Inferential Statistics **First semester, 2020**

Lecture 0: Background on probability

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

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

This lecture is based on [LM18] §2-4 and gives the essential background on Probability Theory that will be useful throughout the course. Proofs are mostly omitted.

0.1 Why and What?

We may come out with a very precise theory about a phenomenon or we may wish to understand more about a certain phenomenon in detail. Whatever is our aim, essentially everything is subject to uncertainty. For instance, if we measure our body temperature at different times on a given day, we do not expect to observe exactly the same values (do we?). Or if we look at the number of telephone calls received from 112, we see that this number varies from day to day.

Whatever is the phenomenon we are studying, to support our theory we need to:

(i) run experiments, which produce data;

(ii) analyse the data and hope that they give credit to our theory.

Statistical inference is the branch statistics which deals with the problem of *inferring* on a theory (or supposition, hypothesis, etc.) from observed data. To accomplish this inference, the theory is translated into a suitable probabilistic model. The data are then used in order to measure their support about this probabilistic model. If the model is supported then we deduce that our theory is supported.

Here are some type of inferential problems.

Example 0.1 *Every washing machine (WM) sold in the UE market must be accompanied by a technical documentation which describes, among other things, the energy consumed during a typical washing cycle. In order to measure the consumed energy, the WM is tested in a laboratory several times and under the same conditions. Using the observed values of energy consumptions the manufacturer wants to* estimate *or learn the most "representative value" of these data. This is a kind of problem that can be solved via estimation.*

Along with this representative value, the manufacturer has also to declare its variability or it has to declare tolerance limits within which the consumed energy is expected to vary with hight probability. This is a kind of problem that can be addressed via confidence intervals.

The manufacturer realises that his WM's consume too much energy with respect to what he expected. His teem of engineers claim that in order to reduce energy consumption, the old WMs motors must be replaced by a newer model. The engineers then run WM's with the old motor and WM's with the new motor measuring the energy consumption in both cases. The problem now is: with the data at hand, is it true that the typical consumption of old motor WM's is greater than the typical consumption of new motor WM's ? This is a kind of problem that can be solved by hypothesis testing.

The basic ingredients in all problems are: data and probability. This lecture and the next one will focus on the latter ingredient. We will see some concepts about summarising data in Lecture 2 and 3 and we will describe the three tools of statistical inference, i.e., estimation, confidence intervals and hypothesis testing, in Lectures 4, 5 and 6, respectively. Lecture 7 introduces some concepts of parametric statistics and the Bayesian approach to statistical inference.

0.2 Set Theory

The theory of probability is intimately related to that of sets, which we briefly outline in this section. A set S is a (well defined) collection of distinct objects denoted by s. The fact that s *belongs to* S is expressed by writing $s \in S$. The negation of this statement is $s \notin S$. We say that A is a subset of S and write $A \subseteq S$, if for every $s \in A$, we have that $s \in S$.

The following definitions are useful for creating new sets from old ones. In particular, the complement of the set A with respect to S is the set $A^c = \{s \in S : s \notin A\}$.

The union of the sets A_j, $j = 1, 2, \ldots, n$ is denoted by $A_1 \cup A_2 \cup \cdots \cup A_n$ or $\cup_{j=1}^n A_j$ and is defined by

$$\cup_{j=1}^n A_j = \{s \in S : s \in A_j, \text{for at least one } j = 1, 2, \ldots, n\}.$$

The intersection of the sets A_j, $j = 1, 2, \ldots, n$ is denoted by $A_1 \cap A_2 \cap \cdots \cap A_n$ or $\cap_{j=1}^n A_j$ and is defined by

$$\cap_{j=1}^n A_j = \{s \in S : s \in A_j, \text{for all } j = 1, 2, \ldots, n\}.$$

Set intersection and set union can be extended also to an infinite (i.e. denumerable or uncountable) collections of sets A_j, $j = 1, 2, \ldots$. The set which contain no elements is called the *empty set* and is denote by $\emptyset$. Two sets A_1 and A_2 are said to be disjoint if $A_1 \cap A_2 = \emptyset$. Sometimes, when dealing with disjoint sets, we will use the shorthand notation $A + B$ in place of $A \cup B$ especially when the union involves more than two disjoint sets. At times we will use the identity $A \cap B^c = A - B$. Two sets A_1 and A_2 are said to be equal if $A_1 \subseteq A_2$ and $A_2 \subseteq A_1$. The sets A_j, $j = 1, 2, \ldots$ are said to be *pairwise disjoint* if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Sets have many interesting properties amongst which we recall **De Morgan's laws**:

$$(\cup_j A_j)^c = \cap_j A_j^c \quad \text{and} \quad (\cap_j A_j)^c = \cup_j A_j^c.$$

The sequence $\{A_n\}, n = 1, 2, \ldots$, is said to be a *monotone sequence* of sets if either

(i) $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$, that is A_n is *increasing*, to be denoted by $A_n\uparrow$

(ii) $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$, that is A_n is *decreasing*, to be denoted by $A_n\downarrow$.

The *limit* of a monotone sequence of sets is defined as follows:

(i) If $A_n\uparrow$, then $\lim_{n\to\infty} A_n = \cup_{n=1}^\infty A_n$

(ii) If $A_n\downarrow$, then $\lim_{n\to\infty} A_n = \cap_{n=1}^\infty A_n$.

More generally, for *any* sequence $\{A_n\}, n = 1, 2, \ldots$, we define the sets

$$\underline{A} = \liminf_{n\to\infty} A_n = \bigcup_{n=1}^\infty \bigcap_{j=n}^\infty A_j \quad \text{and} \quad \overline{A} = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^\infty \bigcup_{j=n}^\infty A_j.$$

The sets $\underline{A}$ and $\overline{A}$ are called the *inferior limit* and *superior limit*, respectively, of the sequence $\{A_n\}$. The sequence $\{A_n\}$ has a *limit* if $\overline{A} = \underline{A}$.

0.3 Probability Theory

By an *experiment* we mean procedure being carried out under a certain set of conditions whereby the procedure can be repeated any number of times under the same set of conditions, and upon the completion of the procedure certain results are observed. An experiment for which the outcome cannot be determined in advance, except that it is known to be one of a set of possible outcomes, is called a *random experiment*. Examples of random experiments are: tossing a coin, rolling a die, recording the number of telephone calls which arrive at a telephone exchange within a specified period of time, the number of defective items in a shipment of many items of the same type, etc. The set of all possible outcomes of an experiment is called the *sample space* and is denoted by $\mathcal{S}$. The elements s of $\mathcal{S}$ are called *sample points* and subsets of $\mathcal{S}$ are called *events*. To be more precise, we are interested on subsets A of $\mathcal{S}$ that form a *σ-field* of subsets in the set $\mathcal{S}$. This σ-filed of subsets, denoted by $\mathcal{A}$, is a collection of subsets (i.e. a set of sets) of $\mathcal{S}$ and is such that:

(i) $\mathcal{A}$ is non empty (non trivial);

(ii) $A \in \mathcal{A}$ implies $A^c \in \mathcal{A}$ (closed under complementation);

(iii) if $A_j \in \mathcal{A}$, for all $j = 1, 2, \ldots$, then $\cup_{j=1}^{\infty} A_j \in \mathcal{A}$ (countably additive).

The sets $\mathcal{S}$ and $\emptyset$ are always events and are called the *certain event* and the *impossible event*, respectively.

Definition 0.1 *A probability function denoted by P is a set function which assigns to each event A a real number denoted by $P(A)$, called the probability of A, and satisfies the following requirements:*

(P1) $P(A)$ *is non-negative, i.e.* $P(A) \geq 0$ *for every event* A.

(P2) P *is normed, i.e.* $P(\mathcal{S}) = 1$.

(P3) P *is σ-additive, i.e. for every collection of pairwise disjoint events $A_j, j = 1, 2, \ldots$, we have* $P(\cup_j A_j) = \sum_j P(A_j)$.

The triple $(\mathcal{S}, \mathcal{A}, P)$ is known as a probability space.

This is the axiomatic definition of probability due to A.N. Kolmogorov.

Here are some properties that can be proved using Definition 0.1.

Theorem 0.1 *On a probability space $(\mathcal{S}, \mathcal{A}, P)$ the following properties are true.*

(a) $P(A^c) = 1 - P(A)$, *for any event* A

(b) $P(\emptyset) = 0$.

(c) For any events A, B, if $A \subseteq B$ then $P(A) \leq P(B)$.

(d) $P(A) \leq 1$ *for any event* A.

(e) For any events A, B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

(f) For any event A_j, $j = 1, 2, \ldots$, $P(\cup_{j=i}^{\infty} A_j) \leq \sum_{j=1}^{\infty} P(A_j)$.

(g) Let $\{A_n\}$ be a sequence of events such that $A_n\uparrow$ or $A_n\downarrow$ as $n \to \infty$. Then

$$P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n).$$

Proof: The proof of (f) and (g) can be accomplished by first noticing that $\cup_{j=1}^{\infty} A_j = A_1 + (A_1^c \cup A_2) + \cdots + (A_1^c \cap \cdots \cap A_{n-1}^c \cap A_n) + \cdots$, where the union on the left hand side involves pairwise disjoint sets. ■

0.3.1 Conditional probability

Sometimes we know that the event B has been observed, i.e. B is realised, and we may wish to calculate the probability of another event A taking into account the fact that B is known to be true. For instance, consider a fair die and suppose that the sides ⚀, ⚁ and ⚂ are painted red and the remaining three sides are printed black. The dice is rolled once and we are asked for the probability of that the upward side is ⚂. Clearly, $P(\{⚂\}) = \frac{1}{6}$. Now suppose that the die is rolled once and we are told that the colour of the upward side is red. The required probability is now $\frac{1}{3}$ because out of the read-coloured faces there is only one ⚂. This latter probability is called conditional probability of ⚂, given the information that the uppermost side was painted red. If we let B stand for the event that ⚂ appears and A for the event that the uppermost side is red, the above-mentioned conditional probability is denoted by $P(B|A)$, which is defined as follows.

Definition 0.2 *Let A be an event such that $P(A) > 0$. Then the conditional probability given A, is the set function denoted by $P(\cdot|A)$ and defined for every event B by*

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

$P(B|A)$ is called the conditional probability of B given A.

The following theorems state some useful properties related to the conditional probability function.

Theorem 0.2 *(Multiplicative Law) Let $A_j, j = 1, 2, \ldots, n$ be events such that $P(\cap_{j=1}^{n-1}) > 0$. Then*

$$P(\cap_{j=1}^{n} A_j) = P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})P(A_{n-1}|A_1 \cap A_2 \cap \cdots \cap A_{n-2}) \cdots P(A_2|A_1)P(A_1).$$

Let $A_j, j = 1, 2, \ldots$, be events such that $A_i \cup A_j = \emptyset$ for all $i \neq j$ and let $\sum_j A_j = \mathcal{S}$. Such a collection of events is called a *partition* of $\mathcal{S}$. Thus we have the following theorem.

Theorem 0.3 *(Law of Total Probability) Let $\{A_j, j = 1, 2, \ldots, \}$ be a partition of $\mathcal{S}$, with $P(A_j) > 0$ for all j. Then for an event $B \in \mathcal{A}$, we have*

$$P(B) = \sum_j P(B|A_j)P(A_j).$$

The Bayes Theorem or the Law of Inverse Probability is founded on the Law of Total Probability but it has its own importance especially in Bayesian statistics as we will see near the end of this course.

Theorem 0.4 *(Bayes Theorem) If $\{A_j, j = 1, 2, \dots, \}$ is a partition of $\mathcal{S}$ and $P(A_j) > 0$ for all j, then*

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_j P(B|A_j)P(A_j)}.$$

The following two simple examples illustrate the properties shown above.

Example 0.2 *An urn contains 10 identical balls, except for the colour, of which five are black, three are red and two are white. Four ball are drawn without replacement. Find the the probability that the first ball is black, the second is red, the third is white and the fourth is black.*

Let A_1 be the event that the first ball is black, A_2 be the event that the second ball is red, A_3 be the event that the third ball is white and A_4 be the event that the fourth ball is black. Then

$$P(A_1 \cap A_2 \cap A_3 \cap A_4) = P(A_4|A_1 \cap A_2 \cap A_3)P(A_3|A_1 \cap A_2)P(A_2|A_1)P(A_1),$$

and using the fact that each of the balls is equally likely to be drawn, we have

$$P(A_1) = \tfrac{5}{10}, \quad P(A_2|A_1) = \tfrac{3}{9}, \quad P(A_3|A_1 \cap A_2) = \tfrac{2}{8}, \quad P(A_4|A_1 \cap \cdots \cap A_1) = \tfrac{4}{7}.$$

Thus the required probability is equal to $\frac{1}{42}$.

Example 0.3 *A multiple choice test question lists five alternative answers, of which only one is correct. If a student has done the homework, then he/she is certain to identify the correct answer; otherwise he/she chooses an answer at random. Let p denote the probability of the event A that the student does the homework and let B be the event the he/she answers the question correctly. Find the expression of the conditional probability $P(A|B)$ in terms of p.*

By noting that A and A^c form a partition of the sample space, i.e. the student either does or doesn't do the homework, by the above results we have

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)} = \frac{1 \cdot p}{1 \cdot p + \frac{1}{5}(1-p)} = \frac{5p}{4p+1}.$$

0.3.2 Independence

For any events A, B with $P(A) > 0$ we defined $P(B|A) = P(A \cap B)/P(A)$. Now since $P(B|A)$ is a real number in the interval $[0, 1]$, by the tricotomy law of real numbers only one of the following possibilities are possible: $P(B|A) < P(B), P(B|A) > P(B)$ or $P(B|A) = P(B)$. As an illustration consider an urn containing 10 balls, seven of which are red and 3 are black. Except for the colour, the balls are identical. Suppose that two balls are drawn successively and without replacement. Then the conditional probability that the second ball is red, given that the first ball was red is $\frac{6}{9}$, whereas the conditional probability that the second ball is red given that the first ball was black is $\frac{7}{9}$. Without any knowledge on the first ball, the probability that the second ball is red equals (by the Low of Total Probability) $\frac{7}{10}\frac{6}{9} + \frac{3}{10}\frac{7}{9} = \frac{7}{10}$. On the other hand, if the balls are drawn with replacement, the probability that the second ball is red, given that the first ball was red is $\frac{7}{10}$. This probability is the same even if the the first ball was black. In other words, knowledge of the event which occurred in the first drawing provides no additional information in the calculation of the event that the second ball is red. These latter events are called *independent*.

Definition 0.3 *The events A, B are said to be statistically (or probabilistically) independent if $P(A \cap B) = P(A)P(B)$.*

Another alternative definition of independence is $P(A|B) = P(A)$, provided $P(B) > 0$. In general, for more than two events we have to resort to the following definition.

Definition 0.4 *The events $A_j, j = 1, 2, \ldots, n$ are said to be mutually or completely independent if*

$$P(A_{j_1} \cap \cdots \cap A_{j_k}) = P(A_{j_1}) \cdots P(A_{j_k}),$$

for any $k = 2, \ldots, n$ and $j_1, \ldots, j_k = 1, 2, \ldots, n$ such that $1 \leq j_1 < j_2 < \cdots < j_k \leq n$. The events are said to be pairwise independent if $P(A_i \cap A_j) = P(A_i)P(A_j)$ for all $i \neq j = 1, 2, \ldots, n$.

Here is an example that illustrates the concept.

Example 0.4 *Let $\mathcal{S} = \{1, 2, 3, 4\}$ and assume that $P(\{1\}) = \cdots = P(\{4\}) = \frac{1}{4}$. Set $A_1 = \{1, 2\}$, $A_2 = \{1, 3\}$, $A_3 = \{1, 4\}$. Then*

$$A_1 \cap A_2 = A_1 \cap A_3 = A_2 \cap A_3 = \{1\}, \quad \textit{and} \quad A_1 \cap A_2 \cap A_3 = \{1\}.$$

Thus $P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = P(A_1 \cap A_2 \cap A_3) = \frac{1}{4}$. Now,

$$\begin{aligned} P(A_1 \cap A_2) &= \tfrac{1}{4} = \tfrac{1}{2} \cdot \tfrac{1}{2} = P(A_1)P(A_2), \\ P(A_1 \cap A_3) &= \tfrac{1}{4} = \tfrac{1}{2} \cdot \tfrac{1}{2} = P(A_1)P(A_3), \\ P(A_2 \cap A_3) &= \tfrac{1}{4} = \tfrac{1}{2} \cdot \tfrac{1}{2} = P(A_2)P(A_3), \end{aligned}$$

but

$$P(A_1 \cap A_2 \cap A_3) = \tfrac{1}{4} \neq \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = P(A_1)P(A_2)P(A_3).$$

Thus, A_1, A_2, A_3 are pairwise independent but are not jointly dependent. As another example, let $\mathcal{S} = \{1, 2, 3, 4, 5\}$ and define P as follows: $P(\{1\}) = \frac{1}{8}$, $P(\{2\}) = P(\{3\}) = P(\{4\}) = \frac{3}{16}$, $P(\{5\}) = \frac{5}{16}$. Consider

$$A_1 = \{1, 2, 3\}, \quad A_2 = \{1, 2, 4\}, \quad A_3 = \{1, 3, 4\}.$$

Then

$$A_1 \cap A_2 = \{1, 2\}, \quad A_1 \cap A_2 \cap A_3 = \{1\}.$$

Thus

$$P(A_1 \cap A_2 \cap A_3) = \tfrac{1}{8} = \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = P(A_1)P(A_2)P(A_3),$$

but

$$P(A_1 \cap A_2) = \tfrac{5}{16} \neq \tfrac{1}{2} \cdot \tfrac{1}{2} = P(A_1)P(A_2).$$

0.4 Random variables and their distributions

Given the triple $(\mathcal{S}, \mathcal{A}, P)$, the sample space $\mathcal{S}$ may be quite an abstract set, and thus it is of little use in real-life applications in which we observe numbers. Indeed, in practical applications we typically observe measurements such as temperature, wind speed, precipitation, consumption of electricity of a certain equipment, life-time of an electronic circuit, etc. We wish thus to be able to characterise these real-valued measurements by means of probability theory. For instance we might want to calculate the probability that the life-time of a certain electronic circuit is at least t, where t is a positive real number. The key to achieve this is through *random variables*, i.e. by suitable transformations or mappings of events of $\mathcal{S}$ into real values $\mathbb{R}$.

We have the following definition of a random variable (r.v.)

Definition 0.5 *A random variable is a function*[1] *which assigns to each sample point $s \in \mathcal{S}$ a real number. r.v.s are denoted by capital letters such as X, Y, Z, etc. For a set $B \subseteq \mathbb{R}$, we denote by $(X \in B)$ the following event (set) in $\mathcal{S}$: $(X \in B) = \{s \in \mathcal{S} : X(s) \in B\}$.*

In the probability space $(\mathcal{S}, \mathcal{A}, P)$, every s (or any event) in $\mathcal{S}$ has a nonnegative probability, thus every real value (or set of values) of the random variable inherit probabilities from P. If the r.v. is called X, we call its probability distribution P_X to distinguish it from the original P defined on $\mathcal{S}$. This probability distribution is defined as follows. Let $B \subseteq \mathbb{R}$, then $P_X(B) = P(X \in B) = P(\{s \in \mathcal{S} : X(s) \in B\})$.

To fix ideas consider the following simple example.

Example 0.5 *Consider throwing a regular die with $\mathcal{S} = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$ once. You win -1\$ (i.e. you loose 1\$) if the it realises $\boxdot, \boxdot, \boxdot$, you win 0\$ if it realises $\boxdot$ and you win 1\$ if it realises $\boxdot$ or $\boxdot$. Let X be the r.v. amount of win in \$. Thus $X = -1, 0, 1$. We see that $X = -1$ realises only if $s \in \{\boxdot, \boxdot, \boxdot\}$, $X = 0$ only if $s \in \{\boxdot\}$ and $X = 1$ if $s \in \{\boxdot, \boxdot\}$. Therefore we have*

$$\begin{aligned} P_X(\{-1\}) = P(X = -1) &= P(\{s \in \mathcal{S} : X(s) = -1\}) \\ &= P(\{\boxdot, \boxdot, \boxdot\}) \\ &= P(\{\boxdot\} \cup \{\boxdot\} \cup \{\boxdot\}) \\ &= P(\{\boxdot\} + P(\{\boxdot\}) + P(\{\boxdot\}) \\ &= \tfrac{3}{6}. \end{aligned}$$

Similarly, we find that $P(X = 0) = \frac{1}{6}$ and $P(X = 1) = \frac{2}{6}$.

An r.v. is called *discrete* if there are countable, i.e. finitely many (as in Example 0.5) or denumerably infinite many points in $\mathbb{R}$, $x_1, x_2, \ldots$, such that $P_X(\{x_j\}) > 0, j \geq 1$ and $\sum_j P_X(\{x_j\}) = \sum_j P(X = x_j) = 1$. Then the function f_X defined on $\mathbb{R}$ by the relationship

$$f_X(x) = \begin{cases} P_X(x) = P(X = x) & \text{if} \quad x = x_j \\ 0 & \text{otherwise} \end{cases},$$

has the properties $f_X(x) \geq 0$ for all x, and $\sum_j f_X(x_j) = 1$. Furthermore $P(X \in B) = \sum_{x_j \in B} f_X(x_j)$. The function $f_X(x)$ is called *probability density function* (p.d.f.) of X.

Suppose now that X is an r.v. which takes values in a an interval I of $\mathbb{R}$ with the following quantification: $P(X = x) = 0$ for every single $x \in I$. This r.v. is called *continuous*. For the type of r.v. that we will be dealing with, such r.v. do possess a function f_X such that $f_X(x) \geq 0$ for all $x \in I$ and $P(X \in B) = \int_B f_X(x)\, dx$, for any subinterval B of I. This function is called again probability density function of X, but notice however that in this case $f_X(x)$ does not represent $P(X = x)$, which is always zero. As rule of thumb to check that a given f_X is a probability density function, it is sufficient to check that

(i) $f_X(x) \geq 0$, for all $x \in \mathcal{X}$

(ii) $\int_{x \in \mathcal{X}} f_X(x)\, dx = 1$,

[1] More technically, we require this function to be a *measurable* mapping between the two sets. However, such technicalities are not relevant for this course since all r.v. that we will meet in this course are measurable.

where $\mathcal{X}$ is the set of all possible values of the r.v. X. For the models considered in this lecture, typically $\mathcal{X}$ is equal to $\mathbb{R}$, or to $\mathbb{R}_{>0}$, the set of positive real numbers or $R_{\geq 0}$ the set of non-negative real numbers. Since any probability density function f_X of an r.v. X can always be defined in such a way that the set $\mathcal{X}$ coincides with $\mathbb{R}$, there is no harm in assuming that $\mathcal{X} = \mathbb{R}$. Thus, hereafter for a continuous r.v. X we assume that the set of all possible values of X is $\mathbb{R}$. Except for some very particular cases, the function f_X, if known, characterises the distribution of X, in the sense that if two r.v. X and Y have the same f_X then $X = Y$.

0.4.1 The cumulative distribution function

If we consider subsets B of $\mathbb{R}$ which are intervals closed on the right, i.e. $B = \{y \in \mathbb{R} : y \leq x\}$ for a given x, then $P_X(B)$ is denoted by $F_X(x)$ and is called the *cumulative distribution function* of X or just *distribution function* (d.f.) of X. To ease notation, we will omit the subscript when no confusion arises. Thus F, is an ordinary point function with values in [0,1], i.e. $F : \mathbb{R} \to [0, 1]$.

Theorem 0.5 *For the distribution function of a r.v. X, the following holds.*

(i) *F is nondecreasing;*

(ii) *F is continuous from the right;*

(iii) *$F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$; we express this by writing $F(-\infty) = 0, F(\infty) = 1$.*

Also the distribution function fully characterises the probability distribution of a r.v. X, that is, there is a bijective mapping between X and its distribution function F_X.

Remark 0.1 *(i) $F(x)$ can be used to find probabilities, such as $P(a < X \leq b)$, for $a < b$,*

$$P(a < X \leq b) = F(b) - F(a).$$

Indeed, $\{-\infty < X \leq b\} = \{-\infty < X \leq a\} + \{a < X \leq b\}$ (recall that "+" here means union of disjoint sets), thus

$$P(-\infty < X \leq b) = F(b) = F(a) - P(a < X \leq b).$$

(ii) If X is discrete, its d.f. is a "step" function and is defined by

$$F(x) = \sum_{x_j \leq x} f(x_j), \quad \text{and} \quad f(x_j) = F(x_j) - F(x_{j-1}),$$

assuming $x_1 \leq x_2 \leq \cdots$.

(iii) If X is continuous, its d.f. is continuous and $F(x) = \int_{-\infty}^{x} f(t)dt$ and $\frac{dF(x)}{dx} = f(x)$, at continuity points of f.

Here is an Example of a discrete r.v.; an example of continuous random variable will be given latter.

Example 0.6 *Consider the r.v. in Example 0.5. Then the d.f. of X is*

$$F(x) = \begin{cases} 0 & \text{if} \quad x < -1 \\ \frac{3}{6} & \text{if} \quad -1 \leq x < 0 \\ \frac{4}{6} & \text{if} \quad 0 \leq x < 1 \\ \frac{6}{6} & \text{if} \quad x \geq 1. \end{cases}$$

Similarly, we find that $P(X = 0) = \frac{1}{6}$ and $P(X = 1) = \frac{2}{6}$. The d.f. and the density function are depicted in Figure 0.1.

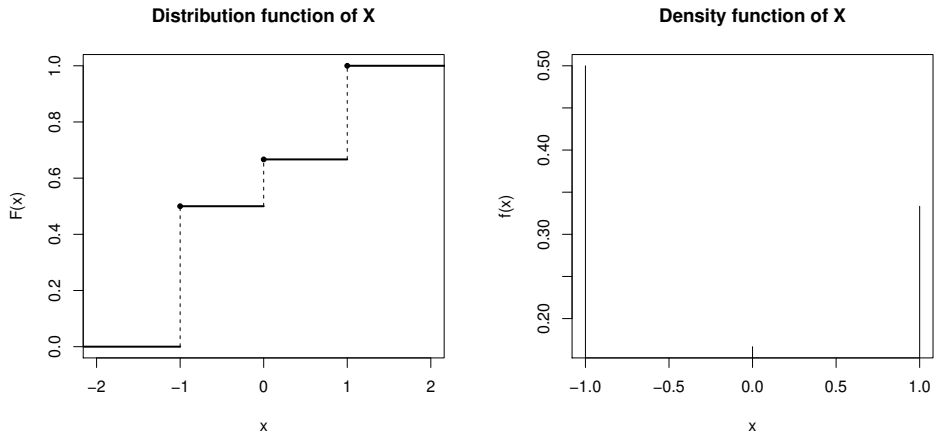


Figure 0.1: Two plots for the r.v. X in Example 0.6. Left: the distribution function of X, the filled dots suggest that the value is included; right: the probability density function.

0.4.2 The quantile function

Let X be an r.v. with d.f. F and consider a number p such that $0 < p < 1$. A pth *quantile* of the *r.v.* X is a number denoted by x_p with the property: $P(X \leq x_p) \geq p$ and $P(X \geq x_p) \leq 1 - p$. For $p = 0.25$ we get a *first quartile* of X, for $p = 0.5$ we get a *median* (or second quartile) of X and for $p = 0.75$ we get the *third quartile*.

If the r.v. X is continuous, then all quantiles are unique and we can talk about *the* pth quantile x_p. In this case, the pth quantile x_p has a more intuitive definition and is given by the solution to the equation $F(x_p) = p$. Another notation for the pth quantile is $F^{-1}(p)$, and for any $0 < p < 1$, $F^{-1}(p) : p \to \mathbb{R}$ is called the *quantile function*. As suggested by the notation, the function $F^{-1}(p)$ is the inverse of F, provided that this is well defined. If F is strictly increasing, which is always true for continuous r.v., then $F(F^{-1}(p)) = p$ and $F^{-1}(F(x)) = x$ for all $p \in (0, 1)$ and $x \in \mathbb{R}$. When the r.v. is not continuous everywhere, then F has jumps and it may be piecewise constant, thus the equation $F(x) = p$ for a given p can have no solutions, exactly one solution or infinitely many solutions. A more general definition of the quantile function that works in all cases is

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}, \quad p \in (0, 1).$$

Example 0.7 *Consider the r.v. in Example 0.6. We find that (see Figure 0.1)* $x_{0.25} = x_{0.5} = -1$ *and* $x_{0.8} = 1$.

0.4.3 Moments

Let X be a r.v. with density function f. Then, for $n = 1, 2, \ldots$, the nth moment of X is denoted by $E(X^n)$ and is defined by

$$E(X^n) = \begin{cases} \sum_x x^n f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

$E(X)$ is also called the expectation of X; an alternative notation for $E(X)$ is μ_X. For an arbitrary constant

c and n as above, the nth *moment of X about c* is denoted by $E[(X-c)^n]$ and is defined by

$$E[(X-c)^n] = \begin{cases} \sum_x (x-c)^n f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-c)^n f(x)dx & \text{if } X \text{ is continuous.} \end{cases}$$

For $c = E(X)$, the moments are called *central moments*. The 2nd central moment of X, is also called the variance of X and also denoted by σ_X^2 or $\text{var}(X)$.

Example 0.8 *Consider the r.v. in Example 0.6. We find that* $E(X) = -1 \cdot \frac{3}{6} + 0 \cdot \frac{1}{6} + 1 \cdot \frac{2}{6} = -\frac{1}{6}$.

Here are some properties for the expectation and the variance that are worth remembering.

Theorem 0.6 *Let $X, X_1, X_2, \ldots, X_n$ be r.v.'s which have finite expectation and finite variance. Then*

(i) *For any constant $a \in \mathbb{R}$, $E(a) = a$ and* $\text{var}(a) = 0$.

(ii) *For any constants $a, b \in \mathbb{R}$, $E(a+bX) = a + bE(X)$ and* $\text{var}(a+bX) = b^2\text{var}(X)$.

(iii) *For any constants $c_1, \ldots, c_n \in \mathbb{R}$, $E(\sum_{j=1}^n c_j X_j) = \sum_{j=1}^n c_j E(X_j)$.*

(iv) $\text{var}(X) = E(X^2) - [E(X)]^2$.

(iv) *If $X \geq 0$ then $E(X) \geq 0$; more generally, if $X \geq X_1$ then $E(X) \geq E(X_1)$.*

(v) $|E(X)| \leq E(|X|)$.

(vi) *For any $\epsilon > 0$, then $P(|X-\mu| < \epsilon) \geq 1 - \frac{\sigma^2}{\epsilon^2}$, where $\mu = E(X), \sigma^2 = \text{var}(X)$; this is known as the Chebyshev inequality.*

0.4.4 Transformation of random variables

Although X and its d.f. F_X could be enough for our purposes, we might however wish to transform the values of X by applying some function $g(\cdot) : \mathbb{R} \to \mathbb{R}$ to get a new r.v. Y. For instance, suppose we are measuring the load of a certain virus that is present in a patient's blood. Instead of reporting x, the number of viruses measured, we may wish to report $y = \log_{10}(x)$. The question is then: given X and its d.f., is Y a r.v. and how to determine F_Y ? The answer is given by the next theorem.

Theorem 0.7 *Let X be a continuous r.v. with d.f. F_X and consider the function $g(x) : \mathbb{R} \to \mathbb{R}$, which is bijective and its inverse is $g^{-1}(y)$. Then $Y = g(X)$ is a r.v. with d.f. F_Y and density function f_Y given respectively by*

$$F_Y(y) = F_X(g^{-1}(y)), \quad \text{and} \quad f_Y(y) = \tfrac{dF_X(g^{-1}(y))}{dy} = f_X(g^{-1}(y)) \left| \tfrac{dg^{-1}(y)}{dy} \right|.$$

If the function $g()$ is surjective but not injective, i.e. there may be $x_1 \neq x_2 \in \mathbb{R}$ such that $g(x_1) = g(x_2) = y$, then the first formula for $f_Y(y)$ can still be used, whereas the formula after the second equality should be applied to subsets of Y where g() is injective.

The transformation function $g(x) = (x-\mu)/\sigma$, where $\mu = E(X), \sigma = \sqrt{\text{var}(X)}$ is of special interest in statistics and is called *standardisation*. This is so because if we let $Z = (X-\mu)/\sigma$ then $E(Z) = 0$ and $\text{var}(Z) = 1$.

Definition 0.6 *Given a r.v. Y with d.f. F_Y and $a, b \in \mathbb{R}$ with $b > 0$, let $T = a + bY$. Then the r.v. T has d.f. $F_{a,b}(y) = F_Y\left(\frac{y-a}{b}\right)$ and the set $\{F_{a,b} : a, b \in \mathbb{R}, b > 0\}$ is called the* location-scale family *of distributions for Y. If $b = 1$, then $F_{a,1} = F_{a,}$ is called* location *family. If $a = 0$, then $F_{0,b} = F_{,b}$ is called* scale *family.*

0.5 Random vectors and their distributions

So far we focused on a single r.v. at time, but reality is multidimensional. For instance, suppose we need to check the performance of a washing machine taken at random from the production line. There are several variables involved, such as the performance of washing cycle say X_1, the duration of the cycle X_2, energy consumption X_3, water consumption X_4, spin speed X_5, etc.

Let $X = (X_1, X_2, \ldots, X_k)$ be a well-behaved vector-valued function of reals, i.e. $X(s) : \mathcal{S} \to \mathbb{R}^k$ on a given a probability triple $(\mathcal{S}, \mathcal{A}, P)$. Then X is called *random* vector and is abbreviated by r.ve.. Likewise in the case of an r.v., the r.ve. X is characterised by its joint d.f defined as $F(x) = P(X \leq x) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_n)$ and by its joint probability density function $f_X(x)$, when it exists. In particular, the r.ve. is called discrete if $P(X = x_j) > 0$, $j = 1, 2, \ldots$, with $\sum_j P(X = x_j) = 1$, and the function $f_X(x) = P(X = x_j)$ for $x = x_j$ and $f_X(x) = 0$ is the p.d.f. of X. Once again $P(X \in B) = \sum_{x_j \in B} f_X(x_j)$ for $B \subseteq \mathbb{R}^k$. The r.ve. X is continuous if $P(X = x) = 0$ for all $x \in J \subseteq \mathbb{R}^k$ but there is a function f_X defined on $\mathbb{R}^k$ such that:

$$f_X(x) \geq 0 \quad \text{for all} \quad x \in \mathbb{R}^k, \quad \text{and} \quad P(X \in J) = \int_J f_X(x) dx_1 dx_2 \cdots dx_k.$$

In the incoming lectures we will see two examples of r.ve.'s: the multinomial distribution which is a discrete r.ve. and the multivariate normal distribution. To make notation easier, in the following sections we specialise our discussion to $k = 2$.

