 1
22 22 1
23 23 4
24 24 3
25 25 1
```

The drinks (1=beer, 2 = water, 3 = Coke, 4 = wine) are four types. So we can define it as `factors`, by calling it `drink_type`:

```
drink_type=factor(Vtr2)
drink_type
 [1] 3 4 1 1 3 4 3 3 1 3 2 1 2 1 2 3 2 3 1 1 1 1 4 3 1
Levels: 1 2 3 4
```

When the command `drink_type` is types on the console, it returns the list and the four levels. To get a frequency table of a categorical variable in R, the `count()` function in the `plyr` package is used. The package `plyr` is installed

```
install.packages('plyr')
```

The `count()` function can then be used:

```
> library(plyr)
```

```
> categ_count<- count(drink_type)
> categ_count
  x freq
1 1 10
2 2 4
3 3 8
4 4 3
```

The function returns the number of each categorical types, (10 people ordered beer, 4 ordered water, 8 ordered Coke and 3 ordered wine). This is defined above as **absolute frequency** and a variable `categ_count` was created to save this numbers.

```
> str(categ_count)
'data.frame': 4 obs. of 2 variables:
 $ x : Factor w/ 4 levels "1","2","3","4": 1 2 3 4
 $ freq: int 10 4 8 3
```

Obviously, the total number is 25, one unit for each customer. Now the relative frequency is each value of absolute frequency divided by the total number of units, as shown above.

```
> rel_freq <- categ_count$freq /25
> rel_freq
[1] 0.40 0.16 0.32 0.12
```

Note that, since `categ_count` is a data.frame and not a single variable, only the property $freq was invoked to be then divided by 25. This operation returned the **relative frequency**.

```
> perc_freq <- (categ_count$freq /25)*100
> perc_freq
[1] 40 16 32 12
```

Finally, the **percentual frequency** can be computed.

To plot the results of the frequency distribution, `ggplot` allows for providing the dataframe and the variable. It automatically recognize the four different classes and plot a bar graph.

