Introduction to Statistics with R

Gabriel Baud-Bovy

Introduction

- Generally speaking, the ultimate goal of every research or scientific analysis is finding relations between variables. The philosophy of science teaches us that there is no other way of representing "meaning" except in terms of relations between some quantities or qualities; either way involves relations between variables.
- The general objective of most studies is therefore to explain the variations of some variable of interest in function of the variations of other variables. This general objective includes more specific goals like
  1. Identifying of variables (or experimental factors) that affect the value of the some other variable
  2. Measuring the strength of the relationship between two variables
  3. Finding a model that predict the values of some variable from the value of other variables

Statistics

- **Statistics** is the “field of study concerned with (1) collection, organization, summarization, and analysis of the data, and (2) the drawing of inferences about a body of data when only a part is observed” (Wayne, 1995, p. 2)
- **Descriptive statistics:** The branch of statistics devoted to the description and summarization of data.
- **Inferential statistics:** The branch of statistics concerned with methods that use a small set of data (sample) to make a decision (inference) about a larger set of data (population).

Descriptive statistics

- Descriptive statistics uses various mathematical formulae (statistics) to summarize the main characteristics of a data set:
  - Central tendency: mean, median
  - Dispersion: variance, standard error, range
  - Distribution: quantiles
- Graphical methods (plots) provide a very powerful way to explore and quickly extract or present information about the data.

Inferential statistics

- The distinction between a population and a sample is essential to inferential statistics:
- **Population:** a population is an entire collection of events in which you are interested (student’s scores, people’s incomes, etc.).
- **Sample:** A subset of the population of interest.
- Population can range from a relatively set of numbers to a very large (all male human beings, all italian students in third grade) or even infinite (all possible drawings that students could theoretically produce) set of numbers.
- Inferential statistics is needed because it is in general impossible to make an ***exhaustive study*** (i.e., observe all elements of a population).

Experiments

- An **experiment** is any process or study which results in the collection of data, the outcome of which is unknown. In statistics, the term is generally restricted to situations in which the researcher has control over some of the conditions under which the experiment takes place.
- Not possible (or much more difficult) to draw inferences or test hypotheses if the experiment has not been well designed.
- Description of an experiment must include a description of the following elements:
  - the experimental units (e.g., subjects or any entity that constitute the focus of the study)
  - the treatments (a description of various experimental factors manipulated by the experimenter)
  - the method used to assign treatments to units (randomization)
  - the measures

Experimental vs. observational study

- The hallmark of the **experimental study** is that the allocation or assignment of individuals is under control of investigator and thus can be randomized. Properly executed experimental studies provide the strongest empirical evidence.
- In an **observational study**, the allocation or assignment of factors is not under control of investigator. Observational studies do not allow to make inferences about **causation** because the mechanism that assigned treatments to units is usually unknown and any difference in responses between treatment groups could be due to other hidden factors rather than to the treatments.
- Observational studies (also known as correlation studies, quasi-experiment or natural experiments) occur when it is impossible (n fields like astronomy, geloy, sociology or political science) or unethical (e.g., risk on human health) to manipulate some factors. They also occur when one analyze the effect of one factor that recorded but not randomized at the moment of the experiment.

External and internal validity

- **External validity (generalizability):** A study is external valid if its conclusions represent the truth for the population to which the results will be applied because both the study population and the reader's population are similar enough in important characteristics. To insure the external validity of our studies, we need to insure that the sample is *representative* of the population of interest. One way of addressing this issue is to select the sample randomly (**random selection**).
- **Internal validity**: In order to insure the internal validity of our studies, we need to has been randomly assign our subjects (once selected) to the treatment groups (**random assignment**). Randomization helps to control that no other factor than the treatment might explain a possible difference between the groups

Haphazard Scheme

- **Simple random sampling** . A sampling procedure that assures that each element in the population has an equal chance of being selected is referred to as simple random sampling. For example, give a number to all elements of the populations and use random number to select the sample.
- **Haphazard Scheme.** Haphazard or other unplanned sampling like taking the first N elements as a sample is not random sampling not random sampling and can lead to biased results.
- **Example.** Say we are testing the effectiveness of a voter education program on high school seniors. If we take all the *volunteers* in a class (haphazard selection scheme), expose them to the program and then compare their voting behavior against those who didn't participate, our results will reflect something other than the effects of the voter education intervention. This is because there are, no doubt, qualities about those volunteers that make them different from students who do not volunteer. In addition, those differences may very well correlate with propensity to vote. In contrast, using a random number generator to select students would ensure that those in the treatment and control groups differ solely due to chance.