0.5.1 Marginal distributions and their moments

Thus let $X = (X_1, X_2)$ be a *bivariate* r.ve. The components X_1 and X_2 are both r.v.'s and their distribution has to be derived from that of $F_X(x)$ as we outline below.

The marginal p.d.f. of X_1, is

$$f_{X_1}(x_1) = \begin{cases} \sum_t P(X_1 = x_1, X_2 = t) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} f_X(x_1, t) dt & \text{if } X \text{ is continuous.} \end{cases}$$

Thus the marginal distribution is obtained by marginalising over the nuisance component by either summation or integration depending on the type of the r.ve.. In the rest of this section we focus on the continuous case; the case of discrete r.ve. is analogous with integrals replaced by sums.

The expectation of X is $\mu = (\mu_1, \mu_2)$, where

$$\mu_1 = \int_{-\infty}^{\infty} t f_{X_1}(t) dt.$$

Furthermore, the variance of X_1 is defined as usual

$$\sigma_1^2 = \int_{-\infty}^{\infty} (t - \mu_1)^2 f_{X_1}(t) dt.$$

We denote by $\text{cov}(X_1, X_2)$ the covariance between X_1 and X_2 which is defined by

$$\begin{aligned}\text{cov}(X_1, X_2) &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E(X_1X_2) - E(X_1)E(X_2) \\ &= \mu_{12} - \mu_1\mu_2 \\ &= \sigma_{12}.\end{aligned}$$

Furthermore, the coefficient correlation denoted by ρ_{12}, is defined by $\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2}\sqrt{\sigma_2^2}}$. Note that $\sigma_{12} \in \mathbb{R}$ and $\rho_{12} \in [-1, 1]$. Note that for any r.v. X which has finite variance, $\text{cov}(X, X) = \text{var}(X)$.

0.5.2 Conditional distributions and moments

Conditional distribution functions are defined similarly to the conditional probability of events (see § 0.3.1). Let $X = (X_1, X_2)$ be an r.ve. with joint p.d.f $f_X(x) = P(X_1 = x_1, X_2 = x_2)$, then we p.d.f. of X_1 given the event $X_2 = x_2$ is defined as

$$P(X_1 = x_1|X_2 = x_2) = \tfrac{P(X_1=x_1, X_2=x_2)}{P(X_2=x_2)}.$$

If X is a continuous r.ve. with p.d.f. $f_X(x)$, then the conditional p.d.f. of X_1 given $X_2 = x_2$ is defined

$$f_{X_1|X_2}(x_1|x_2) = \tfrac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)}.$$

Remark 0.2 *(i) For any fixed x_2, the conditional distributions are proper probability distributions. This can be seen from the fact that, in the discrete case, summing over all possible x for X_1 leaves us with the probability of the event $X_2 = x_2$, i.e. $\sum_t P(X_1 = t, X_2 = x_2) = P(X_2 = x_2)$. A similar observation applies to the continuous case, for which we have*

$$\int_{-\infty}^{\infty} f_X(x)dx_1 = \int_{-\infty}^{\infty} f_X(x_1, x_2)dx_1 = f_{X_2}(x_2).$$

(ii) Note that conditional distributions shown above are functions indexed by x_2. This means that for any s, t in the domain of X_2, if $s \neq t$ then, in general $P(X_1 = x_1|X_2 = s)$ is different from $P(X_1 = x_1|X_2 = t)$. That is, changing the conditioning event could lead to a different conditional distribution.

If conditional distribution admit moments, these are called *conditional moments*. We focus only on the expectation and variance. The *conditional expectation* of X_1 given $X_2 = x_2$ is denoted by $E(X_1|X_2 = x_2)$ or $\mu_{X_1|X_2}$ is defined by

$$E(X_1|X_2 = x_2) = \begin{cases} \sum_t tP(X_1 = t|X_2 = x_2) & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} tf_{X_1|X_2}(t|x_2)dt & \text{if } X \text{ is continuous} \end{cases}.$$

The *conditional variance* of X_1 given $X_2 = x_2$ is denoted by $\text{var}(X_1|X_2 = x_2)$ and is defined by

$$\text{var}(X_1|X_2 = x_2) = \begin{cases} \sum_t (t - \mu_{X_1|X_2})^2 P(X_1 = t|X_2 = x_2) & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} (t - \mu_{X_1|X_2})^2 f_{X_1|X_2}(t|x_2)dt & \text{if } X \text{ is continuous} \end{cases}.$$

0.5.3 Independence of random variables

We remarked in Remark 0.2 that in general the conditional p.d.f. could change if the conditioning events changes. If on the contrary the conditional distribution does not vary with the conditioning event, then the two r.v. involved are said to be *independent*. More concretely, similarly to the concept of independence of events, two random variables X_1 and X_2 are said to be independent if their events $(X_1 = x_1)$ and $(X_2 = x_2)$ are independent, for any admissible x_1, x_2. This is stated formally by the next definition.

Definition 0.7 *Let X_1 and X_2 be two r.v.s with joint d.f. F_X. Then X_1 is independent from X_2 if*

$$P(X_1 = x_1|X_2 = x_2) = P(X_1 = x_1) \quad \textit{and} \quad P(X_2 = x_2|X_1 = x_1) = P(X_2 = x_2), \quad \textit{for all } x_1, x_2,$$

when X_1, X_2 are both discrete. X_1 is independent from X_2 if

$$f_{X_1|X_2}(x_1|x_2) = f_{X_1}(x_1) \quad \textit{and} \quad f_{X_2|X_1}(x_2|x_1) = f_{X_2}(x_2), \quad \textit{for all } x_1, x_2,$$

when X_1, X_2 are both continuous.

Another useful characterisation of independence which applies to discrete as well as continuous random variable is the following.

Theorem 0.8 *The r.v.s $X_1, \ldots, X_n$ are independent if and only if one of the following two (equivalent) conditions hold:*

(i) $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{j=1}^n F_{X_j}(x_j)$, *for all* $x_j \in \mathbb{R}$, $j = 1, \ldots, n$.

(ii) $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{j=1}^n f_{X_j}(x_j)$, *for all* $x_j \in \mathbb{R}$, $j = 1, \ldots, n$.

Furthermore, we have the following properties.

Theorem 0.9 *Let $X_1, \ldots, X_n$ be r.v.s all of the same type, i.e either all discrete or all continuous and having joint d.f. F_X. Then:*

(i) If X_1 and X_2 are independent r.v., then $E(X_1X_2) = E(X_1)E(X_2)$ and $\mathrm{cov}(X_1, X_2) = 0$.

(ii) Given reals a_0, a_1, b_0, b_1 and $X_a = a_0 + a_1X_1, X_b = b_0 + b_1X_2$, then

$$\mathrm{cov}(X_a, X_b) = \mathrm{cov}(a_0 + a_1X_1, b_0 + b_1X_2) = a_1b_1\mathrm{cov}(X_1, X_2).$$

(iii) If $Y = a + bX_1$ for any reals a, b, then $|\mathrm{cov}(X_1, Y)| = \sigma_{X_1}\sigma_Y$.

(iv) Let $\mu_1 = E(X_1), \mu_2 = E(X_2)$, $\sigma_1^2 = \mathrm{var}(X_1), \sigma_2^2 = \mathrm{var}(X_2)$, $\mathrm{cov}(X_1, X_2) = \sigma_{12}$ and set $X = aX_1 + bX_2$, for any real a, b. Then

$$E(X) = a\mu_1 + b\mu_2, \quad \mathrm{var}(X) = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\sigma_{12}.$$

Furthermore, if $\sigma_{12} = 0$ then $\mathrm{var}(X) = a^2\sigma_1^2 + b^2\sigma_2^2$.

(v) More generally if $(b_1, \ldots, b_n) \in \mathbb{R}^n$ and $T = \sum_{j^1}^n b_jX_j$ then

$$E(X) = \sum_{j=1} b_jE(X_j), \quad \mathrm{var}(X) = \sum_{j=1}^n b_j^2\mathrm{var}(X_j) + \sum_{i\neq j} b_jb_j\mathrm{cov}(X_i, X_j).$$

Furthermore, if $\mathrm{cov}(X_i, X_j) = 0$ for all $i \neq j = 1, \ldots, n$, then $\mathrm{var}(X) = \sum_{j=1}^n b_j^2\mathrm{var}(X_j)$.

0.5.4 Further results on random vectors

So far we treated r.ve.'s $X = (X_1, X_2, \ldots, X_n)$ by dealing with its components X_i avoiding matrix and vector notation. However, in some cases the matrix/vector notation leads to more compact notation and thus it is worth knowing. As usual X, is always meant to be column vector, thus in this case it is an $(n \times 1)$ vector. If we let $\mu_i = E(X_i)$, for $i = 1, 2, \ldots, n$ and assuming all $\mu_i < \infty$, then we denote by $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ the column vector of expected value of the vector X. Furthermore, let $\sigma_i^2 = \text{var}(X_i)$, and $\sigma_{ij} = \text{cov}(X_i, X_j)$, $(i, j = 1, 2, \ldots, n)$. Then we define the covariance matrix of X by

$$\Sigma = \text{var}(X) = E\left((X - \mu)(X - \mu)^{\text{T}}\right) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}.$$

For Σ to exists it is sufficient that the diagonal elements are non-zero.

The correlation matrix is denoted by $P = [\rho_{ij}]$, where ρ_{ij} is the element corresponding to the ith row and jth column of P, with $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_i^2 \sigma_j^2}$. In matrix notation, the correlation matrix P can be written as

$$P = \Delta^{-1/2} \Sigma \Delta^{-1/2}, \quad \text{with} \quad \Delta = \text{diag}(\Sigma) = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2).$$

The following properties are useful.

Theorem 0.10 *Let X be an n-dimensional r.ve. with mean vector μ and covariance matrix Σ. Then*

(i) Σ is symmetric, i.e. $\sigma_{ij} = \sigma_{ji}$. Furthermore, Σ is positive semi-definite, i.e. for any $a \in \mathbb{R}^n$, $a^{\text{T}} \Sigma a \geq 0$.

(ii) Let $A = [a_{ij}]$ be a $(p \times n)$ matrix and $b = (b_1, b_2, \ldots, b_p)$ a $(p \times 1)$ vector. If $Y = AX + b$, then Y is a r.ve. and

$$E(Y) = A\mu + b, \quad \text{var}(Y) = A\Sigma A^{\text{T}}$$

0.6 Convergence of random variables

In this section we deal with convergence properties of sequence of random variables. That is, given a sequence of r.v. $\{X_n\}$, we will try to understand what happens to X_n as $n \to \infty$. As we will see, there are different types of convergence which roughly speaking depend on the degree on "niceness" of the distribution of X_n, for all $n \geq 1$.

Hereafter, the notation $X \sim F_X$ means: the r.v. X has distribution F_X.

The following definition introduces a concept that is will be extensively used throughout these handouts.

Definition 0.8 *The r.v. $X_1, \ldots, X_n$ are defined to be identically and independently distributed (abbreviated i.i.d.) if $X_i \sim F_X$ (i.e. X_i are identically distributed), and if $X_1, \ldots, X_n$ are independent (see Theorem 0.8), for all $i = 1, 2, \ldots, n$. In compact notation this is denoted by $X_i \overset{\text{iid}}{\sim} F_X$ for all $i = 1, \ldots, n$.*

0.6.1 Modes of convergence

Definition 0.9 *A sequence of r.v.'s $\{X_n\}$ is said to be convergent in quadratic mean to a r.v. X if*

$$\lim_{n\to\infty} E(X_n - X)^2 = 0.$$

A shorthand notation for this type of convergence is $X_n \xrightarrow{\text{q.m.}} X$.

Definition 0.10 *A sequence of r.v.'s $\{X_n\}$ is said to be almost surely (abbreviated a.s.) convergent to an r.v. X if*

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

This type of convergence is denoted by $X_n \xrightarrow{\text{a.s.}} X$.

Definition 0.11 *A sequence of r.v.'s $\{X_n\}$ is said to be convergent in probability to an r.v. X if for any $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1.$$

This type of convergence is denoted by $X_n \xrightarrow{P} X$.

Definition 0.12 *A sequence of r.v.'s $\{X_n\}$, with each term having d.f. F_n, is said to be convergent in distribution to an r.v. X with d.f. F if*

$$\lim_{n\to\infty} F_n(t) = F(t), \quad \textit{for all } t \textit{ where } F \textit{ is continuous.}$$

This type of convergence is denoted by $X_n \xrightarrow{d} X$.

Here are some relations between the four types of convergence.

Theorem 0.11 *Let $\{X_n\}$ be a sequence of r.v.'s with each term having d.f. F_n, and let X be an r.v. with d.f. F. Then*

(i) If $X_n \xrightarrow{\text{q.m.}} X$ then $X_n \xrightarrow{P} X$.

(ii) If $X_n \xrightarrow{\text{q.m.}} X$ then $X_n \xrightarrow{d} X$.

(iii) If $X_n \xrightarrow{\text{a.s.}} X$ then $X_n \xrightarrow{P} X$.

(iv) If $X_n \xrightarrow{\text{a.s.}} X$ then $X_n \xrightarrow{d} X$.

(v) If $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{d} X$.

0.6.2 The Central Limit Theorem

We now formulate the celebrated Central Limit Theorem (CLT) in its simplest form.

Theorem 0.12 *Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s, with mean μ and variance $\sigma^2 > 0$ both assumed to be finite. Let $\overline{X}_n = \frac{1}{n}\sum_{j=1}^n X_j$, $G_n = P\left[\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq x\right]$ and $\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$. Then*

$$\lim_{n\to\infty} G_n(x) = \Phi(x), \quad \textit{for every } x \in \mathbb{R}.$$

Remark 0.3 *(i) Loosely speaking the CLT says that for large n, the distribution function of the standardised sums $\frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}}$ is close the function $\Phi(x)$, where $S_n = \sum_{j=1}^n X_j$. What "large" and "close" mean is not easy to specify as the value of n and a degree of closeness depend on the nature of the common distribution of the X_n. The experience we'll gain as we proceed will afford some guidance in this regard.*

(ii) Since the CLT involves point-wise convergence of distribution functions, it is also a mode of convergence in distribution. Thus an alternative way of expressing the CLT is by saying that

$$Z_n = \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \xrightarrow{d} Z,$$

where Z is an r.v. with d.f. $\Phi(x)$. Strictly speaking, we should show that $\Phi(x)$ is a d.f. of an r.v.. To this end, we anticipate here that $\Phi(x)$ is the d.f. of a normal random variable, to be introduced in Lecture 1.

0.6.3 Laws of Large Numbers

In this section we focus on certain limit theorems which are known as *laws of large numbers* (LLN). We distinguish two categories of LLN: the strong LLN (SLLN) in which the convergence involved is the a.s convergence, and the weak LLN (WLLN) where the convergence involved is convergence in probability.

Theorem 0.13 *Let $\{X_n\}$ be a sequence of i.i.d. r.v.'s, with finite mean μ, then*

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{\text{a.s.}} \mu, \quad \textit{as } n \to \infty.$$

Theorem 0.14 *Let $\{X_n\}$ be a sequence of i.i.d. r.v.'s, with finite mean μ, then*

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{P} \mu, \quad \textit{as } n \to \infty.$$

0.6.4 The Slutsky Lemma and the delta method

We state now results that concern the convergence of functions of two sequence of r.v.'s generally known as Slutsky's Lemma.

Theorem 0.15 *(Slutsky's Lemma) Let $\{X_n\}$ be a sequence of r.v.'s such that $X_n \xrightarrow{d} X$ and let $\{Y_n\}$ be another sequence of r.v.'s such that $Y_n \xrightarrow{P} c$, with c a fixed constant. Then*

(i) $X_n Y_n \xrightarrow{d} cX$.

(ii) $X_n + Y_n \xrightarrow{d} X + c$.

(iii) $X_n / Y_n \xrightarrow{d} X/c$, *provided* $P(Y_n \neq 0) = 1$, $c \neq 0$

(iv) If $g(\cdot)$ *is a continuous function, then* $g(Y_n) \xrightarrow{P} g(c)$.

Often the distribution of $g(X_n)$ for a fixed n is not easy to derive or it has not an easy expression. The delta method then provides a useful approximation.

Theorem 0.16 *(Delta Method) Let* $X_i \overset{\text{iid}}{\sim} F$ *for* $i = 1, \ldots, n$, *with mean* μ *and variance* $\sigma^2 > 0$ *and let* $g : \mathbb{R} \to \mathbb{R}$ *be differentiable with derivative continuous at* μ, $g'(\mu) \neq 0$. *Then as* $n \to \infty$,

$$\frac{\sqrt{n}(g(\overline{X}_n) - g(\mu))}{\sigma |g'(\mu)|} \xrightarrow{d} Z,$$

where Z *is a r.v. with d.f.* $\Phi(x)$, *the latter defined in Theorem 0.12.*

References

[LM18] Larsen, R. J. and Marx, M. L. (2018) *An Introduction to Mathematical Statistics and its Applications* (6th edition), Pearson Education, Inc., Chapp. 2-4.

Inferential Statistics **First semester, 2020**

Lecture 1: Basic parametric families of random variables

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

This lecture is based on [LM18] §4,§7.3, §10.2 and describes some of the random variables frequently used in statistics.

In this lecture we study several families of random variables and random vectors, and the associated probability distributions, in some detail. Some are classes of discrete distributions and others continuous. Some involve a single parameter and others, two or more parameters. The families to be considered are rather basic, often used as statistical models, to represent variables we observe and measure.

It is always assumed that there is a probability triple $(\mathcal{S}, \mathcal{A}, \mathcal{P})$ for the problem at hand, where $\mathcal{S}$, depending on the problem, could be either countable or uncountable such as the set $\mathbb{R}$.

1.1 Discrete random variables and random vectors

1.1.1 The Binomial distribution

A *Bernoullian experiment* is random experiment with only two possible outcomes: either the event A realises or it does not, i.e. A^c realises. Examples are: the toss of a coin where it could either realise heads (A) or tails (A^c), the roll of a die where numbers are partitioned in even (A) and odd (A^c), etc. We classify the event of interest to us as the "success" (say A) and its complement is called "failure" (A^c). The *Bernoulli r.v.* is defined by

$$X(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{if } s \in A^c \end{cases}.$$

Since A is an event pertaining to some probability space $(\mathcal{S}, \mathcal{A}, \mathcal{P})$ it must have a probability. Let thus $\theta = P(A)$, where $\theta \in (0, 1)$. Then the p.d.f. of the Bernoulli r.v. is defined by

$$P(X = x) = \theta^x (1 - \theta)^{1-x}.$$

We see that $E(X) = \sum_x x P(X = x) = 0 \cdot (1 - \theta) + 1 \cdot \theta = \theta = P(A)$. Thus the expectation of a Bernoulli r.v. equals its success probability. Similarly we find $\text{var}(X) = \theta(1 - \theta)$. We use the notation $X \sim \text{Ber}(\theta)$, to mean that the r.v. X has a Bernoulli distribution with a single parameter θ, called *success probability.*

Now suppose we toss n identical coins, each having probability θ of heads (H) and we ask about the probability of observing 2 heads out of n. If we let E be the event "2 events out of n are head" and denote tails by T, then E realises if

$$E = \{\underbrace{HHT\cdots T}_{n}, \underbrace{HTHT\cdots T}_{n}, \underbrace{THHT\cdots T}_{n}, \ldots\},$$

Now each sample point $s \in E$ involves 2 H's and $n-2$ T's and because the tosses are independent, the probability of say $s = HHT\cdots T$ is

$$P(\{s\}) = P(\{HHT\cdots T\}) = P(H)P(H)P(T)\cdots P(T) = \theta^2(1-\theta)^{n-2},$$

It is clear that for any $s \in E$, $P(\{s\}) = \theta^2(1-\theta)^{n-2}$. Note also that, elements of E are sample points, thus only one can be observed. This means that

$$\begin{aligned} P(E) &= P(\{s_1\}) + P(\{s_2\}) + \cdots + P(\{s_{C_{2,n}}\}) \\ &= C_{2,n}P(\{s_1\}) \\ &= C_{2,n}\theta^2(1-\theta)^{n-2}, \end{aligned}$$

.

where $C_{2,n} = \binom{n}{2} = \frac{n!}{2!(n-2)!}$ is the number of possible orderings of n elements in which there two identical F's and $n-2$ identical T's.

More generally, the probability of getting $y \leq n$ heads in n tosses is given by

$$P(Y = y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad y \in \{0, 1, 2, \ldots, n\}.$$

An r.v. Y with the above probability distribution distribution is called *binomial r.v.* and is denoted by $Y \sim \text{Bin}(n, \theta)$. The Binomial distribution thus also has a single parameter, θ, which has the same interpretation as in the Bernoulli distribution and n is called *index*. The binomial r.v. can also be obtained as the sum of n i.i.d Bernoulli r.v.. Indeed, if $X_i \overset{\text{iid}}{\sim} Ber(\theta)$, for $i = 1, \ldots, n$, then for the r.v. $Y = \sum_{i=1}^n X_i$, it follows that $Y \sim \text{Bin}(n, \theta)$. Using this last fact we have that

$$E(Y) = E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i) = n\theta\,, \quad \text{var}(Y) = n\theta(1-\theta).$$

Remark 1.1 *(i) In calculating the mean and average of the binomial r.v. we used the linearity property of the expectation (L0, Theorem 0.9) and the fact that this r.v. arises as the sum of n i.i.d Bernoulli r.v. with parameter θ.*

(ii) Since the r.v. Y is the sum of n i.i.d. r.v. each of which has finite variance, then we can apply the Central Limit Theorem in order to approximate the d.f. of Y. For instance suppose $Y \sim \text{Bin}(8, 0.5)$ and let us calculate $P(Y \leq 5)$. Applying the definition of the binomial distribution gives

$$P(Y \leq 5) = \sum_{i=0}^{5}\binom{8}{i}0.5^i(1-0.5)^{(8-i)} = 0.8555.$$

However, $P(Y \leq 5) = P\left(\frac{Y-E(Y)}{\sqrt{\text{var}(Y)}} \leq \frac{5-E(Y)}{\sqrt{\text{var}(Y)}}\right)$ and $E(Y) = 8 \cdot 0.5 = 4$, $\text{var}(Y) = 8 \cdot 0.5 \cdot 0.5 = 2$, then $P(Y \leq 5) = P\left(\frac{Y-E(Y)}{\text{var}(Y)} \leq \frac{1}{\sqrt{2}}\right)$. The CLT tells us that, for $n \to \infty$, $P\left(\frac{Y-E(Y)}{\sqrt{\text{var}(Y)}} \leq \frac{1}{\sqrt{2}}\right) \to \Phi\left(\frac{1}{\sqrt{2}}\right)$. Thus an approximation for $P\left(\frac{Y-E(Y)}{\text{var}(Y)} \leq \frac{1}{\sqrt{2}}\right)$ is $\Phi\left(\frac{1}{\sqrt{2}}\right) = 0.7604$. In this case, the inaccuracy of the approximation is due to n which in this case is not high enough. In general, the higher n the better will be the approximation, provided θ is far from the boundary points.

1.1.2 The Negative binomial distribution

Consider a sequence of Bernoullian experiments each with probability of success equal to θ and probability of failure $1-\theta$. Instead of counting the number of successes in a fixed number of trials, we let Y to be the number of failures observed until a fixed number of success $r \in \mathbb{N}$ are obtained. Then Y is an r.v. and is called *negative binomial* r.v., denoted $Y \sim \text{NegBin}(r, \theta)$. It has p.d.f. equal to

$$P(Y = y) = \binom{y+r-1}{y} \theta^r (1-\theta)^y, \quad y \in \mathbb{Z}_{\geq 0}.$$

where $\mathbb{Z}_{\geq 0} = \{0, 1, 2, \ldots, \}$. If $Y \sim \text{NegBin}(1, \theta)$, then Y is also called *geometric r.v.*. Here the number of parameters are r, called *index* and θ the success probability.

For this r.v. it holds that, if $Y_1 \sim \text{NegBin}(n_1, \theta), Y_2 \sim \text{NegBin}(n_2, \theta)$ and $Y = Y_1 + Y_2$, then $Y \sim \text{NegBin}(r, \theta), r = n_1 + n_1$. It holds that $E(Y) = r(1-\theta)/\theta$ and $\text{var}(Y) = r(1-\theta)/\theta^2$.

Remark 1.2 *In the definition of the* $\text{NegBin}(r, \theta)$ *r.v., instead of counting the number of failures, one could take the number of trials performed in order to obtain* r *success. The latter counting approach leads to an alternative definition of the* $\text{NegBin}(r, \theta)$*, in which the range of the r.v. is* $\{r, r+1, r+2, \ldots\}$*. However, we prefer the latter definition since it is the most useful one from a statistical modelling perspective.*

1.1.3 The Possion distribution

An r.v. Y is defined to be Poisson with parameter $\lambda \in \mathbb{R}_{>0}$, written $Y \sim \text{Poi}(\lambda)$, if its p.d.f. is

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y \in \mathbb{Z}_{\geq 0}.$$

The Poisson distribution has thus a single parameter λ, which is called *rate* or *mean* (see below why). This distribution is appropriate for modelling the number of phone calls arriving at a given telephone exchange within a certain period of time, the number of particles emitted by a radioactive source within a certain period of time, etc.

For a $Y \sim \text{Poi}(\lambda)$ r.v. it holds that $E(Y) = \text{var}(Y) = \lambda$.

1.1.4 The multinomial distribution

Consider now *generalised Bernoulli experiments*. Here we have a sequence of n independent experiments, and on each experiments the result is exactly one of the k possibilities $b_1, b_2, \ldots, b_k$. On a given trial let b_i occur with probability θ_i, $i = 1, \ldots, k$, with $\theta_i > 0$ and $\sum_{i=1}^k \theta_i = 1$.

On n independent experiments, we take $\mathcal{S} =$ all k^n ordered sequences of length n with components $b_1, b_2, \ldots, b_k$; for example if $s = b_1 b_3 b_2 b_2 \cdots b_k$, then on trial 1 occurs b_1, on trial 2 occurs b_3, on 3 and 4 occurs b_2 and so on, on the last trial, i.e. nth trial occurs b_k. To the point

$$s = \underbrace{b_1 b_1 \cdots b_1}_{y_1} \underbrace{b_2 \cdots b_2}_{y_2} \cdots \underbrace{b_k \cdots b_k}_{y_k},$$

we assign probability $\theta_1^{y_1}\theta_2^{y_2}\cdots\theta_k^{y_k}$. This is the probability assigned to any sequence having y_i occurrences of b_i, $i = 1, \ldots, k$. The number of such sequences is given by the multinomial coefficient $\frac{n!}{y_1! y_2! \cdots y_k!}$, with $n = \sum_{i=1}^k y_i$. Letting Y_i denote the r.v. which counts the occurrences of b_i, $i = 1, \ldots, k$, we have that

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k) = \frac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k},$$

and the r.ve. $Y = (Y_1, \ldots, Y_k)$ with the above p.d.f. is called multinomial r.v. and is denoted by $Y \sim \text{Mn}(n; \theta_1, \ldots, \theta_k)$. The multinomial distribution has $k-1$ parameters $\theta_1, \ldots, \theta_{k-1}$, because $\theta_k = 1 - (\theta_1 + \ldots + \theta_{k-1})$.

To fix ideas consider the following example.

Example 1.1 *Throw four unbiased dice independently. Find the probability of exactly two 1's and one 2.*

The the trial is given by the throw of a single dice. We first have to figure out k. Since we are only interested in $1, 2$ and everything else that is different from 1 and 2, the possibilities for each single trial are,

$$\begin{array}{lcll} b_1 & = & \text{"1 occurs"} & \theta_1 = \frac{1}{6}, \quad y_1 = 2 \\ b_2 & = & \text{"2 occurs"} & \theta_2 = \frac{1}{6}, \quad y_2 = 1 \\ b_3 & = & \text{"3,4,5, or 6 occurs"} & \theta_3 = \frac{4}{6}, \quad y_3 = 1 \end{array}$$

The required probability is thus

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = \frac{4}{2!1!1!}(1/6)^2(1/6)^1(1/6)^1 = 1/27.$$

The multinomial distribution is useful for modelling categorical variables which can assume one of the k possible values, such as for instance, preference choices against k transport options, or k political parties, etc.

The multinomial distribution has many interesting properties and here are some useful ones.

Theorem 1.1 *Let $Y = (Y_1, \ldots, Y_k) \sim \text{Mn}(n; \theta_1, \ldots, \theta_k)$, then:*

(i) For $k = 2$, the multinomial distribution coincides with the binomial distribution, i.e. $\text{Mn}(n; \theta_1, \theta_2) = \text{Bin}(n, \theta)$.

(ii) Each component of the r.ve. Y is a binomial r.v., i.e $Y_i \sim \text{Bin}(n, \theta_i)$, for all $i = 1, \ldots, k$.

(iii) Every d-subvector $(Y_{i_1}, \ldots, Y_{i_d})$ of Y, $d \leq k$ has a multinomial distribution, where $\{i_1, \ldots, i_d\} \subseteq \{1, 2, \ldots, k\}$.

(iv) If $X \sim \text{Mn}(n_x; \theta_1, \ldots, \theta_k)$ and $Z = Y + X$, then $Z \sim \text{Mn}(n_z; \theta_1, \ldots, \theta_k)$, with $n_z = n + n_x$.

(v) Let $X_1, \ldots, X_k$ be independent Poisson r.v.'s with the d.f. of X_i having parameter λ_i. Then the conditional distribution of $X_1, \ldots, X_k$ given their sum is multinomial:

$$P\left(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k \Big| \sum_{i=1}^{k} x_i = n\right) = \text{Mn}(n; \theta_1, \ldots, \theta_k),$$

where $\theta_i = \lambda_i / \sum_{j=1}^{n} \lambda_j$, $i = 1, \ldots, k$.

(vi) $E(Y) = (n\theta_1, \ldots, n\theta_k)$ and for all $i, j = 1, \ldots, k$

$$\text{cov}(X_i, X_j) = \begin{cases} -n\theta_i\theta_j & \text{if } i \neq j \\ n\theta_i(1 - \theta_j) & \text{if } i = j. \end{cases}$$

1.2 Continuous random variables and random vectors

1.2.1 The normal (Gaussian) distribution

An r.v. Y is said to have *normal distribution* if the p.d.f. is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad y \in \mathbb{R}, \mu \in \mathbb{R}, \sigma^2 > 0,$$

We denote this by $Y \sim \mathrm{N}(\mu, \sigma^2)$, where the mean μ and the variance σ^2 are the parameters of the distribution. For $\mu = 0$ and $\sigma^2 = 1$, the distribution is known as the *standard normal distribution*, denoted by $\mathrm{N}(0, 1)$.

The normal distribution has bell-like shape, and is symmetric around μ. It is a good approximation to the distribution of grades, heights or weights of a large group of individuals, the diameters of hail hitting the ground during a storm, errors in numerous measurements, etc. However, its main significance drives from the Central Limit Theorem. Indeed, if Y is a standard normal r.v., then $\Phi(y)$ is its d.f., i.e.

$$F(y) = P(Y \leq y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = \Phi(y).$$

Many other distributions are based on the normal r.v. as we will see below. Here are some useful properties about this distribution.

Theorem 1.2 *(i) If $Y \sim \mathrm{N}(\mu, \sigma^2)$ and $T = a + bY$, then $T \sim \mathrm{N}(a + b\mu, b^2\sigma^2)$.*

(ii) If $Y \sim \mathrm{N}(\mu, \sigma^2)$ and $Z = \frac{Y-\mu}{\sigma}$, then $Z \sim \mathrm{N}(0, 1)$; that is, a standardised normal r.v. has standard normal distribution.

(iii) Given $Z \sim \mathrm{N}(0, 1)$, then the family of r.v. $\{\mu + \sigma Z : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}\}$ is a location-scale family for Z.

(iv) If $Y_n \sim \mathrm{Bin}(n, \theta)$, then by the CLT $Y_n \xrightarrow{d} \mathrm{N}(n\theta, n\theta(1-\theta))$ (see also Remark 0.3(ii)). This follows because, $Y_n = \sum_{i=1}^{n} X_i$, where $X_i \overset{\text{iid}}{\sim} \mathrm{Ber}(\theta)$, thus $E(Y_n) = n\theta$ and $\mathrm{var}(Y_n) = n\theta(1-\theta)$ and by CLT $\frac{Y_n - E(Y_n)}{\sqrt{\mathrm{var}(Y_n)}} \xrightarrow{d} Z$ thus $\frac{Y_n - n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow{d} Z$, which implies $Y_n \xrightarrow{d} n\theta + \sqrt{n\theta(1-\theta)}Z$. But by (i): $n\theta + \sqrt{n\theta(1-\theta)}Z \sim \mathrm{N}(n\theta, n\theta(1-\theta))$, thus the desired result follows (we used $\xrightarrow{d}$ pretending it to be an $=$, which is legitimate since we are operating under an existing limit operator)

(v) If $Y_n \sim \mathrm{Poi}(n)$, then again $Y_n \xrightarrow{d} \mathrm{N}(n, n)$, because $Y_n = \sum_{i=1}^{n} X_i$, where $Y_i \overset{\text{iid}}{\sim} \mathrm{Poi}(1)$, for all $i = 1, \ldots, n$.

1.2.2 The exponential distribution

An r.v. Y with p.d.f. given by

$$f(y) = \lambda e^{-\lambda y}, \quad \text{for all } y > 0, \quad \lambda > 0,$$

is called *exponential distribution*. We denote this by $Y \sim \mathrm{Exp}(\lambda)$, where λ is the parameter of the distribution called *scale*. The exponential r.v. occurs frequently in waiting time problems.

More specifically, if Y is an r.v. denoting the waiting time between successive occurrences of events following a Poisson distribution, then Y has an exponential distribution. To see this suppose that events occur according

to the Poisson distribution Poi(λ); for instance, particles emitted by a radioactive source with average of particles per unit time equal to λ. Suppose that we have just observed such a particle, and let Y be the r.v. denoting the waiting time until the next particle emission. Since $Y \geq 0$, it follows that $F(y) = 0$ for $y \leq 0$. So let $y > 0$ be the waiting time for the emission of the next item. Then $F(y) = P(Y \leq y) = 1 - P(Y > y)$. Since λ is the average number of particles emitted in a unit of time, in x units of time we expect to see λx particles. But since no particles are observed in $(0, y]$ then $P(Y > y)$ is equal to the probability of observing zero emissions under a Poisson r.v. with expectation λx, i.e. $P(Y > y) = e^{-\lambda y}\frac{(\lambda y)^0}{0!} = e^{-\lambda y}$. That is, $F(y) = 1 - e^{-\lambda y}$, the derivative of which gives $f(y) = \lambda e^{-\lambda y}$. To summarise: we have $f(y) = 0$ for $y \leq 0$ and $f(y) = \lambda e^{-\lambda y}$ for $y > 0$.