```
ggplot(data.frame(drink_type), aes(x=Vtr2)) +
+ geom_bar()
```

The output of this instruction is the figure below:

The figure generated by `ggplot` can be saved as encapsulated post script (.eps), which is a good format to manupulating and printing high quality figures.

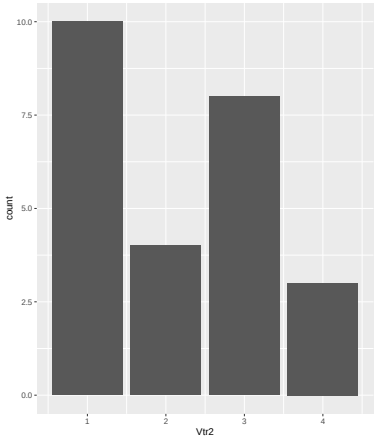


Fig. 9.2 Left: Absolute frequencies plotted with ggplot

9.2 Classes

Now, another example will be shown to introduce the concept of classes. The data shown in the table below describes cumulative annual precipitation (in inches) from 1873 to 1978. (Data are taken from B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York,1993).

The data can be displayed with histograms:

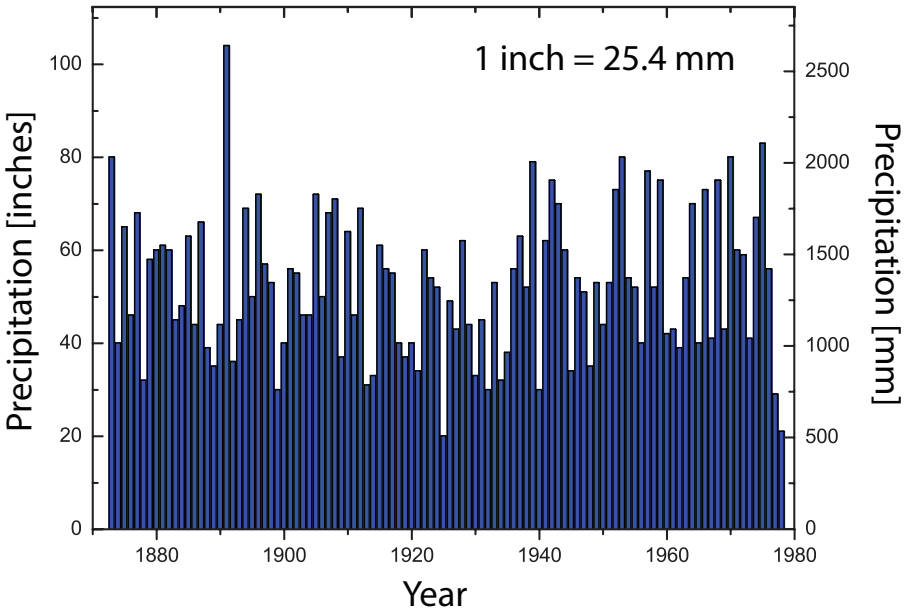


Fig. 9.3 Annual precipitation in Nevada City from 1873 to 1978

Table 9.2 Annual precipitation (inches) at Nevada City from 1873 to 1978

year	inch	year	inch	year	inch	year	inch
1873	80	1900	40	1927	43	1954	54
1874	40	1901	56	1928	62	1955	52
1875	65	1902	55	1929	44	1956	40
1876	46	1903	46	1930	33	1957	77
1877	68	1904	46	1931	45	1958	52
1878	32	1905	72	1932	30	1959	75
1879	58	1906	50	1933	53	1960	42
1880	60	1907	68	1934	32	1961	43
1881	61	1908	71	1935	38	1962	39
1882	60	1909	37	1936	56	1963	54
1883	45	1910	64	1937	63	1964	70
1884	48	1911	46	1938	52	1965	40
1885	63	1912	69	1939	79	1966	73
1886	44	1913	31	1940	30	1967	41
1887	66	1914	33	1941	62	1968	75
1888	39	1915	61	1942	75	1969	43
1889	35	1916	56	1943	70	1970	80
1890	44	1917	55	1944	60	1971	60
1891	104	1918	40	1945	34	1972	59
1892	36	1919	37	1946	54	1973	41
1893	45	1920	40	1947	51	1974	67
1894	69	1921	34	1948	35	1975	83
1895	50	1922	60	1949	53	1976	56
1896	72	1923	54	1950	44	1977	29
1897	57	1924	52	1951	53	1978	21
1898	53	1925	20	1952	73		
1899	30	1926	49	1953	80		

or with points:

A useful method to describe data is to create classes. Classes are simply intervals, classes of ranges) of the variable. For instance in the Table 9.3, classes of 200 mm were created.

Table 9.3 Distribution of frequencies of "Precipitation in Nevada City"

Precipitation (mm)	Absolute freq.	Relative freq.	Percentual freq.
$[200, 400)$	0	0	0
$[400, 600)$	2	0.019	1.9
$[600, 800)$	5	0.047	4.7
$[800, 1000)$	14	0.132	13.2
$[1000, 1200)$	23	0.217	21.7
$[1200, 1400)$	19	0.179	17.9
$[1400, 1600)$	16	0.151	15.1
$[1600, 1800)$	12	0.113	11.3
$[1800, 2000)$	9	0.085	8.5
$[2000, 2200)$	5	0.047	4.7
$[2200, 2400)$	0	0	0
$[2400, 2600)$	0	0	0
$[2600, 2800)$	1	0.009	0.9
Sum	106	1	100

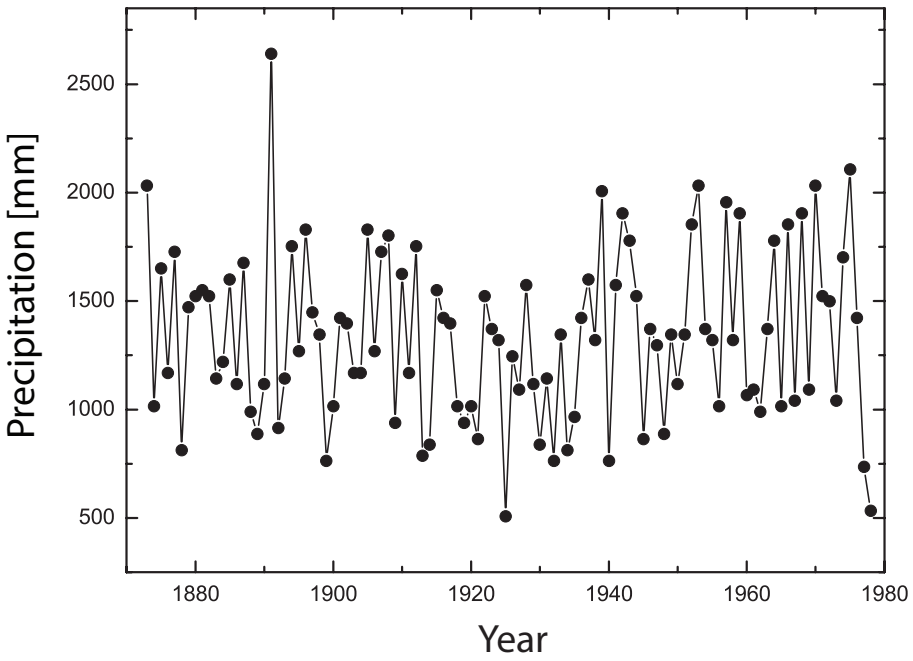


Fig. 9.4 Annual precipitation in Nevada City from 1873 to 1978

In this case, to perform this operation with R, the data file must be imported. This instruction has been shown in the Chapter *Data Management*. In this case the file is called `Nevada_prec.dat` and the separator is a `tab`.

```
#CODE Ch9_1.R
Nevada_prec <- read.table("C:/Users/marco.bittelli.PERSONALE/Documents
/Didattica/R_class_2/exercises/descriptive_stat/Nevada_prec.dat",
sep = "", check.names = FALSE, header = T, na.strings = c("NA", "NAN"))
```

The instruction `read.table` reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file. Now the data in the second column (annual cumulative precipitation) should be cumulated and classified. Before performing this operation, the data are transformed into [mm].