Randomization and Experimental designs

- Assigning randomly treatments to experimental units is fundamental to avoid the effect of co-found factors (internal validity). Nowadays, randomization is achieved using generating random numbers.
- Statistical methods can take advantage of specific features of experimental designs such as pairing or blocking to gain efficiency
  - Blocking is the arrangement of experimental units into groups (blocks) that are similar to one another. Pairing is similar to blocking but involve only two groups and two treatments (one treatment is assigned to one element of the pair and the second treatment is assigned to the second element of the pair). Pairing and blocking reduce known but irrelevant sources of variation between units and thus allows greater precision in the estimation of the source of variation under study.

Take-home message

Satistical methods are useless if the experiment has not been well designed

Statistical methods

Statistical methods I

- The choice of a statistical methods depends first on the scientific hypothesis that one wants to test.
  - For example, one might be interested by the effect of the experimental factors on the central (average) value or on the dispersion (variability) of the dependent variable. One may be interested by the strength of the relationship between two or more variables, etc
- The method depends also on the number and type of data collected:
  - Number of dependent variables => Univariate versus Multivariate method
  - Categorical (or discrete) data versus continuous data => see later.

Statistical methods II

- **Univariate methods** assume that there is only one variable of interest. All other variables are used to explain variations of this variable.
  - Example of methods: Analysis of variance (ANOVA), simple and multiple regression, etc.
- **Multivariate methods** are used when two or more variables are necessary to characterize (e.g., x and y coordinates of the final position of a pointing movement, set of EEG, voxels in an MR image).
  - Examples: MANOVA, principal component analysis (PCA), multivariate discriminant analysis (MDA), etc.
- **Repeated-measures** can be analyzed either with (1) with univariate methods if time, space or any other within-subject factor are viewed as independent variables or (2) with multivariate methods if the whole records is considered at once.

Statistical methods III

		Independent variable	
		Continuous	Discrete
Dependent variable	Continuous	regression methods	ANOVA and t tests
	Discrete	logistic regression, log-linear models	Tables of contingency (e.g., chi square test)

- Discrete variables as dependent variables are usually counts or proportions. Discrete variables as independent variables define groups.
- Distinction is not always strict. For example, ANOVA has been used to analyze counts or proportions under some conditions.
- There are more general theoretical frameworks that encompass several of these methods. For example, Generalized Linear Models (GLM) include linear regression, multiple regression, ANOVA and logistic regression as special cases.

Probability Distributions

The concept of variable

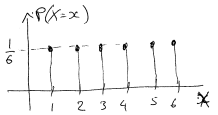
- A **variable** is a quantity that can take different values
  - A ***discrete variable*** is one which may take on only a countable number of distinct values such as 0, 1, 2, 3, 4, or blue, green, red, …
  - A ***continuous variable*** can assume any numerical values (e.g, tree height, body weight).
- The main characteristics of a variable are its ***distribution***, its ***central value*** and its ***dispersion***.
- The term dependent and independent variables terms apply best to experimental studies where the experimenter controls the "independent variable" in order to assess its effect on the "dependent" variable. More generally, the independent variables are variables that are used to "explain" the variation of the dependent variable.

Random variable

- A ***random variable*** is a variable with a probability distribution. The ***probability distribution*** specifies how the values of the random variable are distributed (see probability textbook for technical definition). For example, a gaussian random variable has a gaussian distribution, etc.
- A random variable can be either discrete or continuous:
  - Discrete random variables have discrete probability distributions
  - Continuous random variables have a continuous probability distributions

Discrete probability distribution

- A ***discrete probability distribution*** specifies the probability $P(X=x_i)$ for each value x_i of a discrete random variable X.
- Example of events that have a discrete probability distribution:
  - throw of a coin (2 possible values),
  - throw of a dice (6 possible values),
  - wining number in the lottery,
  - sum of the throw of two dices,
  - number of heads in N throw of a coin, etc.