The exponential distribution is "memoryless". In an application in which the event is failure of an electronic component of an equipment, this means that how long the electronic component will be working does not depend on how long it has been working already. In other words, the future is independent of the past, in that the future unfoldment of the process, as seen from any point in time, does not depend on what has gone on before. To show this memoryless property, consider the following conditional probability

$$P(Y > r + y | Y > r) = \frac{P(Y>r+y)}{P(Y>r)} = \frac{e^{-\lambda(r+y)}}{e^{-\lambda r}} = e^{-\lambda y},$$

which is the same as $P(Y > y)$. The future looks the same no matter when you start watching the process, and no matter hwo much you may have learned about the previous failures.

For $Y \sim \text{Exp}(\lambda)$ it holds that $E(Y) = \frac{1}{\lambda}$ and $\text{var}(Y) = \frac{1}{\lambda^2}$.

1.2.3 The gamma distribution

An r.v. Y with p.d.f. given by

$$f(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}, \quad y > 0, \quad \alpha > 0, \quad \lambda > 0$$

is called *gamma distribution*, where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-\lambda y} dy$ is the *gamma function*. We denote this by $Y \sim \text{Ga}(\alpha, \lambda)$, where α, λ are the parameters of the distribution, α is called the *shape parameter* and λ is called the *scale*.

Remark 1.3 *When working with the gamma distribution the following properties of the gamma function are worth remembering.*

(i) $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, $\alpha > 1$

(ii) $\Gamma(1) = \int_0^\infty e^{-y} dy = 1$

(iii) *For* $n \in \mathbb{N}$, $\Gamma(n) = (n-1)(n-1)\cdots 1 = (n-1)!$.

(iv) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ *and* $\Gamma\left(\frac{3}{2}\right) = \frac{1}{\sqrt{\pi}}$.

Here are some useful properties of the gamma distribution

Theorem 1.3 *(i) If* $Y \sim \text{Ga}(\alpha, \lambda)$ *then* $E(Y) = \frac{\alpha}{\lambda}, \text{var}(Y) = \frac{\alpha}{\lambda^2}$.

(ii) If $Y \sim \text{Ga}(\alpha, \lambda)$ *then* $kY \sim \text{Ga}(\alpha, \frac{\lambda}{k})$, $k > 0$.

(iii) If $Y_1 \sim \text{Ga}(\alpha_1, \lambda)$ *and* $Y_2 \sim \text{Ga}(\alpha_2, \lambda)$*, with* Y_1 *independent from* Y_2 *and* $Y = Y_1 + Y_2$*, then*

$$Y \sim \text{Ga}(\alpha_1 + \alpha_2, \lambda).$$

(iv) If $Y_i \overset{\text{iid}}{\sim} \text{Ga}(\alpha_0, \lambda)$*,* $i = 1, \ldots, n$*, by the CLT we have that*

$$\frac{\sum_{i=1}^n Y_i - n\frac{\alpha_0}{\lambda}}{\frac{\sqrt{n\alpha_0}}{\lambda}} \xrightarrow{d} \text{N}(0,1)$$

and thus $\sum_{i=1}^n Y_i \xrightarrow{d} \text{N}\left(\frac{n\alpha_0}{\lambda}, \frac{n\alpha_0}{\lambda^2}\right)$*: note that*

$$\sum_{i=1}^n Y_i \sim \text{Ga}(n\alpha_0, \lambda)$$

by induction from (ii). While the former result holds by the CLT, thus is a limiting property as $n \to \infty$*, the latter is a property of the gamma distribution thus holds exactly. Another way to express this is to say that* $\text{Ga}(n\alpha_0, \lambda) \dot{\sim} \text{N}\left(\frac{n\alpha_0}{\lambda}, \frac{n\alpha_0}{\lambda^2}\right)$*, where the symbol " $\dot{\sim}$ " means that the two distributions are approximately equal for large* n*.*

1.2.4 The Weibull distribution

A Weibull r.v. results from a power of an exponential r.v.. Suppose $X \sim \text{Exp}(1/\beta)$, so that $E(X) = \beta$ and let $Y = X^{1/\alpha}$ for $\alpha > 0$. Then the r.v. Y is said to have Weibull distribution and has d.f.

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(X^{1/\alpha} \le y) \\ &= P(X \le y^\alpha) \\ &= F_X(y^\alpha) \\ &= 1 - e^{-\frac{y^\alpha}{\beta}}. \end{aligned}$$

Differentiating the d.f. we get the p.d.f

$$f(y) = \tfrac{\alpha}{\beta} y^{\alpha-1} e^{-\frac{y^\alpha}{\beta}}, \quad y > 0.$$

We denote it by $Y \sim \text{Wei}(\alpha, \beta)$ where α (the shape) and β (the rate) are the parameters of the distribution. The Weibull distribution is extensively used for modelling life time data as an alternative to the exponential and gamma distributions.

1.2.5 The chi-square distribution

The r.v. $Y \sim \text{Ga}(\nu/2, 1/2)$ with $\nu \in \mathbb{N}$ is known as the χ^2 (chi-square) distribution, denoted χ^2_ν (or $Y \sim \chi^2_\nu$ if we wish to express that the r.v. Y follows a chi-square distribution) where the parameter ν is called *degrees of freedom*. Thus, the χ^2_ν distribution is nothing more than a special case of the gamma and like the latter it has many interesting properties. Here are some of them

Theorem 1.4 *(i) If* $Y \sim \chi^2_\nu$*, then* $E(Y) = \nu, \text{var}(Y) = 2\nu$*.*

(ii) If $X \sim \mathrm{N}(0,1)$ *and* $Y = X^2$ *then* $Y \sim \chi^2_1$.

(iii) If $X_i \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$, $i = 1, \ldots, n$ *and* $Y = \sum_{i=1}^n X_i^2$ *then* $Y \sim \chi^2_n$.

As we will see in the incoming lectures, the chi-square distribution arises as sampling distribution to many functions of r.v. used in inferential statistics.

1.2.6 The t-Student distribution

Let $Z \sim \mathrm{N}(0,1)$ and $U \sim \chi^2_\nu$, where Z and U are independent, then the r.v. $T = \frac{Z}{\sqrt{\frac{U}{\nu}}}$ is called t-Student r.v. with degrees of freedom $\nu \in \mathbb{N}$. This distribution is denoted by t_ν, has domain equal to $\mathbb{R}$ and it holds that $E(T) = 0$ whenever $\nu > 1$ and $\mathrm{var}(T) = \frac{\nu}{\nu-2}$, whenever $\nu > 2$.

The t-Student distribution is symmetric around 0 and has a bell-like shape, just like the normal distribution. With respect to the latter, the t-Student has heavier tails, at least for low values of ν. It can be shown that $t_\nu \to \mathrm{N}(0,1)$ as $\nu \to \infty$.

Also the the t-Student distribution arises as sampling distribution of many functions of r.v. used in inferential statistics, but it is also used as a flexible statistical model since it can acommodate extremely low and extremely large observations which under the normal model are unlikely.

1.2.7 The F distribution

Let $U_1 \sim \chi^2_{\nu_1}$ and $U_2 \sim \chi^2_{\nu_2}$ with U_1 independent from U_2. Then the r.v.

$$Y = \frac{U_1/\nu_1}{U_2/\nu_2},$$

is said to have F distribution with $\nu_1, \nu_2 \in \mathbb{N}$ *degrees of freedom*, denoted $Y \sim F_{\nu_1,\nu_2}$

Here are some useful properties.

Theorem 1.5 *(i) If* $Y \sim F_{n_1,n_2}$, *then* $\frac{1}{Y} \sim F_{n_2,n_1}$

(ii) If $T \sim t_\nu$, *then* $T^2 = \frac{Z^2}{\frac{U}{\nu}} \sim F_{1,\nu}$

Also the the F distribution arises as sampling distribution for many functions of r.v. used in inferential statistics.

1.2.8 The multivariate normal distribution

This is our second probability distribution for random vectors. Just like the multinomial distribution can be seen as a multivariate extension of the binomial distribution, the multivariate normal distribution extends the normal distribution in the random vector case.

There are different ways for introducing the multivariate normal distribution. Here we follow a bottom-up approach in which we construct the multivariate normal from the standard normal r.v. which we have

already encountered. Let $Z_i \overset{\text{iid}}{\sim} \mathrm{N}(0,1), i = 1, \ldots, p$, then by what we have see in L0, the r.ve. $(Z_1, \ldots, Z_p)$ has joint p.d.f

$$\begin{aligned} f_{Z_1,\ldots,Z_p}(z_1, \ldots, z_p) &= \prod_{i=1}^{n} f_{Z_i}(z_i) \\ &= \prod_{i=1}^{n} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_i^2} \\ &= \tfrac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{p} z_i^2} \\ &= \tfrac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2} z^{\mathrm{T}} z}, \quad z = (z_1, z_2, \ldots, z_p). \end{aligned}$$

We say then that the r.ve $Z = (Z_1, \ldots, Z_p)$ has *standard multivariate normal distribution*, denoted by $Z \sim \mathrm{N}_p(0, I_p)$, where I_p is the unit diagonal matrix and 0 here denotes the p vector of zeros. Note that by construction $E(Z_i) = 0$ and $\mathrm{cov}(Z_i, Z_j) = 0$ for $i \neq j$ and $\mathrm{cov}(Z_i, Z_i) = \mathrm{var}(Z_i) = 1$, for all $i, j = 1, \ldots, p$.

But we are not done yet, what we constructed so far is only <u>one</u> multivariate normal distribution, i.e. the standard one, and we aim at the <u>family of multivariate normals</u>. To this end, let A be $(p \times p)$ matrix for which A^{-1} exists and let $\mu \in \mathbb{R}^p$. Consider the function $g : \mathbb{R}^p \to \mathbb{R}^p$ given by $g(Z) = Y = AZ + \mu$ which has inverse $g^{-1} : \mathbb{R} \to \mathbb{R}$ given by $g^{-1}(Y) = Z = A^{-1}(Y - \mu)$. Then by §0.4.4 of L0, we have

$$f_Y(y) = f_Z(g^{-1}(y)) \left| \tfrac{dg^{-1}(y)}{dy} \right|,$$

where the now $y \in \mathbb{R}^p$ and $\left| \frac{dg^{-1}(y)}{dy} \right|$ denotes the determinant of the Jacobian matrix associated with the transformation. In particular, it holds that

$$\tfrac{dg^{-1}(y)}{dy} = A^{-1}.$$

Putting all the pieces together and letting $\Sigma = AA^{\mathrm{T}}$, such that $|\Sigma| = |AA^T| = |A|^2$ and $|\Sigma|^{-1/2} = |A|^{-1}$, we have the p.d.f.

$$\begin{aligned} f_Y(y) &= f_Z(g^{-1}(y)) \left| \tfrac{dg^{-1}(y)}{dy} \right| \\ &= \tfrac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}\left[A^{-1}(y-\mu)\right]^{\mathrm{T}}\left[A^{-1}(y-\mu)\right]} |A^{-1}| \\ &= \tfrac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{1/2}} e^{-\frac{1}{2}(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)}. \end{aligned}$$

It follows that $E(Y) = \mu$ and $\mathrm{cov}(Y_i, Y_j) = \sigma_{ij}$, where σ_{ij} is the elements on the ith row and jth column of Σ. The r.ve. Y with the above distribution is called the *p-variate normal distribution* with mean vector μ and covariance matrix Σ, and denoted by $Y \sim \mathrm{N}_p(\mu, \Sigma)$.

Here are some useful properties of the multivariate normal distribution.

Theorem 1.6 *Let $Y \sim \mathrm{N}_p(\mu, \Sigma)$. Then*

(i) If $X = BY + b$, with B a $(p \times p)$ matrix and $b \in \mathbb{R}^p$, then $X \sim \mathrm{N}_p(B\mu + b, B\Sigma B^{\mathrm{T}})$.

(ii) $Y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, where $\mu_i = E(Y_i)$ and $\sigma_i^2 = \mathrm{var}(Y_i)$, $i = 1, \ldots, p$.

(iii) All the conditional distributions involving components of Y are normal with suitable parameters.

1.3 Generating random variates

So far we have been concerned with mathematical properties of r.v. and their probability distributions. However, the study of mathematical properties per se may not be fully satisfactory for at least two reasons. Firstly, there are many r.v.'s which have p.d.f.'s or d.f.'s with complicated expressions, and sometimes are not even available. In this case, mathematical properties such as expectations are typically impossible to study analytically. Secondly, we may want to check how good is a given probability model as an approximation of a given set of observed experimental data. In both cases, we can still achieve our goal if we could generate random draws from the given distribution.

There are many methods for generating random draws from a given distribution function, here we limit ourselves to the *inverse transform sampling* method. Before explaining the method,we need to introduce another continuous r.v., the *uniform distribution*. An r.v. X is said to have the uniform distribution in the interval $[0, 1]$ if it has density

$$f(x) = 1_{[0,1]}(x), \quad x \in [0, 1]$$

where $1_{[a,b]}(x)$ is the indicator function which takes 1 if $x \in [a, b]$ and 0 otherwise. We denote this by $X \sim \mathrm{Unif}(0, 1)$, and we say that the r.v. X is uniform between 0 and 1. Although it is not of direct interest as an approximating model for observed data, the uniform distribution is of vital importance in statistics and probability since it is the hart of many algorithms for simulating random variables. In fact, the uniform distribution is at the heart of the inverse transform sampling method.

Theorem 1.7 *Let Y be an r.v. with d.f. F, assumed to be continuous with quantile function F^{-1}, and let $U \sim \mathrm{Unif}(0, 1)$. Define the r.v. $X = F^{-1}(U)$, Then for any $y \in \mathbb{R}$, we have*

$$\begin{aligned} P(X \leq y) &= P(F^{-1}(U) \leq y) \\ &= P(U \leq F(y)) \\ &= \int_0^{F(y)} 1_{[0,1]}(x)dx \\ &= F(y). \end{aligned}$$

Where the last equality holds because $F(y) \in [0, 1]$.

This theorem is valid for any cumulative distribution function, and tells us that if we want to generate random variates Y, which has d.f. F, then it is enough to generate random variates U and then transform them by applying $F^{-1}(U)$.

Example 1.2 *Let $Y \sim \mathrm{Exp}(1)$. Then*

$$F(y) = \int_0^y e^{-t}\, dt = 1 - e^{-y}.$$

It follows that $F^{-1}(t) = -\log(1 - t) = \log(1/(1 - t))$. Therefore by Theorem 1.7, if $U \sim \mathrm{Unif}(0, 1)$, then $F^{-1}(U) = \log(1/(1 - U))$. It follows that $\log(1/(1 - U)) \sim \mathrm{Exp}(1)$, as for Y.

References

[LM18] Larsen, R. J. and Marx, M. L. (2018) *An Introduction to Mathematical Statistics and its Applications* (6th edition), Pearson Education, Inc., §2, §3, §4, §7.3, §10.2.

Practice Lecture 1: Basics of R and probability distributions

Erlis Ruli (ruli@stat.unipd.it)

16 October 2020

1 R Basics

The R (R Core Team, 2020) programming language and environment is designed for statistical analysis. It is open, free (see https://www.R-project.org/) and is written and maintained by a very active community of statisticians. A major design feature is extendability. R makes it very straightforward to code up statistical methods in a way that is easy to distribute and for others to use. The first place to look for information on getting started with R is http://cran.r-project.org/manuals.html. I will assume that you have installed R, and have at least discovered the function `q()` for quitting R.

You might find it easier to work with R if you use an IDE. For this I recommend R Studio (https://rstudio.com/); the free Desktop edition is enough.

The following command creates the vector named "a", and assigns to this vector the numbers $1, 2, \ldots, 10$.

```
> a <- 1:10
```

The symbol "<-" is an assignment symbol, the colon ":" is a function that creates regular sequence of integers and the symbols "<" on the left is the R prompt, just like the $ in a unix terminal. Square brackets are used for subsetting vector elements (In R, indexing of objects starts from 1 and not from 0 as for instance it happens in C). For instance the third element of a, that is a_3 is

```
> a[3]
```

```
## [1] 3
```

The symbol "## [1]" says that the R output has only one row and the value delivered in this case is 3. We can select more than one element at time by, for instance the first three elements of the vector a are

```
> a[1:3]
```

```
## [1] 1 2 3
```

Elements of a such as the 2nd, 3rd and 7th can be extracted by

```
> a[c(2,3,7)]
```

```
## [1] 2 3 7
```

where we used the function `c`, which stands for "concatenate". We could also have used the concatenate function to define a, such as

```
> a <- c(1,2,3,4,5,6,7,8,9,10)
```

But in this case, it is not very convenient.

Matrices are built by the function `matrix`. For example, here is a 5×2 matrix

```
> A <- matrix(c(1:10), ncol=2)
```

We will use the terms "R function" and "R command" almost interchangeably. Here we are actually creating internally a vector such as the vector a above and then we reshape it to fill a 5×2 matrix named A. The command `matrix` has other options that for the moment we ignore.

The manual of an R command, say `diag` we can be checked by the command `?diag`. Try it by typing `?matrix` at the prompt.

Matrix subsetting is done via square brackets with double index notation

```
> # (1,2) element of A is
> A[1,2]
```

```
## [1] 6
```

We see here also how to place comments in R, i.e. any text preceded by the "#".

2 Generating random variates from a given distribution

2.1 The uniform

Let us start from the uniform random variable (r.v.). Recall that, if $U \sim \text{Unif}(0,1)$, then U has p.d.f. equal to one in the interval [0,1] and zero elsewhere. Here is the plot of the p.d.f. of a uniform r.v.

```
> par(mfrow = c(1,2))
> plot(dunif, xlim=c(-1,2),
+      ylab="f(x)", main = "The p.d.f of the Unif(0,1) r.v.")
> plot(punif, xlim=c(-1,2),
+      ylab="F(x)", main = "The d.f of the Unif(0,1) r.v.")
```

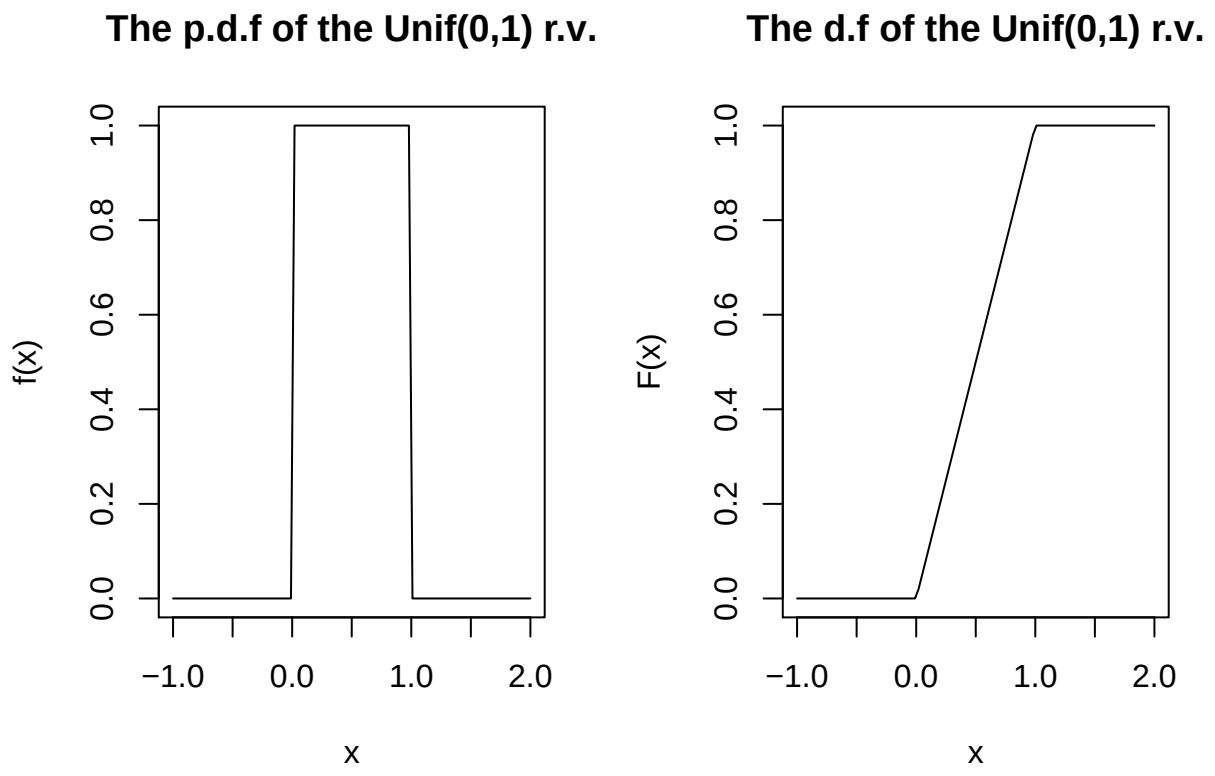


Figure 1: The p.d.f (left) and d.f. of the Unif(0,1) r.v..

To create the figure on the left, we used the function `dunif` which is R's built-in function for the p.d.f. of the

Unif$(0,1)$ r.v. Then we used the function `plot` which is a generic `R` function for creating figures of many types. Built-in `R` function for dealing with probability distibutions can be categorised into four groups:

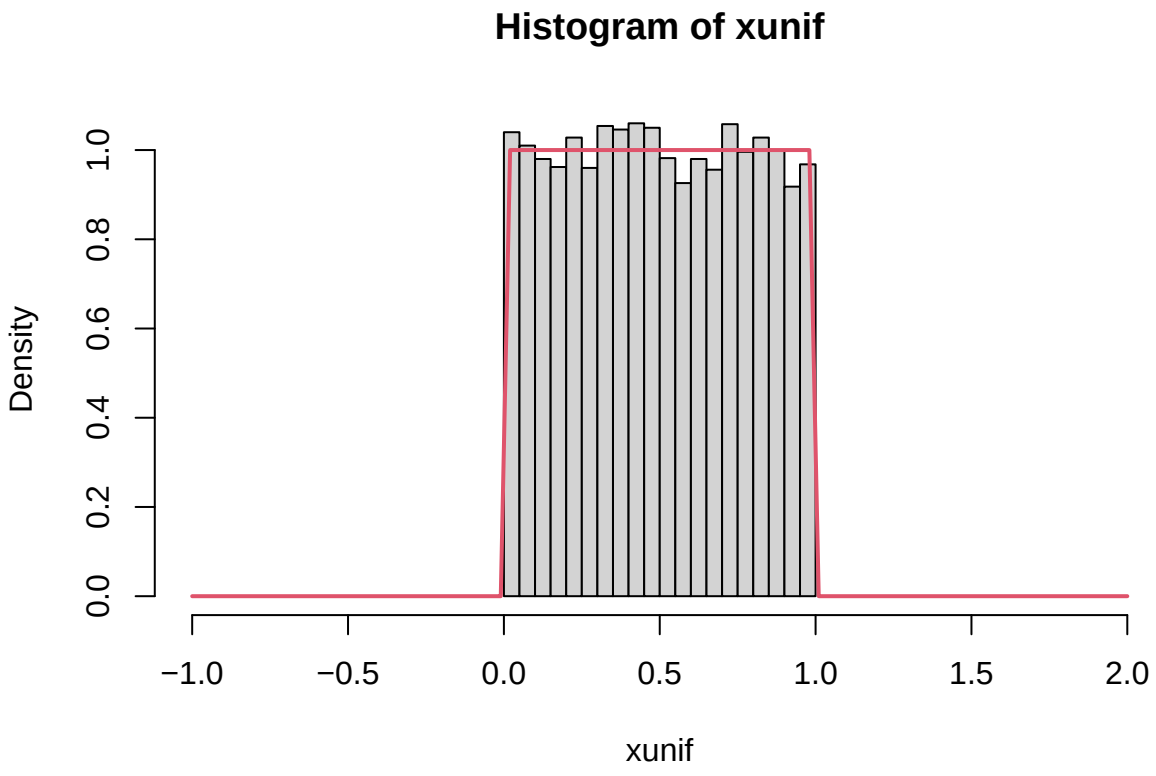
- functions starting by the letter `d`, e.g. `dunif`, dnorm, etc., give the probability density function of the r.v.;
- functions starting by the letter `r`, e.g. `runif`, `rnorm`, etc., are used for generating random draws;
- functions starting by the letter `p`, e.g. `punif`, `pnorm`, etc., give the distribution function;
- functions starting by the letter `q`, e.g. `qunif`, `qnorm`, etc., give the quantile function.

We now use the `runif` command to draw random samples of size 10^4. To make sure you get the figure below use the same random seed as mine.

```
> set.seed(2020)
> xunif <- runif(1e+4)
```

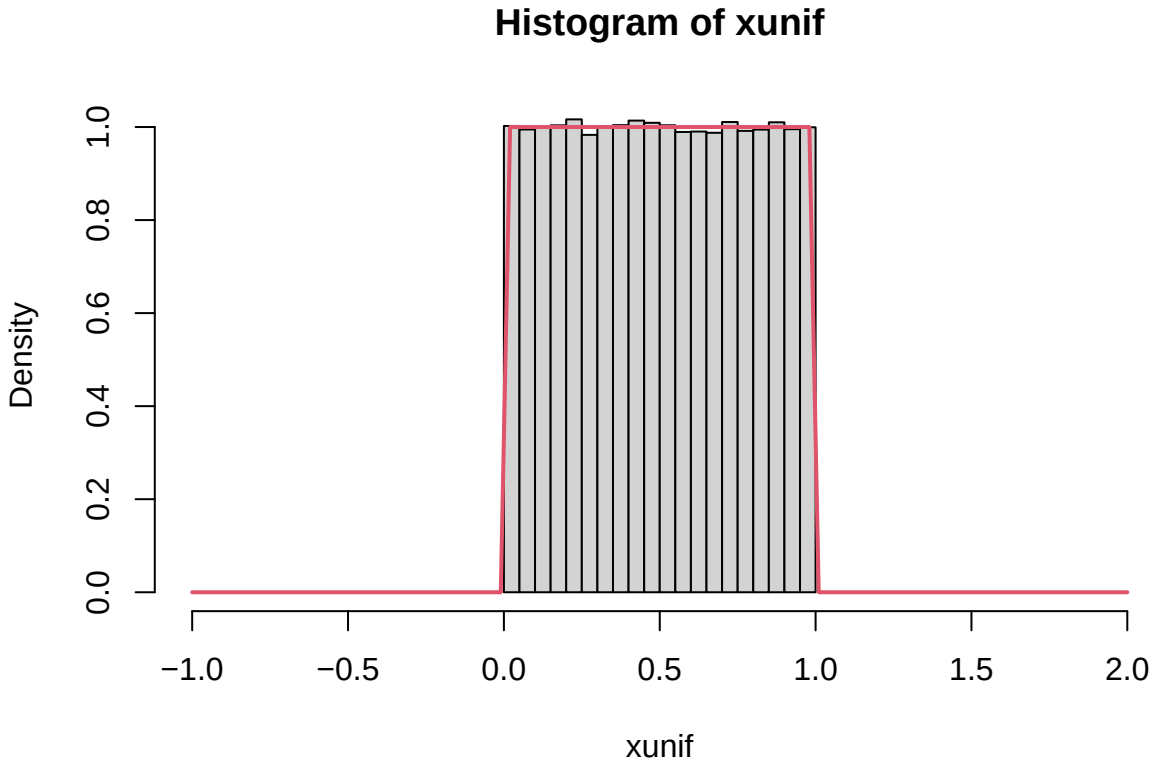
Let's see how uniform are our draws by constructing a scaled histogram (using option `freq=FALSE`) and then let's compare the latter with the the p.d.f. of the Unif$(0,1)$ r.v. We will see the histogram in the next Lecture, but for the time being...

```
> hist(xunif, xlim=c(-1,2), freq = FALSE)
> plot(dunif, xlim=c(-1,2), add=TRUE, col=2, lwd=2)
```



The histogram seem quite close to the p.d.f.. Changing the random seed or simply not using a random seed produces a slightly different histograms, all the times we generate new random samples. However, let us now increase the sample size

```
> set.seed(2020)
> xunif <- runif(3e+5)
> hist(xunif, xlim=c(-1,2), freq = FALSE)
> plot(dunif, xlim=c(-1,2), add=TRUE, col=2, lwd=2)
```



The histogram is now extremely close to the p.d.f.. Thus the fealing is that for the sample size going to infinity the histogram tends to converge to the p.d.f. thus the sample we have got by runif is actually a random sample which has distribution U(0,1).

2.2 Monte Carlo approximations

Approximations of features of a probability distribution by means of simulations are generally known as Monte Carlo approximations. For instance, above we approximated the density of the Unif(0,1) distribution by a histogram. This approximations are very useful when the involved probability distribution is difficult to work with.

Here is how this method works for approximating features of a distribution such as the expected value or the probability that the r.v. is whithin a given interval. For simplicity, we focus here on the Unif(0,1) distribution but note that the technique we will see applies to any probability distribution.

First note that the Unif(0,1) distribution has expected value

$$E(X) = \int_0^1 x 1_{[0,1]}(x) dx = \frac{1}{2}.$$

Then we generate N samples $x_1, \ldots, x_N$ from the Unif(0,1) distribution. This sample is also called Monte Carlo sample. Finally, we replace the integral by the sum divided by N, i.e.

$$\int_0^1 x 1_{[0,1]}(x) dx \approx \frac{1}{N} \sum_{i=1}^{n} x_i.$$

Thanks to the LLN, as $N \to \infty$ the average $\frac{1}{N} \sum_{i=1}^{n} x_i$ will converge to the true value $\mu = \frac{1}{2}$.

As another example, suppose we wish to compute $P(1/2 < X < 2/3)$, when $X \sim \text{Unif}(0, 1)$. Again this can be computed exactly by

$$P(1/2 < X < 2/3) = \int_{1/2}^{2/3} 1_{[0,1]}(x)dx = \frac{1}{6}.$$

To figure out how to approximate this quantity via a Monte Carlo method, first write this probability in the form of an expectation. In particular, let $1_{(1/2,2/3)}(x)$ be an indicator function which takes value 1 if $x \in (1/2, 2/3)$ and zero otherwise. Then we have that

$$\begin{aligned} P(1/2 < X < 2/3) &= \int_{1/2}^{2/3} 1_{[0,1]}(x)dx \\ &= \int_0^1 1_{(1/2,2/3)}(x)1_{[0,1]}(x)dx \\ &= E[1_{(1/2,2/3)}(X)]. \end{aligned}$$

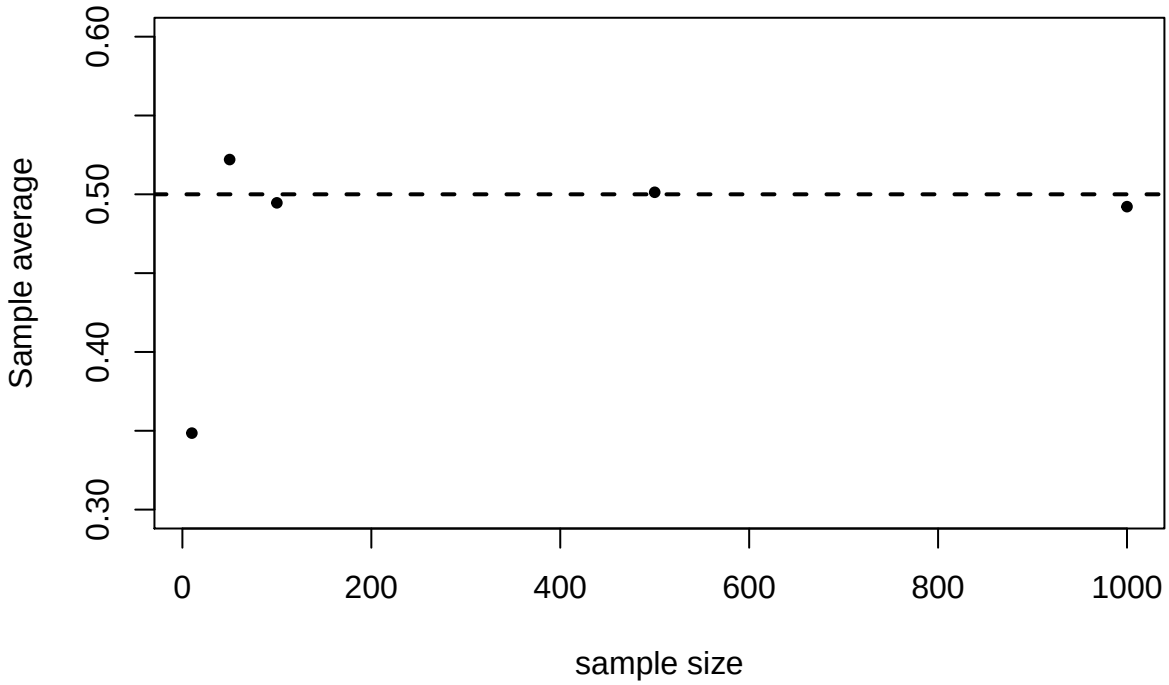
Thus to approximate the expectation we again draw N samples $x_1, \ldots, x_N$ from the Unif(0,1) distribution and replace the integral by the sum divided by N

$$E[1_{(1/2,2/3)}(X)] \approx \frac{1}{N}\sum_{i=1}^{N} 1_{(1/2,2/2)}(x_i),$$

Note that $1_{(1/2,2/2)}(x_i)$ gives only 1's and 0's thus the sum over N gives the propoprtion or the relative frequency of 1's. Thanks to the LLN, as $N \to \infty$ this relative frequency will converge to the true probability $\frac{1}{6}$.

As a numerical illustration let's see how closer we can get to the expected value of the Unif(0,1) as N increases. By the (S)LLN we know that the average converges to the true mean for the sample size going to infinity. Let's see if that's the case with our random draws.

```
> set.seed(2020)
> xunif1 <- runif(10)
> set.seed(2020)
> xunif2 <- runif(50)
> set.seed(2020)
> xunif3 <- runif(1e+2)
> set.seed(2020)
> xunif4 <- runif(5e+2)
> set.seed(2020)
> xunif5 <- runif(1e+3)
> averages <- c(mean(xunif1), mean(xunif2), mean(xunif3),mean(xunif4),mean(xunif5))
> plot(x = c(10,50,1e+2,5e+2,1e+3), y=averages, ylim=c(0.3, 0.6), pch=20,
+      ylab = "Sample average",
+      xlab="sample size")
> abline(h = 1/2, lwd=2, lty=2)
```



We note that the sample averages get closer to 1/2, as the sample size increases, just as we suspected.

3 The Gaussian distribution

Gaussian or normal random draws can be performed by the command `rnorm`. For instance, 10^3 draws from the standard normal distribution can be obtained by

```
> set.seed(2020)
> # generate 1000 r. draws from the N(0,1) dist
> xnorm <- rnorm(1e+3)
```

If we want random draws from say N(1/2,2) then we have two options: either we resort to the options of the `rnorm`, i.e.