```
Nevada_prec$Prec_mm = Nevada_prec$Prec * 25.4
```

Here a new variable `Prec_mm` has been created. This can be inquired with the instruction `str()`.

```
str(Nevada_prec)
'data.frame': 106 obs. of 3 variables:
 $ Year : int 1873 1874 1875 1876 1877 1878 ...
 $ Prec : int 80 40 65 46 68 32 58 60 61 60 ...
 $ Prec_mm: num 2032 1016 1651 1168 1727 ...
```

Data can be plotted in any moment, with a simple instruction like

```
> plot(Nevada_prec$Prec_mm)
```

It is possible to collect data into **classes**, of type $x_1 \dashv x_{i+1}$ or $x_1 \vdash x_{i+1}$, in case it is preferred to include the lower limit instead of the upper limit by specifying `right= FALSE` as an additional argument of the function.

```
classes<-table(cut(Nevada_prec$Prec_mm, breaks=c(0,200,400,600,800,1000,
1200,1400,1600,1800,2000,2200,2400,2600,2800)),right=TRUE )
```

The output of this statement is

```
> classes
(0,200] (200,400] (400,600] (600,800] ...
0 0 2 5 ...
```

which are the number of events belonging to the different classes. Also notice that R printed the upper value as a close interval as specified. Note that classes are obtained if they start from zero.

There is another method to obtain classes, by using the function `hist()`:

```
histogram=hist(Nevada_prec$Prec_mm, c(0,200,400,600,800,1000,
1200,1400,1600,1800,2000,2200,2400,2600,2800),plot=TRUE)
```

The structure of the variable is then:

```
str(histogram)
List of 6
 $ breaks : num [1:15] 0 200 400 600 800 1000 1200 1400 1600 1800 ...
 $ counts : int [1:14] 0 0 2 5 14 23 19 16 12 9 ...
 $ density : num [1:14] 0.00 0.00 9.43e-05 2.36e-04 6.60e-04 ...
 $ mids : num [1:14] 100 300 500 700 900 1100 1300 1500 1700 1900 ...
 $ xname : chr "Nevada_prec$Prec_mm"
 $ equidist: logi TRUE
 - attr(*, "class")= chr "histogram"
```

where the `breaks` are specified, then the number of samples `count`ed within each class, the density, etc.

The classification by classes is useful to understand the frequency of given events. In the graph below, the occurrence of years with precipitation comprised between 400 and 600 mm.
Often it is useful to normalize the distribution to obtain a better representation of the occurrence of events.

9.2.1 Tabular and Graphical representation

The values for the cumulative frequencies are presented in the table below

A very useful representation of the cumulative frequencies is the use of cumulative graphs. How many units of the ensemble have a value smaller than x^* ? In the 20%

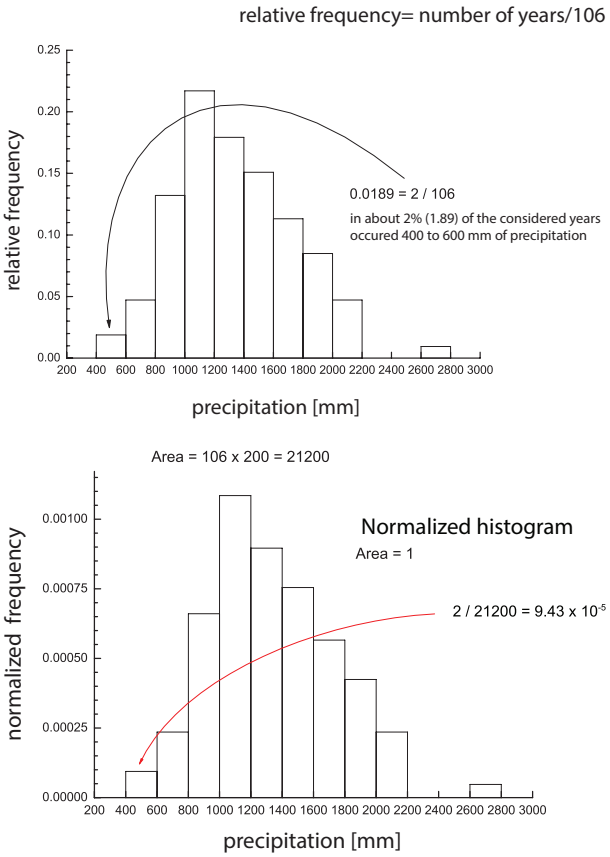


Table 9.4 Distribution of frequencies for "precipitation at Nevada City"

precipitazioni (mm)	n_i	N_i	F_i	P_i
$[200, 400)$	0	0	0	0
$[400, 600)$	2	2 + 0 = 2	0.02	1.89
$[600, 800)$	5	2 + 5 = 7	0.07	6.61
$[800, 1000)$	14	7 + 14 = 21	0.20	19.82
$[1000, 1200)$	23	21 + 23 = 44	0.42	41.51
$[1200, 1400)$	19	44 + 19 = 63	0.59	59.43
$[1400, 1600)$	16	63 + 16 = 79	0.75	74.53
$[1600, 1800)$	12	79 + 12 = 91	0.86	85.85
$[1800, 2000)$	9	91 + 9 = 100	0.94	94.34
$[2000, 2200)$	5	100 + 5 = 105	0.99	99.06
$[2200, 2400)$	0	105 + 0 = 105	0.99	99.06
$[2400, 2600)$	0	105 + 0 = 105	0.99	99.06
$[2600, 2800)$	1	105 + 1 = 106	1	100

(19.8%) of the years there are precipitation less than 1000 [mm].In the 6% of the years precipitation are *not* less than 2000 mm. For 94.34% of the years, precipitation are < di 2000 mm, therefore ≥ 2000 mm: 100 − 94.34 ≈ 6). In the figure below, cumulative frequencies are plotted:

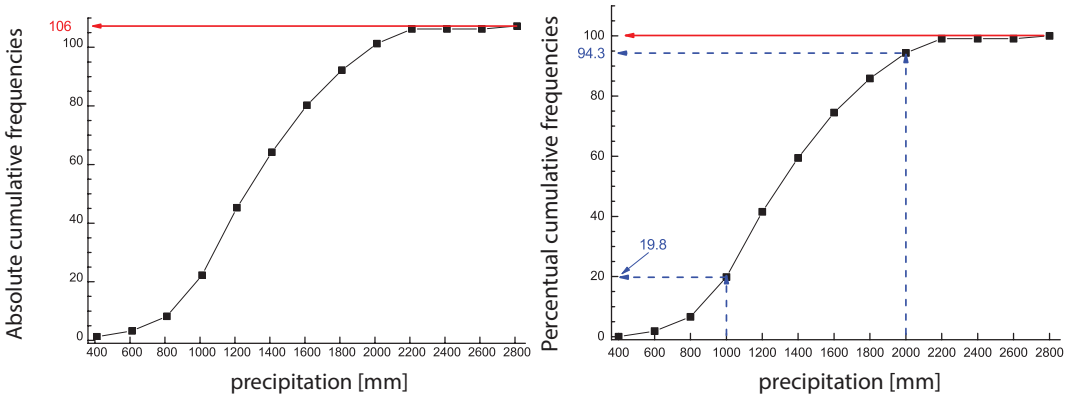


Fig. 9.5 Absolute (left) and relative (right) cumulative precipitation at the Nevada city experimental station (Nevada, USA).

9.3 Cumulative curves and frequencies

A cumulative curve is often important to understand the behaviour of a variable. In this example, daily precipitation data collected in the Emilia Romagna region (Italy) from 1961 to 2018, are used (**?**). The original data set is available at the Eraclito Project website.

As described above data can be represented using frequencies distribution and classes to obtain a first idea at the distribution. A very useful representation is a cumulative curve, frequency and distribution. For instance, cumulative precipitation curves as function of time, provide an idea of the seasons where precipitation occurs, as well as the total amount of precipitation over a given period. For instance, if the cumulative curve is steep in the period March−June and Sept−December, it means that a large amount of precipitation occurs in the spring and fall. This is typical of Mediterranean climates. On the other hand, in tropical areas large amount of precipitation occurs in the summer months due to the Monsoon, determining a steep cumulative curve during summer.

9.3.1 Absolute Cumulative Frequency

The **absolute cumulative frequency** of a mode is

$$N_i = n_1 + n_2 + \cdots + n_i$$

is given by the summation of the individual data. The code below shows how to compute and plot a cumulative curve for data from the experimental station of Cadriano (Bologna, Italy) for the year 1961. The code below reads the data from an excel file. It parse the date value as a date format and define variables to select the time period to use to compute the cumulative curve. A `while` loop is written to loop over the period and compute the cumulative curve.

```
#CODE Ch9_2.R
library(xts) ### provides functionality for working with time series
library(lubridate) ### manage dates
library(dplyr)
library(readxl) ### import excel file
#The excel file Prec_ER_daily.xlsx is opened, the structure of the
#newly created dataframe is analysed and the data are visualized.

setwd("~/Didattica/R_class_4/exercises/Ch9_statistical_inference")
Prec_ER_daily <- read_excel("Prec_ER_daily.xlsx")

# Only data from the Cadriano station are analyzed
Prec_ER_daily_cadriano <- Prec_ER_daily[,c("date","cadriano")]
prec_cadriano<-Prec_ER_daily_cadriano$cadriano
str(Prec_ER_daily_cadriano)

#Daily cumulative curve

date<-as.Date(Prec_ER_daily_cadriano$date)
start <- date[1]
end <- date[365]
start
end
theDate
as.Date(theDate)
theDate<-start
theDate
vec_date<-vector()
precip<-vector()
cumprec_vec<-vector()
prec_daily<-0
cumprec<-0

i<-1

while (theDate <= end){
       vec_date[i]<-theDate
       prec_daily=prec_cadriano[i]
       cumprec<-cumprec + prec_cadriano[i]
       cumprec_vec[i]<-cumprec
       output<-c(i,prec_daily,cumprec)
       print(output)
       theDate <-theDate + days(1)
       i<-i+1
}
```

