Basic Statistics in R

(development version)

Fernando C.Barbi*

February 2017

Contents

*fcbarbi@gmail.com

1 Marginal, Conditional and Joint Probabilities

1.1 Definitions

The **probability** of an event A is the relative rate of its occurrence in the sample. Suppose that the outcomes of an event can be of two types and the number of each type in the sample are A and B.

We can calculate the relative occurrence of A by dividing the occurrences of A by the total number of events $A + B$, or $P(A) = A/(A + B)$. This probability is know as **marginal**: one dimension is chosen, $P(A)$, regardless of any other, its probability is calculated with no dependency on other events.

The **joint probability** is the most detailed information and combines all dimensions, it is noted as $P(A, B)$. For each state of A we have a particular probability that is combined with each and every state of dimension B.

The **conditional probability** is used when one or more dimensions are determined a priori, as a condition. It is used to answers questions such as "given A which is the probability of B ?" and they are noted as $P(B|A)$.

1.2 A visual representation in two dimensions

To illustrate with an example, suppose we model wages (w) as dependent on sex (s) and education (e). Each of these criteria can be understood as a "dimension" of the dataset so that we can tabulate the groups of individuals surveyed according to them.

Assume (a, b, c, d) are the number of surveyed individuals distributed as in table 1 and that the total number of individuals in the survey is $T = a + b + c + d$.

	s=Female	s=Male
e=High	a	c
e=Low	b	d

Table 1: Joint probability distribution

- Note that $p(s, e)$ returns a set of values (a discrete distribution) that is $(a/T, b/T, c/T, d/T)$ while $p(s = F, e = H)$, or $p(F, H)$, returns simply a scalar a/T.
- The marginal probability $p(e = H)$ is a scalar $(a+c)/T$ because $p(H) = p(H, F) + p(H, M)$.

- If we condition on High experience $p(s = F|e = H) = (a/T)/((a + c)/T) = a/(a + c)$.
- Check that $p(s = F, e = H) = p(s = F|e = H)p(e = H)$.
- Summing up all groups $a/T + b/T + c/T + d/T = 1$ results in 100% of the sample.
- The rule can be generalized $p(A, B, C) = p(A|B, C)p(B|C)p(C)$.

You can answer many questions with this joint distribution, such as:

1. Which is the average wage of women with high experience?

$$E(w|s = F, e = H)$$

2. Which is the average wage of women?

$$E(w|s = F) = E(w|s = F, e = H)p(e = H)+E(w|s = F, e = L)p(e = L)$$

3. Which is the average wage of highly experienced individuals?

$$E(w|e = H) = E(w|e = H, s = F)p(s = F)+E(w|e = H, s = M)p(s = M)$$

4. Which is the average wage in the sample? You can get it by vertical (sex) or by horizontal (experience) aggregation as in

$$E(w) = E(w|s = F)p(s = F) + E(w|s = M)p(s = M) \text{ or}$$
$$E(w) = E(w|e = H)p(e = H) + E(w|e = L)p(e = L)$$

1.3 Application.Wage gender gap analysis

To practice we use the `Wages` dataset from package `Ecdat`. There is data on experience (exp) measures in years of full-time work experience, a factor for marriage status (married is "yes" or "no"), a factor with levels (sex is "male" or "female"), years of education (ed), a factor for black individuals (black is "yes" or "no") and the logarithm of wage (lwage).

```
data(Wages,package="Ecdat")
> head(Wages)
  exp wks bluecol ind south smsa married  sex union ed black   lwage
1   3  32      no   0   yes   no     yes male    no  9    no 5.56068
2   4  43      no   0   yes   no     yes male    no  9    no 5.72031
3   5  40      no   0   yes   no     yes male    no  9    no 5.99645
...
```

If you prefer to work with numerical dummies you can easily convert factors. To create a new column `d_black`

```
Wages$d_black <- ifelse(Wages$black=="no",0,1)
```

To get the average wage for male workers

```
mean(Wages[Wages$sex=="male",]$lwage)
```

We can add another condition to get the average wage for male workers that have at least 12 years of education

```
mean(Wages[Wages$sex=="male" & Wages$ed>12,]$lwage)
```

1.4 Data tables to handle big databases

Using `data.table` the commands get shorter (and execution is much faster...)

```
require(data.table)
Wages <- as.data.table(Wages)
# create the dummy with the d_ prefix
Wages$d_black <- 0
Wages[black=="yes",d_black:=1]
mean(Wages[sex=="female" & d_black==1 & ed>12,]$lwage)
```

Now try to answer these questions and assume that high educated individuals have 12 or more years of education (they are in the upper quantile):

1. Do married women study less than unmarried women?

   ```
   mean(Wages[sex=="female" & married=="yes",]$ed)
   mean(Wages[sex=="female" & married=="no",]$ed)
   ```

2. Is there a gender wage gap among highly educated individuals?

   ```
   mean(Wages[sex=="male"   & ed>=12,]$lwage)
   mean(Wages[sex=="female" & ed>=12,]$lwage)
   ```

3. Is the wage wage gender gap relatively bigger in the South than in the North?

```
wMN <- mean(Wages[sex=="male" & south=="no",]$lwage)
wMS <- mean(Wages[sex=="male" & south=="yes",]$lwage)
wFN <- mean(Wages[sex=="female" & south=="no",]$lwage)
wFS <- mean(Wages[sex=="female" & south=="yes",]$lwage)
(wMN-wFN)/wFN # relative wage gender gap in the North
(wMS-wFS)/wFS # relative wage gender gap in the South
```

4. Do more educated workers' have more homogeneous wages than the less educated workers?

   ```
   var(Wages[ed>12,]$lwage)
   var(Wages[ed<12,]$lwage)
   ```

To visualize aggregate data you can build a table with the conditional expectations for wages by sex, and sex and color

```
aggregate(Wages$lwage,list(Wages$sex),mean)
aggregate(Wages$lwage,list(Wages$sex,Wages$black),mean)
```

One last comment: is this sample representative of the total population?

```
> table(Wages$sex,Wages$married)
           no  yes
  female  457   12
  male    316 3380
```

Women are *under-represented* with 469 individuals when compared to 3696 male subjects.It is well-known that the number of individuals from both sexes in human populations are very close, with a slight difference in favor of female individuals. This can be corrected if the analyst *resample* for 469 male individuals at random. But there is another issue: it appears that women in this sample tend to be unmarried while the majority of men are married. This may impact the results if there is a systematic practice to pay smaller wages to married women.

2 Bayes Theorem

The Bayes Theorem states that the joint distribution $f(X,Y)$ is the product of the conditional $f(X|Y)$ by the marginal probability $f(Y)$ and that these can be interchanged

$$f(X,Y) = f(X|Y)f(Y) = f(Y|X)f(X) = f(X \cap Y)$$

Suppose we define the probability "universe" C, $p(C) = 1$, and segment it in k groups C_1,...,C_k that do not overlap so that

$$p(C) = p(C \cap C_1) + \cdots + p(C \cap C_k) = p(C_1 \cup C_2 \cup \ldots C_k)$$

The Bayes theorem states that

$$p(C \cap C_j) = p(C_j)p(C|C_j) \quad \vee j = 1, ..., k$$

We can state the **law of total probability** as

$$p(C) = p(C_1)p(C|C_1) + \cdots + p(C_k)p(C|C_k) = \sum_{i=1}^{k} p(C_i)p(C|C_i)$$

Now we can restate the Bayes theorem in more general terms

$$p(C_j|C) = \frac{p(C \cap C_j)}{p(C)} = \frac{p(C_j)p(C|C_j)}{\sum_{i=1}^{k} p(C_i)p(C|C_i)}$$

2.1 *Application.* A typical Bayes problem

Suppose there are two bowls C_1 and C_2 with red (R) and blue (B) balls in each so that C_1 has 3 red and 7 blue balls while C_2 has 8 red and 2 blue balls. Given that one randomly selected ball is red which is the probability that it came from C_1?