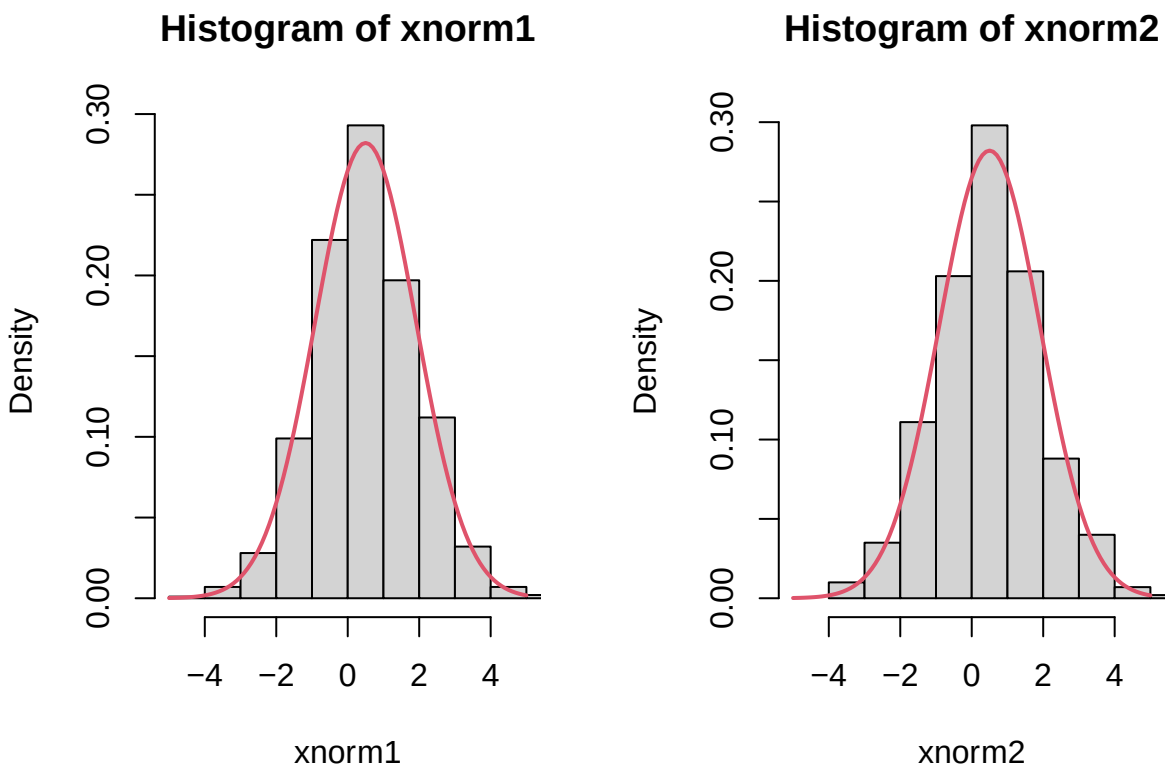
```
> # generate 1000 r. draws from the N(1/2,3) dist
> xnorm1 <- rnorm(1e+3, mean = 1/2, sd = sqrt(2))
```

where note that `R` the option `sd` is used to specify the standard deviation of the required normal distribution. The second alternative is to resort to the properties of the normal distribution (see L1) and thus

```
> mu <- 1/2
> st.dev <- sqrt(2)
> xnorm2 <- mu + st.dev*xnorm
```

We can comapre the random draws using a histogram with the true p.d.f. as below

```
> par(mfrow = c(1,2))
> hist(xnorm1, xlim=c(-5,5), freq = FALSE)
> plot(function(x) dnorm(x, mean = 1/2, sd=sqrt(2)),
+      xlim=c(-5,5), add=TRUE, col=2, lwd=2)
> hist(xnorm2, xlim=c(-5,5), freq = FALSE)
> plot(function(x) dnorm(x, mean = 1/2, sd=sqrt(2)),
+      xlim=c(-5,5), add=TRUE, col=2, lwd=2)
```



Here we see also the use of the command `function`. This command is used to define user-defined functions. For instance, suppose we want to define the function $f(x) = (x^3 + \log(x))/x$. The `R` code for this is

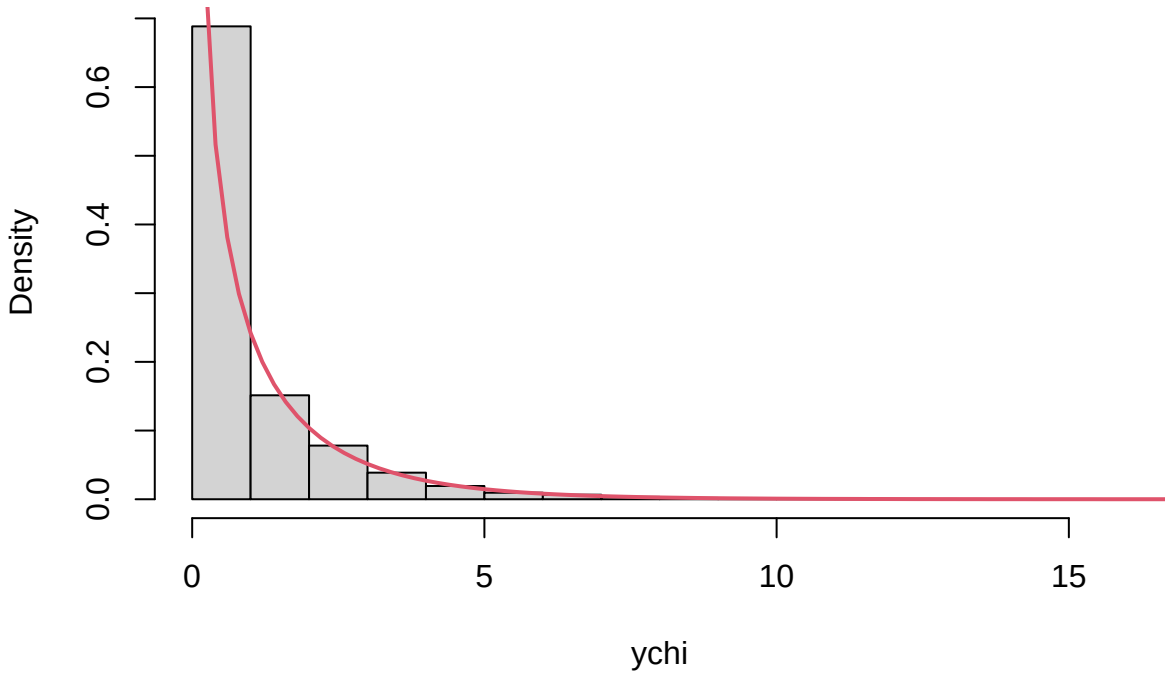
```
> fx <- function(x) {
+   out1 <- x^3 + log(x)
+   out2 <- out1/x
+   return(out2)
+ }
```

Functions in `R` can have more than one arguments. These can be of any type, vectors, matrices, lists, etc. and the return objects can be of any type. The only constraint is that functions cannot return multiple objects.

4 Transformation of a r.v.

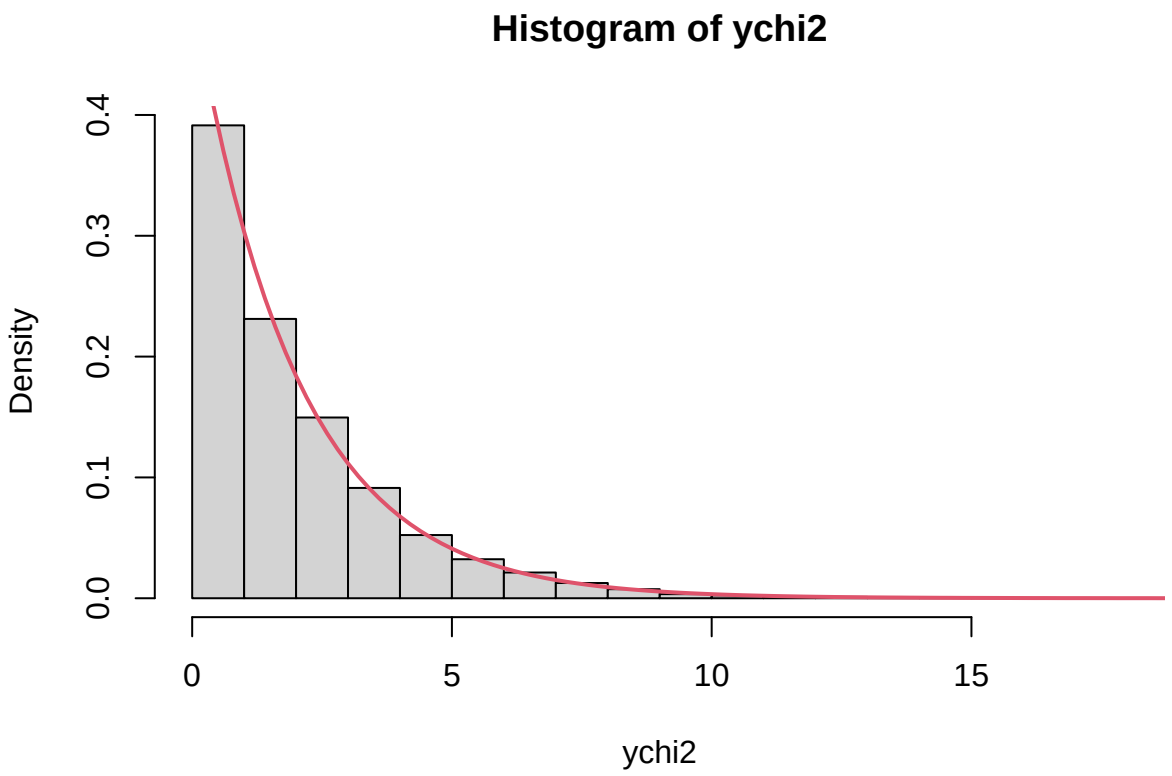
From Lecture 1 we know that if $Z \sim \mathrm{N}(0,1)$, then $Y = Z^2 \sim \chi^2_1$. Let's check this by simulation. The idea is to generate a large sample from the $\mathrm{N}(0,1)$ distribution, square all the values and then compare these squarred values by means of an histogram with the theoretical distribution, i.e. χ^2_1.

```
> set.seed(2020)
> znorm <- rnorm(1e+4)
> ychi <- znorm^2
> hist(ychi, freq = FALSE)
> plot(function(x) dchisq(x, df=1), add=TRUE, lwd=2, col=2, xlim=c(0,20))
```



We see that the histogram is quite close to the theoretical p.d.f., given by the function `dchisq`. Te be convinced it is not just luck, lets try sums of two r.v.. Recall that if $Z_1 \sim \mathrm{N}(0,1), Z_2 \sim \mathrm{N}(0,1)$ and Z_1 is independent from Z_2, then $Y = Z_1^2 + Z_2^2 \sim \chi_2^2$. So we have

```
> set.seed(2020)
> znorm1 <- rnorm(1e+4)
> znorm2 <- rnorm(1e+4)
> ychi2 <- znorm1^2+ znorm2^2
> hist(ychi2, freq = FALSE)
> plot(function(x) dchisq(x, df=2), add=TRUE, lwd=2, col=2, xlim=c(0,20))
```



This tells us that if we wanted to generate values from the χ^2_2 distribution, we can generate two random draws from the $N(0,1)$, square each of them and then sum. Of course we do not need to take such a long way, since `R` has the function `rchisq` which we can be use to generate samples from χ^2_p, for any p.

Exercise: Generate 10^4 random draws from $N(0,1)$, take the exponential and plot the relative histogram. This distribution is called log-normal (see its p.d.f in the help of the comand `dlnorm` or at https://en.wikipedia.org/wiki/Log-normal_distribution). Make sure that the histogram is close to the theoretical distribution. Mathematically, you are required to show that if $Z \sim N(0,1)$ then $Y = \exp(Z) \sim \text{logN}(\mu, \sigma)$ with parameters μ, σ to be defined from those of Z.

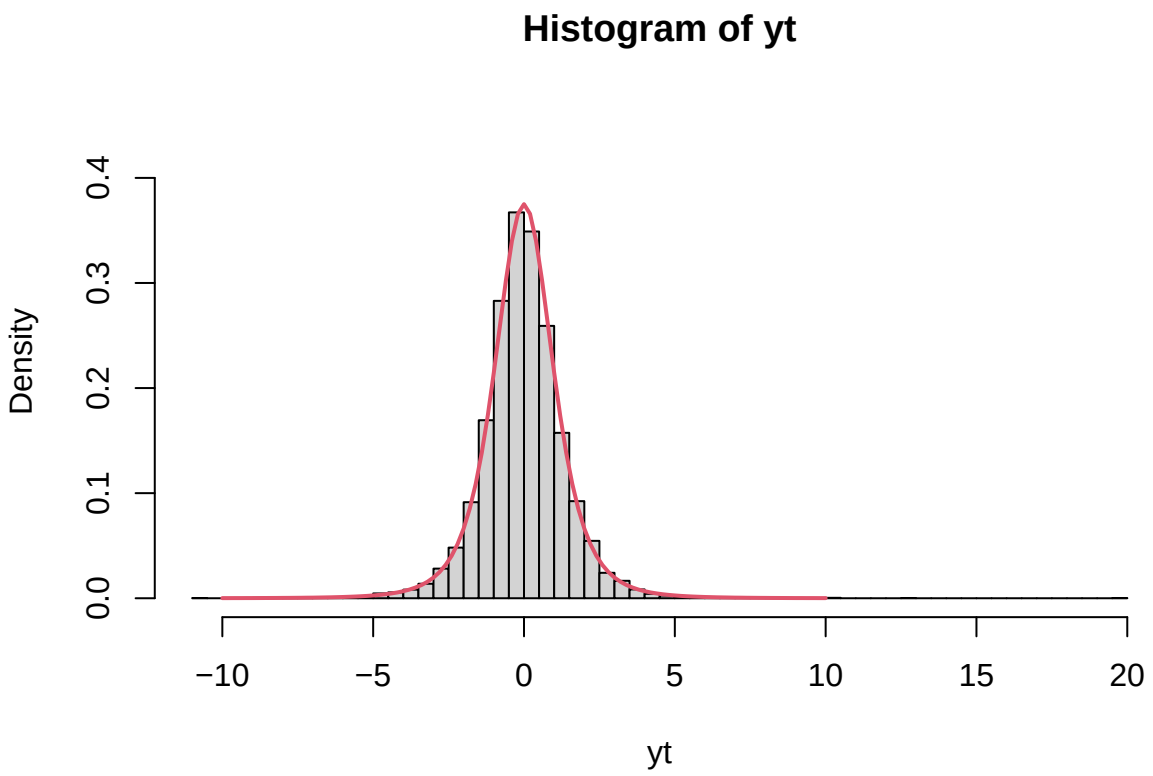
5 Generating t-Student random variates

The t-Student r.v. with ν degrees of freedom is defined (see L1) by $T = \frac{Z}{\sqrt{X/\nu}}$, where $Z \sim \text{N}(0,1)$ and $X \sim \chi^2_\nu$. This definition immediately gives us a way for generating t-Student random samples. That is

```
> # drawing from t-Student distribution with nu=4
> # using its definition
> set.seed(2020)
> nu <- 4
> znorm <- rnorm(1e+4)
> xchi <- rchisq(1e+4, df=4)
> yt <- znorm/sqrt(xchi/nu)
```

Now we compare the drawn samples with the theoretical p.d.f.

```
> hist(yt, freq = FALSE, breaks = 50, ylim=c(0,0.45))
> plot(function(x) dt(x, df=nu), add=TRUE, lwd=2, col=2, xlim=c(-10,10))
```



To have a clearer picture, we increased the number of bars with respect to the default option, by specifying the option `breaks=50`. Of course, `R` has a built-in function for generating t-Student random draws, it's called `rt`. Above we used the function `dt`, which gives the p.d.f of the a t-Student r.v..

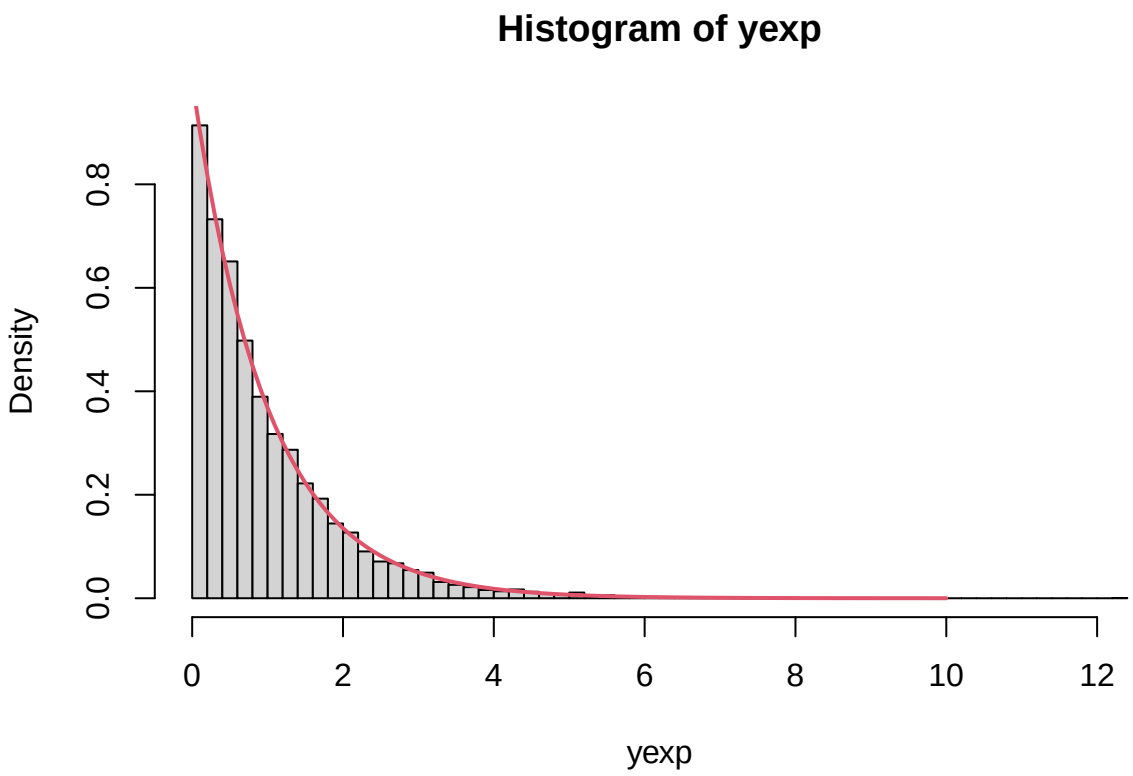
6 The inverse transform sampling method

We illustrate the inverse transform sampling (ITS) method by the example of an Exp(1) r.v.. We saw that the d.f. for this r.v. is $F(x) = 1 - \exp(-x)$ and thus the quantile function is $F^{-1}(p) = -\log(1-u)$. Here is the code for drawing 10^4 samples with the ITS method.

```
> set.seed(2020)
> unif <- runif(1e+4)
> yexp <- -log(1-unif)
> # alternatively we can also use
> # yexp1 <- -log(unif)
> # becasue U ~ Unif(0,1), then also 1-U ~ Unif(0,1)
```

Again, we comapre the distribution of the samples with the theoretical one by means of a histogram.

```
> hist(yexp, freq = FALSE, breaks = 50)
> plot(dexp, add=TRUE, lwd=2, col=2, xlim=c(0,10))
```



Of course, R has a built-in function for generating Exp(1) random draws, it's called `rexp`.

Exercise: Generate 10^4 samples from the Exp(3/2) distribution by the ITS method.

7 The Central Limit Theorem

7.1 Convergence of binomial r.v.

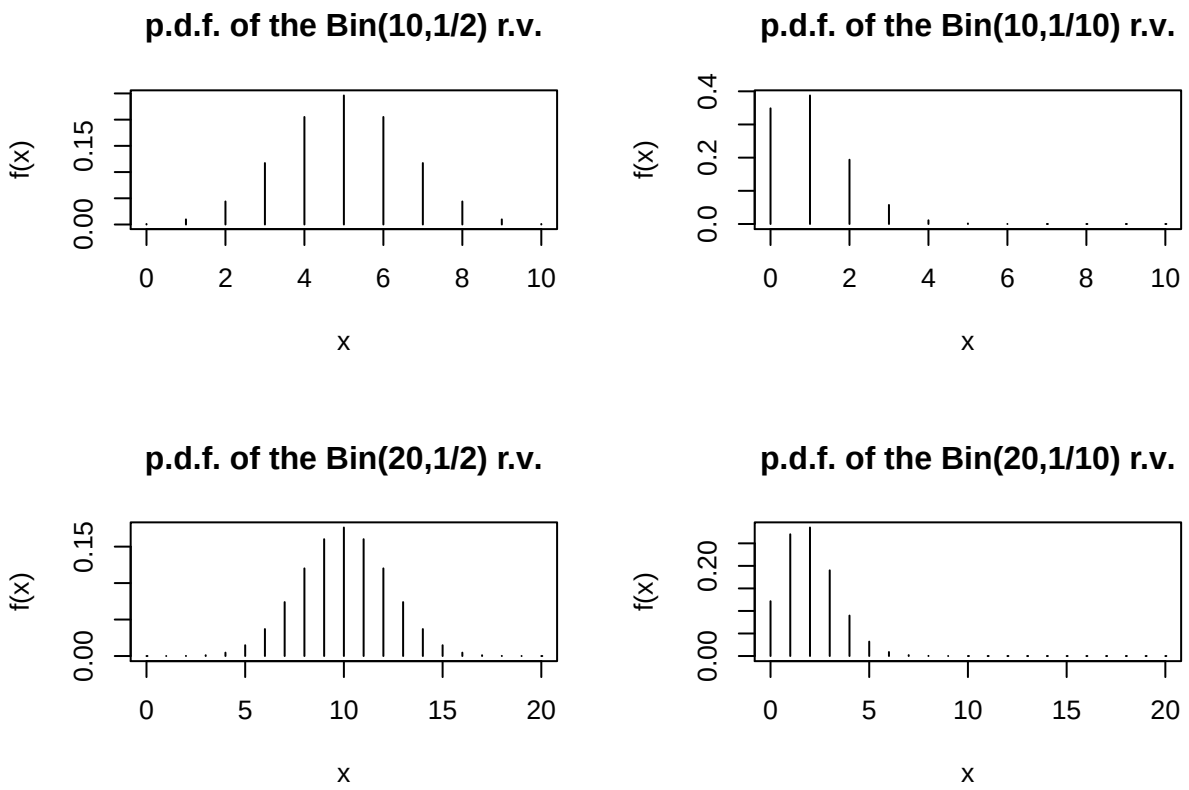
From theory we know that if $Y \sim \text{Bin}(n, \theta)$, then $\frac{Y-E(Y)}{\sqrt{var(Y)}} = \frac{Y-n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow{d} \text{N}(0,1)$. Recall also that the CLT applies here because Y is the sum of n Bernoulli r.v..

Let us see how this works in practice. Consider the following four scenarios

- $n = 10, \theta = 1/2$
- $n = 10, \theta = 1/10$
- $n = 20, \theta = 1/2$
- $n = 20, \theta = 1/10$

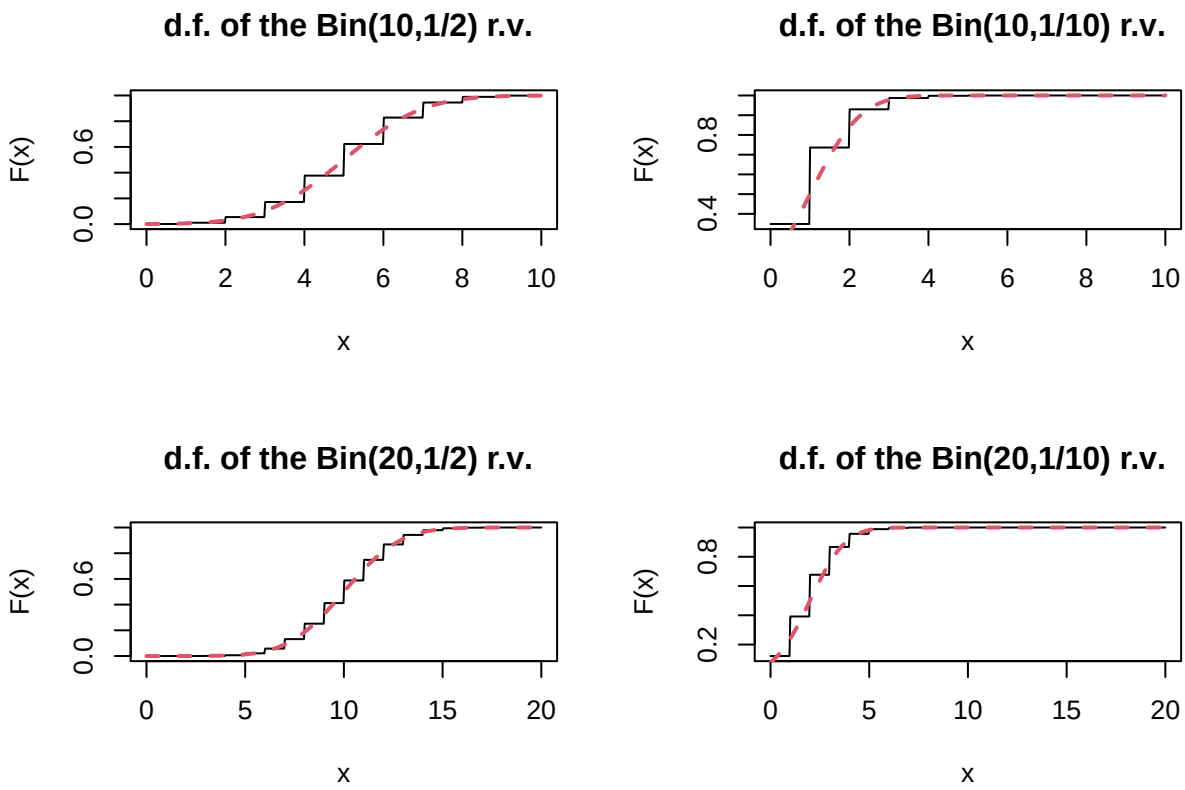
We consider first the probability density functions for each scenario. As we see from the plots of the p.d.f. the closer the success probability to the boundary of the parameter space, the skewed is the p.d.f.

```
> par(mfrow=c(2,2))
> plot(x = 0:10, dbinom(0:10, size=10, prob=1/2), type="h",
+      ylab="f(x)", xlab="x", main="p.d.f. of the Bin(10,1/2) r.v.")
> plot(x = 0:10, dbinom(0:10, size=10, prob=1/10), type="h",
+      ylab="f(x)", xlab="x", main="p.d.f. of the Bin(10,1/10) r.v.")
> plot(x = 0:20, dbinom(0:20, size=20, prob=1/2), type="h",
+      ylab="f(x)", xlab="x", main="p.d.f. of the Bin(20,1/2) r.v.")
> plot(x = 0:20, dbinom(0:20, size=20, prob=1/10), type="h",
+      ylab="f(x)", xlab="x", main="p.d.f. of the Bin(20,1/10) r.v.")
```



In the following plot we show the d.f. of the binomial r.v. and the d.f. of the the normal distbution as implied by the CLT

```
> par(mfrow = c(2,2))
> plot(function(x) pbinom(x, size=10, prob=1/2),xlim=c(0,10), n=500,
+      ylab="F(x)", main = " d.f. of the Bin(10,1/2) r.v.")
> plot(function(x) pnorm(x, mean=10*1/2, sd=sqrt(10*1/4)),xlim=c(0,10),
+      add=TRUE, col=2, lwd=2, lty=2)
>
> plot(function(x) pbinom(x, size=10, prob=1/10),xlim=c(0,10), n=500,
+      ylab="F(x)", main = " d.f. of the Bin(10,1/10) r.v.")
> plot(function(x) pnorm(x, mean=10*1/10, sd=sqrt(10*9/90)),xlim=c(0,10),
+      add=TRUE, col=2, lwd=2, lty=2)
>
>
> plot(function(x) pbinom(x, size=20, prob=1/2),xlim=c(0,20), n=500,
+      ylab="F(x)", main = " d.f. of the Bin(20,1/2) r.v.")
> plot(function(x) pnorm(x, mean=20*1/2, sd=sqrt(20*1/4)),xlim=c(0,20),
+      add=TRUE, col=2, lwd=2, lty=2)
>
> plot(function(x) pbinom(x, size=20, prob=1/10),xlim=c(0,20), n=500,
+      ylab="F(x)", main = " d.f. of the Bin(20,1/10) r.v.")
> plot(function(x) pnorm(x, mean=20*1/10, sd=sqrt(20*9/90)),xlim=c(0,20),
+      add=TRUE, col=2, lwd=2, lty=2)
```



As we expected, the larger the sample size, the closer is the $Bin(n, \theta)$ d.f. to the normal d.f. For values of θ near the boundary points, i.e. $\theta \approx 0$ or $\theta \approx 1$, the normal approximation is less accurate.

8 The multivariate normal distribution

For obvious geometrical reasons we focus here on the bivariate normal distribution. Thus we have a $X = (X_1, X_2)$.

We first define an R function that computes the density of the bivariate normal distribution

```
> # real-valued function with arguments:
> # x: vector (2 x 1), all other argumens are scalars
> dbvnorm <- function(x, mu1, mu2, sigma1.2, sigma2.2, sigma12){
+   # build the covariance matrix Sigma
+   Sigma = matrix(c(sigma1.2, sigma12, sigma12,sigma2.2),ncol=2)
+
+   # build the vector of means
+   mu = c(mu1, mu2)
+
+   # compute a first part the of p.d.f, aka the "kernel"
+   out1 = exp(-0.5 * t(x-mu) %*% solve(Sigma) %*% (x-mu))
+
+   # scale the kernel by the normalising constant
+   out2 = out1/((2*pi)^{2/2} * sqrt(det(Sigma)))
+
+   # return the output
+   return(out2)
+
+   # you can also omit print() by writing just
```

```
+   # out2
+ }
```

We have tow options for drawing a bivariate distribution, either via a perspective plot or via countur levels. In the following we see both approaches.

Firstly, we need to construct a regular bivariate grid and evaluete our newly defined function over it. We have to re-arrange the grid values in order to have as input an $(m \times 2)$ matrix. The output will be an $(m \times 1)$ vector which has to be re-arranged into a square matrix.

```
> mu1 <- 1
> mu2 <- 2
>
> # reguular univ. grid on 1st dimension
> x1 <- seq(-2,5, len=100)
> head(x1)
```

```
## [1] -2.000000 -1.929293 -1.858586 -1.787879 -1.717172 -1.646465
```

```
> # reguular univ. grid on 2nd dimension
> x2 <- seq(-1, 6, len=100)
> head(x2)
```

```
## [1] -1.0000000 -0.9292929 -0.8585859 -0.7878788 -0.7171717 -0.6464646
```

```
> # regular bivariate grid expanded row-wise
> x1x2 <- expand.grid(x1, x2)
> head(x1x2)
```

```
##        Var1 Var2
## 1 -2.000000   -1
## 2 -1.929293   -1
## 3 -1.858586   -1
## 4 -1.787879   -1
## 5 -1.717172   -1
## 6 -1.646465   -1
```

We fix also $\Sigma = [\sigma_{ij}]$ by setting with $\text{var}(X_1) = \sigma_1^2 = 1$, $\text{var}(X_2) = \sigma_2^2 = 2$ and $\text{cov}(X_1, X_2) = \sigma_{12} = 1$. Thus

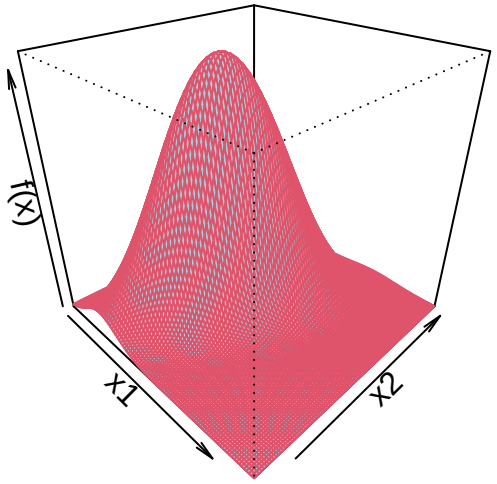
```
> sigma1.2 <- 1
> sigma2.2 <- 2
> sigma12 <- 1
> Sigma <- matrix(c(1,1,1,2),ncol=2)
```

Now we are ready to evaluate the bivariate normal p.d.f. over the grid. For this, we can either use a for/while loop or use `apply` function. We take the former approach here.

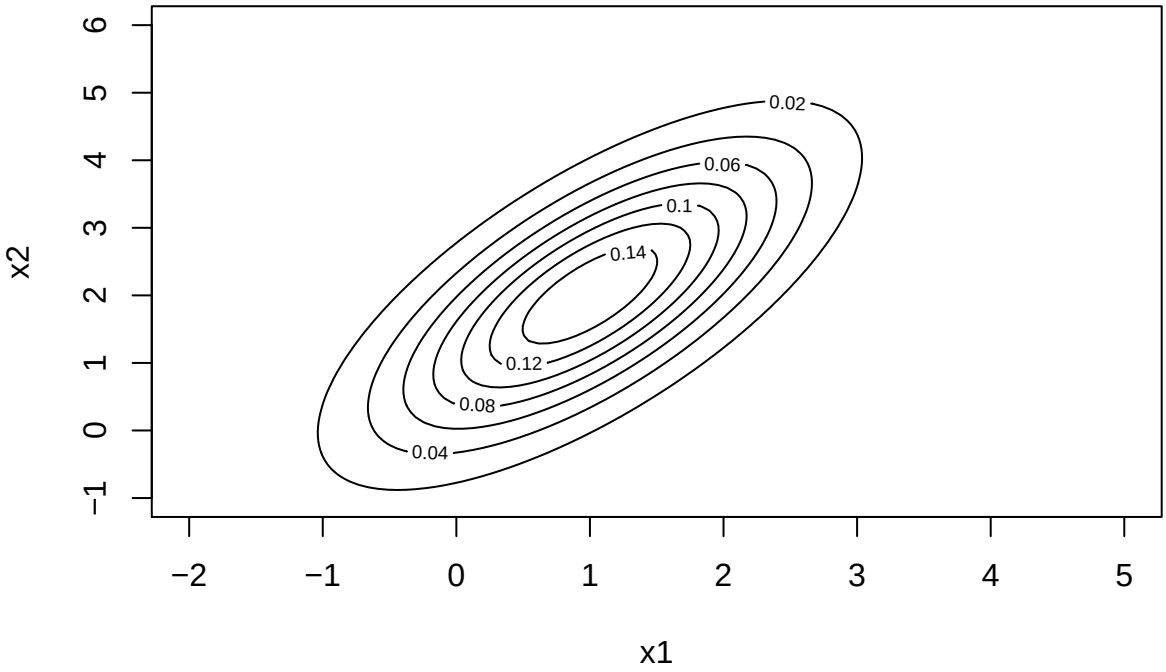
```
> z.pdf <- apply(x1x2,MARGIN = 1,
+                function(x) dbvnorm(x,
+                                    mu1=mu1,
+                                    mu2=mu2,
+                                    sigma1.2=sigma1.2,
+                                    sigma2.2=sigma2.2,
+                                    sigma12=sigma12))
> # we reorder the object to  100 x 100 matrix
> z.pdf.mat <- matrix(z.pdf, ncol=100, nrow=100, byrow = F)
```

Now we are ready for the plots

```
> persp(x1,x2,z.pdf.mat, zlab = "f(x)",
+         col="lightblue",theta=45,phi=30,
+   xlab = "x1", ylab = "x2", shade=0.01, border = 2)
```




```
> contour(x1,x2,z.pdf.mat, xlab="x1", ylab="x2")
```



Wee see that the contours are ellipses with main principal axis having positive slope. The center of the ellipse is at the point μ.