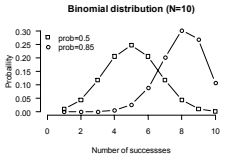
- Some properties of discrete probability distributions

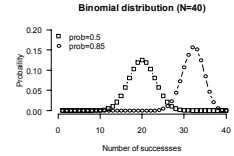
$$0 \le P(X = x_i) \le 1$$

$$\sum_i P(X = x_i) = 1$$

- Example of discrete probability distributions:
  - Discrete uniform distribution
  - Bernouilli distribution
  - Binomial distribution

Binomial distribution





- Number of successes over N trials given a probability of success p
- For small N, the binomial distrubition is asymmetric when p≠0.5.
- For large N, the distribution becomes more symmetric.
- The mean (expected value) and standard deviation of the binomial distribution are

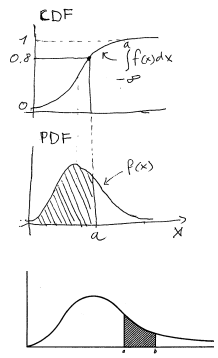
Continuous probability distribution

- The ***probability density function*** **(PDF)** is a function $f(x)$ that specifies the density of probability for each value x of the variable X.
- The ***cumulative density probability function*** **(CDF)** defines the probability P that the random variable X has a value smaller than a

$$P(X \le a) = \int_{-\infty}^{a} f(x)dx$$

- The probability $P(a<X<b)$ that the random variable X takes a value between a and b is

$$P(a \le X \le b) = \int_{a}^{b} f(x)dx$$

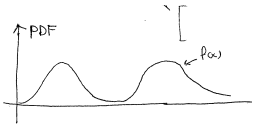


Continuous probability distribution

- A continuous probability distribution can have almost any shape as long as the area under the curve is equal to one:

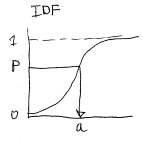
$$P(-\infty \le X \le +\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$

- **Exercise.** Draw the CDF corresponding to the following bimodal probability function:



The inverse density function

- The **inverse density probability function** (IDF) is the inverse of cumulative probability distribution: given a probability 0≤P≤1, it will give the value a such that P(X≤a).



- The inverse probability distribution is used to compute
  - interval of confidences,
  - critical values in hypothesis testing,
  - the point of subjective equality (PSE) in psychophysics,
  - etc.

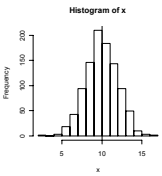
Continuous probability distributions

- Main continuous probability distributions:
  - The uniform distribution
  - The normal distribution
  - The Student distribution
  - The Chi-square distribution
  - The Fisher distribution

Continuous probability distribution

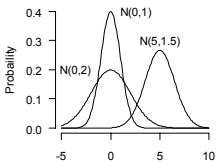
function	density function	cumulative function	inverse function	random generation
unifnorm	dunif	punif	qunif	runif
gaussian	dnorm	pnorm	qnorm	rnorm
student	dt	pt	qt	rt
binomial	dbinom	pbinom	qbinom	rbinom
Fisher	df	pf	qf	rf

```
# generate 1000 random numbers from a gaussian
# distribution with mean = 10 and sd=2
x<-rnorm(1000,mean=10,sd=2)
mean(x)
[1] 10.05171
sd(x)
[1] 2.039555
# make an histogram
hist(x)
# Probability that normally distributed random
# variable is larger than 0
pnorm(0,mean=0,sd=1)
[1] 0.5
# critical value u such that the prob of a normally
# distributed variable Z is smaller than u is 0.95
qnorm(0.95,mean=0,sd=1)
[1] 1.644854
```



Notes:
- All these functions are vectorialized
- Use set.seed to set to replicate series of random numbers

The normal distribution



- The PDF of the normal distribution $N(\mu,\sigma)$ is:

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

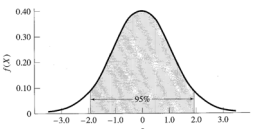
where μ is the mean and σ the standard deviation .

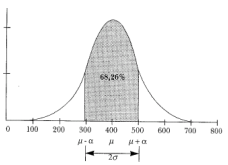
- The standard normal distribution N(0,1) has a mean of 0 and a standard deviation of 1.

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

- It is possible to a transform variable X that has a normal distribution N(μ,σ) into another variable Z that has standard normal distribution.
- In real case, the theoretical values μ and σ are not known but estimated by computing the sample mean m and standard deviation s.