```
#plotting
plot(as.Date(vec_date),cumprec_vec,type="l",xlab="Time [day]"
,ylab="Cumulative precipitation [mm]")
```

Figure 9.6 depicts cumulative precipitation for the Cadriano experimental station. As described above, in the year 1961, about 200 mm of precipitation occurred between January−March, about 400 mm in the period April−October and 600 mm in the period October−January. Clearly, there is variability among different years, and estimators must be used to derive more detailed information.

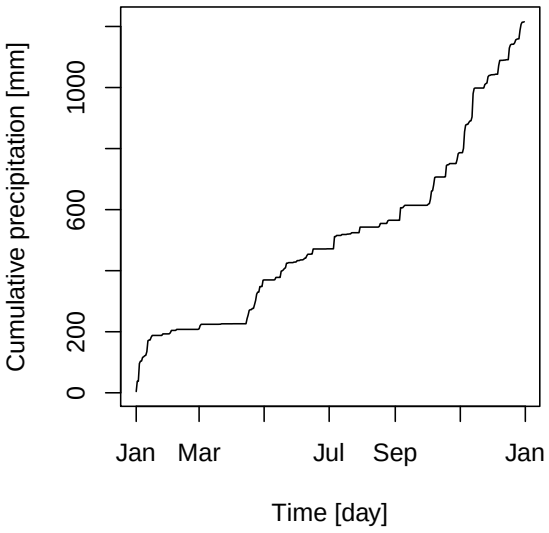


Fig. 9.6 Cumulative precipitation at the Cadriano (Bologna, Italy) experimental station for the year 1961.

9.3.2 Relative Cumulative Frequency

The data above are plotted as absolute values. Another useful way of plotting these data is to use a relative cumulative curve. The **relative cumulative frequency** of a mode is

$$F_i = f_1 + f_2 + \cdots + f_i$$

```
#CODE Ch9_9_2.R
library(xts) ### provides functionality for working with time series
library(lubridate) ### manage dates
library(dplyr)
library(readxl) ### import excel file
#The excel file Prec_ER_daily.xlsx is opened, the structure of the
#newly created dataframe is analysed and the data are visualized.

setwd("~/Didattica/R_class_4/exercises/Ch9_statistical_inference")
Prec_ER_daily <- read_excel("OpenDataFiles/data/Prec_ER_daily.xlsx")

# Only data from the Cadriano station are analyzed
Prec_ER_daily_cadriano <- Prec_ER_daily[,c("date","cadriano")]
prec_cadriano<-Prec_ER_daily_cadriano$cadriano
str(Prec_ER_daily_cadriano)

#Daily cumulative curve
max_prec_value<-sum(Prec_ER_daily_cadriano$cadriano[1:365])

date<-as.Date(Prec_ER_daily_cadriano$date)
start <- date[1]
end <- date[365]
start
end
theDate
as.Date(theDate)
theDate<-start
theDate
vec_date<-vector()
precip<-vector()
prec_daily<-0
cumprec<-0
cumprecrel<-0
cumprec_vec<-vector()
cumprec_vec_rel<-vector()
i<-1

while (theDate <= end){
       vec_date[i]<-theDate
       prec_daily=prec_cadriano[i]
       cumprec<-cumprec + prec_cadriano[i]
       cumprec_vec[i]<-cumprec
       cumprecrel<-cumprec/max_prec_value
       cumprec_vec_rel[i]<-cumprecrel
       output<-c(i,prec_daily,cumprecrel)
```

```
        print(output)
        theDate <-theDate + days(1)
        i<-i+1
}

#plotting
plot(as.Date(vec_date),cumprec_vec_rel,type="l",
xlab="Time [day]",ylab="Cumulative precipitation [mm]")
```

Figure 9.7 depicts relative cumulative precipitation for the Cadriano experimental station.

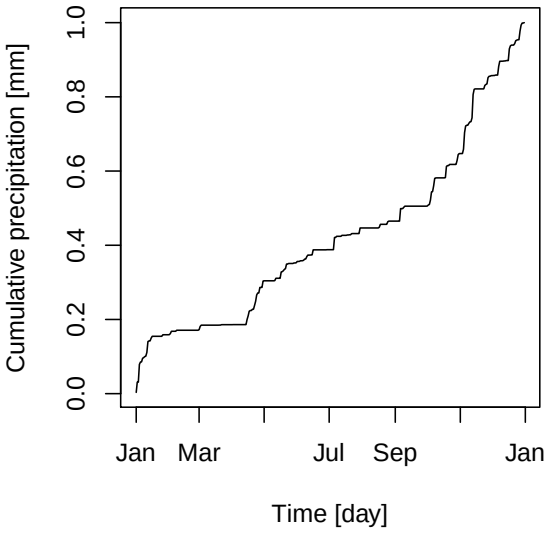


Fig. 9.7 Relative cumulative precipitation at the Cadriano (Bologna, Italy) experimental station for the year 1961.

9.3.3 Percentual Cumulative Frequency

The **percentual cumulative frequency** of a mode is

$$P_i = p_1 + p_2 + \cdots + p_i$$

where the data plotted above are expressed as percentage instead of fraction.

9.4 Measures of Central Tendency

The simplest statistical inference problem is **Point estimation**, where a single value (statistic) from the sample data to estimate a population parameter.

In general, define θ as a parameter of the population described by random variables X. X is unknown (the distribution is unknown), and θ is unknown. The quantity θ is called *parameter* (of interest) and it is a constant of the population. θ is a measure of the distribution of one or more character of the population, for instance the mean μ or the variance σ^2.

9.4.1 Means

The arithmetic mean is defined as the sum of the measurements divided by the total number of measurements. Usually, the *population mean* is denoted by the greek letter μ, while the *sample mean* is denoted by the symbol $\bar{x}$. The population mean is computed over the complete set of measurements, while the sample mean is computed over a subset sample of the measurements.

$$\bar{x} = \frac{\sum_i x_i}{n} \tag{9.1}$$

where $\bar{x}$ is the arithmetic mean, x_i is each individual measurement and n is the total number of measurements. In the example below, the vector `x` is defined, containing a number n of measurements and to the variable `am` (which stands for arithmetic mean) is assigned the output of the function `mean(x)`. The function `mean()` is an internal function of R, that computes the arithmetic mean. Note that the total number of elements (`n`) is read by the function as elements of the vectors. Finally, the result is printed on screen.