Tip: This classic Bayes problem asks for $p(C_1|R)$ and it can be easily solved with this joint distribution table ($T = 3 + 7 + 8 + 2 = 20$):

	C_1	C_2
Red	3	8
Blue	7	2

Can you complete the exercise?

3 Distribution and Density function

3.1 Density, cumulative and quantile function

- The **probability density functions** (aka. PDF) $f(x)$ is defined as

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \quad \text{so that} \quad f(x) \geq 0 \vee x$$

- The **(cumulative) distribution function** $F(x)$ (aka. CDF) is defined as the area under the density curve

$$F(x) = \int_{-\infty}^{x} f(x)dx \quad \text{or} \quad \frac{dF(x)}{dx} = f(x)$$

- The **quantile function** is the inverse of the distribution function, $q(x) = F^{-1}(x)$. Note that no matter the distribution the quantile function is always distributed as $q(x) \sim U(0,1)$ where $U(0,1)$ is the uniform distribution between 0 and 1, that will be seen later.

- The **probability mass function** (aka. PMF) of the random variable X to be in the interval (a,b) is a scalar given by

$$p(a < X < b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$$

- For a random variable $f(x)$, $X \sim f(x)$, the **expected value** for continuous and discrete distributions are

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \quad \text{and} \quad E(X) = \sum_{i=1}^{N} x_i f(x_i)$$

The intuition for this result is that each observation x_i is weighted according to its probability $f(x_i)$ to get to the expected value.

4 Useful Distributions

The convention in R is to name the distribution X as `dX()` for the density, `pX()` for the distribution function, `qX()` for the quantile function and `rX()` for the random generator.

4.1 Uniform

The Uniform distribution density function for $x \sim U(a, b)$ is given by

$$f(x) = \frac{1}{b-a} \text{ for } x \in (a, b)$$
$$f(x) = 0 \text{ elsewhere}$$

To generate 10 random values from $a = 0$ to $b = 1$ use

```
runif(n, min = 0, max = 1)
```

Note that if you need 10 integer values, 0 or 1, you have to use

```
sample(c(0,1),10,replace=TRUE)
```

The `replace=TRUE` parameter tells the function to replace the draw to the sample because we have only 2 observations and we ask for 10 draws. If the sample is bigger than the number of draws to extract there is no need for replacement.

4.2 Normal or Gaussian

The Normal distribution density function for $x \sim N(\mu, \sigma^2)$ is given by

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The key moments are: mean $E(x) = \mu$ and variance $Var(x) = \sigma^2$.

Any Gaussian distribution $X \sim N(\mu, \sigma^2)$ can be converted to a standard distribution $Z \sim N(0, 1)$ by doing

$$Z = \frac{X - \mu}{\sigma}$$

The standard **R** package offers `scale()` to perform this mapping that we reproduce with function `standard()`. We print results from both for comparison

```
x <- rnorm(1e3,mean=83,sd=12)
standard <- function(x) { (x-mean(x))/sd(x) }
z1 <- standard(x)
z2 <- scale(x)
head(cbind(z1,z2))
```

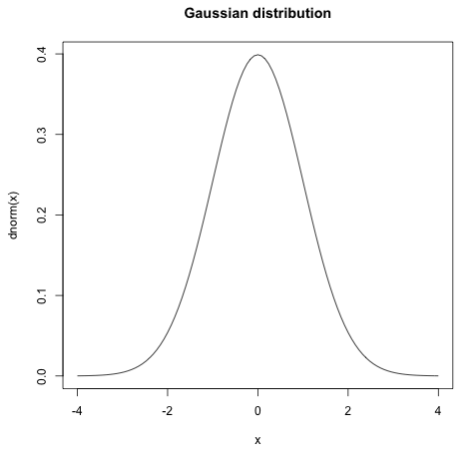
*Application.*Calculate the values of a standard normal variable so that the area under the Gaussian density curve represents 95% of the total area. Assume symmetry so that there is 2.5% of the area left out on each side of the curve.

```
qnorm(0.025) # -1.959964
qnorm(1-0.025) # 1.959964
pnorm(1.96)-pnorm(-1.96) # 0.9500042
```

The critical values are for the 95% CI are -1.96 and 1.96. We will use this later on. Now we draw the density curve for the Gaussian distribution with

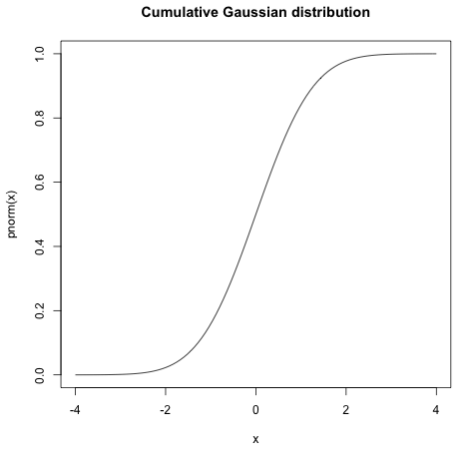
```
# plots the normal density curve between -4 and 4
curve(dnorm(x),-4,4,main="Gaussian distribution")
```

Figure 1: Gaussian distribution




```
# plots the cumulative probability curve between -4 and 4
curve(pnorm(x),-4,4,main="Cumulative Gaussian distribution")
```

Figure 2: Cumulative Gaussian distribution



4.3 Student-t

The Student-t distribution is derived from the Gaussian. Assume N has a standard normal distribution $N \sim N(0,1)$ and Q has a chi-square distribution with r degrees of freedom, $Q \sim \chi^2(r)$, we can define T with a Student-t distribution given by

$$T = \frac{N}{\sqrt{Q/r}}$$

The key moments for $x \sim t(r)$ are: mean $E(x) = 0$ for $r > 1$, variance $V(x) = r/(r-2)$ for $r > 2$, skewness $S(x) = 0$ for $r > 3$, kurtosis $K(x) = 3+6/(r-4)$ for $r > 4$. These conditions are imposed to assure the existence of each moment.

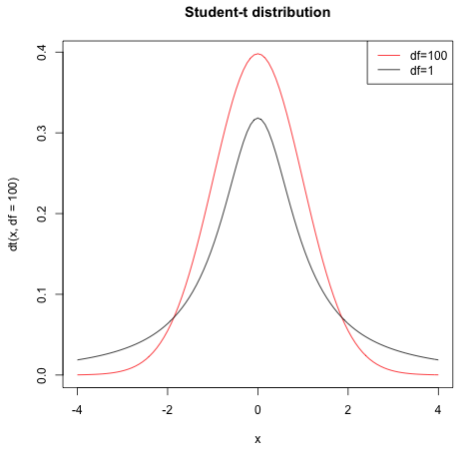
To find the critical values for a 95% confidence interval for this symmetric distribution we proceed as we did before.

```
qt(0.025,df=100)   # -1.983972
qt(1-0.025,df=100) # 1.983972
pt(1.98,df=100)-pt(-1.98,df=100) # 0.9495484
```

To compare Student distributions with different degrees of freedom we plot figure 3

```
curve(dt(x,df=100),-4,4,add=F,col="red",main="Student-t")
curve(dt(x,df=1),-4,4,add=T,col="black",ylab="")
```

Figure 3: Student t distribution



4.4 Chi-square

The Chi-square distribution with r degrees of freedom is derived as the sum of r square standard normal $N(0,1)$ distributions, noted as $Q \sim \chi^2(r)$, distributed as

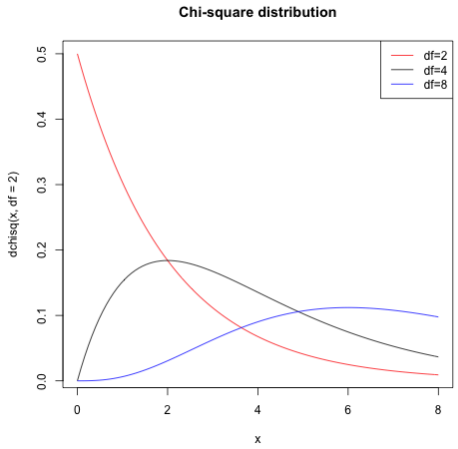
$$Q = \sum_{i=1}^{r} N_i^2$$

```
qchisq(0.975,df=10) # 20.48318
qchisq(1-0.975,df=10) # 3.246973
pchisq(20.5,df=10)-pchisq(3.25,df=10) # 0.95005
```