The normal distribution





- The standard normal distribution ($Z \sim N(0,1)$)

u	Interval $[-u,u]$	Probability $Pr(-u \le Z \le u)$
1	[-1,1]	0.6826
1.96	[-1.96,1.96]	0.95
2.576	[-2.576,2.576]	0.99

- The normal distribution (e.g, $X \sim N(400,100)$)

u	Interval $[\mu-u\sigma,\mu+u\sigma]$	Probability X in the interval
1	[300,500]	0.6826
1.96	[204,596]	0.95
2.576	[142.4,657.6]	0.99

- Proof:

$$\Pr(\mu - u\sigma \le X \le \mu + u\sigma) = \Pr\left(-u \le \frac{X-\mu}{\sigma} \le u\right) = \Pr(-u \le Z \le u)$$

Exercise.

- The function qnorm(u,mean,sd) gives the probability Pr(X<u) for a normally distributed variable with mean μ and standard deviation sigma.
- Let's assume that the variable X is normally distributed with a mean μ=15 and a standard deviation sigma=5. Use the function qnorm to compute the probability that the variable X takes
  1. a value smaller than 5
  2. a value between 10 and 20
  3. a value larger than 20

  Answers: 1) 0.0228, 2) 0.8413-0.1587=0.6826, 3) 1-0.8413=0.1587
- Use the function pnorm to find out an interval centered on the mean such that the variable X has 80% of chance of taking a value inside this interval.

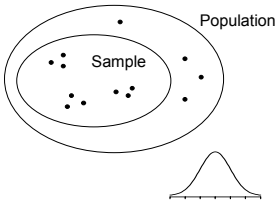
  Answer: [8.59,21.41]

Central Limit Theorem

- The normal distribution plays a very important role of statistics for both practical and theoretical reasons.
- The mathematical expression of the normal probability distribution has nice properties.
- The ***Central Limit Theorem*** states that a sum of random variables with finite variance tend approximately normally distributed (i.e. following a normal or Gaussian distribution).
- The Central Limit Theorem is one theoretical reason that motivates the use of normal distribution to model noise in statistical models (since noise can be arguably viewed as the summed effect of many unknown processes)

Estimation

Population and Sample



- **Population:** the entire collection of values in which you are interested (student's scores, people's incomes, etc.).
- **Sample:** A subset of the population of interest.
- The probability distribution of the population is usually unknown. Often, the shape of the population distribution is assumed (e.g., normal distribution) and only the parameters of the assumed distribution are unknown.

Parameter and Statistic

- A **parameter** is a value, usually unknown (and which therefore has to be estimated), used to represent a certain population characteristic. For example, the population mean is a parameter that is often used to indicate the average value of a quantity.
- A **statistic** is a quantity that is calculated from a sample of data.
  - It is possible to draw more than one sample from the same population and the value of a statistic will in general vary from sample to sample. For example, the average value in a sample is a statistic. The average values in more than one sample, drawn from the same population, will not necessarily be equal.

Estimator

- **Estimator**: a statistic used to estimate the unknown values of a parameter of the corresponding population.
  - Within a population, a parameter is a fixed value which does not vary. In contrast, the value of the estimator computed from a sample will be different for each sample drawn from the population.
  - Parameters are often assigned Greek letters (e.g. μ), whereas statistics or estimators are assigned Roman letters (e.g. m).
  - **Example.** The average of a sample (or sample mean) m is used to give information about the overall mean in the population (or theoretical mean μ) from which that sample was drawn.

Estimators (normal distribution)

- Mean $m = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Variance $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2$
- Standard Deviation $s = \sqrt{s^2}$

See Chapter 3 of Bertolini & Nardi 1998

- 2nd formula to compute the variance

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2$$
$$= \frac{1}{N-1}\left(\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} m^2 - 2m\sum_{i=1}^{N} x_i\right)$$
$$= \frac{1}{N-1}\left(\sum_{i=1}^{N} x_i^2 - Nm^2\right)$$
$$= \frac{1}{N-1}\left(\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2\right)$$

R functions

Functions operating on vectors

function	description
sum(x)	sum
mean(x)	mean
var(x)	variance
sd(x)	standard deviation
cor(x,y)	correlation

Functions operating on matrices

function	description
rowSums(x) colSums(x)	sum
colMeans(x) rowMeans(x)	mean

Examples

```
# sum elements of a vector
x<-c(2,5,4,2)
sum(x)
[1] 13
# missing values
x<-c(2,5,NA,2)
mean(x,na.rm=TRUE)
[1] 3

# matrix
x<-matrix(1:6,2,3)
x
[,1] [,2] [,3]
[1,]   1    3    5
[2,]   2    4    6
colSums(x)
[1]  3  7 11
rowMeans(x)
[1] 3 4
```