Inferential Statistics **First semester, 2020**

Lecture 2: Descriptive statistics and statistical models

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

In this lecture we introduce some basic tools of descriptive statistics and we introduce the concept of statistical model.

In Probability Theory, we know (almost) everything about the d.f. F_X of an r.v. X, i.e. we know the value of its parameter θ (or vector of parameters), and we are interested in calculating $P(X \in B)$ for some subset $B \subseteq \mathbb{R}$. In Statistics, we observe data (i.e. numbers), say, $x_1, x_2, \ldots, x_n$, which we assume to be generated from some distribution F_X, at some parameter value $\theta = \theta_0$. Thus, we know the family to which F_X belongs to (i.e. F_X could be exponential or normal or gamma, etc.) but the value of θ under which the data are generated, i.e. θ_0, is unknown.

The aim of statistics is thus to guess θ_0. For instance, Nature gives you the sample: 0.318, 1.765, 0.259, 0.450, 0.730, 0.235, 0.017, 1.010, 1.418, 0.480 and it reveals you that the sample was generated from $\text{Exp}(\lambda_0)$. Assuming Nature never lies, with statistics we can produce a guess for λ_0. In particular, statistics gives us the tools we need for producing a guess for λ_0 with *theoretical guarantees*. By the way, producing a guess for λ_0 is called *parametric estimation*, since we are trying to estimate/guess the value of a parameter of a distribution from the observed data.

The set of values $x_1, \ldots, x_n$ are called the *observed sample* and the easiest situation is that with x_i generated from the same distribution and independently of each other. This type of sample is called *i.i.d. sample*. Hence, in practice, we observe $x_1, \ldots, x_n$ an i.i.d. sample, which is assumed to be made of n independent realisations of the r.v. $X \sim F_\theta$. Equivalently, we can say that $x_1, \ldots, x_n$ is an i.i.d. sample, where each x_i is a realisation of $X_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$. Here we changed notation by using F_θ in place of the usual notation F_X. This is to highlight the fact that we know the family F_X but, θ is unknown and it is the object of interest.

Note the difference between $X_1, \ldots, X_n$ and $x_1, \ldots, x_n$. The former is a vector of r.v.'s, i.e. an r.ve., thus it is a function, whereas x_i is a number; often we call $X_1, \ldots, X_n$ the *random sample*. To further clarify the difference between X_i and x_i, think about the difference between a voltmeter and a voltage, the voltmeter, i.e. X_i, produces measurements of voltages x_i. When dealing with samples, being them observed or random, the overall number of samples n is called *sample* size.

2.1 Descriptive statistics

Descriptive statistics are numerical and graphical tools used in order to extract useful information from a given sample. We introduce briefly some the most widely used numerical measures and graphical tools in order to summarise a sample. The latter can be a random sample $X_1, \ldots, X_n$ or an observed sample $x_1, \ldots, x_n$. In Section 2.2 we deal with univariate samples, i.e. X_i's are r.v. and x_i are scalar numbers. In Section 2.3 we deal with bivariate samples, i.e $X_i = (X_{i1}, X_{i2})$ and $x_i = (x_{i1}, x_{i2})$.

2.2 Univariate samples

Location (or *centrality*) and *dispersion* (or *spread*) are important numerical properties of a distribution. Typical measures of location are the expected value (expectation for short) and the median. For symmetric distributions, the expectation and the median coincide. When the distribution has an elongated right tail (respectively, left tail), i.e. the distribution is skewed to the right (resp., to the left), the expectation is greater (resp. lower) than the median. Typical measures of dispersion are the variance (or the standard deviation), the *interquartile range* and the *median absolute deviation from the median*. In this section we define some of them and we will study their properties latter.

Let $X_1, \ldots, X_n$ be r.v.'s with common distribution F_θ, then we define the *sample average* by

$$\overline{X} = \tfrac{X_1+\ldots+X_n}{n},$$

and the *sample variance* by

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

Note that $\overline{X}$ and S^2 are themselves r.v.'s. If $x_1, \ldots, x_n$ is an observed sample then we define the *observed sample average* and the *observed sample variance* by respectively

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

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

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

In the following we limit our selves to the case of observed samples $x_1, \ldots, x_n$. The development with random samples is analogous; just replace x_i by X_i and x by X, when necessary.

Given the observed sample $x_1, \ldots, x_n$, the *ordered observations* (or observed ordered statistics) are $x_{(1)} \leq x_{(2)} \leq \ldots x_{(n)}$, where $x_{(i)}$ is called the ith observed ordered statistics; in particular, $x_{(1)}$ is the observed minimum, $x_{(n)}$ is the observed maximum, and so on. For instance, if the observed sample is : 1.1, 0.5, 0.4, 3, 2.2, the observed ordered statistics are: 0.4, 0.5, 1.1, 2.2, 3.0. In this case, $x_1 = 1.1 = x_{(3)}$, $x_2 = 0.5 = x_{(2)}$ and so on.

The *observed lower quartile* is defined as $x_{(\lceil n/4 \rceil)}$, the *observed upper quartile* is defined by $x_{(\lceil 3n/4 \rceil)}$, where $\lceil u \rceil$ denotes the smallest integer greater than u. The observed sample median is defined by

$$\mathrm{Me}_x = \mathrm{Me}(x_1, \ldots, x_n) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \quad \text{odd}, \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & n \quad \text{even}. \end{cases}$$

The *sample inter quartile range* (IQR) is defined by $\mathrm{IQR}_x = x_{(\lceil 3n/4 \rceil)} - x_{(\lceil n/4 \rceil)}$ and the *sample median absolute deviation from the median* (MAD) is defined by

$$\mathrm{MAD}_x = \mathrm{MAD}(x_1, \ldots, x_n) = \mathrm{Me}(|x_1 - \mathrm{Me}_x|, \ldots, |x_n - \mathrm{Me}_x|).$$

The sample median is often preferred to the sample average as a measure of location when there are observations which are too far from the bulk of the data. In this case also the sample IQR or the sample MAD are preferred to the sample variance (or sample standard deviation).

Remark 2.1 *Note the the importance of the key word sample used above. For instance, think about the sample variance S^2. We saw in L0, that the variance σ^2 measures the dispersion (or spread) of a given r.v. X. On the other hand, S^2 also deals with dispersion, but for a random sample $X_1, \ldots, X_n$. Thus, σ^2 is a feature of the population, that is, a feature of the set of all possible units, which could be infinite. On the other hand, S^2 is a feature of the sample of n units and, as we will see in the incoming Lectures, $S^2 \to \sigma^2$ as $n \to \infty$.*

Remember, one of aims of statistics is to learn about or infer on the features of the population, for instance its variance σ^2, by means of features of the sample.

2.2.1 Histograms

We can get an idea about the shape of the distribution by comparing measures of location and dispersion, but often graphical representation of the data give us a clearer picture. A simple technique to obtain an idea about the probability density of the population through an observed sample is the *histogram*. Here we consider only its observed sample version.

For a partition $a_0 < a_1 < a_2 < \cdots < a_m$ that covers the range of data $x_1, \ldots, x_n$, the histogram is the function that, on each interval $(a_{j-1}, a_j]$, takes on the value equal to the number of sample points x_i belonging to that interval divided by the length of the interval, $j = 1, \ldots, m$. Often the *scaled* version is considered in which the counts are divided by the sample size n. The scaled histogram is useful when we want to compare the distribution of the observed sample with a theoretical one, because like a theoretical probability distribution, also the area under the scaled histogram sums to one. In $x \in (a_{j-1}, a_j]$, the scaled histogram is given by

$$h_n(x) = \tfrac{1}{n(a_j - a_{j-1})} \sum_{i=1}^{n} 1_{(a_{j-1}, a_j]}(x_i),$$

where the indicator function $1_{(a_{j-1}, a_j]}(x_i)$ equals 1 if $x \in (a_{j-1}, a_j]$ and 0 otherwise. A scaled histogram can provide a good impression of the density of the data, provided the the partition $a_0 < a_1 < a_2 < \cdots < a_m$ has been chosen well and that the sample size n is not too small.

2.2.2 Empirical distribution function

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

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

where $\mathcal{I}_{X_i}(x)$ is a Bernoulli r.v. which assumes value 1 in the event $X_i \leq x$ and 0 otherwise. $F_n(x)$ is a step function that increases by n^{-1} at each observation. If the observed sample $x_1, \ldots, x_n$ is used, then the observed empirical distribution function is defined by

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

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

Just like the scaled histogram aims at approximating the p.d.f. of the distribution F that has generated the data, the empirical distribution function approximates the d.f. F itself. In practice, the EDF should be preferred because it:

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

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

2.2.3 Boxplots

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

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

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

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

2.2.4 QQ-plots

The *QQ-plot* or *quantile-quantile plot* is another graphical method that is commonly used for comparing a given distribution F with a given sample $x_1, \ldots, x_n$. To build this plot, we first get the ordered statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, and the QQ-plot is obtained by plotting the set of values

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

Here F is a user-supplied d.f.. The use of $j/(n+1)$ quantile of F is due to the following fact. It can be shown that, for an i.i.d. random sample $X_1, \ldots, X_n$, the jth *ordered statistic* $X_{(j)}$ is an r.v. and if $U_1, \ldots, U_n$ are such that $U_i \overset{\text{iid}}{\sim} \text{Unif}(0,1)$, then $E(X_{(j)}) = E\left[F^{-1}\left(U_{(j)}\right)\right] \doteq F^{-1}\left(\frac{j}{n+1}\right)$. Hence the plotting positions $F^{-1}(j/(n+1))$ are approximate expected order statistics, justifying their use in QQ-plots.

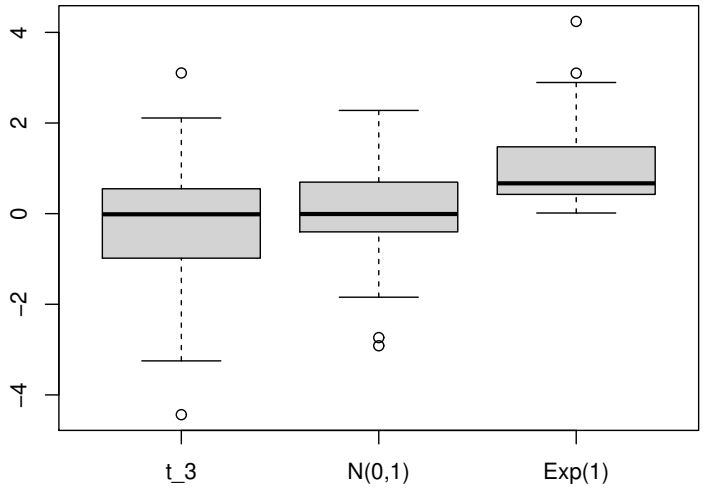


Figure 2.1: Boxplots of samples of size 50 from the t-Student distribution with 3 degrees of freedom (left), the standard normal distribution (middle), and the exponential distribution with unit rate (right).

If the points of a QQ-plot lie along the $y = x$ line, then this indicates that observations $x_1, \ldots, x_n$ are compatible with the distribution F. Typically F is chosen to be the normal distribution. Deviations from the half-plane line could be of various types, for instance the points could be rotated, shifted, or could have a curved shape. Shifts and rotations with respect to the $y = x$ line indicate differences in terms of location and scale, respectively. Whereas U-shaped or S-shaped QQ-plots indicate differences in terms of asymmetry or in terms of the length of the tails, respectively.

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

2.3 Bivariate samples

Observed data for each unit sample may arise also as vectors. For instance, an urban meteorological stations located in a city, can provide real-time measured data about: temperature, pressure, humidity, rain fall, wind speed, solar irradiation, and concentration of various pollutants.

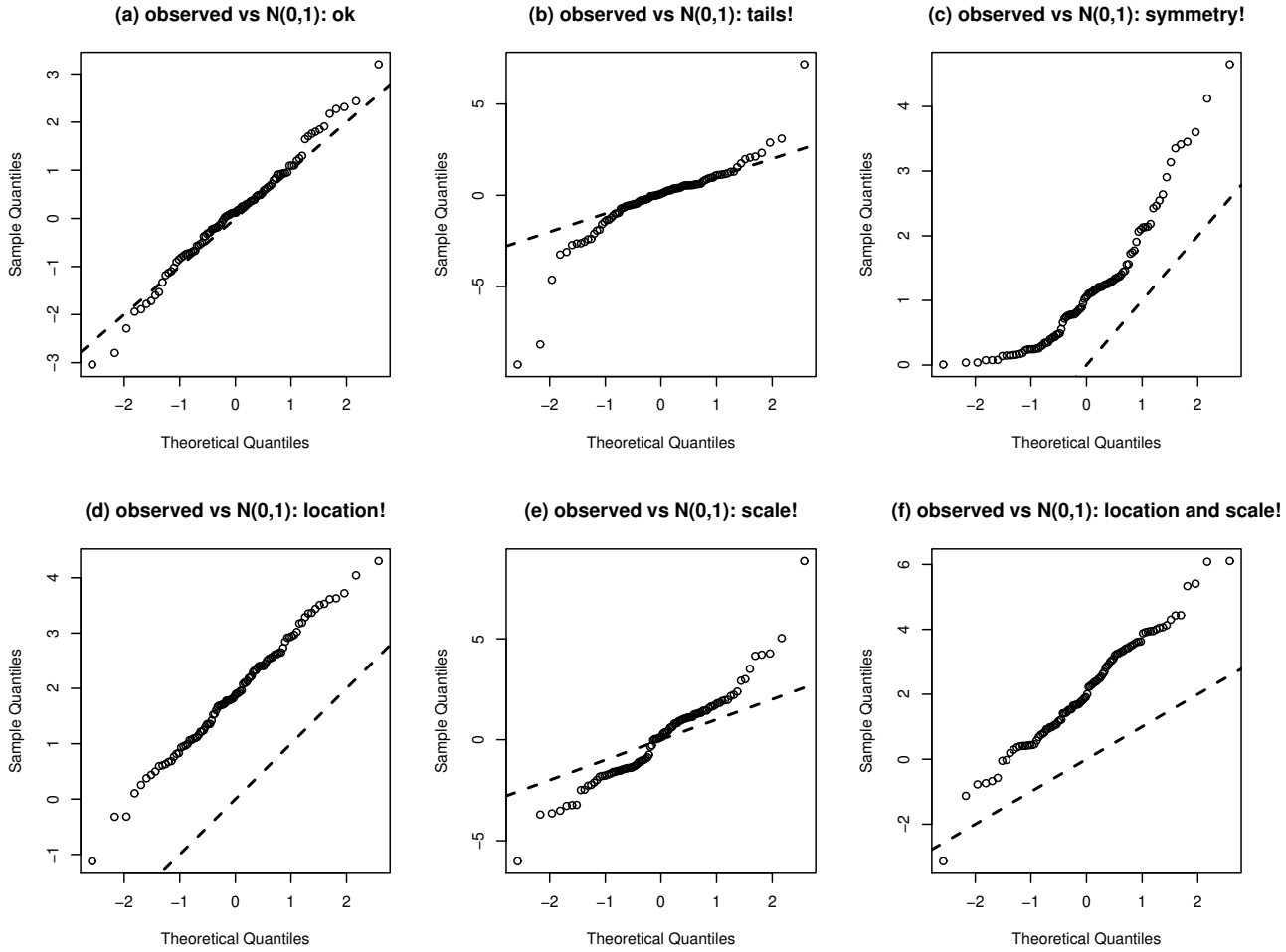


Figure 2.2: QQ-plots of six different observed samples each of size 100, with respect to the standard normal distribution.

Here we focus on two variables at time, thus let $(x_1, y_1), \ldots, (x_n, y_n)$ be the observed sample of size n obtained as realisations of the r.ve.'s (X_i, Y_i), $i = 1, \ldots, n$. Then for the random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, the *sample covariance* between the two variables is defined by

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

and the *sample correlation coefficient* is defined by

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

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

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

Again note that S_{XY}, R_{XY} are random variables wheres s_{xy}, r_{xy} are numbers. Both s_{xy} and r_{xy} as well as their random versions are measures of linear association between the two variables involved. s_{xy} ranges over the set of real numbers whereas $r_{xy} \in [-1, 1]$. The higher r_{xy} the higher is the degree of linear association between two variables. A correlation coefficient equal to zero implies that there is no linear association between two variables, though the variables could be related in some non linear fashion as in Figure 2.3. Here the pairs of data are shown by means of a graphical method *scatter plot*. Only the plot on the left shows a strong (linear) correlation. The plot on the right shows a clear quadratic relation between x_i and y_i^2 but the correlation coefficient is essentially zero.

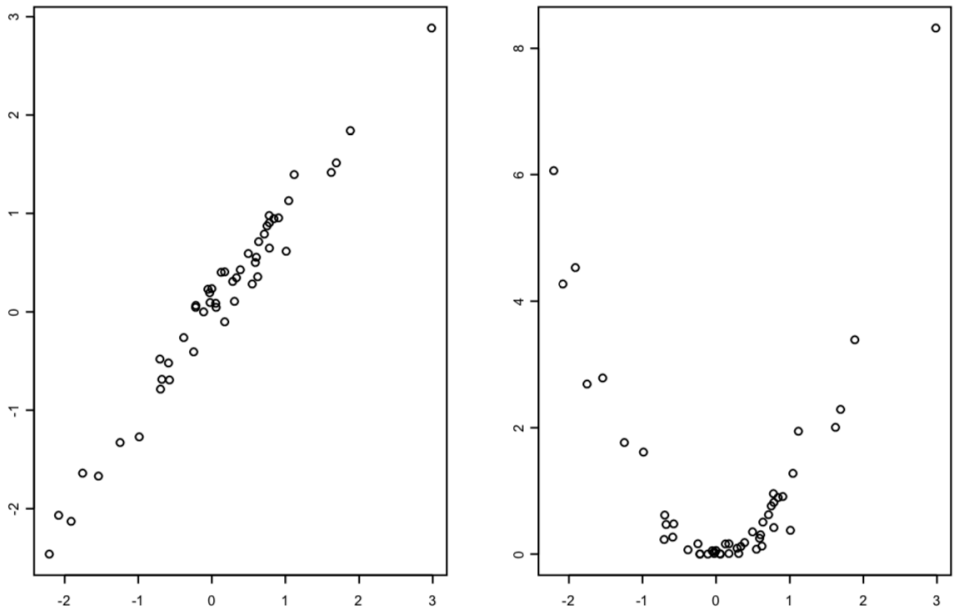


Figure 2.3: Examples of scatter plots. On the left it is shown (x_i, y_i) with $r_{xy} = 0.95$. On the right it is shown (x_i, y_i^2) for which $r_{xy^2} = -0.05$.

2.4 Statistical models

In this section we introduce the concept of parametric statistical model. The adjective "parametric" is due to the fact that the statistical model to be introduced in the next two subsections is based on a probability distribution function which depends (i.e. is indexed) by a parameter, the latter laying in a finite-dimensional Euclidean space. The qualification parametric is essential in general because as we will see near the end of this course, there are also nonparametric statistical models. For the sake of brevity, here after we write statistical models when referring to parametric statistical models.

A statistical model can be defined as a probabilistic representation of a physical system with the aim of modelling its current behaviour and predicting its future behaviour. Typically, we have $x_1, \ldots, x_n$ measurements of some characteristics of the physical system and based on these measurements we might be interested in

computing a measure of location or some measure of dispersion. One approach for doing this is to assume that each observation is a realisation of an r.v. belonging to a given family with parameters to be "learned" from the observations. The aim is then to learn or estimate the parameter of this model. Here is an example.

Example 2.1 *Every washing machine (WM) sold in the UE market must be accompanied by technical documentation which describes, among other things, the energy consumed during a typical washing cycle. In order to measure the consumed energy, the WM is tested in a laboratory, say, n times under the same conditions. The resulting values of energy consumptions are $x_1, \ldots, x_n$. It is of interest to know both a measure of location for these data and their spread. A reasonable approach to this problem is to assume that each observation x_i is a realisation of a r.v. with distribution $\mathrm{N}(\mu, \sigma^2)$, where μ and σ^2 are to be found. The normal assumption for measurements such as energy is very reasonable. This is because the consumed energy of an WM is the sum of the energy consumed by its components (i.e. electric circuits, motor, heater, etc.). Thus by the CLT, the sum will be approximately normally distributed.*

We focus on statistical models based on the assumption of random samples being identically and independently, i.e. i.i.d, distributed and statistical models based only on the assumption of independence. It is worth stressing that the assumption of data being realisations of identically and independently distributed r.v. or of independent but not identically distributed r.v. is not mathematically or statistically verifiable and is founded on the way the data are collected. For instance, in Example 2.1 it is reasonable to assume that the data are i.i.d. since the performance of a washing cycle has no consequences on that of the next cycle. On the other hand, in a particularly cold winter day, the water temperature in the laboratory could be lower than the temperature in summer days. Thus it seems reasonable to assume that the amount of energy consumed by the WM depends on the environment temperature. In the latter case the measurements may be independent but not identically distributed.

2.4.1 Identically and independently distributed r.v.

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with $X_i \sim F_\theta$, or more compactly $Y_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$, where the distribution F_θ, is indexed by the unknown parameter θ. Then, since Y_i are i.i.d., their joint distribution function is

$$F_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} F_\theta(y_i)$$

and the joint probability density function is

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} f_\theta(y_i).$$

A *statistical model* is a family of joint d.f. or joint p.d.f. for the r.v. $Y_1, \ldots, Y_n$ indexed by the parameter θ, that is the set

$$\{F_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$$

or

$$\{f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}.$$

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

Example 2.2 *Let* $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$ *where* $\mu \in \mathbb{R}$ *and* $\sigma^2 \in \mathbb{R}_{>0}$ *are unknown parameters. The joint p.d.f. of these r.v.'s is*

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

and the statistical model

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

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

Example 2.3 *Let* $X_i \overset{\text{iid}}{\sim} \mathrm{Poi}(\lambda)$, $i = 1, \ldots, n$ *where* $\lambda \in \mathbb{R}_{>0}$ *is the unknown parameter. The joint p.d.f. of these r.v.'s is*

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

and the statistical model

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

is called the Poisson model, where the unknown parameter is λ.

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

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

and the statistical model

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

is called the Bernoulli model, where the unknown parameter is θ.

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

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

Example 2.5 *A manufacturer (P1) of motors for washing machines (WM) claims that his next generation motors (called NGM1) are energetically more efficient, while achieving the same speed as the previous sold, i.e. old, version motors (OM). Which one should we buy?*

In order to verify this claim, two WM are taken, one is equipped with NGM1 and the other is equipped with OM. Suppose that the WM with OM is tested n times, whereas the WM with NGM2 is tested m times. Let $X_1, \ldots, X_n$ be the r.v.'s denoting the energy consumption of the WM with OM and let $Y_1, \ldots, Y_m$ be the r.v.'s denoting the energy consumption of the WM with NGM1. Measurements are reasonably independent from one other. Furthermore, we expect the WM with OM to possibly behave differently from the WM with NGM1. Thus the distribution of X_i could be "different" from that of Y_j. By different we mean that they could have different parameter values but both d.f.'s must belong to the same family of distributions. Then a reasonable assumption is $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu_x, \sigma_x^2), i = 1, \ldots, n$ and $Y_j \overset{\text{iid}}{\sim} \mathrm{N}(\mu_y, \sigma_y^2), j = 1, \ldots, m$. Thus the assumed distribution depends on the type of motor the WM is equipped with.

Under these assumptions, the statistical model is

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

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

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

Example 2.6 *Another manufacturer (P2) of motors for washing machines (WM) also claims that his next generation motors (called NGM2) are energetically more efficient, while achieving the same speed as previous sold, i.e. old, version motors (OM). P2 also claims that NGM2 are more efficient than NGM1. Which one should we buy?*

In order to verify this claim, three WM are taken, one is equipped with the OM, one with NGM1 and one with NGM2. Suppose that the WM with OM is tested n times, the WM with NGM1 is tested m times and the WM with NGM2 is tested q times. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be as in Example 2.5 and let $Z_1, \ldots, Z_q$ be the r.v.'s denoting the energy consumption of the WM with NGM2. Again, measurements are reasonably independent from one other. Furthermore, we expect the WMs with OM, NGM1 and NGM2 to behave differently from each other. In addition to those of Example 2.5, it is also reasonable to assume that $Z_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu_z, \sigma_z^2), i = 1, \ldots, n$. Again, the assumed distribution depends on the type of motor the WM is equipped with.

Under these assumptions, the statistical model is

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

where the unknown parameters which are to be learned from the data are $(\mu_x, \mu_y, \mu_z, \sigma_x^2, \sigma_y^2, \sigma_z^2)$. *In these type of problems we might be interested to know if* $\mu_x = \mu_y = \mu_z$ *while assuming that* $\sigma_x^2 = \sigma_y^2 = \sigma_z^2$ *(called homoscedasticity assumption). This is again an hypothesis testing problem called analysis of variance or just ANOVA, and we will see more about it the incoming lectures.*

The next example extends Example 2.6 to a situation in which the r.v.'s vary continuously with some other fixed quantity.

Example 2.7 *Suppose we are measuring some quantity which result is due in part to a variable under our control and in part due to randomness. For instances, we might be interested to know how the lifetime of WM motors varies with the environment humidity. Or suppose that you are studying a device for removing arsenic (which has negative health effects on human beings) from drinkable water but you know that the effectiveness of the removal depends on the pH of water; so it is of interest to assess how arsenic removal changes with water pH.*

Let $Y_1 \ldots, Y_n$ *be r.v.'s with* $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$. *Thus we assume that the* n *r.v.'s have different means* μ_i *but the same variance* $\sigma^2 > 0$*; all parameter are unknown. Furthermore, let* $\mu_i = \alpha + \beta t_i$ *where* $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ *are unknown parameters and* t_i *is a non-random variable. For instance, in the problem of WM motor lifetime,* t_i *is the environment humidity at ith measurement (or day); in the problem of arsenic removal from water,* t_i *could be the pH of the water at the ith water sample. Using the properties of the normal distribution (assuming again that* t_i *are non stochastic) we could also write*

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

or in a more commonly used form

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

This is because, given r.v.'s $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, *then* $\alpha + \beta t_i + \epsilon_i \sim \mathrm{N}(\alpha + \beta t_i, \sigma^2)$ *so* $\alpha + \beta t_i + \epsilon_i$ *has the same distribution as* Y_i, $i = 1, \ldots, n$.

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

The statistical model in this case can be determined easily thanks to the assumption that measurements are independent of each other, and is

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

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

This problem is known as (simple) linear regression and word "linear" is due to the linear equation $\alpha + \beta t_i$. *The word "simple" it is because it deals with a single predictor variable* t_i. *If there are more than one predictor variables say* w_i, v_i, *etc., then we could also include them into the regression model which then would be called* multiple regression model.

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

Example 2.8 *A polar station located in a remote area of the northern Polar region is powered by electricity generated by an outdoor generator. To deal with freezing outdoor temperatures, the generator has a special electric system made of 7 switches which run in parallel. The generator stops working if 5 or more switches breakdown. It is predicted that the next night will be exceptionally freezing, with predicted temperature equal to $-40°C$. Given this temperature, researchers living the in the polar station would like to know what is the probability that the generator will stop working.*

Let $t_1, \ldots, t_n$ be temperature levels registered in n occasions in the past in the same location of the polar station and let $Y_1 \ldots, Y_n$ be r.v.'s with $Y_i \sim \text{Bin}(7, \theta_i)$ which represent the number of failed switches when there are 7 switches overall. Thus we assume that the n r.v.'s have different success probability θ_i and index equal to 7. Furthermore, it is reasonable to assume that the behaviour of the switches will depend on the temperature, thus we assume that the success probability θ_i is a function of temperature t_i, by $\theta_i = \frac{e^{\alpha+\beta t_i}}{1+e^{\alpha+\beta t_i}}$ where $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ are unknown parameters. Thus in a more compact form the model we built is

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

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

The statistical model is then

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

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

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

This model is known as the logistic regression model.

In general, building or designing a good statistical model for a problem at hand may be a difficult task and it requires some patience and experience. Most importantly, it requires deep knowledge about statistical models. The latter is typically acquired through statistical modelling courses and is outside the scope of this course. Nevertheless, the tools of statistical inference, which are the main aim of this course, are paramount to this deeper knowledge about statistical models.

References

[LM18] Larsen, R. J. and Marx, M. L. (2018) *An Introduction to Mathematical Statistics and its Applications* (6th edition), Pearson Education, Inc.

Practice Lecture 2: Descriptive statistics and models

Erlis Ruli (ruli@stat.unipd.it)

20 October 2020

1 Univariate samples

To illustrate univariate summaries let us use a real dataset. There are many datasets that come shipped with R, here we consider the `cars` dataset. The latter reports speed and stopping time for 50 car models.

```
> head(cars)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

```
> summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

We see that the average speed for this observed sample is equalt o 15.4 very close the the median, whereas the average distance is 42, slightly different from the median, which in this case is 36. The command `summary` gives also the observed ordered statistics: average, lower quartile, median, upper quartile and the maximum.

The sample variance is computed by

```
> var(cars$speed)
```

```
## [1] 27.95918
```

```
> var(cars$dist)
```

```
## [1] 664.0608
```

We can also compute the median of a sample by

```
> median(cars$speed)
```

```
## [1] 15
```

```
> median(cars$dist)
```

```
## [1] 36
```

The observed quartiles can be obtained by

```
> quantile(cars$speed, probs = 1/4)# lower quarile
```

```
## 25%
##  12
```

```
> quantile(cars$speed, probs = 3/4) # upper quartile
```

```
## 75%
##  19
```

```
> quantile(cars$speed, probs = 2/4) # the median
```

```
## 50%
##  15
```

By the quantile function we can thus compute the *observed sample quantile* of any level $p \in (0, 1)$. There are many possible definitions of the observed sample quantile. One of them is the following. Let $\hat{x}_p$ denote the observed sample quantile, then for a sample $x_1, \ldots, x_n$, we define

$$\hat{x}_p = \begin{cases} (x_{(pn)} + x_{(pn+1)})/2 & \text{if} \quad pn \in \mathbb{N} \\ x_{(\lceil pn \rceil)} & \text{if} \quad pn \notin \mathbb{N} \end{cases}$$

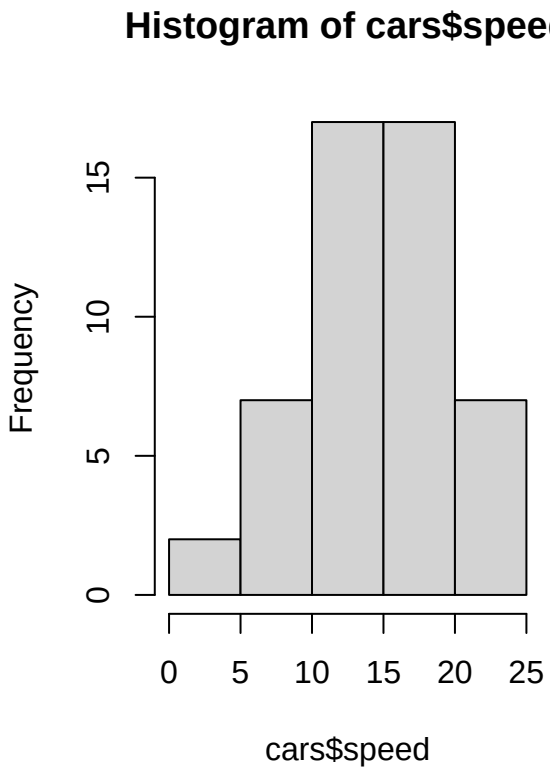
$\hat{x}_p$ extends the observed quartiles to a more general division of the distribution.

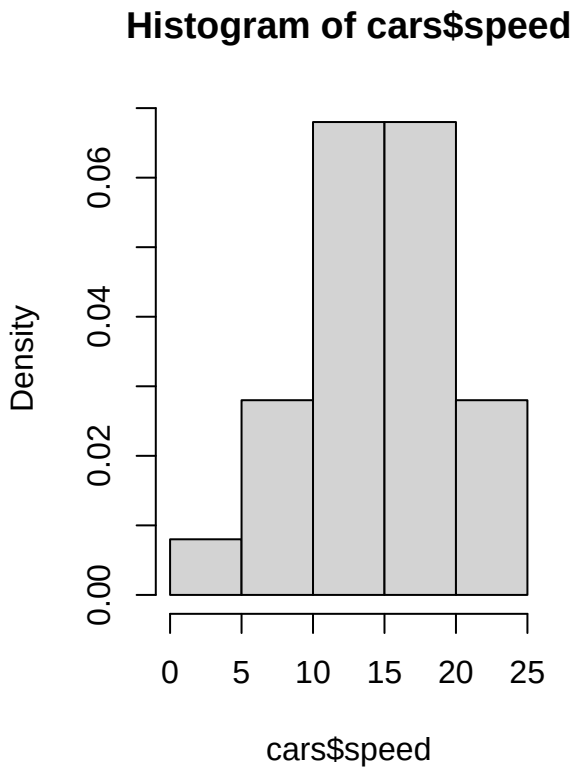
There are other ways of computing observed quantiles, according to the definition used (see the help page of the function `quantile`) which may give different answers. Partly, this is also due to the fact that we may have ties in the observed values. However, in practice such differences are negligible, especially for large samples. This issue does not arise when we compute the quantile of a random variable since its definition (based on the infimum) guarantees us a unique solution. In the above example we used `R`'s default method; see the help page for more details.