We plot different degrees of freedom curves for comparison as in figure 4

```
curve(dchisq(x,df=2),0,8,add=F,col="red",main="Chi-square")
curve(dchisq(x,df=4),0,8,add=T,col="black")
curve(dchisq(x,df=8),0,8,add=T,col="blue")
```

Figure 4: Chi-Square distribution



4.5 F Distribution

The F distribution is composed as a ratio of Chi-square distributions: if Q_1 and Q_2 are independent Chi-square variables with r_1 and r_2 degrees of freedom, $Q_1 \sim \chi(r_1)$ and $Q_2 \sim \chi(r_2)$, then

$$F = \frac{Q_1/r_1}{Q_2/r_2}$$

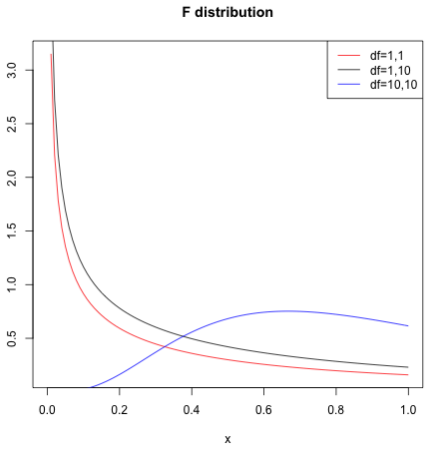
```
qf(0.975,  df1=10,df2=10) # 3.716792
qf(1-0.975,df1=10,df2=10) # 0.2690492
pf(3.71,df1=10,df2=10)-pf(0.26,df1=10,df2=10) # 0.9525199
```

We plot different degrees of freedom curves for comparison as in figure 5

```
curve(df(x,df1=1,df2=1),add=F,col="red",main="F distribution",ylab="")
curve(df(x,df1=1,df2=10),add=T,col="black",ylab="")
curve(df(x,df1=10,df2=10),add=T,col="blue",ylab="")
```

Note that values are always positive and the distribution is not symmetric.
It is used to compare models, as will be explain later in section 11.

Figure 5: F distribution



5 Distribution Moments

5.1 Mean, median, mode

Some population moments to keep in mind are: mean (E), variance (Var), covariance (Cov), skewness (S) and kurtosis (K)

$$E(x) = \frac{1}{T}\sum_{i=1}^{T} x_i = \bar{x} \qquad S(x) = \frac{E(x_i - \bar{x})^3}{\sigma^3} \qquad K(x) = \frac{E(x_i - \bar{x})^4}{\sigma^4}$$

$$Var(x) = \frac{1}{T}\sum_{i=1}^{T}(x_i - \bar{x})^2 = \sigma^2 \quad Cov(x,y) = \frac{1}{T}\sum_{i=1}^{T}(x_i - \bar{x})(y_i - \bar{y}) = \sigma_{xy}$$

Note that covariances are symmetric, $\sigma_{xy} = \sigma_{yx}$. The sample variance $s_{\mathrm{b}}^2(x)$ requires a degrees-of-freedom correction to become unbiased $s^2(x)$

$$s^2(x) = \frac{T}{T-1}s_{\mathrm{b}}^2 = \frac{1}{T-1}\sum_{i=1}^{T}(x_i - \bar{x})^2$$

To show it we calculate the bias of the uncorrected estimator s_{b}^2

$$\begin{aligned}
\mathrm{E}\left[\sigma^2 - s_{\mathrm{b}}^2\right] &= E\left[\frac{1}{T}\sum_{i=1}^{T}(x_i-\mu)^2 - \frac{1}{T}\sum_{i=1}^{T}(x_i-\overline{x})^2\right] \\
&= \frac{1}{T}\,\mathrm{E}\left[\sum_{i=1}^{T}\left((x_i^2 - 2x_i\mu + \mu^2) - (x_i^2 - 2x_i\overline{x} + \overline{x}^2)\right)\right] \\
&= \mathrm{E}\left[\mu^2 - 2\overline{x}\mu + \overline{x}^2\right] \\
&= \mathrm{E}\left[(\overline{x}-\mu)^2\right] = \mathrm{Var}(\overline{x}) = \frac{\sigma^2}{T} \\
\mathrm{E}\left[s_{\mathrm{b}}^2\right] &= \sigma^2 - \frac{\sigma^2}{T} = \frac{T-1}{T}\sigma^2
\end{aligned}$$

*Application.*Generate a normal random sample with 100 observations and calculate mean, standard error, skewness and kurtosis.

```
N <- 100
x <- rnorm(N)
mu <- sum(x)/N; mu
se <- sqrt(sum((x-mu)^2)/(N-1));se # standard error
sd(x) # compare with se
skew <- sum((x-mu)^3)/(N*se^3); skew
kurt <- sum((x-mu)^4)/(N*se^4); kurt
# To check: install.packages("timeDate")
timeDate::skewness(x)
timeDate::kurtosis(x,method="moment")
```

The **median** is the value that divides the distribution in two samples of the same size. If the distribution is symmetric (normal and t-student) the median is equal to the mean. *Application.* Test that the median is more robust (varies less) to outliers than the mean.

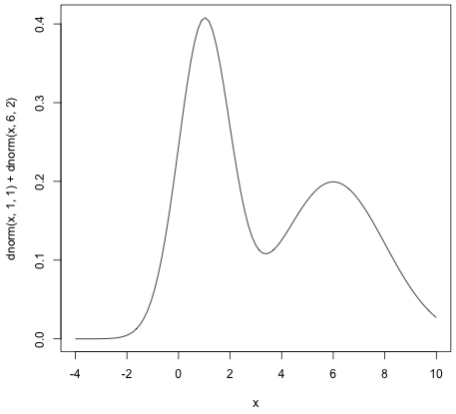
```
set.seed(222)
x1 <- rnorm(100)
x2 <- x1
x2[10] <- -100
x2[90] <- +200
cat("mean varied ",abs(mean(x2)-mean(x1))) # 1.001038
cat("median varied ",abs(median(x2)-median(x1))) # 0
```

The **mode** is any value that occurs most frequently in the set. In a *multi-modal* distribution there are more than one mode: the mixing of two normal distributions suggests a "sugarloaf"[1] profile

[1] `https://en.wikipedia.org/wiki/Sugarloaf_Mountain`

```
curve(dnorm(x,1,1)+dnorm(x,6,2),-4,10)
```

Figure 6: Multimodal distribution



5.2 Correlation

The **Pearson correlation** is measure of the strength of association of two variables so that $\rho_{XY} \in (-1, +1)$ that makes them comparable

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

The statistical significance of the Pearson correlation coefficient ρ obtained from a N observations sample can be tested using a Student t distribution, the statistic is

$$t = \frac{\rho}{\sqrt{\frac{1-\rho^2}{N-2}}} \sim t_{df=N-2}$$

```
set.seed(222)
X <- rnorm(1e3)
Y1 <- 0.005*X
Y2 <- 0.005*X + rnorm(1e3)*0.5

rho1 <- cor(X,Y1); rho1
pval1 <- cor.test(X,Y1)$p.value; pval1

rho2 <- cor(X,Y2); rho2
pval2 <- cor.test(X,Y2)$p.value; pval2
```

The null hypothesis is $H0 : \rho = 0$ for the test that shows that $\rho_1 = 1$ (significant) and $\rho_2 = 0.04625252$ (not significant).

The **standard deviation** (SD) is a measure of dispersion of the data in the DGP (data gerenating process aka. the "theoretical" distribution) and is usually noted σ. Variance is the square of SD, σ^2.

Once a sample is collected the dispersion of the sample is identified as the **standard error** (SE) and noted $\hat{\sigma}$ as it is an estimator of the true parameter σ. The sample is an "empirical" distribution and there is no obligation that it comes from a parametric distribution.