```
> x<- c(5.24, 5.55, 4.69, 4.39, 6.87, 5.15 ,
4.61, 5.20, 5.49, 4.81,2.74, 3.50, 5.19, 5.40,
3.81, 6.49, 6.34, 4.45, 5.10, 3.17)
> am<- mean(x)
> am
[1] 4.9095
>
```

Note that if the data have no values (NA), it must be specified in the mean as

```
> x<- c(5.24, 5.55, 4.69, 4.39, 6.87, 5.15 ,
4.61, 5.20, NA, 4.81,2.74, 3.50, 5.19, 5.40,
3.81, 6.49, 6.34, 4.45, NA, 3.17)
```

```
> am<- mean(x,na.rm=TRUE)
> am
[1] 4.9095
>
```

It means that the NA (not available data) are removed from the computation of the mean. This must be used also for other parameters like variance and standard deviation.

There are other means that can be used. Important ones are the **geometric**, the **harmonic** and the **logarithmic** means. The **geometric mean** is a mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values:

$$\left(\prod_{i=1}^{n} x_n\right)^{1/n} = \sqrt[n]{x_1 x_2 x_n} \tag{9.2}$$

This mean is used when the elements of the sample ranges several orders of magnitude and the arithmetic mean is not a representative value of central tendency. If we have two properties, the first one takes bigger numbers (from 300 to 450), while the second takes small numbers (from 6 to 3). In the first case it is an increase in 50%, while in the second case it is a decrease in 50%.

With the use of an arithmetic mean, the first one has a much higher weight, with the mean going from 153 to 227, because the first number is increasing but the second number, although it is decreasing, has little effect.

```
> x<-c(300,6)
> y<-c(450,3)
> mean(x)
[1] 153
> mean(y)
[1] 226.5
```

With the geometric mean the second variable is weighted more heavily, with the average going from 42.4 to 36.7.