1.1 Histograms

We have seen histrograms also in P1. The command to be used is `hist`, its option `breaks` controls the number of points in the partition and the option `freq` controls the type of histogram.

```
> par(mfrow=c(1,2))
> # a non scaled histogram
> hist(cars$speed)
> # a scaled histogram
> hist(cars$speed, freq = FALSE)
```

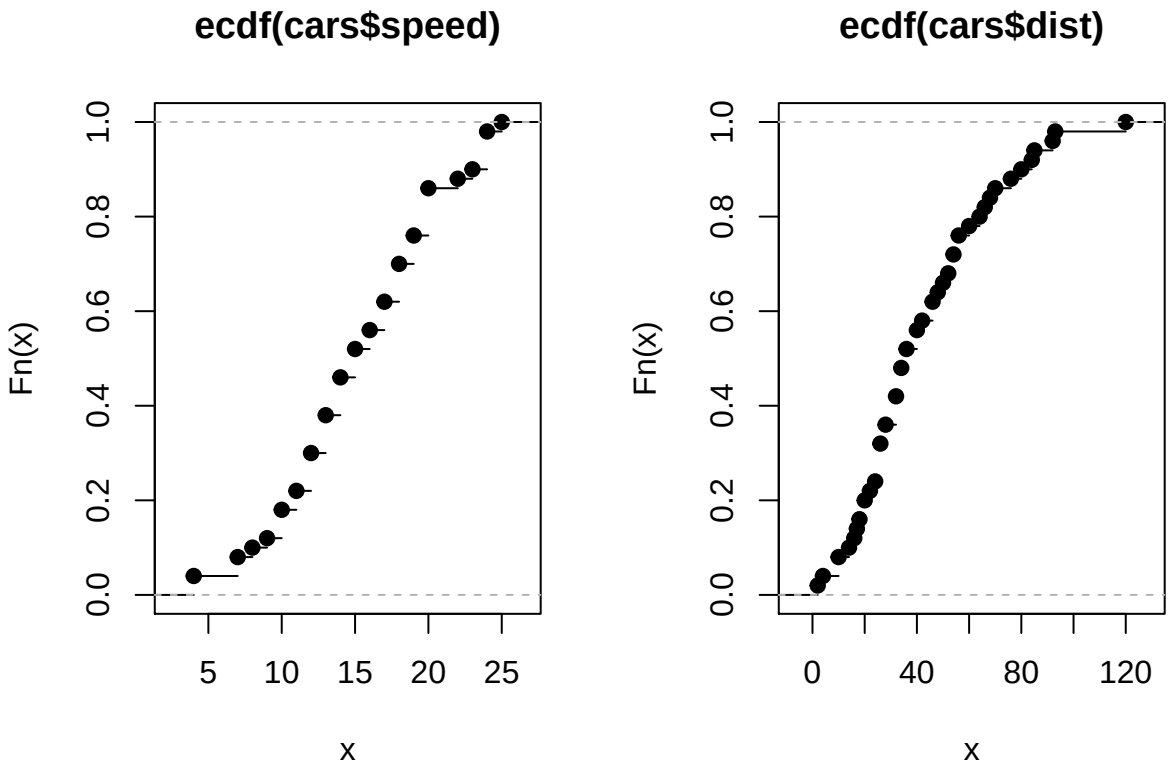




1.2 Empirical distribution function

The empirical distribution function is computed with

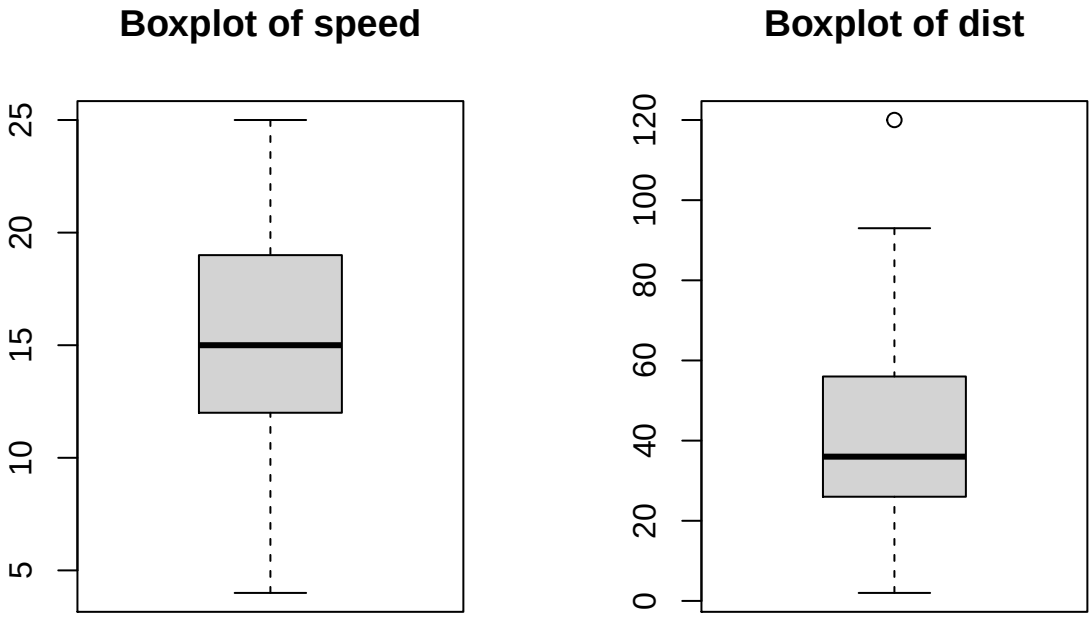
```
> par(mfrow=c(1,2))
> plot(ecdf(cars$speed))
> plot(ecdf(cars$dist))
```



1.3 Boxplots

Boxplots are computed with

```
> par(mfrow=c(1,2))
> boxplot(cars$speed, main="Boxplot of speed")
> boxplot(cars$dist, main="Boxplot of dist")
```

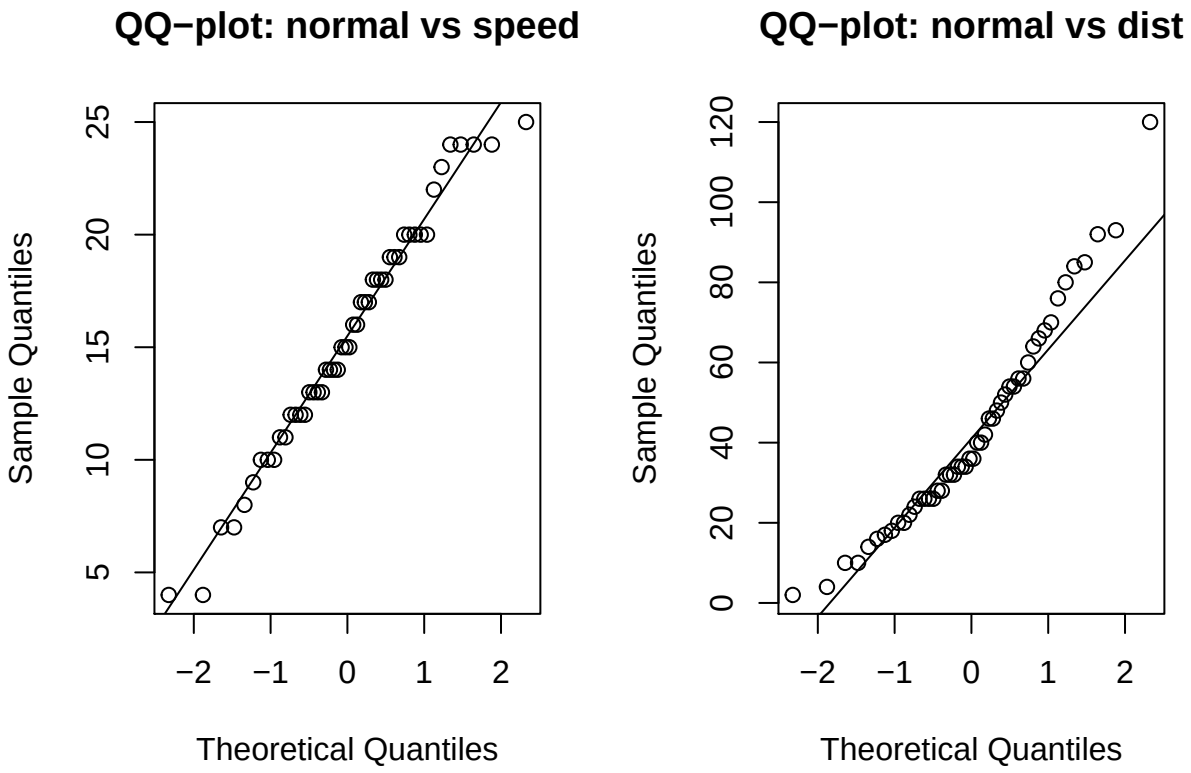


For the variable `dist` we notice an unusual observation, i.e. an outlier.

1.4 QQ-plots

To build QQ-plots agains the normal quantiles we can use the following commands

```
> par(mfrow=c(1,2))
> qqnorm(cars$speed,main="QQ-plot: normal vs speed")
> qqline(cars$speed)
>
> qqnorm(cars$dist,main="QQ-plot: normal vs dist")
> qqline(cars$dist)
```



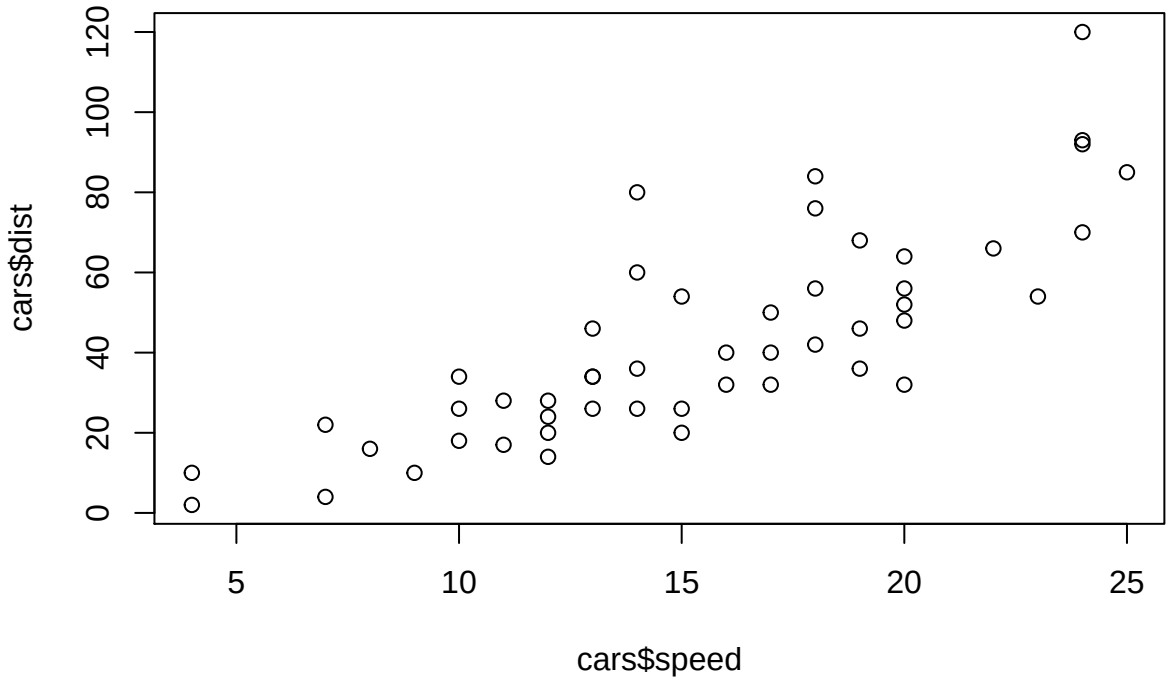
We see that the quantiles of `speed` are quite close to those of the normal distribution whereas those of `dist` are quite different. This suggests that the normal model could be a good model for `speed` but perhaps not for `dist`. In particular, we notice that `dist` has a skewed distribution. For the half-plane line we used the command `qqline` which is a robust version of the $y = x$ line.

Exercise For the variable speed only, try to build the QQ-plot "by hand". For this, suppose that F is N(15.4, 28). The quantile function of the normal distribution is given by the command `qnorm`.

2 Bivariate samples

From the scatter plot of the two variables shown below we notice that there is a positive linear association between the two variables.

```
> plot(cars$speed, cars$dist)
```



Indeed, the sample covariance and sample correlation indices which are

```
> # covariance
> cov(cars$speed, cars$dist)
```

```
## [1] 109.9469
```

```
> # correlation: option 1
> cor(cars$speed, cars$dist)
```

```
## [1] 0.8068949
```

```
> # correlation: option 2
> cov(cars$speed, cars$dist)/
+   sqrt(var(cars$speed)*var(cars$dist))
```

```
## [1] 0.8068949
```

Inferential Statistics **First semester, 2020**

Lecture 3: The likelihood function

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

The likelihood function is one of the most important tools of statistical inference, both in the frequentist and in the Bayesian setting. In this lecture we introduce the likelihood function from a pure descriptive perspective and we illustrate its use by means of practical examples. In the incoming lectures we will see how to use the likelihood function for conducting inference on the parameters of a statistical model. Near the end of this course we will see how the likelihood function is used in the Bayesian setting.

3.1 The likelihood function

Definition 3.1 *Let $Y_1, \ldots, Y_n$ be a random sample of size n, with $Y_i \sim F_\theta$, $\theta \in \Theta$ independently for each $i = 1, \ldots, n$ and let $f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) = \prod_{i=1}^n f(y_i; \theta)$ be the joint probability distribution of the sample. For a fixed sample $y_1, \ldots, y_n$, the likelihood function is denoted by $L(\theta)$ and is the function*

$$L(\theta) : \Theta \to \mathbb{R}_{>0},$$

with

$$L(\theta) = L(\theta; y_1, \ldots, y_n) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) = \prod_{i=1}^n f(y_i; \theta).$$

Thus the likelihood function is obtained by holding fixed $y_1, \ldots, y_n$ and letting θ vary in Θ, the space of allowable values for θ. The set Θ is also called the parameter space and its characteristics (discrete or continuous, finite or infinite, etc.) depend on the d.f. F_θ.

For a given parameter value, say $\theta = \theta_0$, the likelihood function can be interpreted as the probability of observing a sample like the one actually observed. Different parameter values lead to different probabilities, i.e. likelihoods. The higher is this probability the more likely is to observe data such as those actually observed, under the model F_θ. That is, if $L(\theta_1) > L(\theta_2)$, then θ_1 is to be preferred to θ_2 since the former is more likely to generate the observed data. Thus, it is reasonable to look for the value(or values) of θ which maximises (or maximise) $L(\theta)$; note that the likelihood function typically takes on very small values.

We start illustrating the likelihood function by an example in which the parameter space is finite.

Example 3.1 *A manufacturer of coins for the gambling industry produces three types of coins, with faces named W (Win) and L (Loose). Coins of type U1 are such that $P(\{W\}) = 1/3$, coins of type U2 are such that $P(W) = 1/4$ and coins of type F are such that $P(W) = 1/2$.*

Nature picks a coin at random from the three types, without revealing to us the type, and tosses the coin three times. If we denote by $\theta = P(\{W\})$, then by straightforward probability calculus we get $P(\{WWW\}) = \theta^3$, $P(\{LLL\}) = (1-\theta)^3$, $P(\{WWL\}) = \theta^2(1-\theta)$ and $P(\{WLL\}) = \theta(1-\theta)^2$. Table 3.1 gives the sample points and the associated probabilities for this experiment, for each value of θ.

sample	Type of coins		
	$\theta = 1/4$ (type U2)	$\theta = 1/3$ (type U1)	$\theta = 1/2$ (Type F)
WWW	0.0156	0.0370	0.125(•)
WWL	0.0469	0.0741	0.125(•)
WLW	0.0469	0.0741	0.125(•)
LWW	0.0469	0.0741	0.125(•)
WLL	0.1406	0.1482(•)	0.125
LWL	0.1406	0.1482(•)	0.125
LLW	0.1406	0.1482(•)	0.125
LLL	0.4219(•)	0.2963	0.125

Table 3.1: Likelihood function of θ for the problem of tossing three coins.

Reading the table column-wise we get the usual probability distribution of the simple events, i.e. summation of rows of a given column gives 1. Hence in the table we see three different probability distributions. On the other hand, the likelihood function is obtained by reading the table row-wise; thus in this case we have four different likelihood functions.

The likelihood functions are read as follows. Suppose, we observed $\{WWW\}$. Then for $\theta = 1/4$ the likelihood, i.e. the probability, of observing this sample is 0.0156. On the other hand, for $\theta = 1/2$ such a probability is equal to 0.125. Thus, since the observed sample $\{WWW\}$ is more likely to be observed if $\theta = 1/2$ than for other values of θ, we can say, i.e. infer, that with the above observed sample, the coin chosen by Nature is more likely to be of type F.

On the contrary if we observed $\{LLL\}$ then it is more likely that the coin is of type U2.

Remark 3.1

(i) Whatever is our deduction, i.e. inference, from the observed sample and the likelihood function, we will never know Nature's choice. That is, we will never know the truth. However, as we will see below, with increasing amounts of data we can be increasingly more confident about our deduction.

(ii) In practice we have only one observed sample, or few observed samples if we have enough funds, but certainly we cannot observe the complete sample space. Thus, in practice, we can analyse only a single row (or few rows) of the above table, that is, we can only deal with one likelihood function.

(iii) On the contrary to point (ii), with simulated data we do know the truth and in this case we can study more or less exactly the behaviour of maximum of the likelihood function. For instance, in this example, we see that the the maximum of the likelihood function has discrete distribution with support $\{1/4, 1/3, 1/2\}$.

(iv) Often the likelihood function can be factored as $L(\theta) = w(y_1, \ldots, y_n)q(\theta; y_1, \ldots, y_n)$, where $w(y_1, \ldots, y_n)$ is a function of the sample and $q(\theta; y_1, \ldots, y_n)$ is a function of both the parameter θ and the sample $y_1, \ldots, y_n$ or a function of it. If this is the case, then we can discard $w(y_1, \ldots, y_n)$ and write $L(\theta) \propto q(\theta; y_1, \ldots, y_n)$, where the symbol "$\propto$" means "proportional to".

(v) Since $L(\theta) \in \mathbb{R}_{>0}$, then it is often easier to work with its natural logarithm, which we denote by $\ell(\theta)$, thus $\ell(\theta) = \log L(\theta)$ and call it the log-likelihood function. If the remark (iv) applies, then we can write $\ell(\theta) = \log q(\theta; y_1, \ldots, y_n) + \text{const.}$, where const. means not a function of θ.

Exercise. *Let $\widehat{\theta}$ denote the value of θ at which the maximum of the likelihood function is achieved. Remark 3.1(iii) and Table 3.1 suggest that $\widehat{\theta}$ is a r.v.. Determine its distribution.*

In most practical problems the parameter space Θ is infinite and uncountable, i.e. $\Theta \subseteq \mathbb{R}^d$. Here are some examples of this kind.

Example 3.2 *Let $Y_1, \ldots, Y_n$ be a random sample as in Example 2.3 (Lecture 2). Then the likelihood function is*

$$L(\lambda) \propto e^{-n\lambda}\lambda^{\sum_{i=1}^n y_i},$$

and the log-likelihood is

$$\ell(\lambda) = -n\lambda + \sum_{i=1}^n y_i \log(\lambda) + \text{const.}.$$

Example 3.3 *Let $Y_1, \ldots, Y_n$ be a random sample as in Example 2.4 (Lecture 2). Then the likelihood function is*

$$L(\theta) \propto \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i},$$

and the log-likelihood is

$$\ell(\lambda) = \sum_{i=1}^n y_i \log(\theta) + \left(n - \sum_{i=1}^{y_i}\right)\log(1-\theta) + \text{const.}.$$

Example 3.4 *Let $Y_1, \ldots, Y_n$ be a random sample as in Example 2.2 (Lecture 2). Then the likelihood function is*

$$L(\mu, \sigma^2) \propto \tfrac{1}{(\sigma^2)n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2}$$

and the log-likelihood function is

$$\begin{aligned}
\ell(\mu, \sigma^2) &= -\tfrac{n}{2}\log(\sigma^2) - \tfrac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2 + \text{const.} \\
&= -\tfrac{n}{2}\log(\sigma^2) + \tfrac{n\mu^2}{2\sigma^2} - \tfrac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 + \overline{y}\tfrac{n\mu}{\sigma^2} + \text{const.}.
\end{aligned}$$

Note that here the likelihood function is $L(\mu, \sigma^2) : \mathbb{R} \times \mathbb{R}_{>2} \to \mathbb{R}_{>0}$.

Example 3.5 *Let $Y_1, \ldots, Y_n$ be an i.i.d. random sample with $Y_i \sim \text{Wei}(\alpha, \beta)$. The joint distribution of this random sample is given by the product of their marginal distributions, thus the statistical model is*

$$\left\{ \tfrac{\alpha^n}{\beta^n}\left(\prod_{i=1}^n y_i\right)^{\alpha-1} e^{-\sum_{i=1}^n y_i^\alpha/\beta} : (\alpha, \beta) \in \mathbb{R}^2_{>0} \right\}.$$

It follows that the likelihood function is

$$L(\alpha,\beta) \propto \tfrac{\alpha^n}{\beta^n}\left(\prod_{i=1}^{n} y_i\right)^{\alpha} e^{-\sum_{i=1}^{n} y_i^{\alpha}/\beta}$$

and the log-likelihood function is

$$\ell(\alpha,\beta) = n(\log\alpha - \log\beta) + \sum_{i=1}^{n}\log(y_i)\alpha - \sum_{i=1}^{n} y_i^{\alpha}/\beta + \text{const.}.$$

We may wish to compare two different likelihood functions, perhaps based on different models or perhaps using different sets of data. In any case, when plotting the likelihood function it is useful to plot a scaled version of it called *relative likelihood* defined by

$$RL(\theta) = \tfrac{L(\theta)}{L(\widehat{\theta})}.$$

where $\widehat{\theta}$ is the value of θ at which $L(\theta)$ achieves its maximum. Note that $0 < RL(\theta) \leq 1$ by definition since $L(\theta) \leq L(\widehat{\theta})$.

3.2 The observed information

Assume we wish to evaluate the conformity of the products of a production line with respect to quality standards and n products are taken at random. Then the statistical formulation of this problem is to consider $Y_1, \ldots, Y_n$ an i.i.d. random sample with $Y_i \sim \text{Ber}(\theta)$, with θ unknown. Owing to time constraints, it was decided to take only $n = 10$ samples, and the observed sample $(y_1, \ldots, y_n)$ resulted

$$(0,1,1,1,0,1,1,0,1,1),$$

where 1 indicates that the product conforms with the quality standards and 0 indicates that the product does not conform. With this observed sample, the likelihood and the log-likelihood functions are respectively

$$L(\theta) = \theta^7(1-\theta)^3, \quad \text{and} \quad \ell(\theta) = 7\log\theta + 3\log(1-\theta).$$

The relative likelihood function is shown in Figure 3.1 by the thick black curve. From the plot of the relative likelihood we note that there is a θ such that the observed sample has the highest probability of being observed, that is, the likelihood function has a maximum. Let us locate this maximum. Taking the first derivative of the log-likelihood we get

$$\ell'(\theta) = \tfrac{d\ell(\theta)}{d\theta} = \tfrac{7}{\theta} - \tfrac{3}{1-\theta}.$$

Solving $\ell'(\theta) = 0$ in θ we obtain the solution to this equation being equal to $\widehat{\theta} = 7/10$. This solution is a global maximum since the function is strictly concave (the second derivative is negative everywhere).

Now suppose we increase the sample size to 50 and the observed sample is such that there are thirty five 1's and fifteen zeros. The relative likelihood function for this second sample is also shown in Figure 3.1 in red dashed.

From this figure we notice that the dashed (red) curve is narrower than the thick (black) one. For instance, while with a sample size equal to 10, values of $\theta \in (0.2, 1)$ are plausible since the likelihood is much higher than for $\theta \notin (0.2, 1)$, for the red curve, plausible values of θ are found in much narrow intervals, say $(0.5, 0.85)$. Thus, although the most plausible value of θ is identical in both cases, the range of plausible values of θ with $n = 10$ is wider than that with $n = 50$.

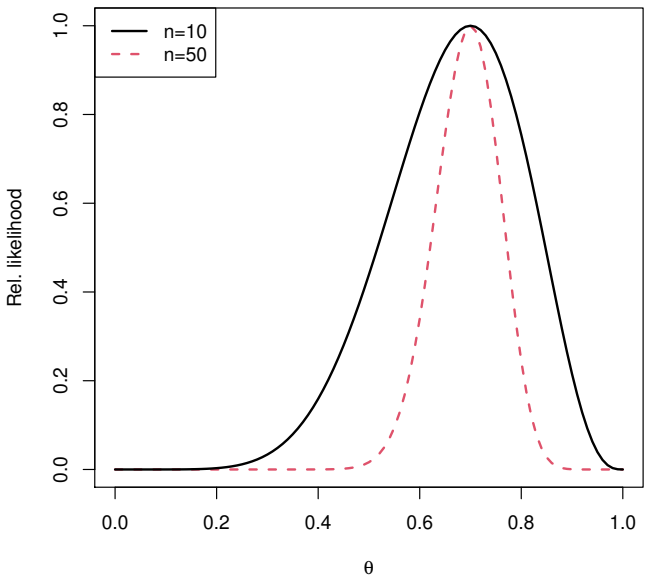


Figure 3.1: Relative likelihood functions for two observed Bernoulli random samples.

To capture this idea it is useful to introduce the concept of observed information, which we define below.

Definition 3.2 *Let $\ell(\theta)$ be the log-likelihood then the observed information is a function of θ defined by*

$$J_n(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2}.$$

If θ is a vector-valued parameter with $\theta \in \mathbb{R}^d$, then the information is a $d \times d$ matrix and is defined by

$$J_n(\theta) = -\frac{d^2\ell(\theta)}{d\theta d\theta^{\mathrm{T}}}.$$

$J_n(\theta)$ is a function of θ and it quantifies the amount of data in a given likelihood function. Typically, the higher is the sample size, the higher is the observed information, i.e. if n_1 and n_2 are sizes of two samples and $1 < n_1 < n_2$ then $J_{n_1} \leq J_{n_2}$. For instance, in the above example with $n = 10$ the information is $\frac{3}{(1-\theta)^2} + \frac{7}{x^2}$, which is lower than $\frac{15}{(1-\theta)^2} + \frac{35}{x^2}$, the observed information with $n = 50$.

3.3 Some computational issues

Assuming that $L(\theta)$ is differentiable, to locate its maximum we proceed by finding $\widehat{\theta}$, the solution in θ to the equation

$$\frac{d\ell(\theta)}{d\theta} = 0.$$

This is the *first-order condition*. Then we check that $J_n(\widehat{\theta}) > 0$ or that $J_n(\widehat{\theta})$ is positive definite; this is the *second-order condition*. With this two conditions we make sure that $\widehat{\theta}$ is at least a *local* maximum. In the example of Section 3.2 we were able to compute $\widehat{\theta}$ by this procedure analytically. However, the analytic solution of $\frac{d\ell(\theta)}{d\theta} = 0$ is not always feasible.

For instance, consider the observed sample of $n = 10$ waiting times in minutes at a regional telephone exchange for informations about infection by the SARS-COV-2 virus:

$$5.1, 7.4, 10.9, 21.3, 12.3, 15.4, 25.4, 18.2, 17.4, 22.5.$$

A distribution that is often used for modelling lifetime, duration or survival data is $\text{Wei}(\alpha, \beta)$. With the Weibull model and assuming that the observed sample is i.i.d., we have the log-likelihood function shown in Figure 3.2 by means of contours (see also Example 3.5). Also shown in the figure is the point in the parameter space with maximum likelihood, i.e. $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta}) = (2.8, 17.6)$.

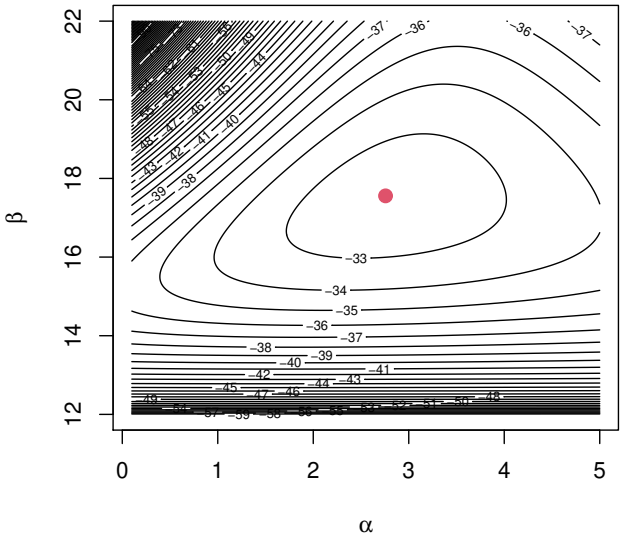


Figure 3.2: Contour plot of the likelihood function under the Weibull model based on an observed sample of size $n = 10$. The red dot indicates the maximum of the likelihood with coordinates $(\widehat{\alpha}, \widehat{\beta}) = (2.8, 17.6)$.

To find the point on the parameter space at which the maximum of the likelihood is achieved, that is $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta})$, we need to solve the first-order condition. The latter leads to the non nonlinear system:

$$\begin{cases} \frac{\partial \ell(\alpha,\beta)}{\partial \alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n} \log y_i - \frac{\sum_{i=1}^{n} y_i^{\alpha} \log y_i}{\beta} = 0, \\ \frac{\partial \ell(\alpha,\beta)}{\partial \beta} = -\frac{n}{\beta} + \frac{\sum_{i=1}^{n} y_i^{\alpha}}{\beta^2} = 0. \end{cases}$$

From the second equation we get

$$\widehat{\beta}(\alpha) = \tfrac{1}{n} \sum_{i=1}^{n} y_i^{\alpha}.$$

Replacing β by $\widehat{\beta}(\alpha)$ in the first equation we obtain

$$g(\alpha) = \sum_{i=1}^{n} \log y_i + \tfrac{n}{\alpha} - \tfrac{n \sum_{i=1}^{n} y_i^{\alpha} \log y_i}{\sum_{i=1}^{n} y_i^{\alpha}} = 0.$$

The equation $g(\alpha) = 0$ cannot be solved explicitly, but it is possible to solve it by an iterative numerical method such as Newton-Raphson method. Specifically, we can define a sequence $\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots$ such that

$$\widehat{\alpha}_m = \widehat{\alpha}_{m-1} - \tfrac{g(\widehat{\alpha}_{m-1})}{g'(\widehat{\alpha}_{m-1})},$$

where $\alpha_0 > 0$ is an initial value, $g'(\alpha)$ is the derivative of $g(\alpha)$ and $\widehat{\alpha}_m \to \widehat{\alpha}$ as $m \to \infty$. Once we have $\widehat{\alpha}$, then we set $\widehat{\beta} = \frac{1}{n} \sum_{i=1}^{n} y_i^{\widehat{\alpha}}$ and so $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta})$.

This approach is used in order to compute the coordinates of the red point in Figure 3.2. Note that, to assure that $\widehat{\theta}$ is a local maximum we also need to check the second-order condition, i.e. the matrix $J_n(\widehat{\theta})$ must be positive definite.

References

[LM18] Larsen, R. J. and Marx, M. L. (2018) *An Introduction to Mathematical Statistics and its Applications* (6th edition), Pearson Education, Inc..

Practice Lecture 3: The likelihood function

Erlis Ruli (ruli@stat.unipd.it)

02 November 2020

1 The Bernoulli model

Consider again the example in Lecture 3, Section 3.2. In which we assume that the observed sample

$$(0, 1, 1, 1, 0, 1, 1, 0, 1, 1)$$

is a realisation of the random sample $X_i \overset{i.i.d.}{\sim} \text{Ber}(\theta)$, where θ is unknown parameter. We show how to define the likelihood function in `R` and how to plot it.