The **standard error of the mean** (SEM) is the standard deviation of the mean of random samples drawn from the original population, $\overline{X}$, and is equal to $\sigma/\sqrt{T}$ where T is the size of the sample. The SEM quantifies the uncertainty in the estimation of the mean while SE indicates dispersion of the data from the mean.

The estimator $\overline{X}$ for the true parameters for the mean μ of a sample with T observations is

$$\overline{X} = \frac{1}{T}\sum_{i=1}^{T} x_i$$

The variance for $\overline{X}$ is σ^2/T. To check this just assume the true variance to be $Var(x_i) = \sigma^2$ and that x_i are independently drawn so that $Var(x_1 + \cdots + x_T) = T\sigma^2$ (note that there are no covariance terms).

$$\begin{aligned} Var(\overline{X}) &= Var\left(\frac{x_1 + \cdots + x_T}{T}\right) = \frac{Var(x_1 + \cdots + x_T)}{T^2} \\ &= \frac{T\sigma^2}{T^2} = \frac{\sigma^2}{T} \end{aligned}$$

Therefore, if $X \sim N(\mu, \sigma^2)$ then $\overline{X} \sim N(\mu, \sigma^2/T)$.

```
# theoretical distribution is N(mu,sigma^2)
N <- 1e3
mu <- 20      # theoretical distribution mean
sigma <- 5    # theoretical distribution standard deviation
set.seed(222) # generate empirical distributions
x   <- replicate( 1e4, mean( rnorm(N, mean=mu, sd=sigma) ) )
se  <- sd(x);se          #  0.157952
sem <- sigma/sqrt(N);sem #  0.158113
mean(x)                  # 20.00055
```

Another frequently used measure of dispersion is the **Coefficient of Variation** of X defined as $CV(X) = V(X)/E(X)^2$. It scales variance in terms of the mean making two random variables directly comparable. The example below shows that $x1$ varies significantly more than $x2$.

```
set.seed(222)
x1 <- rnorm(N,2,4)
x2 <- rnorm(N,2*1e4,1e3)
cv1 <- var(x1)/mean(x1)^2; cv1 # 4.204661
cv2 <- var(x2)/mean(x2)^2; cv2 # 0.002460
cv1/cv2 #  1709.196
```

5.3 Application.Table of sample statistics

To capture the *metadata* (data on data) for the study sample we can use several tools. In interactive mode, it's quick to summarize data

```
require(Ecdat)
data(Wages)
summary(Wages)
```

but this is usually not enough, more moments are needed so we can use

```
require(psych)
describe(Wages)
```

As an alternative to loading the library, once it is installed it can be invoked by using `::` as in

```
Wages_stats <- psych::describe(Wages,range=FALSE)
```

To save the metada in a CSV file to export to other softwares

```
write.csv(Wages_stats,"Wages_stats.csv")
```

To save the table in Excel format you have to install `openxlsx` package

```
require(openxlsx)
write.xlsx(Wages_stats, file = "Wages_stats.xlsx", row.names=TRUE)
```

To save the table in LaTeXformat, we need package `xtable` installed

```
print(xtable::xtable(Wages_stats),file="Wages.tex",type="latex")
```

The file "Wages.tex" contains the code to generate table 2. You should have a look at the package `psych` documentation. We just draw your attention to columns with * to indicate these are factors, variables expressing categories that may have to be treated into numeric before use. Also note the columns `sd` for standard deviation and `se` for standard-error of the mean.

Table 2: Descriptive Statistics for Wages

	vars	n	mean	sd	skew	kurtosis	se
exp	1	4165.00	19.85	10.97	0.40	-0.93	0.17
wks	2	4165.00	46.81	5.13	-2.89	11.91	0.08
bluecol*	3	4165.00	1.51	0.50	-0.04	-2.00	0.01
ind	4	4165.00	0.40	0.49	0.43	-1.82	0.01
south*	5	4165.00	1.29	0.45	0.92	-1.15	0.01
smsa*	6	4165.00	1.65	0.48	-0.65	-1.58	0.01
married*	7	4165.00	1.81	0.39	-1.62	0.61	0.01
sex*	8	4165.00	1.89	0.32	-2.45	4.00	0.00
union*	9	4165.00	1.36	0.48	0.57	-1.68	0.01
ed	10	4165.00	12.85	2.79	-0.26	-0.29	0.04
black*	11	4165.00	1.07	0.26	3.30	8.91	0.00
lwage	12	4165.00	6.68	0.46	-0.04	0.52	0.01

6 Independent and identically distributed samples

6.1 Independent random variables

The notion of **independence** of two random variables X and Y is related to the fact that they come from different Data Generating Processes (DGPs). The intuition is that you can't understand X by studying Y because there is no information on X in Y. We use this concept when formulating the **strict exogeneity** hypothesis on the linear model $y = \beta X + \varepsilon$, that is: $E(X|\varepsilon) = 0$ is a necessary condition to get a consistent estimator for β. In practise this is a very strong condition that can be relaxed to a **weak exogeneity** condition represented by the absence of correlation, $E(X\varepsilon) = 0$.

We say that two random variables are **identically distributed** if they come from DGP's with the same moments (means and variance). In practice, though, truly **independent and identically distributed** (iid) samples are difficult to generate. In the lab we can try to generate them by running separate randomization processes but the final outcome depends on the quality of the sampler inside the sofware. **R** uses a very good one but even so we find some correlation in samples generated from very different DGPs.

If X and Y "come from" or "are derived from" the same DGP, clearly they are not independent.

```
set.seed(222)
XY <- rnorm(1e6)
X <- sample(XY,1e3)
Y <- sample(XY,1e3)
t.test(X,Y) # p-value = 0.211
```

```
cor(X,Y)    # 0.04612209
```

We will show later how to test for the same origin, for now just assume that the t test does not reject that X and Y are from the same DGP. Now if we run different randomization processes (notice that both mean and variance are different in the example below) we should get correlation zero but we don't. We can only reject by t-testing the samples that they come different DGP's.

```
set.seed(222)
X <- rnorm(1e3,mean=5,sd=1.5)
Y <- rnorm(1e3,mean=2,sd=0.5)
t.test(X,Y) # p-value < 2.2e-16
cor(X,Y)    # 0.03625168
```

Of course, we should also study the statistical significance (a concept we later explain when talking about tests) of the correlations we just found, but this is not critical for the point we want to make here: even taking all precautions to generate different, independent samples, we get pretty close correlations in the two examples above. This is to show that iid samples are more of a theoretical hypothesis than a practically achievable target.

In terms of notation, $iid(\mu, \sigma^2)$ means that the sample can come from any distribution, assumed to be iid, as long as it has mean μ and variance σ^2. This is a relaxed condition on the usual $N(\mu, \sigma^2)$. As a general rule, the less restrictive are your hypothesis the more general (and useful) are the results.

6.2 Conditional independence

To say that "X is independent of Y", $X \perp Y$, means that there is no information on X that can be obtained from Y. To say that "X is independent of Y given Z", $X \perp Y|Z$, means that all the information about X contained in Y comes from Z, so X and Z are not independent and if you take out all that Z brings to Y there is nothing left to explain X.

```
N <- 1e3
set.seed(222)
Z <- rnorm(N)
X <- 2*Z + rnorm(N,sd=0.1)
Y <- 0.5*Z + rnorm(N)
cor(X,Z)         # 0.9987634
cor(X,Y)         # 0.4582502
a <- cor(Y,Z); a # 0.4593239
cor(Y-a*Z,X)     # 0.0550044
```

To construct X we use Z and introduce a little noise to break the perfect collinearity. Because of the noise a will not be 0.5 but should be very close to it. The notion to retain here is that if you omit Z it seems that Y explains X but this happens through a **common factor** Z. One overestimates the correlation between X and Y when Z is omitted; this is the **omitted variable bias** that plagues incomplete econometric models. The code `cor(Y-a*Z,X)` expresses the idea that *controlling* for Z the correlation between Y and X is non-existent.

```
m1 <- lm(X~Y)
m2 <- lm(X~Y+Z)
stargazer::stargazer(m1,m2,out="omit.tex")
```

Table 3: Omitted variable bias

	Dependent variable:	
	X	
	(1)	(2)
Y	0.820***	−0.001
	(0.050)	(0.003)
Z		2.004***
		(0.004)
Constant	0.026	−0.001
	(0.056)	(0.003)
Observations	1,000	1,000
R^2	0.210	0.998
Adjusted R^2	0.209	0.998
Note:	*p<0.1; **p<0.05; ***p<0.01	

In table 3 model (1) presents a significant correlation that is not true, it is called **spurious**, and (2) shows the true relationship, with estimation of the Z coefficient at 2.004 that is very close to the true value of 2. The fitting of the second model as measured by the adjusted R^2 is also much better than the first. Note that the coefficient for Y in the second model is not statistically significant and by putting models side by side we can visualize the biasing process of this variable.