Notes:
- Look at the apply function to apply any function to a colum or row of a matrix
- Look at the tapply and aggregate functions to apply a function to group of observation

Exercise.

codice soggetto	peso in kg	scarto dalla media	quadrato degli scarti	codice soggetto	peso in kg	scarto dalla media	quadrato degli scarti
1	29	–0,5	2,25	13	22	–8,5	72,25
2	26	–4,5	20,25	14	32	1,5	2,25
3	38	7,5	56,25	15	27	–3,5	12,25
4	25	–5,5	30,25	16	29	–1,5	2,25
5	31	0,5	0,25	17	35	4,5	20,25
6	30	–0,5	0,25	18	28	–2,5	6,25
7	35	4,5	20,25	19	29	–1,5	2,25
8	42	11,5	132,25	20	31	0,5	0,25
9	33	2,5	6,25	21	30	–0,5	0,25
10	31	0,5	0,25	22	27	–3,5	12,25
11	33	2,5	6,25	23	30	–0,5	0,25
12	24	–6,5	42,25	24	35	4,5	20,25
				totale	722	0	468

- Compute the mean and standard deviation of the weight of the 24 babies (see Table 3.4 in Chapter 3 of Bertolini & Nardi, 1998) with both formulae.
  - Answer: sum=732, average = 30.5, standard deviation = 4.51

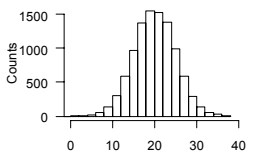
Sampling distribution

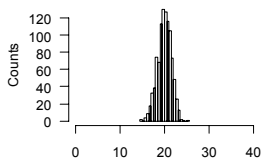
- **Sampling distribution**: probability distribution of a statistic (or of an estimator).
  - Statistic or estimator have a distribution since their value depend on the sample.
  - The sampling distribution usually depends on the population distribution AND on the size of the sample.

Sampling distribution of the mean

- The sampling distribution of the mean is simply the distribution of the means of an infinite number of samples drawn under certain specified conditions.
- If the measurements follow a normal distribution N(μ,σ), then
  - the mean *m* of a sample of N measurements will also be normally distributed with the mean μ and standard deviation σ/√N. $m \propto N\left(\mu, \frac{\sigma}{\sqrt{N}}\right)$
  - The score z follows the standard normal distribution. $z = \frac{m-\mu}{\frac{\sigma}{\sqrt{N}}} \propto N(0,1)$
- **Exercise.** Create an artificial data set with 1000 normally distributed random number (mean=50, SD=10). Divide the data set into 100 samples of size N=10 and compute the mean for each sample. Plot the distribution of the parent population and the distribution of the means.

Exercise





- Generate 10000 values from a normal distribution with μ=20 and σ=5 corresponding to 1000 samples of size 10:

  y<-rnorm(10000,20,5)
- Make an histogram of the the 10000 random values

  hist(y)
- Compute the mean and standard deviation of the sample
- Compute the mean for each sample (note: You can use tapply or aggregate)
- Make an histogram of the 1000 mean values.
- Compute the standard deviation of the means is close to the predicted value.

Confidence intervals

- When an experiment is repeated several times, some variation in the value of the estimated parameter is expected (remember that the sample is randomly selected plus there might be random errors in the measures or noise in the process under study, etc.).
- The ***confidence interval*** is an estimated range of values which is likely to include the true (unknown) value of the population parameter. In other words, the confidence interval defines a range of values that has a high probability (typically 95%) of encompassing the true (population) of the parameter. The width of the confidence interval gives us some idea about how uncertain we are about the unknown parameter.
- The **confidence level** is the probability (expressed as a percentage) that the true but unknown value of the parameter is inside confidence interval. The confidence intervals are usually calculated so that this percentage is 95%, but we can produce 90%, 99%, 99.9%, confidence intervals for the unknown parameter. If one repeats the experiment several times and estimates the confidence interval each time, the **confidence level** is also the percentage of computed intervals that include the true parameter.