```
> library(EnvStats)
> x<-c(300,6)
> y<-c(450,3)
> geoMean(x)
[1] 42.42
> geoMean(y)
[1] 36.74
```

In general the geometric mean is a more accurate estimator when the progression is multiplicative instead of additive. A geometrical analogy could be drawn if we plot the numbers on a linear scale for values that are close, and on a exponential curve. In the second case the geometric mean will return a more accurate representation of the distribution of numbers.

Another useful mean is the **harmonic mean**:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_n}} = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n} \right)^{-1} \tag{9.3}$$

Since the harmonic mean is based on reciprocal of numbers, it is useful when dealing with ratios and relationship among ratios. Other measures of position are the **median** and the **mode**.

The **median** is the value separating the higher half from the lower half of a data sample. It is the value in the middle of the distribution. So for odd number of elements it is (1,3,4,**6**,8,9,11). For even number of elements, it is the mean between the two middle number(1,3,4,**6,8**,9,11,13), therefore $6 + 8/2 = 7$. Therefore the median value is the central value, which is the value having half the observations smaller and half larger than it.

The **mode** is the value of y at which the frequency distribution is maximum. For symmetrical distributions, the median, mean and mode coincide.

The means are also referred to each class.

9.4.2 Means for classes

The arithmetic mean is a quantitative property that can be subdivided in classes:

$$\bar{x} = (1/n) \sum_{i=1}^{N} c_i n_i \tag{9.4}$$

where c_i is the central value of the i-th class and n_i is the absolute frequency. If $c_i = \bar{x}_i$, there are no approximations.

An example is presented, with the following classes: $[0, 5), [5, 10), [10, 15)$, therefore with $N = 3$. The absolute frequencies are n_i: $3, 5, 2$. The central values are $(5 + 0)/2 = 2.5, (5 + 10)/2 = 7.5, (10 + 15)/2 = 12.5$. The mean is $\approx [(3 \times 2.5) + (5 \times 7.5) + (2 \times 12.5)]/(3 + 5 + 2) = 7$. Given n data in N classes, of numerosity $n_1, \ldots, n_N$; x_{ij}, with i-th data of the j-th class.

The mean of the j-th class is:

$$\bar{x}_j = (1/n_j) \sum_{i=1}^{n_j} x_{ij} \; (j = 1, \ldots, N) \tag{9.5}$$

The mean of the n data is:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{N} \sum_{i=1}^{n_j} x_{ij} = \frac{1}{n} \sum_{j=1}^{N} n_j \bar{x}_j = \sum_{j=1}^{N} f_j \bar{x}_j$$

The difference (residuals) from the mean is $x_i - \bar{x}$, the sum of the residuals $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$ and $\sum_{i=1}^{n} (x_i - a)^2 =$ minimum for $a = \bar{x}$ namely, the mean is the value that is *closer* to all the observations.

9.4.3 Median

9.4.4 Mode

9.5 Measures of Variability

It is important to use indicators of variability in a distribution of data. The simplest but useful measurement is the **range**, which defines the interval between the minimum and maximum value of the distribution. Many different measures can be obtained by measuring the deviations $y - \bar{y}$. The first that would come to mind is the mean deviation. However, if deviations have opposite signs, the total mean deviation could be zero. Therefore, a possibility is to ignore the minus sign and compute the absolute values. However, a more easily interpreted function of the deviations is the **variance**, which is the sum of the squared deviations of the measurements from their mean.

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n - 1} \tag{9.6}$$

Also for the measures of variability, different symbols are used, where s^2 is the **sample variance** and σ^2 is the **population variance**. The use of $(n - 1)$ is not arbitrary, since it makes an unbiased estimator of the population variance. If we were to draw a very large number of samples, each of size n, from the population of interest and we compute s^2 for each sample, the average sample variance would equal the population variance σ^2. Had we have divided by n in the definition of the sample variance (s^2), the average sample variance computed from a large number of samples would be less than the population variance, hence s^2 would tend to underestimate σ^2. Ideally, the population variance σ^2, would be computed as

$$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n} \tag{9.7}$$

but often the entire population is unknown, so we replace the population mean with its best estimate that is the sample mean, as shown in eqn 9.6. The problem is that subtracting $\bar{y}$ from the sum makes the sum as small as it possibly could be. Since the value of $\bar{y}$ is centered around the distribution, while μ can be any value. Therefore, the population variance is likely to be larger than the sample variance because deviations of sample values from μ are likely to be larger than deviations from $\bar{y}$. For this reason, the sample variance is divided by $(n - 1)$ to allow for larger values of sample variance. It can be proved mathematically that using $(n - 1)$ allows for an unbiased estimation of population variance. However, if possible, it is generally more useful to know the mean and variance of the population rather than that of the sample.

The **sample coviariance** is a measure of the relationship between two paired datasets:

$$\mathrm{Cov}_{x,y} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{9.8}$$

9.6 Coefficient of variation

The coefficient if variation is a adimensional number:

$$CV = \frac{\sigma}{\overline{x}} \tag{9.9}$$

it can also be expressed in percentage:

$$CV = \frac{\sigma}{\overline{x}} \times 100 \tag{9.10}$$

it is used when all the values of the distribution are positive. It compares the variability among variables of different nature. For instance, the length and weight of a group of cats have been measured. The average length is 45 cm and the standard deviation is 5 cm $\Rightarrow CV_l \approx 11\%$. After weighting them, the mean is 6.5 Kg and the standard deviation is 1.5 Kg $\Rightarrow CV_p \approx 23\%$. Therefore, with respect to the mean, the weight of cats is more variable than their length.

9.7 Quantiles

Quantiles are cutting points that divide the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample.

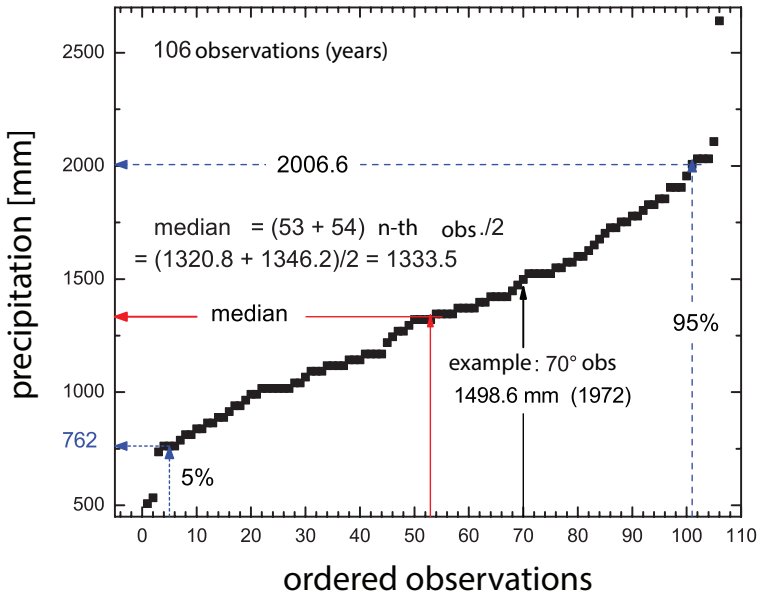
For instance quartiles are the three cut points that divide a dataset into four equal−sized groups (0−25, 25−50, 50−75 and 75−100%). A decile is a cut point that divide the distribution in ten intervals.

The precipitation data are ordered in incremental values as shown in the table below.

Table 9.5

	anno	inch	mm		ordinati (mm)
1	1873	80	2032	1	508
2	1874	40	1016	2	533.4
3	1875	65	1651	3	736.6
4	1876	46	1168.4	4	762
5	1877	68	1727.2	5	762
6	1878	32	812.8	6	762
7	1879	58	1473.2	7	787.4
8	1880	60	1524	8	812.8
9	1881	61	1549.4	9	812.8
10	1882	60	1524	10	838.2
...	...	...	...	...	...
97	1969	43	1092.2	97	1905
98	1970	80	2032	98	1905
99	1971	60	1524	99	1905
100	1972	59	1498.6	100	1955.8
101	1973	41	1041.4	101	2006.6
102	1974	67	1701.8	102	2032
103	1975	83	2108.2	103	2032
104	1976	56	1422.4	104	2032
105	1977	29	736.6	105	2108.2
106	1978	21	533.4	106	2641.6

The ordered data are plotted. The graph depicts the median value, computed as described above, the value of the 70^{th} observation, and the 5 and 95 % quantiles.



The instruction to order data is shown below using the function `sort()`.