7 Expected Value, Variance and Covariance

7.1 Expected Value

Some useful expected value properties are

$$\begin{aligned} E(c) &= c \\ E(aX) &= aE(X) \\ E(aX + bY) &= aE(X) + bE(Y) \end{aligned}$$

7.2 Some useful inequalities: Jensen, Markov and Chebyshev

1. The **Jensen's inequality** provides a way to characterize a function `g()` as *concave* or *convex* by testing the relation between $E(g(x))$ and $g(E(x))$. if `g()` if concave ($g''(x) \leq 0$) we have

   $$E(g(x)) \leq g(E(x))$$

   The inequality changes sign for the convex relationship. We can test that `log()` is a concave function

   ```
   x <- runif(1e2,1,1000)
   mean( log(x) ) # 5.912128
   log( mean(x) ) # 6.190685
   ```

2. The **Markov's inequality** provides a lower bound for a probability irrespective of the distribution

   $$p(X > a) \leq \frac{E|X|}{a} \quad \text{for } \vee\, a > 0$$

   Note that $p(X \leq a)$ is `pnorm(a)` and $p(X \geq a)$ is `1-pnorm(a)` so we verify

   ```
   set.seed(222); x <- rnorm(1e3)
   a <- 2
   1-pnorm(a)      # 0.0227501
   mean(abs(x))/a # 0.3979126
   ```

3. The **Chebyshev's inequality** extends Markov's and assumes a Gaussian distribution to provide smaller intervals

   $$p(|X - \mu| > a) \leq \frac{\sigma^2}{a^2} \quad \text{if } X \sim N(\mu, \sigma^2)$$

```
1-pnorm(a)  # 0.0227501
var(x)/a^2  # 0.2466364
```

We verify the inequality but most important is to note that the value of σ^2/a^2 is always smaller than $E|X|/a$. You can try resampling many times to confirm that. We might say that Chebyshev's provides a more efficient boundary than Markov's.

7.3 Variance and Covariance

Variance and covariance can be written as

$$\text{Var}(X) = E(X^2) - E(X)^2 \quad \text{and} \quad \text{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

This is easily checked if you "open" the expression

$$\frac{1}{T}\sum_{i=1}^{T}(x_i - \bar{x})^2 = \frac{1}{T}\sum_{i=1}^{T}(x_i^2 + \bar{x}^2 - 2x_i\bar{x}) = E(X^2) + E(X)^2 - 2E(X)^2 = E(X^2) - E(X)^2$$

An alternate form for the covariance is

$$\begin{aligned}\text{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y \\ &= E(XY) - \mu_X\mu_Y\end{aligned}$$

If X and Y are centered, $\mu_X = \mu_Y = 0$, than $\text{Cov}(X,Y) = E(XY)$.

Some useful variance properties:

$$\begin{aligned}V(c) &= 0 \\ V(aX) &= a^2V(X) \\ V(aX+bY) &= a^2V(X) + b^2V(Y) + 2ab\,\text{Cov}(X,Y)\end{aligned}$$

where a, b and c are constants, and X and Y are random variables with any distribution.

The **Law of Iterated Expectation** (LIE) tells the consequences of constraining to smaller information sets (through unconditional expected value)

$$E(E(X \mid Y)) = E(E(X) \mid Y) = E(X)$$

We address the concept of conditional expectations ahead.

Finally, another useful result is the **Law of Total Variance** (LTV)

$$Var(X) = E(V(X \mid Y)) + V(E(X \mid Y))$$

This decomposition is used to explain the "Bias-Variance trade-off" in econometric models. The idea is that the more complex is a model (with more variables) while bias is reduced variance increases, that's the **trade-off between bias and variance**.

8 Confidence Intervals

A 95% confidence interval (CI) provides the boundaries (aka. the *critical values*) for 95 out of 100 draws from the sample. The difference between 1 (certainty) and the confidence interval is known as the **significance level** (α = 1-CI). The confidence interval for a $\alpha = 5\%$ significance of a normally distributed random variable can be constructed as

$$0.95 = P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{T}} \leq 1.96\right) \tag{1}$$

$$= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{T}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{T}}\right) \tag{2}$$

The boundaries of the CI are known as the **critical values** and can be easily calculated with

```
qnorm(0.05/2) # -1.959964
qnorm(1-0.05/2) # 1.959964
```

8.1 Application.Function to calculate confidence intervals

The target is to build a function to calculate the CI that receives either the moments of a theoretical distribution or the observations of an empirical one.

We build a flexible function `ci()` to calculate intervals for a Gaussian distribution that can be called with an empiric distribution or with the first two moments (mean and standard deviation) of a theoretical Gaussian distribution. By the way, keep in mind that these boundaries are as good as the assumption that the empirical distribution follows a Gaussian.

```
ci <- function( size=1, mean=0, se=1, alpha=.05, x=NULL ){
 if (!is.null(x)) {
   mean = mean(x,na.rm=TRUE)
   se   = sd(x,na.rm=TRUE)
   size = sum(!is.na(x))
 }
 sem <- se/sqrt(size)
 cilo <- mean + qnorm(alpha/2)*sem
 cihi <- mean + qnorm(1-alpha/2)*sem
 c(cilo,mean,cihi)
}
set.seed(222); X <- rnorm(1e2,4,1)
ci(x=X)      # 3.834210 4.021980 4.209749
ci(1e2,4,4)  # 3.216014 4.000000 4.783986
```

We use the code `sum(!is.na(x))` instead of the more usual `length(x)` to measure sample size as it discards all `NA` observations, the effective size is calculated by summing up all the non missing observations. It's a subtle difference that prevents wrong estimations if the sample happens to have missing observations.

We can test the impact of the sample size in the boundaries of the interval: the formula indicates that the bigger the sample the smaller is the dispersion of the mean. We test with a sample with 100 observations that $10,000$.

```
Xci <- ci(x=X)
sum( Xci[1]>=X & X<=Xci[3] ) # 38/100 = 38%
# increase sample size
set.seed(222); Y <- rnorm(1e4,4,1)
Yci <- ci(x=Y)
sum( Yci[1]>=Y & Y<=Yci[3] ) # 4926/10^4 = 49.26%
```

The small sample led to only 38% of the values in the interval while in the larger this performance grew to 49.26%. When the sample grows to 10^8 the inclusion rate grows very little to 49.98%, so there seems to be convergence. Can you explain why? Tip: Calculate `pnorm(4,mean=4,sd=1)`.

Note that we usually assume a **two-sided** test, so that 5% is actually split in 2 intervals of 2.5% that correspond to the area under on each side of the density curve where the null hypothesis is rejected. We explain more about two-sided tests later when we discuss hypothesis testing.

9 Treatment of Outliers

Outliers could be intuitively defined as observations that are "significantly different" than the rest of the sample. If you try to create ane econometric model with the outliers in the sample you may get poor results since they tend to increase the variation in data, a phenomenon called *heterocedasticity*.

9.1 Outliers in cross-section data

For cross-section data it is possible to "clean" the dataset by taking out observations that are too discrepant. A word of caution is in order. This process should not be done without a prior analysis of the origin and structure of the outliers: if the analyst observes that outliers have a common pattern (eg. they all tend to happen on one side of the distribution) further investigation should be conducted on the data collection process before deciding on exclusion. Suppose the outliers all come from a surveyor that had a mis-calibrated device, it might be better to find the average bias this surveyor introduced and adjust the corresponding observation subtracting the bias and preserving the outliers in the sample.