Confidence interval of the sample mean (I)

- For example, let's compute the mean *m* of a sample of size and assume that *the standard deviation σ of the population is known*. According to the central limit theorem, the sample mean *m* is normally distributed with an true (unknown) mean μ and a standard deviation of $\sigma_m = \sigma/\text{sqrt}(N)$.
- From a previous slide, we know the probability

$$\Pr(\mu - 1.96\sigma_m \le m \le \mu + 1.96\sigma_m) = \Pr\left(-1.96 \le Z = \frac{m-\mu}{\sigma_m} \le 1.96\right) = 0.95$$

- It can easily be shown that probability that the sample mean m is in the interval $[\mu - u\sigma_m, \mu + u\sigma_m]$ is equal to the probability that the true (unknown) mean is in the interval $[m - u\sigma_m, m + u\sigma_m]$:

$$\Pr(\mu - u\sigma_m \le m \le \mu + u\sigma_m) = \Pr(m - u\sigma_m \le \mu \le m + u\sigma_m)$$

Therefore, if $u = 1.96$,

$$\Pr(m - 1.96\sigma_m \le \mu \le m + 1.96\sigma_m) = 0.95$$

- The interval $[m\text{-}1.96\sigma_m, m\text{+}1.96\sigma_m]$ is the 95% confidence interval for the sample mean because there is 95% of chance that the true mean is inside this interval. Similarly, $[m\text{-}2.576\sigma_m, m\text{+}2.576\sigma_m]$ is the 99% confidence interval, etc.

Student distribution

- Let's assume a sample of N normally distributed observations. In general, the variance (or standard deviation) of the population is *unknown*.
- If one assume that the true value of the mean of the distribution is μ, then it can be shown that the statistic *t* follows a distribution of Student with N-1 degrees of freedom (*m* and s_m are the estimated mean and standard deviation of the distribution of the mean, *s* is the estimated standard deviation of the parent population).

$$t = \frac{m-\mu}{s_m} = \frac{m-\mu}{\frac{s}{\sqrt{N}}} \propto t(N-1)$$

- The Student distribution is wider when the sample size (N) is smaller which correspon to the fact that we have less confidence in t estimate *s* of the standard deviation when t sample is small. For large N, student's distribution is undistinguishable from the normal distribution

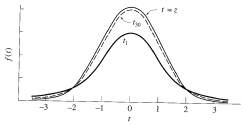


FIGURE 7·5 *t distribution for 1, 30, and ∞ degrees of freedom*

Confidence interval of the sample mean (II)

- In general, the standard deviation σ of the population is *unknown*. In this case, we have seen that the statistic

$$t = \frac{m-\mu}{s_m}$$

where $s_m = s/\text{sqrt}(N)$ follows a distribution of Student with N-1 degrees of freedom

- From a table of probability for the distribution of student, we can retrieve the value u such that

$$\Pr(-u \le t \le u) = 0.95$$

for example, u=2.023 for N=40 (i.e, N-1=39 degrees of freedom). From the definition of t, we also have

$$\Pr(-u \le t \le u) = \Pr\left(-u \le \frac{m-\mu}{s_m} \le u\right) = \Pr(m - us_m \le \mu \le m + us_m)$$

- Therefore, the interval $[m\text{-}2.023\ s_m, m\text{+}2.023\ s_m]$ is the 95% confidence interval for the sample mean when the standard deviation of the population is unknown and needs to be estimated.

Exercise. Confidence interval.

- Compute the 95% confidence interval for the RT of the Sternberg dataset.

For N-1=299, the value of the distribution of Student that corresponds to a 95% interval of confidence is 1.9679.

```
> m<-mean(stern$rt)
> s<-sd(stern$rt)
> m+qt(c(0.025,0.975),299)*s/sqrt(300)
[1] 58.78175 61.73825
or
> fit<-lm(rt~1,stern)
> confint(fit,level=0.95)
                2.5 %   97.5 %
(Intercept) 58.78175 61.73825
```

$$m - 1.9679 s_m = m - 1.9679 \frac{s}{\sqrt{N}}$$
$$= 60.26 - 1.9679 \frac{13.0106}{\sqrt{300}}$$
$$= 60.26 - 1.9679 \times 0.7512$$
$$= 60.26 - 1.4783 = 58.7818$$
$$m + 1.9679 s_m = 60.26 + 1.4783 = 61.7383$$

- Compute the 95% interval of confidence for the weight of the babies

Answer: [27.5,32.7]