Call `yobs` the vector with the observed sample

```
> yobs <- c(0,1,1,1,0,1,1,0,1,1)
```

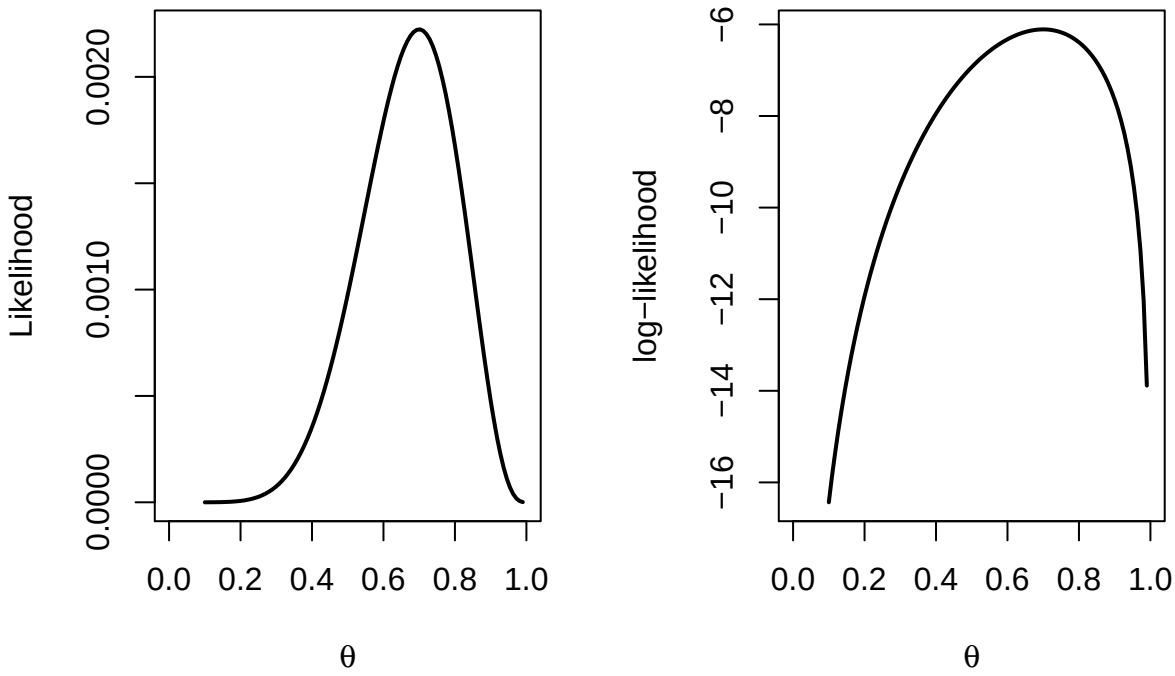
To define a likelihood function in `R` we have two options. The first option is to define it using the mathematical definition. The second option is to resort to the `R`'s implementation of the p.d.f. of the involved r.v.'s. Whenever possible, option 2 is better, since it is less error-prone.

```
> # likelihood function: option 1
> Lik1.ber <- function(theta, y){
+   n = length(y)
+   vi = theta^y  * (1-theta)^(1-y)
+   return(prod(vi))
+ }
>
> # log-likelihood function: option 1
> lLik1.ber <- function(theta, y){
+   n = length(y)
+   vi = y * log(theta)  + (1-y)*log(1-theta)
+   return(sum(vi))
+ }
>
>
> # likelihood function: option 2
> Lik2.ber <- function(theta, y){
+   vi = dbinom(y, size=1, prob = theta)
+   return(prod(vi))
+ }
```

Here is the plot of the likelihood function and of the log-likelihood function

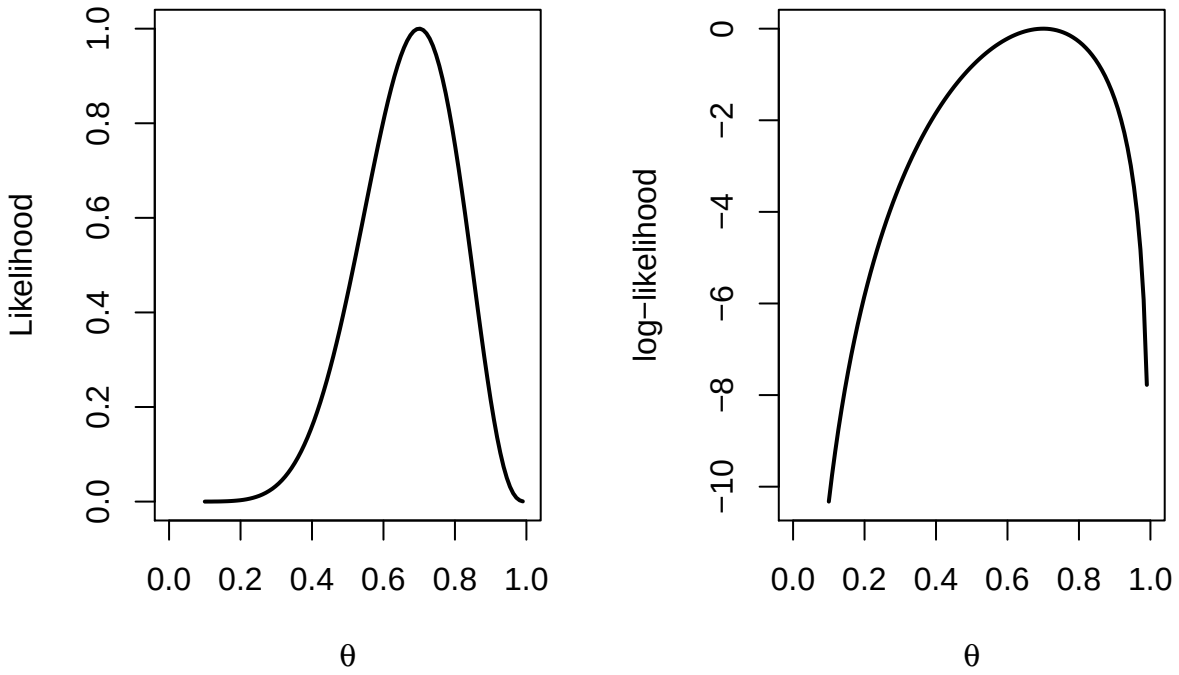
```
> par(mfrow=c(1,2))
> theta.gr <- seq(0.1,0.99, len=100)
> fLik <- sapply(theta.gr, Lik1.ber,  y=yobs)
> flLik <- sapply(theta.gr, lLik1.ber,  y=yobs)
> plot(theta.gr, fLik, xlim=c(0,1),lwd=2,
```

```
+      xlab=expression(theta),
+      type="l",ylab="Likelihood")
> plot(theta.gr, flLik, xlim=c(0,1),lwd=2,
+      xlab=expression(theta),
+      type="l", ylab="log-likelihood")
```



To plot the relative likelihood $RL(\theta)$ (and the relative log-likelihood, i.e. $\log RL(\theta)$) function we must scale the likelihood by its maximum, thus a quick way to do that is

```
> par(mfrow=c(1,2))
> plot(theta.gr, fLik/max(fLik), xlim=c(0,1),lwd=2,
+      xlab=expression(theta),
+      type="l",ylab="Likelihood")
> plot(theta.gr, flLik-max(flLik), xlim=c(0,1),lwd=2,
+      xlab=expression(theta),
+      type="l", ylab="log-likelihood")
```



Notice that while the maximum of the relative likelihood function is 1, the maximum value of the relative log-liekelihood is $\log 1 = 0$.

We saw in Section 3.2 that the maximum of the likelihood function is reached at the value $\hat{\theta} = 0.7$. Let's see now how to compute this maximum numerically. For doing this we use the log-likelihood function since the latter is numerically more stable.

```
> # note: by default optimize performs numerical minimisation,
> # thus we use the option maximum=TRUE
> optimize(f = lLik1.ber, interval = c(0.1, 0.99),
+          y = yobs,
+          maximum = TRUE)
```

```
## $maximum
## [1] 0.7000096
##
## $objective
## [1] -6.108643
```

This numerical result agrees with the analytical solution.

R also has routines for approximating numerical derivatives of the given function. These are provided by an external package that has to be installed and loaded. For instance, if we wished to compute the observed information, i.e. the negative of the second derivative of the log-likelihood, evaluated at $\theta = 0.7$ we could do the following.

```
> # if not already installed
> #install.packages("numDeriv")
>
> # after installation, load the package
> library(numDeriv)
>
> # check first that the first-order cond. is satisfied
> grad(lLik1.ber,0.7, y=yobs)
```

```
## [1] -2.71399e-11
```

```
> # observed information at theta=0.7
> -hessian(lLik1.ber,0.7, y=yobs)
```

```
##           [,1]
## [1,] 47.61905
```

```
> # check that this agrees with the analytic version
> 3/(1-0.7)^2 + 7/0.7^2
```

```
## [1] 47.61905
```

In the code above we also showed how to compute the first derivative of the log-likelihood function, by means of the `grad` function.

2 The Weibull model

Consider the example in Lecture 3, Section 3.3. In which we assume that the observed sample

$$(5.1, 7.4, 10.9, 21.3, 12.3, 15.4, 25.4, 18.2, 17.4, 22.5)$$

is a realisation of the random sample $X_i \overset{i.i.d.}{\sim} \text{Wei}(\alpha, \beta)$, where α, β are the unknown parameters. We show how to define the likelihood function in `R` and how to plot it. For numerical stability reasons it is better to work with a slightly differnt version of the Weibull distribution, which p.d.f. is given by

$$f(y; \alpha, \lambda) = \frac{\alpha}{\lambda^\alpha} \left(\frac{y}{\lambda}\right)^{\alpha-1} e^{-(y/\lambda)^\alpha}.$$

This version of the Weibull distribution correponds to the one given in the Lecture 1 and Lecture 3 with $\beta = \lambda^\alpha$. Furthermore, this version is also the one implemented in `R`.

Set the observed sample in the memory of `R`

```
> yobs=c(5.1, 7.4, 10.9, 21.3, 12.3, 15.4,
+        25.4, 18.2, 17.4, 22.5)
```

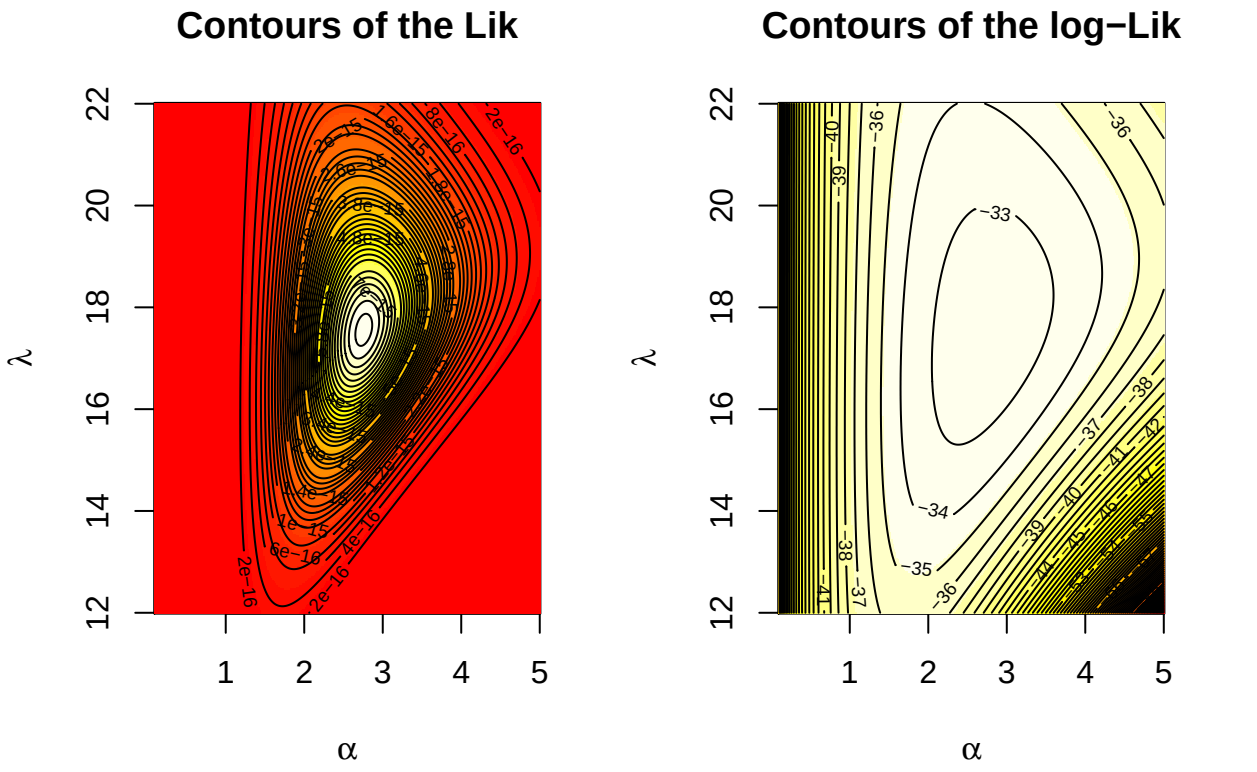
To define the likelihood function, this time we use option 2.

```
> # likelihood function
> Lik.wei <- function(theta, y){
+   # using dwibull(x, alpha, beta)
+   vi = dweibull(y, shape = theta[1], scale = theta[2])
+   return(prod(vi))
+ }
>
> # log-likelihood function
> lLik.wei <- function(theta, y){
+   n = length(y)
+   vi = dweibull(y, shape = theta[1], scale = theta[2], log = TRUE)
+
+   return(sum(vi))
+ }
```

Now we plot the likelihood surface. Actually we consider two versions, one plots the contour levels of the likelihood function and the other plots the contour levels of the log-likelihood function.

```
> # set a regular grid for alpha and beta
> theta1 <- seq(0.1, 5, len=300)
> theta2 <- seq(12,22, len=300)
```

```
>
> # build the bivariate grid
> theta12 <- expand.grid(theta1,theta2)
>
> # evaluate the log-Lik over the grid
> fLik <-apply(theta12,1,Lik.wei,y=yobs)
> flLik <- apply(theta12,1,lLik.wei,y=yobs)
>
> # transform the previous objects in matrices
> Lzmat <- matrix(fLik, ncol=300, byrow = F)
> lLzmat <-matrix(flLik, ncol=300, byrow = F)
>
> par(mfrow = c(1,2))
> image(theta1,theta2, -Lzmat, col = heat.colors(30, rev = TRUE),
+       xlab=expression(alpha),
+         ylab=expression(lambda), main="Contours of the Lik")
> contour(theta1,theta2, Lzmat, nlevels = 30, add = TRUE)
> image(theta1,theta2, -lLzmat,col = heat.colors(30, rev = TRUE),
+       xlab=expression(alpha),
+         ylab=expression(lambda),main="Contours of the log-Lik")
> contour(theta1,theta2, lLzmat, nlevels = 100, add=TRUE)
```



In Lecture 3 we saw that the maximum the likelihood (or log-likelihood function) is reached at the point $\hat{\theta}$ with coordinates $(2.8, 17.6)$. To get this value, we first solved $\partial\ell(\alpha,\lambda)/\partial\lambda = 0$ in λ and found the partial solution for a fixed α, i.e. $\hat{\lambda}(\alpha) = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^{\alpha}\right)^{1/\alpha}$. Then, replacing λ by $\hat{\lambda}(\alpha)$ in $\partial\ell(\alpha,\lambda)/\partial\alpha = 0$ gives the equation

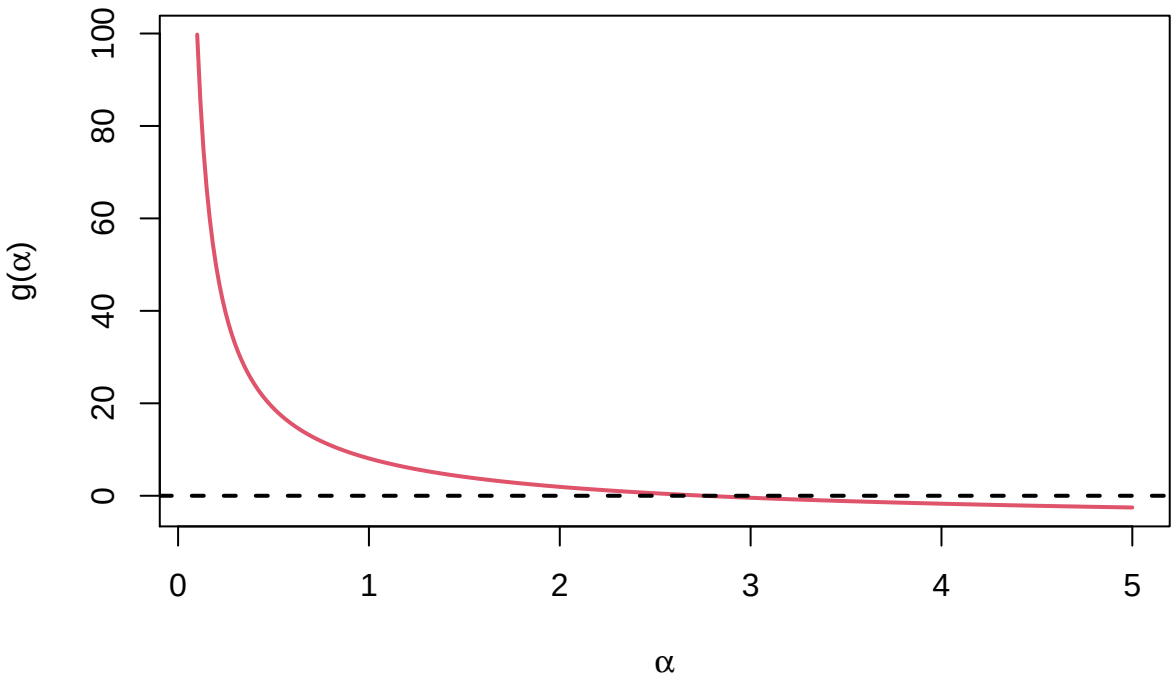
$$g(\alpha) = \sum_{i=1}^{n} \log y_i + \frac{n}{\alpha} - \frac{n\sum_{i=1}^{n} y_i^{\alpha}\log y_i}{\sum_{i=1}^{n} y_i^{\alpha}} = 0$$

Let' define g first.

```
> # define the function g(alpha)
> g.a <- function(a, y){
+   n = length(y)
+   oo = sum(log(y)) + n/a - n*sum(y^a * log(y))/sum(y^a)
+
+   return(oo)
+ }
```

Here is how the function looks like

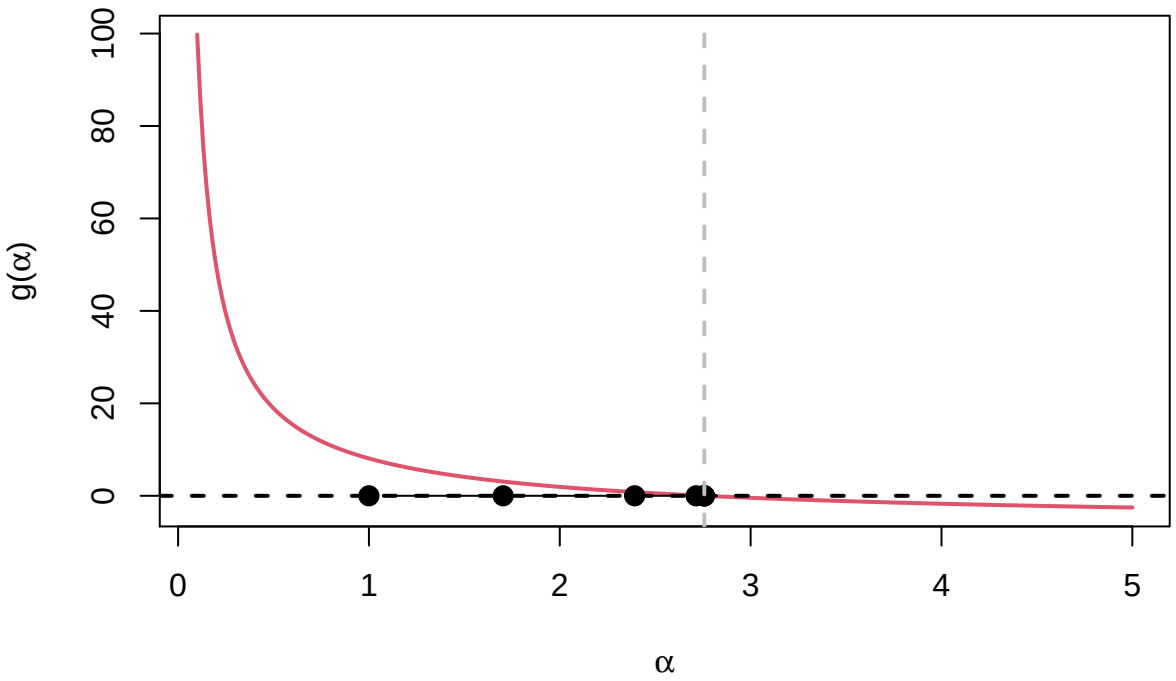
```
> ga <- sapply(theta1, g.a, y=yobs)
> plot(theta1, ga, xlim=c(0.1, 5),
+      xlab=expression(alpha), ylab=expression(paste("g(",alpha,")")),
+      type="l", lwd=2, col=2)
> abline(h=0, lwd=2, lty=2)
```



Since this cannot be solved analytically, let's try to get the first terms of the sequence given by the Newton-Raphson method, in which we approximate $g'(\alpha)$ numerically

```
> theta1. <- theta1
> # set the starting point alpha0
> theta0 <- 1
>
> # this is a numericall approximation of dg(a)/da
> g.a.prime <- function(a,y) {
+   dd <- grad(function(x) g.a(x, y), a)
+ }
>
>
> # compute the first 7 iterations
> theta1 <- theta0 - g.a(theta0, yobs)/g.a.prime(theta0, yobs)
> theta2 <- theta1 - g.a(theta1, yobs)/g.a.prime(theta1, yobs)
```

```
> theta3 <- theta2 - g.a(theta2, yobs)/g.a.prime(theta2, yobs)
> theta4 <- theta3 - g.a(theta3, yobs)/g.a.prime(theta3, yobs)
> theta5 <- theta4 - g.a(theta4, yobs)/g.a.prime(theta4, yobs)
> theta6 <- theta5 - g.a(theta5, yobs)/g.a.prime(theta5, yobs)
> theta7 <- theta6 - g.a(theta6, yobs)/g.a.prime(theta6, yobs)
>
> plot(theta1., ga, xlim=c(0.1, 5),
+      xlab=expression(alpha), ylab=expression(paste("g(",alpha,")")),
+      type="l", lwd=2, col=2)
> abline(h=0, lwd=2, lty=2)
>
> points(y = rep(0,8),
+      x = c(theta0,theta1, theta2,theta3,theta4,theta5,theta6,theta7),
+      xlab="iteration",
+      ylab = "hat.alpha at iteration", pch=20, cex=2,
+      type="b", main="Newton-Raphson iterations")
> abline(v = 2.75717, lty=2, lwd=2, col="grey")
```



From the plot we wee that at iteration 7 the Newton-Raphson algorithm has already reached convergence since the sequence of $\hat{\alpha}_i$ seem to have converged to the the horizontal line at the point 2.75717, thus we conclude that $\hat{\alpha} = 2.75717$. Obviously we need to a criteria for stopping the sequence, there are many possibilities to accomplish this. However, instead of doing this by hand, in order to find the solution of $g(\alpha) = 0$, we will use the command `uniroot` as follows

```
> (oo <- uniroot(g.a, interval = c(0.1, 10), y=yobs))
```

```
## $root
## [1] 2.757155
##
## $f.root
## [1] -9.16928e-07
##
## $iter
```

```
## [1] 7
##
## $init.it
## [1] NA
##
## $estim.prec
## [1] 8.872539e-05
```

Form the output of the command we see that the solution, i.e. the root of $g(\alpha) = 0$, is found to be $\hat{\alpha} = 2.757154$. Plugging in $\hat{\alpha}$ in $\hat{\lambda}(\alpha)$ we get $\hat{\lambda}(\hat{\alpha}) = \hat{\lambda}$, i.e.

```
> hat.alpha <- oo$root
> hat.lambda <- (mean(yobs^hat.alpha))^(1/hat.alpha)
```

Note that $\hat{\beta} = \hat{\lambda}^{\hat{\alpha}}$.

To compute the maximum of the likelihood function we can also use a hill-climbing approach. However, since we have more than one parameter, the numerical optimisation routine `optimize` cannot be used. This time we use `optim` with the option `L-BFGS-B` which is the bound and low-memory version of the Broyden–Fletcher–Goldfarb–Shanno algorithm.

```
> # we let the legthy output of optim be placed
> # in the object oo and not printed on the screen
> # par = c(2,16): is the starting value of the optimizer
> # fn = function(x) -lLik.wei(x,yobs) redefines the log-likelihood function
> #  multiplying it by -1.
> oo <- optim(par = c(2,1),
+             fn = function(x) -lLik.wei(x,yobs),
+             method="L-BFGS-B",
+             lower = c(0.5,0.1),
+             upper = c(4,100),
+             hessian = TRUE) # with this option we get the J_n matrix
> oo
```

```
## $par
## [1]  2.757154 17.556445
##
## $value
## [1] 32.42347
##
## $counts
## function gradient
##       30       30
##
## $convergence
## [1] 0
##
## $message
## [1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"
##
## $hessian
##            [,1]       [,2]
## [1,]  2.1202361 -0.2200252
## [2,] -0.2200252  0.2466318
```

The result of optim agrees with the one provided previously. The object `oo` also reports the Hessian matrix at the maximum. Since we miminised the negative log-likelihood function, this Hessian matrix is also the

observed information matrix evaluated at $\hat{\theta}$, i.e. $J_n(\hat{\theta})$. To check the second-order condition it is sufficient o check that the eigenvalues of $J_n(\hat{\theta})$ are all positive.

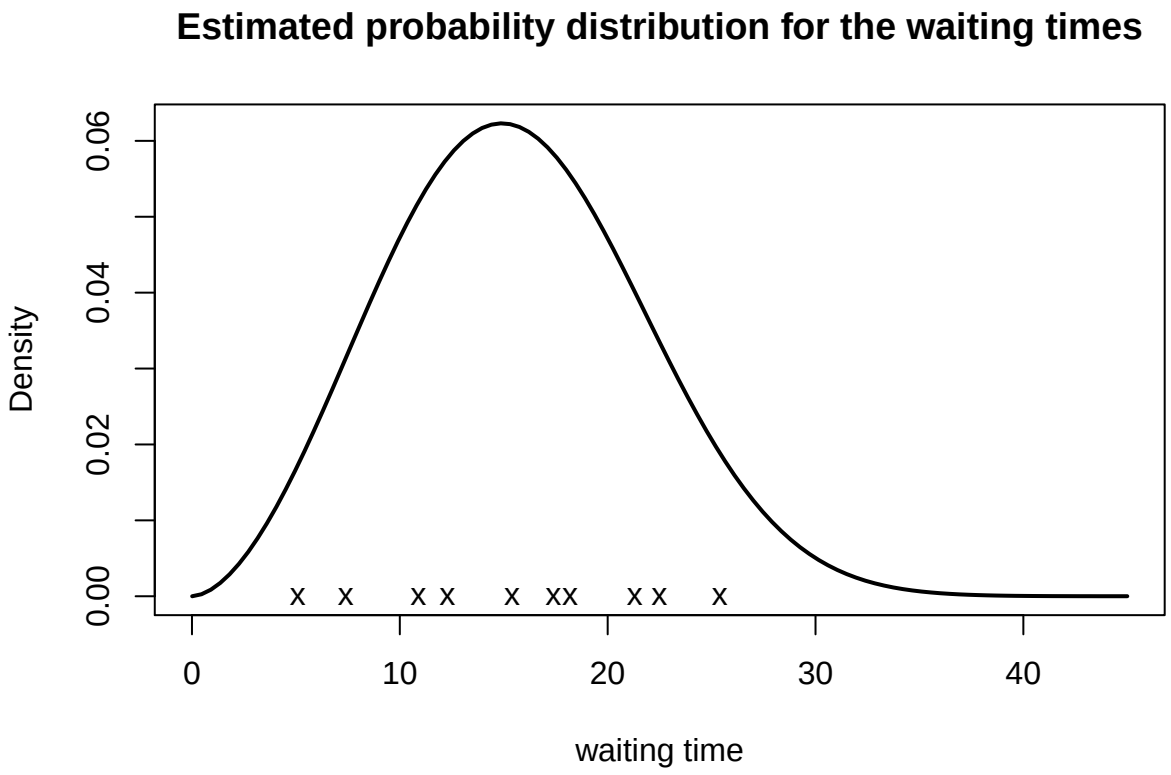
```
> eigen(oo$hessian)
```

```
## eigen() decomposition
## $values
## [1] 2.1457278 0.2211401
##
## $vectors
##            [,1]       [,2]
## [1,] -0.9933553 -0.1150880
## [2,]  0.1150880 -0.9933553
```

Thus we conclude that the value $\hat{\theta}$ found is a local maximum.

We have thus found that the most likely model is the Wei$(\hat{\alpha}, \hat{\lambda})$, where $\hat{\alpha} = 2.76$ and $\hat{\lambda} = 17.56$. Let's plot this most likely model and let's compute some summaries from it.

```
> plot(function(x) dweibull(x, shape = oo$par[1],
+                           scale = oo$par[2]),
+      xlim=c(0.01, 45), lwd=2,
+      ylab="Density", xlab="waiting time",
+      main= "Estimated probability distribution for the waiting times")
> points(x=yobs,y=rep(0,length(yobs)), pch="x")
```



For instance the estimated average weiting time is

```
> oo$par[2]*gamma(1+1/oo$par[1])
```

```
## [1] 15.62419
```

which is very close but not exactly equal to the empirical average

```
> mean(yobs)
```

```
## [1] 15.59
```

The probability of waiting more than 30 minutes is

```
> 1-pweibull(30, shape = oo$par[1], scale = oo$par[2])
```

```
## [1] 0.0125161
```

Whereas the estimated probability of waiting between 5 and 15 minutes is

```
> pweibull(15, shape = oo$par[1], scale = oo$par[2])-
+   pweibull(5, shape = oo$par[1], scale = oo$par[2])
```

```
## [1] 0.4460469
```

Exercise

Assume now that $X_i \overset{iid}{\sim} \text{Ga}(\alpha, \beta)$, $i = 1, \ldots, n$. Repeat all the computation performed in Section 2.

Inferential Statistics **First semester, 2020**

Lecture 4: Point estimation

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

Here is an example of estimation problem inspired by a real-life application.

A manufacturer is going to introduce a new line of washing machines (WM) on the market. Before that, the manufacturer has to quantify the amount of energy consumed by the WM's under the cotton 40°C washing program. A teem of engineers decides to take at random n WM's and measures for each of them the energy consumed under the specified washing program. The collected data are $y_1, \ldots, y_n$, where y_i is the energy consumed by the ith machine, for $i = 1, \ldots, n$. A reasonable model for this problem is to set $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, and the problem then translates in how to *estimate* the mean μ and the variance σ^2. An estimate for these parameters leads to a specific probability model for the population of WM's of the new line and the average amount of energy consumed by the WM's can be set equal to the estimate of μ.

Simplifying things to the extreme, the problem of statistical inference can be stated as follows. Assuming the variable of interest has distribution F_θ, the aim is to learn and verify hypothesis about θ. A random sample $Y_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$ of size n is taken and the implied statistical model is thus

$$\{F_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) : \theta \in \Theta\}.$$

Nature picks $\theta = \theta_0$ and with this value (or vector, depending on the model) generates the observed sample $y_1, \ldots, y_n$. For instance, F_θ could be the exponential distribution and a possible Nature's choice could be $\lambda_0 = 1$. We are given the model and the observed sample, and we are asked to guess θ_0. That is, we are required to produce a guess for θ_0 which is theoretically guaranteed to be close, or approximately equal to θ_0. We have already seen in Lecture 3 how to produce a possible guess (we denoted it by $\widehat{\theta}$) using the likelihood function, but we have no idea if this guess has theoretical guarantees.

More formally, such a guess is called *estimate* and the tool that produces estimates is called *estimator*. Estimation theory deals with methods for building estimators, and is the objective of this Lecture.

4.1 Statistics

Definition 4.1 *Let $Y_1, \ldots, Y_n$ be a random sample with distribution $Y_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$, indexed by θ. Consider a function*

$$T_n = T(Y_1, \ldots, Y_n),$$

with $T_n : \mathbb{R}^n \to \mathbb{R}^d$ not depending on other unknown quantities. Then T_n is called a sample statistic or simply a statistic.

A statistic is a function applied to the random sample $Y_1, \ldots, Y_n$, thus (by Lecture 1) it is an r.v. if $d = 1$, and it is an r.ve. if $d > 1$. The probability distribution of a statistic is called *sampling distribution*.

The statistic T_n evaluated at the observed sample is denoted by

$$t_n = T(y_1, \ldots, y_n),$$

and is called *observed statistic*. An observed statistic is thus a fixed number if $d = 1$ or a vector of numbers if $d > 1$.

Examples of statistics are the sample mean $\overline{Y}$, the sample variance S_Y^2, the sample median Me_Y, etc., the empirical distribution function F_n is also a statistic. Examples of observed statistics are the observed sample average $\overline{y}$, the observed sample variance s_y^2; the observed empirical distribution function $\widehat{F}_n$ is also an observed statistics as is the histogram. Statistics can also be obtained by combining other statistics, for instance $T_n = (\overline{Y}, S_Y^2)$ is a bivariate statistic.

The probability distribution of the a statistic T_n depends on F_θ but also on the function $T(\cdot)$ and in some cases it can be determined analytically. Here are some notable results about some statistics and their distributions.

Theorem 4.1 *Let* $Y_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$ *with* $E(Y_i) = \mu < \infty$ *and* $\text{var}(Y_i) = \sigma^2 < \infty$. *Then*

(i) $E(\overline{Y}) = \mu$ *and* $\text{var}(\overline{Y}) = \frac{\sigma^2}{n}$.

(ii) $E(S_Y^2) = \sigma^2$ *and* $\text{var}(S_Y^2) = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}$, *where* $\mu_k = E(Y_1^k)$, *is the kth moment of* Y_1 *(see L0, p.9).*

Theorem 4.1(i) thus tells us that whatever is the distribution of the random sample, assuming the latter has finite mean μ and finite variance σ^2, the sample mean has expected value equal to μ and variance equal to σ^2 divided by n. A similar result is given also for the sample variance.

To see why (i) holds note that

$$\begin{aligned} E(\overline{Y}) &= E\left(\tfrac{\sum_{i=1}^n Y_i}{n}\right) \\ &= \tfrac{1}{n} E(Y_1 + \ldots + Y_n) \\ &= \tfrac{1}{n}(E(Y_1) + \ldots + E(Y_n)) \\ &= \tfrac{1}{n}(\mu + \ldots + \mu) = \mu. \end{aligned}$$

Furthermore

$$\begin{aligned} \text{var}(\overline{Y}) &= E\left[(\overline{Y} - E(\overline{Y}))^2\right] \\ &= E\left[\left(\tfrac{\sum_{i=1}^n Y_i}{n} - \mu\right)^2\right] \\ &= \tfrac{1}{n^2} E\left[\left(\sum_{i=1}^n Y_i - n\mu\right)^2\right] \\ &= \tfrac{1}{n^2} E\left[(Y_1 - \mu)^2 + \cdots + (Y_n - \mu)^2 + 2(Y_1 - \mu)(Y_2 - \mu) + 2(Y_1 - \mu)(Y_3 - \mu) + \cdots + 2(Y_{n-1} - \mu)(Y_n - \mu)\right] \\ &= \tfrac{1}{n^2}\{E[(Y_1 - \mu)^2) + \cdots + E[(Y_n - \mu)^2]\} \\ &= \tfrac{1}{n^2} n\sigma^2 = \tfrac{\sigma^2}{n}. \end{aligned}$$

where we have used the fact that $\text{cov}(Y_i, Y_j) = 0$ since Y_i and Y_j are independent (by Theorem 0.9(i), Lecture 0). Part (ii), and in particular the second result is more tedious to derive.

Remark 4.1

(i) In general $\overline{Y}$ and S_Y^2 are not stochastically independent. However, as we will see in the next theorem, if the random sample is taken from the normal distribution then these two statistics are independent.

(ii) Since $\overline{Y}$ is a sum of r.v. with finite mean and variance, then by the Central Limit Theorem (CLT) we have that

$$\frac{\sqrt{n}(\overline{Y}-\mu)}{\sigma} \xrightarrow{d} \text{N}(0,1), \quad \text{as } n \to \infty.$$

The next theorem is similar except that the random sample is assumed to be normally distributed.

Theorem 4.2 *Let $Y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $i = 1, \ldots, n$ then*

(i) $\overline{Y} \sim N(\mu, \sigma^2/n)$.

(ii) $\frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2_{n-1}$.

(iii) $\overline{Y}$ *is independent from* S_Y^2.

(iv) $\dfrac{\frac{\sqrt{n}(\overline{Y}-\mu)/\sigma}{\sqrt{\frac{(n-1)S_Y^2}{\sigma^2}}}}{\frac{n-1}{n-1}} = \frac{\sqrt{n}(\overline{Y}-\mu)}{\sqrt{S_Y^2}} \sim t_{n-1}$.

(v) $\text{var}(S_Y^2) = 2\frac{\sigma^4}{n-1}$.