If the decision is for the exclusion of outliers researchers usually target quantiles 1% and 99% in cross-section samples. This can be done through a *non-parametric* approach (that does not assume any underlying distribution for data) using function `quantile()`. To calculate the cutoff value for the quantiles of any empirical distribution

```
set.seed(222)
x <- rnorm(1e3)
xlo <- quantile(x,0.01) # -2.161609
xhi <- quantile(x,0.99) # +2.282738
sum( x<=xlo | x>=xhi ) # 20
```

There were 20 observations on the extreme quantiles out of 1000 observations of a random distribution. Could this loss be predicted by theory? Certainly, just consider dropping 2% of the sample, in this case $0.02 * 1000 = 20$. A final note: these cutoff values are not symmetric, as we would expect, because coming from an empirical distribution they are specific to this draw of 1000 observations from a standard normal random distribution. They are slightly different from the critical values of a 98% confidence interval (taking 1% of each side of this symmetric distribution) for a normal distribution, as predicted by theory. The theoretical values are

```
qnorm(0.01) # -2.326348
qnorm(0.99) # +2.326348
```

9.2 A word on outliers in time series

In time series analysis one should not discard outliers since data is usually auto-regressively correlated (the observation in t relies on the value observed at time $t-1$) so that important information on the structure of the process can be lost if we drop specific observations in time. Consider the process x_t

$$x_t = \theta x_{t-1} + \delta d_t + \varepsilon_t$$

In this case it would be better to use an *intervention variable* d_t to control for a "hic" in data at $t = T_0$ so that the analyst would construct a "dummy" variable with 0's and 1's attributing $d_t = 0$ everywhere except for $d_{T_0} = 1$.

Suppose we have 100 observations of an stationary auto-regressive process with and the outlier happens from the observations 50 to 53. Package `dynlm` estimates dynamic linear models.

```
require(dynlm)
set.seed(222)

outlier <- 50:53 # position of the outliers
N <- 100         # sample size
x <- rep(NA,N)   # initialize the observed values
x[1] <- 0
eps <- rnorm(N,mean=0,sd=0.1) # gaussian residuals

for (i in 2:N) x[i] <- 0.5*x[i-1] + eps[i]
x[outlier] <- 0.5
x <- zooreg(x[1:N]); plot(x)

# estimate AR(1) model
summary( dynlm(x~L(x,1)) )
```

The results suggest that 38% of the variation in data is explained and the coefficient estimate is 0.62 while we know the true value is 0.5.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.007913   0.012451   0.636    0.527
L(x, 1)     0.620787   0.080647   7.698 1.17e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1222 on 97 degrees of freedom
Multiple R-squared:  0.3792,Adjusted R-squared:  0.3728
F-statistic: 59.25 on 1 and 97 DF,  p-value: 1.172e-11
```

Now we create the dummy with 1's in the positions 50 to 53 and 0's elsewhere.

```
# create dummy
dum <- rep(0,N)
dum[outlier] <- 1
dum <- zooreg(dum)

# estimate AR(1) model with dummy
summary( dynlm(x~L(x)+dum) )
```

Now with the dummy the model improves fitting to explain 54% of variation in data and the coefficient 0.41 got closer to the true value of 0.5.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.001128   0.010860  -0.104    0.917
L(x)         0.416935   0.077877   5.354 5.87e-07 ***
dum          0.351919   0.060229   5.843 7.01e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1055 on 96 degrees of freedom
Multiple R-squared:  0.5421,Adjusted R-squared:  0.5325
F-statistic: 56.82 on 2 and 96 DF,  p-value: < 2.2e-16
```

10 Convergence and the Central Limit Theorem

Assume that

$$X \sim N(\mu, \sigma^2) \quad \text{and} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where n is the sample size and $X = (X_1, X_2, \ldots, X_n)$.

10.1 Strong Law of Large Numbers

The **Strong Law of Large Numbers** (SLLN) states that the sample mean converges to the true mean with certainty (ie. probability=1):

$$\overline{X} \to \mu$$

10.2 Weak Law of Large Numbers

The **Weak Law of Large Numbers** (WLLN) states that as the sample size grows the average converges to the true parameter for the mean of the distribution, any distribution:

$$\lim_{n\to\infty} P(|\overline{X} - \mu| > \varepsilon) \to 0$$

This result comes from the Chebyshev inequality (seen in section 7). We can simulate and visualize the process of convergence as the sample size grows.

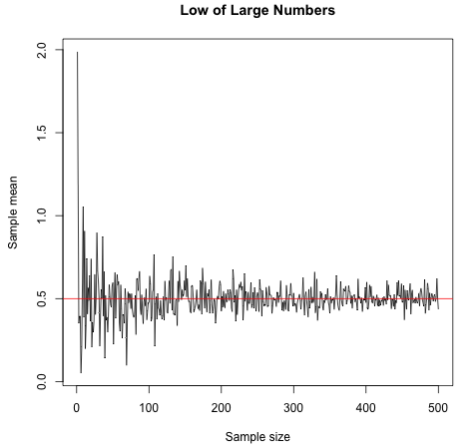
```
N <- 500
mu <- 0.5
means <- rep(NA,N)
set.seed(222)
for (i in 1:N) means[i] <- mean(rnorm(i,mean=mu))
plot(means, type='l',
     main = "Low of Large Numbers",
     xlab="Sample size", ylab="Sample mean")
abline(h = mu, col="red")
```

Figure 7 shows the convergence of the mean estimators for the true value $\mu = 0.5$.

10.3 Central Limit Theorem

The **Central Limit Theorem** states that for any sequence of random variables X_1,...X_n from any distribution with mean μ and variance σ^2 the

Figure 7: Convergence of the mean estimators



random variable

$$Y_n = \frac{\sum_{i=1}^{n}(X_i - n\mu)}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \to N(0,1)$$

The CLT states that Y_n converges in distribution to standard normal distribution $N(0,1)$ as $n \to \infty$.

10.4 Application.Check the convergence predicted by CLT

Check that by increasing the degrees of freedom of a t-Student distribution the critical value converges to the one corresponding to $N(0,1)$

```
qnorm(.975) # 1.959
qt(.975, df = c(10,100,1e3,1e4)) # 2.228 1.983 1.962 1.960
```

We can think of another test to show that the Student-t distribution tends to the $N(0,1)$ as the number of degrees-of-freedom grows to infinity by recalling that the Student-t distribution is composed out of a standard normal and a χ^2 distribution, $T(r) = N(0,1)/\sqrt{Q(r)/r}$, so that $\lim_{r\to\infty} Q(r)/r = 1$

```
# lim Q(r)/r = 1 as r goes to infinity
Qr_r <- function(x) { mean(rchisq(1e3,df=x))/x }
as.numeric( lapply( c(1,1e2,1e3,1e4,1e5,1e6), Qr_r ) )
[1] 0.9770687 1.0043828 1.0032461 1.0000710 0.9998624 1.0000790
```

The `lapply` function simply runs the function `Qr_r` against a vector of increasing degrees of freedom. It's clear that the result converges to 1.

11 Hypothesis Testing

In all tests presented here the samples are assumed to be iid. There is a class of tests ("paired" tests) that uses correlated samples but we are not treating them in this text.

11.1 The Mechanics of Hypothesis Testing

To perform a statistical test you need:

1. a null hypothesis H_0
2. an alternative hypothesis H_a
3. a test statistic, and
4. a reference distribution

If the statistic falls outside of the "non-rejection" region[2] of the reference distribution, we reject the null hypothesis and consider the alternative. If the statistic falls inside the "non-rejection" zone we "don't reject" the null hypothesis. The non-rejection zone is marked by the critical value(s), as will be explained.

The expression "don't reject" is more precise than saying that we "accept" the null hypothesis, although there is no practical difference. The idea of "accepting" a null hypothesis somehow conveys the idea that you are certain of it. You just have to be aware that you may be making a mistake (type I or II) and how likely it might be. As this text is not targeted to statisticians we may eventually relax on the terminology to make it accessible to a broader audience. Check in table 4 the two kinds of errors that can happen.