```
NevadaPrec <- read.table("C:/Users/marco.bittelli.PERSONALE/
Documents/Didattica/R_class_3/exercises/descriptive_stat
/data/Nevada_prec.dat",sep = "", check.names = FALSE,
header = TRUE, na.strings = c("NA", "NAN"))

NevadaPrec$Prec_mm = NevadaPrec$Prec * 25.4

NevadaPrec$Year <- as.numeric(NevadaPrec$Year)

str(NevadaPrec)

#sorting data in increasing order
NevadaPrec$Prec_mm_sorted<- sort(NevadaPrec$Prec_mm,
decreasing = FALSE, na.last = NA)
NevadaPrec$Prec_mm_sorted
```

The quantiles are computed with the instruction below.

```
#compute quantiles
quantile(NevadaPrec$Prec_mm_sorted,probs = seq(0, 1, 0.05))
0% 5% 10% 15% 20% 25% 30% 35% ...
```

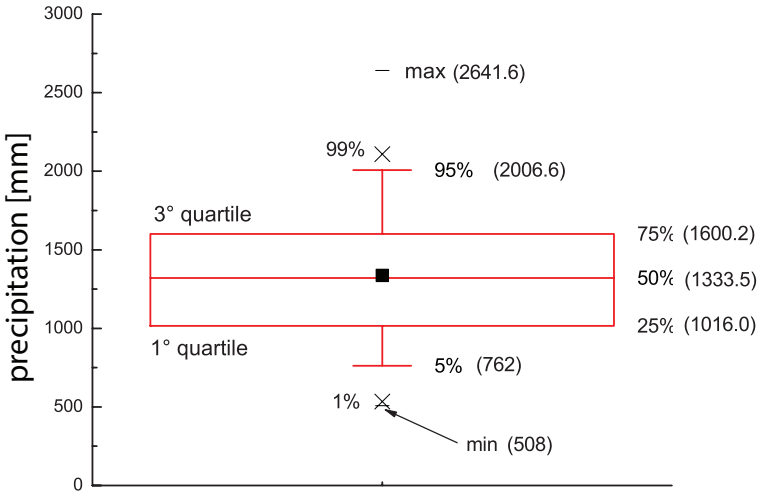
```
508.00 768.35 850.90 933.45 1016.00 1022.35 1092.20 1136.65 ...
```

9.8 The box plot

The box plot is a useful graph to represent the data. It provides information about the simmetry of the distribution and incorporates numerical measures of central tendency. The graph belows show the an *histogram* by classes (left) and a *box plot* on the right. The box plot provided information about the variability of the distribution as well. The box plot uses the median and **hinges** of a distribution. Hinges are very similar to quartiles, but depending on the distribution they may differ very slightly from the first and third quartiles. The instruction below shows how to plot the box plot of the precipitation data.

```
#plot the box plot
boxplot(NevadaPrec$Prec_mm_sorted)
```

The **box plot** (also called **skeletal box plot**) is constructed by drawing a box between the lower and upper quartiles, with a solid line drawn accross the box to locate the median. Usually a straight line is drawn to connect the box to the largest value and another straight line is drawn to connect the box the smallest value.



9.9 Exercises

1. Download daily precipitation from ARPAE web site Dexter 3. Import the data, format it to be parsed into dates and run frequency analysis
2. Separate the daily precipitation data in classes. Compute mean and standard deviation. Plot the cumulative graph and discuss the results.
3. Plot the box plot for the data.