Decision	H_0 is True	H_0 is False
Reject H_0	Error Type I	Correct (Test Power)
Don't Reject H_0	Correct	Error Type II

Table 4: Hypothesis testing decision table

[2]The non-rejection region for a 5% significance level of a normal distribution in a two-sided test is $(-1.96, 1.96)$.

We should always specify the null and the alternative hypothesis to tell if it's a one-side or two-side test. If the $H_0 : \beta = 0$ and $H_a : \beta \neq 0$ we have a two-sided test since there is no requirement that β be positive or negative but if $H_0 = \beta > 0$ and $H_a : \beta \leq 0$ than it's a one-sided test and the critical value changes from `qnorm(0.975)=1.96` to `qnorm(0.95)=1.64`. To summarize:

- **Error Type I**: known as a *false positive* is the error of rejecting a null hypothesis when it is actually true. It occurs when we are observing a difference when in truth there is none (or more specifically - no statistically significant difference).
- **Error Type II**: also known as a *false negative*: the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. It is the error of failing to accept an alternative hypothesis when you don't have adequate power. It happens when we fail to observe a difference when it exists.
- **p-value**: the maximum probability of committing error type I
- **test power**: is the probability of rejecting a false null hypothesis. It is usually noted as β and it is directly associated to the smallest difference that we want to capture in the null hypothesis.
- **ROC curve**: ROC (Receiver Operating Characteristic) curves are a graphical representation of the contrast between true positive rates and the false positive rate at various thresholds. They show the trade-off between the sensitivity of the model (true positives) and the probability it will trigger a false alarm (false positives). Check R Package **ROCR** (**?**).

11.2 Student-t Test

To test if an estimator is significant we formulate the null and alternative hypotheses that $H_0 : \beta = 0$ and $H_a : \beta \neq 0$. The statistics is

$$t = \frac{\hat{\beta}}{\sqrt{\hat{\sigma}^2/N}} \sim t_{N-1}$$

and the reference distribution is Student-t with $N - 1$ degrees of freedom, where N is the sample size. The critical values are calculated using the Student-t as a reference so that the boundaries for the non-rejection zone of a 95% confidence interval are

$$\overline{x} + t_{N-1}^{0.025} * sem \leq \mu \leq \overline{x} + t_{N-1}^{0.975} * sem$$

where $sem = \hat{\sigma}/\sqrt{N}$ and $\hat{\sigma}$ is an estimate for the true parameter σ.

```
# H0: beta1=0 and Ha: bhat1<>0
set.seed(222)
bhat1 <- rnorm(100,mean=5,sd=1)
mean(bhat1) # 5.02198
t.test(bhat1) # p-value < 2.2e-16
```

The p-value is located in the rejection region (it's smaller than 2.5%) so we conclude, with 95% confidence, that the true value of the parameter β_1 is not zero. Note that the null hypothesis is done for β_1 while the statistic is done using the estimator $\hat{\beta}_1$. We assume that the estimator represents correctly the parameter (aka. "consistent"). To test that the sample mean against a specific value, believed to be the true parameter, we do

```
t.test(bhat1,mu=5)
```

This test can also be used to compare two sample estimators $H_0 : \beta_1 = \beta_2$ and $H_a : \beta_1 \neq \beta_2$ using the statistics

$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2)}{\sqrt{\hat{\sigma}^2/N}} \sim t_{N_1+N_2-2}$$

if both variances are assumed to be equal (even if sample sizes N_1 and N_2 are different) or

$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2)}{\sqrt{\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2}} \sim t_m$$

if variances are assumed to be different. By default `t.test()` assumes that variances are different. This test, also know as **Welch's t test**, has m degrees of freedom where

$$m = \frac{\left(\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}\right)^2}{\frac{\hat{\sigma}_1^4}{N_1^2(N_1-1)} + \frac{\hat{\sigma}_2^4}{N_2^2(N_2-1)}}$$

Let's try by simulating another sample with same mean but higher variance

```
# H0: bhat1=bhat2 and Ha: bhat1<>bhat2
set.seed(222)
bhat2 <- rnorm(100,mean=5,sd=3)
mean(bhat2) # 5.065939
```

The expected values for each estimator are close but not equal, 5.02198 and 5.065939. We want to know if they come from the same generating process already knowing that they come from different processes that differ only in variance.

```
t.test(bhat1,bhat2) # p-value = 0.8849
```

The p-value is located in the "non-reject" region so we conclude, with 95% confidence, that $\hat{\beta}_1$ and $\hat{\beta}_2$ come from the same data generating process (DGP). This result seems fair but remember: the two DGPs differ in variance. How can we capture this difference?

11.3 F Test

To test if the variance is the same, we can use an F test to compare the variances of two samples from normal populations and the null hypothesis is that true ratio of variances is equal to 1.

```
var.test(bhat1,bhat2) # p-value < 2.2e-16
```

The null hypothesis is rejected so the variances are not the same. Behind the scene we are testing the ratio of the two variances from samples that may differ in size, N_1 and N_2

$$H_0 : F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = 1 \quad H_a : \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \neq 1$$

while the F statistic follows a F distribution with $N_1 - 1$ and $N_2 - 1$ degrees of freedom. The p-value for the test is

```
vr <- var(bhat1)/var(bhat2) # 0.1111111
pf(vr,df1=99,df2=99) # 5.413684e-24
```

where `vr` is the variance ratio. The confidence interval for this test is

```
alpha <- 0.05
vrlo <- vr/qf(alpha/2,99,99)   # 0.0747601
vrhi <- vr/qf(1-alpha/2,99,99) # 0.1651371
```

11.4 Jarque-Bera Normality Test

The **Jarque-Bera** (JB) test checks if an empirical distribution approximates a gaussian distribution. The test statistic is obtained from the Skewness (S) and Kurtosis (K) of the empirical distribution and is calculated as

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \sim \chi^2_{df=2}$$

The JB statistic asymptotically has a chi-squared distribution with two degrees of freedom. The null hypothesis (H_0) is that data comes from a normal

distribution, so it is a joint hypothesis of the skewness is zero and the *excess kurtosis* is zero, $H_0 : S = 0$ and $(K - 3) = 0$. The kurtosis for a gaussian distribution is 3, so the use of excess kurtosis is a way to measure the difference in relation to the gaussian.

The JB test of time series x requires package `tseries`

```
library(tseries)
set.seed(222)
jarque.bera.test( rnorm(100,mean=0,sd=1) ) # p-value = 0.1156
jarque.bera.test( rnorm(100,mean=0,sd=5) ) # p-value = 0.0486
```

In the first case the test statistic falls in the non-reject region, so we don't reject that, at 95% confidence, the sample has a normal distribution, but in the second case the p-value is very small and suggests that we should reject H_0 so the sample would not come from a normal distribution. JB test is looses power in small samples so you might want to proceed with caution and try other tests.

11.5 Shapiro-Wilks Normality Test

The **Shapiro-Wilks** test is another test for normality in which the null-hypothesis is also that the population is normally distributed

```
set.seed(222)
shapiro.test(rnorm(100, mean=0, sd=5)) # p-value = 0.07387
shapiro.test(rnorm(10 , mean=0, sd=5)) # p-value = 0.00133
```

For a big sample of 100 observations the Shapiro-Wilks test does not reject the null (at 5% significance level) while it rejects it for a smaller sample based on the same data generating process.

11.6 Kruskal-Wallis test

To compare multiple groups when the data in these groups do not follow a normal distribution, you can use the Kruskal-Wallis test that assume the null that the location parameters of the distributions are the same in each group

```
N <- 1e3
set.seed(222)
u <- runif(N,-4,4)
n <- rnorm(N)
```

```
t <- rt(N,df=N-1)
kruskal.test( list(u,n,t) )
```

The $p - value = 0.4845$ tells that we can't reject that these samples come from the same distribution.

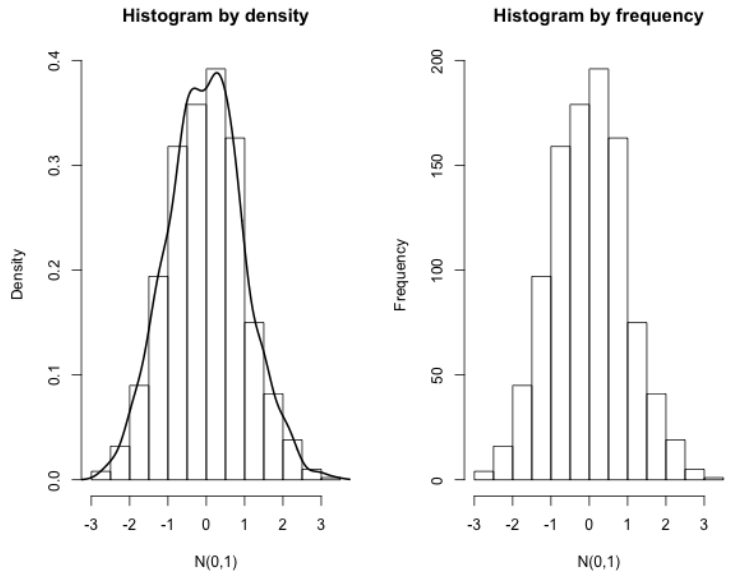
12 Data visualization

12.1 Native plots

12.1.1 Histogram

```
op<-par(mfrow=c(1,2))
set.seed(222)
x <- rnorm(100,mean=0,sd=1)
hist(x, freq=FALSE, breaks=10,main="Histogram by density",xlab="N(0,1)")
lines(density(x),col="black",lwd=2)
hist(x, freq=TRUE,main="Histogram by frequency",xlab="N(0,1)")
par(op)
```

Figure 8: Histogram for a Gaussian distribution

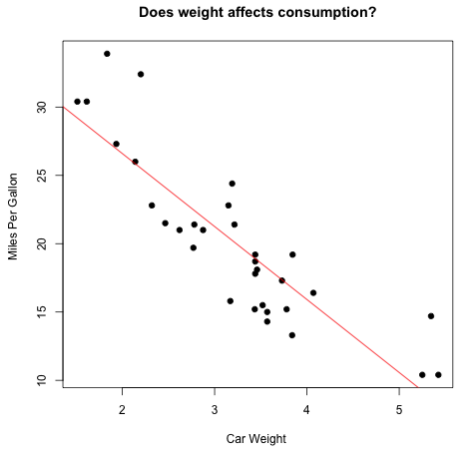


12.1.2 Scatter Plot

The scatter plot is useful to see correlations in data.

```
attach(mtcars)
plot(wt, mpg, main="Scatterplot Example",
   xlab="Car Weight ", ylab="Miles Per Gallon ", pch=19)
abline(lm(mpg~wt), col="red") # regression line
```

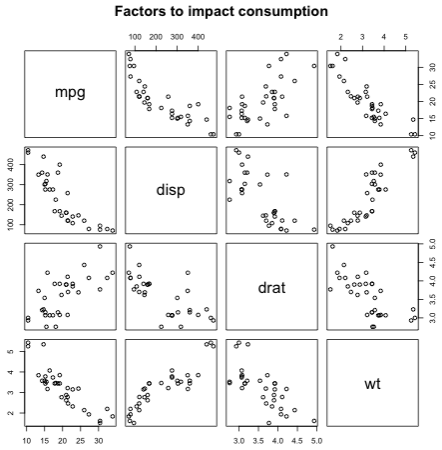
Figure 9: Scatter plot to correlate two random variables



We can also build a scatterplot matrix to visually inspect different correlations.

```
pairs(~mpg+disp+drat+wt,data=mtcars,
        main="Factors to impact consumption")
```

Figure 10: Matrix to correlate many random variables

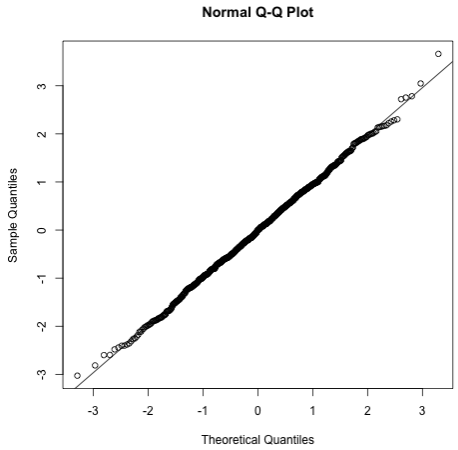


12.1.3 Quantile-Quantile plot

You can also visually inspect the residuals for normality by using a Quantile-Quantile (QQ) plot:

```
x <- rnorm(1000)
qqnorm(x)
qqline(x)
```

Figure 11: QQ plot of a Gaussian distribution



The empiric distribution is compared to a reference distribution: the closeness to the central line shows how close (or different) are these two distributions. Student-t distributions

```
x <- rt(1000,df=5)
qqnorm(x,main="Student t QQ-Plot")
qqline(x)
```

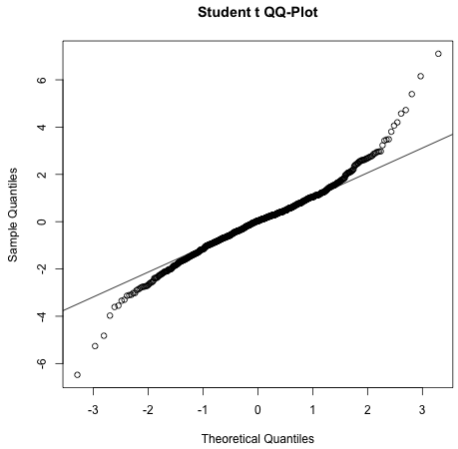
Note how the extremes (tails) of the distribution get out of line. Student-t are known to have "fait tails" when compared to gaussian distributions.

12.1.4 Box Plot

The box plot shows the relative position of the mean and the dispersion for each sample.

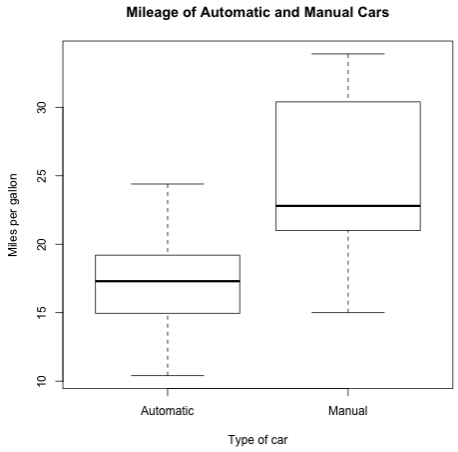
```
data(mtcars)

# name the groups
mtcars$am2 <- factor(mtcars$am, levels = c(0,1),
 labels = c("Automatic", "Manual"))

attach(mtcars)
```

Figure 12: QQ plot of a Student-t distribution




```
boxplot(mpg~am2,main = "Mileage of Automatic and Manual Cars",
 xlab = "Type of car (atuomatic/manual)", ylab = "Miles per gallon")
```

Figure 13: Boxplot of Gaussian distributions



13 Distributions for bayesian analysis

13.1 Beta distribution

$$Beta(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

$$\begin{aligned} f(x,\alpha,\beta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{1}{Beta(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} \end{aligned}$$

13.2 Gamma distribution

$$f(x,\alpha,\beta) = \frac{\beta^{\alpha}x^{\alpha-1}e^{-x\beta}}{\Gamma(\alpha)}$$

where

$$\Gamma(t) = \int_0^{\infty} x^{t-1}e{-}xdx$$

```
curve(dgamma(x,shape=1, rate=0.5),0,10,col="blue")
curve(dgamma(x,shape=1, rate=2),0,10,col="red",add=TRUE)
curve(dgamma(x,shape=5, rate=2),0,10,col="green",add=TRUE)
curve(dgamma(x,shape=10, rate=2),0,10,col="black",add=TRUE)
```

13.3 Inverse Gamma distribution