Theorem 4.2(i) tells us that if the random sample has a normal distribution, then also the sample mean has exactly a normal distribution. Furthermore, the point (ii) tells us that the the "scaled" sample variance has a chi-square distribution with $n-1$ degrees of freedom and point (iii) tells us that, contrary to the general case, under the normal distribution, $\overline{Y}$ is independent from S_Y^2. Point (iv) says that the statistic $\frac{\sqrt{n}(\overline{Y}-\mu)}{\sqrt{S_Y^2}}$, follows the t-Student distribution with $n-1$ degrees of freedom. These results are the building blocks of many confidence intervals and hypothesis testing procedures as we will se in the incoming lectures. There are other notable results about statistics and their distributions, but these will be discussed in the incoming lectures.

4.2 Properties of estimators

An *estimator* is a sample statistic which aim is to provide estimates of an unknown parameter; thus if T_n is a statistics, then it is an estimator. On the other hand, the statistic evaluated at the observed sample, i.e $t_n = T(y_1, \ldots, y_n)$, is a number, thus it is an *estimate*. To simplify notation and keep the two things separately, we denote the estimator by $\widehat{\theta}_n = T(Y_1, \ldots, Y_n)$, instead of T_n. We adopt the – usual and perhaps confusing – convention in which $\widehat{\theta}_n$ denotes also the estimate. Thus, $\widehat{\theta}_n$ refers simultaneously to two different objects: estimator and estimate. This is only apparently confusing since in a given context, it will be clear to which object the notation is referring to. Sometimes we omit n and write simply $\widehat{\theta}$.

While all estimators are statistics, not all estimators are useful for a problem at hand. Furthermore, for a given problem there may be many estimators available and we need to have criteria for choosing the best among them. So we begin by presenting some criteria and then take up some ways of deriving estimators that are apt to be good.

4.2.1 Sufficiency

Roughly speaking, the random sample $Y_1, \ldots, Y_n$ which is assumed to be generated from the model F_θ, contains useful information about the unknown parameter θ but some information is redundant. A statistic, and thus an estimator, transforms the random sample in a random vector of lower dimension and such a transformation may lead to a loss of information about θ. Such a loss of information however will not happen if the the statistic is *sufficient*.

The statistic T_n is sufficient for θ, the parameter of the distribution F_θ, if and only if the conditional distribution of the random sample $Y_1, \ldots, Y_n$ given the value of T_n does not depend on θ, i.e. if the p.d.f.

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n | T_n) \text{ does not depend on } \theta.$$

A useful criterion for checking that a given statistic T_n is sufficient is based on the likelihood function. Indeed, T_n is sufficient if and only if the likelihood function is of the form

$$L(\theta) \propto g(T(y_1, \ldots, y_n); \theta),$$

where $g(\cdot)$ is any positive real-valued function. Here are some examples.

Example 4.1 *Let $Y_1, \ldots, Y_n$ be a random sample with $Y_i \overset{\text{iid}}{\sim} Geo(\theta)$, where $Geo(\theta)$ is the geometric r.v. (see Lecture 1). The joint distribution of the random sample is then*

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} \theta(1-\theta)^{y_i} = \theta^n (1-\theta)^{\sum_{i=1}^n y_i}.$$

For a fixed observed sample, this is the likelihood function and it depends on the y_i's only through the value of their sum. So $\sum_{i=1}^n y_i$ is a sufficient statistic.

The likelihood function criterion is not a good tool for showing that a particular statistic is not sufficient, since it is not easy to see that a likelihood cannot be expressed as a function of θ and only the statistic. In this respect the following theorem may be useful.

Theorem 4.3 *If a sufficient statistic T is a function of the statistic U, then U is sufficient. If T is a function of U and U is not sufficient, neither is T.*

4.2.2 Unbiasedness

Let $\widehat{\theta}$ be an estimator for θ, based on the random sample $Y_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$. The bias of an estimator is defined by

$$\text{bias}(\widehat{\theta}; \theta) = E(\widehat{\theta}) - \theta.$$

We say that $\widehat{\theta}$ is *unbiased* if $E(\widehat{\theta}) = \theta$ or equivalently if $\text{bias}(\widehat{\theta};\theta) = 0$. Unbiasedness used to receive much attention but these days is considered less important; many interesting estimators used in practice are biased. An estimator with vanishing bias as the sample size increases, i.e. with the property $\text{bias}(\widehat{\theta}_n;\theta) \to 0$ as $n \to \infty$, is called *asymptotically unbiased*.

4.2.3 Efficiency

In principle, the lower the bias, the better is the estimator. However, unbiasedness alone is not enough for judging the performance of an estimator. Indeed it is possible to build unbiased estimators which are useless; see below for an example. Another reason is that unbiasedness tells us nothing about the variability of $\widehat{\theta}$ with respect to the true parameter value θ. For this reason, the *mean squared error* (MSE) is a more sound criterion for judging the performance of the estimator. The MSE is defined by

$$\text{MSE}(\widehat{\theta};\theta) = E(\widehat{\theta} - \theta)^2 = \text{var}(\widehat{\theta}) + [\text{bias}(\widehat{\theta};\theta)]^2.$$

Thus if the estimator $\widehat{\theta}$ is unbiased, $\text{MSE}(\widehat{\theta};\theta)$ is just the variance of the estimator.

If $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are two competing estimators for θ and $\text{MSE}(\widehat{\theta}_1;\theta) < \text{MSE}(\widehat{\theta}_2;\theta)$ then $\widehat{\theta}_1$ is said to be *relatively more efficient* than $\widehat{\theta}_2$. The *relative efficiency* of $\widehat{\theta}_1$ with respect to $\widehat{\theta}_2$ is defined by the ratio

$$\text{eff}(\widehat{\theta}_1, \widehat{\theta}_2;\theta) = \frac{\text{MSE}(\widehat{\theta}_2;\theta)}{\text{MSE}(\widehat{\theta}_1;\theta)}.$$

Thus, if $\text{eff}(\widehat{\theta}_1, \widehat{\theta}_2;\theta) > 1$ ($\text{eff}(\widehat{\theta}_1, \widehat{\theta}_2;\theta) < 1$), θ_1 is relatively more (less) efficient than θ_2 and instead of choosing estimators with lowest or zero bias, we prefer to select estimators with lowest MSE or with highest relative efficiency. Here is an example.

Example 4.2 *By Theorem 4.1(i) we know that $E(\overline{Y}) = \mu$, thus the sample mean is an unbiased estimator of the true mean. Let us denote by $\theta = (\mu, \sigma^2)$. It follows that $\text{MSE}(\overline{Y};\theta) = \text{var}(\overline{Y}) = \sigma^2/n$. Consider the alternative statistic $T_n = Y_1$ the first observation in the sample is also unbiased, because $E(Y_1) = \mu$. However, its mean squared error is $\text{var}(Y_1) = \sigma^2$, which is larger than that of $\overline{Y}$ when $n > 1$.*

Remark 4.2 *Just like the bias also the MSE, and thus also the efficiency, may depend on θ. Indeed, an estimator that has lowest MSE for a certain unknown parameter value, may not have lowest MSE for all possible parameter values. For instance, Figure 4.1 shows the MSE as a function of θ for two hypothetical estimators. From this figure we can conclude that for $\theta < 0$, Estimator 1 is better, i.e. it is more efficient, than Estimator 2; for $\theta > 0$ Estimator 2 is better than Estimator 1, and for $\theta = 0$ the two estimators are equally good, at least in terms of efficiency.*

Given a class of estimators we can always choose the one with highest efficiency, but there may be estimators outside the given class which could be more efficient. The question thus is:

is there a lower bound for the variance of an estimator ? That is, is there a *most efficient* estimator among all estimators?

In this generality the answer is negative; but in the rather wide (but still restricted) class of unbiased estimators, there is a lower bound of variances – a "best" variance – which can sometimes be achieved. This is stated formally in the next theorem.

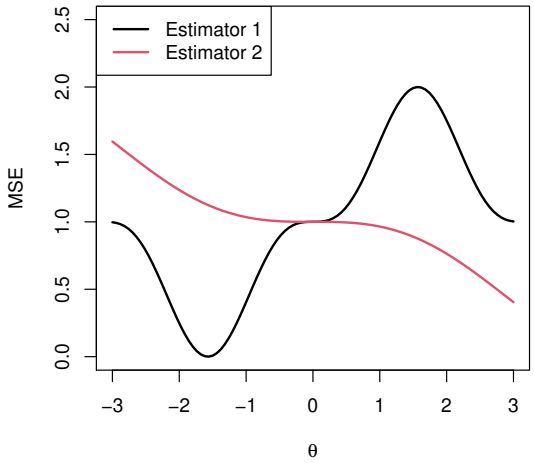


Figure 4.1: MSE as a function of θ for two hypothetical estimators.

Theorem 4.4 (Cramér-Rao's inequality) *Consider the random sample* $Y_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$ *and let and* $f(y;\theta)$ *be the p.d.f. of* Y_i*'s. Furthermore, let* $\widehat{\theta}_n = T(Y_1, \ldots, Y_n)$ *be an unbiased estimator for* θ*. Then*

$$\text{var}(\widehat{\theta}_n) \geq \frac{1}{I_n(\theta)},$$

where $I_n(\theta) = nE\left[\left(\frac{d \log f(Y_1;\theta)}{d\theta}\right)^2\right]$.

If there exists an estimator $\widehat{\theta}_n$ such that $\text{var}(\widehat{\theta}_n) = \frac{1}{I_n(\theta)}$, then $\widehat{\theta}_n$ is called *efficient*. For such an unbiased estimator we define the efficiency by

$$\text{eff}(\widehat{\theta}_n;\theta) = \frac{1}{\text{var}(\widehat{\theta}_n) I_n(\theta)}.$$

When the efficiency of $\widehat{\theta}_n$, defined for each n tends to a finite limit as $n \to \infty$, this limit is called *asymptotic efficiency*:

$$\lim_{n\to\infty} \text{eff}(\widehat{\theta}_n;\theta) = \lim_{n\to\infty} \frac{1}{\text{var}(\widehat{\theta}_n) I_n(\theta)}.$$

4.2.4 Consistency

The essence of the notion of consistency of an estimator is that when the estimator is applied to the whole population as the sample, it produces the true value of the parameter being estimated. More formally, if $\widehat{\theta}_n$ is an estimator for θ defined for every sample size n, then $\widehat{\theta}_n$ is *consistent* if for any $\epsilon > 0$

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| > \epsilon) = 0,$$

or more compactly if $\widehat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$. Here is an example.

Example 4.3 *Let* $Y_i \overset{\text{iid}}{\sim} \text{N}(\mu, 1)$, $i = 1, \ldots, n$. *Let* $\widehat{\mu}_n = \overline{Y}$ *be an estimator for* μ. *Note that in this case it is known that* $\sigma^2 = 1$. *We know that* $\overline{Y} \sim \text{N}(\mu, 1/n)$, *so for any* $\epsilon > 0$,

$$\begin{aligned} \lim_{n\to\infty} P(|\widehat{\mu}_n - \mu| > \epsilon) &= \lim_{n\to\infty} \{1 - [\Phi(\sqrt{n}\epsilon) - \Phi(-\sqrt{n}\epsilon)]\} \\ &= 1 - 1 + 0 = 0. \end{aligned}$$

Thus $\widehat{\mu}_n = \overline{Y}$ *is consistent. Here we used the fact that* $\lim_{n\to\infty} \Phi(\sqrt{n}\epsilon) = 1$ *since* $\Phi(\cdot)$ *is a d.f. (see Theorem 0.5(iii), Lecture 0).*

The following result connects bias and efficiency with consistency and provides an easier way for checking the consistency of an estimator.

Theorem 4.5 *For an estimator* $\widehat{\theta}_n$, *if* $\lim_{n\to\infty} \text{bias}(\widehat{\theta}_n; \theta) = 0$ *and* $\lim_{n\to\infty} \text{var}(\widehat{\theta}_n) = 0$, *then* $\widehat{\theta}_n$ *is consistent.*

Before closing this section we give some further examples.

Example 4.4 *Let* $Y_1, \ldots, Y_n$ *be an i.i.d random sample with* $E(Y_i) = \mu$ *and* $\text{var}(Y_i) = \sigma^2$, $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}$, $i = 1, \ldots, n$. *We wish to estimate* μ, *while* σ^2 *is known.*

Since $E(\overline{Y}) = \mu$ *and* $\text{var}(\overline{Y}) = \sigma^2/n$, *the estimator* $\widehat{\mu}_n = \overline{Y}$ *for* μ *is consistent, for any value of* σ^2.

By the CLT we know that for large n, $\overline{Y} \xrightarrow{d} \text{N}(\mu, \sigma^2/n)$. *Thus for large* n *we can use* $\text{N}(\mu, \sigma^2/n)$ *as an approximation to the sampling distribution of* $\overline{Y}$. *Nevertheless, If* $Y_i \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, *then* $\overline{Y} \sim \text{N}(\mu, \sigma^2/n)$ *and* $\overline{Y}$ *is efficient for* μ *(check!).*

Example 4.5 *Let* $Y_1, \ldots, Y_n$ *be an i.i.d random sample with* $E(Y_i) = \mu$ *and* $\text{var}(Y_i) = \sigma^2$, $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}$, $i = 1, \ldots, n$. *We wish to estimate* σ^2, *while* μ *is known.*

Since we know the mean of $Y_i's$ *is* μ, *then a reasonable estimator for* σ^2 *is* $\widehat{\sigma^2}_\mu = \frac{\sum_{i=1}^n (Y_i - \mu)^2}{n}$. *It can be proved that* $E\left(\widehat{\sigma^2}_\mu\right) = \sigma^2$ *and that* $\widehat{\sigma^2}_\mu$ *is consistent for* σ^2.

If $Y_i \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, *then* $\frac{n\widehat{\sigma^2}_\mu}{\sigma^2} \sim \chi^2_n$.

Example 4.6 *Let* $Y_1, \ldots, Y_n$ *be an i.i.d random sample with* $E(Y_i) = \mu$ *and* $\text{var}(Y_i) = \sigma^2$, $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}$, $i = 1, \ldots, n$. *We wish to estimate* σ^2 *but this time* μ *is unknown. In this case,* μ *is also called nuisance parameter, while* σ^2 *is the parameter of interest.*

Since μ *is unknown, we cannot use* $\widehat{\sigma^2}_\mu$ *as an estimator for* σ^2. *However, we can use the sample variance* S^2. *We have already proved that* $E(S^2) = \sigma^2$ *(see Theorem 4.1) so* S^2 *is an unbiased estimator for* σ^2. *Furthermore, from Theorem 4.1 (ii) we have that* $\lim_{n\to\infty} \text{var}(S^2) = \frac{2\sigma^4}{n-1}$, *so* S^2 *is consistent.*

Recall from Theorem 4.2, that if $Y_i \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, *then* $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Example 4.7 *Let* $Y_i \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, $i = 1, \ldots, n$ *where the success probability* θ *is unknown. Note that* $\theta = E(Y_i)$, *thus for estimating* θ *we can use the same estimator as in Example 4.4. Let* $\widehat{\theta} = \overline{Y}$. *In this particular case we have that*

$$\overline{Y} = \tfrac{1}{n} \sum_{i=1}^n Y_i = \tfrac{1}{n} \sum_{i=1}^n 1_{Y_i = 1}\,,$$

where $1_{Y_i=1}$ is the indicator function which takes 1 if $Y_i = 1$ and 0 otherwise. The estimator $\widehat{\theta}$ can also be interpreted as the proportion of Y_i's that are equal to 1, i.e. the sample proportion of successes or simply the sample proportion. By the same results of Example 4.4 we can conclude that $\widehat{\theta}$ is consistent. Furthermore, in this case $\text{var}(\widehat{\theta}) = \frac{1}{I_n(\theta)}$, so $\widehat{\theta}$ is efficient.

The sampling distribution of $\widehat{\theta}$ can be determined exactly, since $n\widehat{\theta} = \sum_{i=1}^n Y_i$ thus

$$n\widehat{\theta} \sim \text{Bin}(n, \theta).$$

Otherwise, if n is high enough, we can appeal to the CLT to get the approximate sampling distribution

$$\widehat{\theta} \dot{\sim} \text{N}\left(\theta, \tfrac{\theta(1-\theta)}{n}\right).$$

4.3 Methods for building estimators

So far we studied the performance of estimators that have suggested themselves on intuitive grounds. If intuition does not suggest an estimator (and even when it does), it helps to have methods for "deriving" estimators. In the following we illustrate three of them.

4.3.1 Method of moments

To estimate a single function of parameters the method of moments is to use that same function of the corresponding sample moments. In particular, the mean of a distribution is estimated as the sample mean and a population variance is estimated as $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2$. To estimate k parameters, the method of moments is first to express those parameters in terms of the first k population moments and then to replace those population moments with the corresponding sample moments.

Here are some examples

Example 4.8 *The variance of an r.v. Y, assuming it exists, is $\sigma^2 = E(Y^2) - [E(Y)]^2$. Letting $Y_1, \ldots, Y_n$ be a random sample, the previous function applied to sample moments is $\overline{Y^2} - (\overline{Y})^2$, which is equal to $(n-1)S^2/n = \widehat{\sigma^2}$. Thus $\widehat{\sigma^2}$ is the method of moments estimator for σ^2.*

Example 4.9 *Given the random sample $Y_i \overset{\text{iid}}{\sim} \text{Ga}(\alpha, \lambda)$, $i = 1, \ldots, n$ let's estimate the parameters α and λ. For this, recall that for $Y \sim \text{Ga}(\alpha, \lambda)$, the expectation is $E(Y) = \alpha/\lambda$, and the variance is $\text{var}(Y) = \alpha/\lambda^2 = E(Y^2) - [E(Y)]^2$. To get the method of moments estimators we replace $E(Y)$ by $\overline{Y}$ and $E(Y^2)$ by $\overline{Y^2} = \sum_{i=1}^n Y_i^2$ to get*

$$\widehat{\lambda}_{\text{MM}} = \tfrac{n\overline{Y}}{(n-1)S^2}, \quad \widehat{\alpha}_{\text{MM}} = \tfrac{n(\overline{Y})^2}{(n-1)S^2}.$$

The method of moments gives estimators that are consistent and asymptotically normal (thanks to the CLT) but typically not efficient. However, the usefulness of this method stems from the fact that often the estimators are easy to calculate and can thus be used as starting points for better estimators such as those obtained by the method of maximum likelihood.

4.3.2 Method of least squares

In some situations the random sample $Y_1, \ldots, Y_n$ may be expressed as a "signal plus noise" relation, such as

$$Y_i = g_i(\theta) + \epsilon_i, \quad i = 1, \ldots, n,$$

where the signal $g_i(\theta)$ is a known deterministic function up the the unknown parameter θ and ϵ_i is the noise, which is a random component. For instance, Y_i could be the measurement of some physical quantity which it is believed to be equal to $g_i(\theta)$ plus some noise due to accidental errors or because of pure random fluctuations of the phenomenon under study; this noise is represented by the r.v. ϵ_i. It is reasonable to assume that for a large number of measurements, the negative fluctuations compensates the positive ones, thus $E(\epsilon_i) = 0$. Furthermore, it is reasonable to assume that the random fluctuations do have some variance, e.g. $\mathrm{var}(\epsilon_i) = \sigma^2$ and the fluctuations are stochastically independent, all other things held equal, i.e. $\mathrm{cov}(e_i, e_j) = 0$, $i, j = 1, \ldots, n$, $i \neq j$.

The method of *least squares* (LS) consists in estimating θ through the estimator

$$\widehat{\theta}_{\mathrm{LS}} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (Y_i - g_i(\theta))^2.$$

Here it is an example.

Example 4.10 *Suppose $Y_1, \ldots, Y_n$ are bacterial counts measured at time points $t_1, \ldots, t_n$ in a culture of cells and we aim at studying their growth rate in time. A possible model for this problem is the following linear multiple regression model*

$$\begin{aligned} Y_i &= g(t_i; \theta) + \epsilon_i, \\ g(t_i; \theta) &= \theta_1 + \theta_2 t_i + \theta_3 t_i^2, \quad i = 1, \ldots, n, \end{aligned}$$

where $\theta \in \mathbb{R}^3$. The solution in θ to the linear system

$$\frac{d}{d\theta} \sum_{i=1}^{n} (Y_i - g_i(\theta))^2 = 0,$$

is the LS estimator for θ. If we let $Y = (Y_1, \ldots, Y_n)$ and $X = [1_n \,|\, T \,|\, T^2]$ be an $n \times 3$ matrix, where $1_n = (1, \ldots, 1)$ is the all-ones vector of dimension n, $T = (t_1, \ldots, t_n)$ and $T^2 = (t_1^2, \ldots, t_n^2)$, then the LS estimator can be defined compactly by

$$\widehat{\theta}_{\mathrm{LS}} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Y.$$

Under suitable conditions the LS estimator is unbiased, asymptotically efficient and asymptotically normally distributed. However, it use is limited only to those problems which can be stated as a signal plus noise relation.

4.3.3 Method of maximum likelihood

In Lecture 3, we defined $\widehat{\theta}_n \in \Theta$ as the point at which the maximum of the likelihood function is achieved. More formally, the point

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} L(\theta),$$

is called a *maximum likelihood estimate* (MLE) of θ. Note that different observed samples may lead to a different MLE's $\widehat{\theta}_n$, thus for a random sample $Y_1, \ldots, Y_n$, $\widehat{\theta}_n$ is also an r.v. or an r.ve., depending on the dimension of θ. Furthermore, $\widehat{\theta}_n$ does not depend on θ, i.e. it is a statistic, so it is an estimator. In the random sample case, $\widehat{\theta}_n$ is called a *maximum likelihood estimator* (MLE). Use of the symbol $\widehat{\theta}_n$ for both observed sample and random sample cases is unfortunate but widespread, so we adhere to this use convention.

As it was pointed out in Lecture 3, it is often easier to work with the natural logarithm of the likelihood function $L(\theta)$, i.e. the log-likelihood function $\ell(\theta)$. We illustrate the method of maximum likelihood by means of examples.

Example 4.11 *Consider Example 3.2 in Lecture 3. The log-likelihood function is*

$$\ell(\lambda) = -n\lambda + \sum_{i=1}^{n} y_i \log(\lambda) + \text{const.}.$$

From the first-order condition $d\ell(\lambda)/d\lambda = 0$ we get solution $\widehat{\lambda} = \overline{y}$. Furthermore, we have that $d^2\ell(\lambda)/d\lambda^2 < 0$ for all $\lambda > 0$ thus $\widehat{\lambda}$ is actually a global maximum thus $\widehat{\lambda} = \overline{y}$ is the MLE of λ.

Example 4.12 *Consider Example 3.4 in Lecture 3. The log-likelihood function is*

$$\ell(\mu, \sigma^2) = -\tfrac{n}{2}\log(\sigma^2) + \tfrac{n\mu^2}{2\sigma^2} - \tfrac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \overline{y}\tfrac{n\mu}{\sigma^2} + \text{const.}$$

By solving the first order conditions we find $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma^2})$, where $\widehat{\mu} = \overline{y}$ and $\widehat{\sigma^2} = \sum_{i=1}^{n}(y_i - \overline{y})^2/n$. Furthermore, after some algebra, it can be shown that the observed information matrix $J(\mu, \sigma^2)$ at $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma}^2)$ is positive definite, thus $\widehat{\theta}$ is at least a local maximum.

Actually, in this case, $\widehat{\theta}$ is a global maximum. In general multi-parameter problems, showing that $\widehat{\theta}$ is a global maximum could be a formidable task, if possible at all, but in this case it is immediate. First note that for any $\sigma^2 > 0$,

$$\tfrac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_{i=1}^{n}(y_i-\overline{y})^2/\sigma^2} \geq \tfrac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_{i=1}^{n}(y_i-\mu)^2/\sigma^2}, \quad \textit{for any } \mu \in \mathbb{R}.$$

Thus to check that $\widehat{\theta}$ is a global maximum it suffices to check that $\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_{i=1}^{n}(y_i-\overline{y})^2/\sigma^2}$ achieves its global maximum at $\sigma^2 = n^{-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$. This can be done by straightforward univariate calculus.

The estimators $\widehat{\mu} = \overline{Y}$ and $\widehat{\sigma^2} = n^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ are the MLE's of μ and σ^2, respectively; alternatively we can say that $\widehat{\theta} = \left(\overline{Y}, n^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right)$ is the MLE of $\theta = (\mu, \sigma^2)$.

Example 4.13 *The method of maximum likelihood is often applied to problems involving multinomial probabilities. In particular, suppose that $Y_1, \ldots, Y_m$, the counts of the m possibilities or cells $b_1, b_2, \ldots, b_m$ follow a multinomial distribution with total count $n = \sum_{i=1}^{m} Y_i$ and cell probabilities $\theta_1, \ldots, \theta_m$. Our aim is to estimate the parameter vector $\theta = (\theta_1, \ldots, \theta_m)$ from the observed cell counts $y_1, \ldots, y_m$. The joint p.d.f is*

$$f_{Y_1,\ldots,Y_m}(y_1,\ldots,y_m;\theta_1,\ldots,\theta_m) = \tfrac{n!}{\prod_{i=1}^m y_i!}\prod_{j=1}^n \theta_j^{y_j}, \quad \theta_j > 0, \quad \sum_{j=1}^m \theta_j = 1,$$

from which, at the fixed observed sample, we get the log-likelihood function

$$\ell(\theta_1,\ldots,\theta_m) = \log n! - \sum_{i=1}^m \log y_i! + \sum_{i=1}^m y_i \log\theta_i.$$

To maximise this function subject to the constraint $\sum_{j=1}^m \theta_j = 1$*, we introduce a Lagrange multiplier and maximise the extended log-likelihood function*

$$\Lambda(\theta_1,\ldots,\theta_m;\lambda) = \ell(\theta_1,\ldots,\theta_m) + \lambda\left(\sum_{i=1}^n \theta_i - 1\right).$$

Setting the partial derivatives equal to zero, we have the following system of equations:

$$\theta_j = -\tfrac{y_j}{\lambda}, \quad j = 1,\ldots,m.$$

Summing both sides of this equation we get $\sum_i \theta_i = 1 = -\frac{n}{\lambda}$*, thus* $\lambda = -n$*. Therefore,* $\widehat{\theta}_i = y_i/n$*,* $i = 1,\ldots,n$*. The log-likelihood function is strictly concave thus* $\widehat{\theta} = (y_1/n,\ldots,y_m/n)$ *is a global maximum. So* $\widehat{\theta}$ *is the MLE of* θ*.*

In some situations, such as those frequently occurring in genetics, the multinomial cell probabilities $\theta_1,\ldots,\theta_m$ *are functions of other unknown parameters* τ*; that is* $\theta_i = \theta_i(\tau)$*. In such cases the log-likelihood of* τ *is*

$$\ell(\tau) = \log n! - \sum_{i=1}^m \log y_i! + \sum_{i=1}^m y_i \log\theta_i(\tau).$$

Here is a concrete example.

If gene frequencies are in equilibrium, the genotypes AA*,* Aa *and* aa *occur in a population with frequencies* $(1-\tau)^2$*,* $2\tau(1-\tau)$ *and* τ^2*, according to the Hardy-Winberg law. In a sample of Chinese population of Hong Kong in 1937, blood types occur with the following frequencies, where* M *and* N *are erythrocyte antigens*

	Blood Type			
	M	*MN*	*N*	Total
Frequency	342	500	187	1029

If we denote by y_1, y_2, y_3 *the counts in the three cels and let* $n = 1029$*, then the log-likelihood of* τ *is*

$$\begin{aligned}\ell(\tau) &= \log n! - \sum_{i=1}^3 \log y_i! + y_1\log(1-\tau)^2 + y_2\log[2\tau(1-\tau)] + y_3\log\tau^2 \\ &= \log n! - \sum_{i=1}^3 \log y_i! + (2y_1+y_2)\log(1-\tau) + (2y_3+y_2)\log\tau + y_2\log 2.\end{aligned}$$

This time there is no need to incorporate the constraint that the cell probabilities sum to 1, since the functional form of $\theta_i(\tau)$ is such that $\sum_i \theta(\tau) = 1$. Setting the derivative equal to zero, we have

$$-\tfrac{2y_1+y_2}{1-\tau} + \tfrac{2y_3+y_2}{\tau} = 0.$$

Solving this we obtain the MLE

$$\widehat{\tau} = \tfrac{2y_3+y_2}{2n}.$$

The method of maximum likelihood so widely used that is considered the gold standard method of estimation in parametric statistical models. One of the reason of this widespread usage is that, under certain rather broad conditions, the MLE $\widehat{\theta}_n$ possesses many of the above properties of estimators, as n diverges. We study more closely the properties of the MLE in the next section

4.4 Properties of the maximum likelihood estimator

So far we were mostly concerned about two features of the distribution of an estimator, location and scale. Sometimes however these two features are not enough and it is useful to know the whole d.f. of an estimator, i.e. its sampling distribution. Unfortunately, the exact sampling distribution of many practical estimators is difficult or even impossible to calculate analytically. Nevertheless, there are two viable alternatives to this:

- simulation, e.g. the method of *bootstrap*;
- asymptotic approximations, e.g. Central Limit Theorem, etc.

We will touch upon the bootstrap near the end of this course. As far as asymptotic approximations are concerned, we saw in some of the examples of Section 4.2 that sometimes the sampling distribution of $\widehat{\theta}_n$ can be determined exactly, for each n and in other cases it can be approximated by the CLT for large n.

The sampling distributions of estimators can also be approximated via simulation or via asymptotic approximations. In the following we concentrate on the MLE and study some of its most relevant properties and we spend some time explaining what these properties mean and why they are good things.

Theorem 4.6 *Under suitable regularity condition on F_θ the following results hold.*

(i) If T_n is a sufficient statistic for θ, then the MLE of θ is a function of the data through T_n.

(ii) The MLE is equivariant, i.e. if $\widehat{\theta}_n$ is the MLE of θ then $g(\widehat{\theta}_n)$ is the MLE of $g(\theta)$.

(iii) The MLE is consistent, i.e. $\widehat{\theta}_n \xrightarrow{P} \theta_0$, where θ_0 is the true parameter value.

(iv) The MLE is asymptotically efficient: roughly, this means that among all well-behaved estimators, the MLE has the smallest variance, at least for large n.

(v) The MLE is asymptotically normal: $(\widehat{\theta}_n - \theta_0)/\mathrm{se}(\widehat{\theta}_n) \xrightarrow{d} N(0,1)$ as $n \to \infty$, where $\mathrm{se}(\widehat{\theta}) = \sqrt{\mathrm{var}(\widehat{\theta}_n)}$ and can be often computed or approximated analytically.

Property (i) tells us that the MLE is a sufficient statistic, whenever there exists one. To see why this is the case, it T_n is sufficient, by the likelihood factorisation criterion we have

$$L(\theta) = q(\theta; t_n) q(y_1, \ldots, y_n)$$

Since the maximum of $L(\theta)$ is also the maximum of $q(\theta; t_n)$ then, it is clear that $\widehat{\theta}$ will depend on the observed sample $y_1, \ldots, y_n$ through $t_n = T(y_1, \ldots, y_n)$ so, the MLE is sufficient.

To see why property (*ii*) is true, for simplicity we restrict ourselves to a function $g : \Theta \to \mathcal{T}$ being bijective. Let $h = g^{-1} : \mathcal{T} \to \Theta$ denote the inverse of g. Then $\theta = h(\tau)$ and $\tau = g(\theta)$, so we have that

$$L(\theta) = L(h(\tau)),$$

This means that the likelihood seen as function of θ is identical to the likelihood seen as function of τ. But then $L(\widehat{\theta}) = L(h(\widehat{\tau}))$, so $\widehat{\theta} = h(\widehat{\tau})$ and thus $\widehat{\tau} = g(\widehat{\theta})$. Here is an example on this point.

Example 4.14 *Let $Y_i \overset{\text{iid}}{\sim} \text{Poi}(\theta)$, $i = 1, \ldots, n$ be a random sample and consider the observed sample $(y_1, \ldots, y_n)$. Our aim is to estimate e^θ from this observed sample. Given $L(\theta)$ the likelihood function of θ, the MLE of θ is found to be $\widehat{\theta} = \overline{y}$. Our aim is to estimate $\tau = e^\theta$ and by the equivariance principle thus we readily have that the MLE of τ is $\widehat{\tau} = e^{\widehat{\theta}} = e^{\overline{y}}$.*

Alternatively, we can ignore the aforementioned principle and try to get the MLE of τ by maximising the log-likelihood function with respect to τ. For this we have first to reparametrise *the likelihood function in terms of τ. To reparametrise in terms of τ note first that τ is related to θ through the function $g(\theta)$ given by exponential function. Thus the inverse of g is $\theta = \log \tau$. Thus $L(\theta) = L(\log \tau)$. Taking the logarithm in on the right hand side of the last equality we get*

$$\ell(\log \tau) = -n \log \tau + \sum_{i=1}^{n} y_i \log \log \tau + \textit{const.}.$$

Maximising $\ell(\log \tau)$ with respect to τ we get the maximum $\widehat{\tau} = e^{\overline{y}}$. Thus the MLE of τ is again $\widehat{\tau} = e^{\overline{y}}$.

To see why property (*iii*) is true, recall from Lecture 3 that $L(\theta; y_1, \ldots, y_n)$ was defined for a fixed observed sample $y_1, \ldots, y_n$. If we have another observed sample, say $y_1^*, \ldots, y_n^*$, from the same statistical model, then we would have the likelihood function $L(\theta; y_1^*, \ldots, y_n^*)$. If the two observed samples are different, or they lead to different sufficient statistics if they are available, then the two likelihood functions may be different. This implies that also $\widehat{\theta}(y_1, \ldots, y_n)$ the maximiser of $L(\theta; y_1, \ldots, y_n)$ may be different from $\widehat{\theta}(y_1^*, \ldots, y_n^*)$ the maximiser of $L(\theta; y_1^*, \ldots, y_n^*)$. Thus $\widehat{\theta}$ is an r.v. (or r.ve.) whereas $L(\theta; Y_1, \ldots, Y_n)$ is a random function[1].

Therefore, consider maximising the average random likelihood function

$$\tfrac{1}{n}\ell(\theta; Y_1, \ldots, Y_n) = \tfrac{1}{n} \sum_{i=1}^{n} f(Y_i; \theta).$$

As n tends to infinity, the law of large numbers implies that

$$\begin{aligned} \tfrac{1}{n}\ell(\theta; Y_1, \ldots, Y_n) \xrightarrow{P} E\left(\tfrac{1}{n} \sum_{i=1}^{n} \log f(Y_i; \theta) \right) &= E(\log f(Y_1; \theta)) \\ &= \int \log f(t; \theta) f(t; \theta_0)\, dt. \end{aligned}$$

[1] Again remember that in practical applications we deal only with one observed sample and thus we only have a single likelihood function from which we compute an MLE (here E=estimate!) of θ.

Here we used θ_0 for the true value under which the sample is generated, to distinguish it from θ, the argument of the likelihood function. It is thus plausible that for large n, the θ that maximises $\ell(\theta; Y_1, \ldots, Y_n)$ should be close to the θ that maximises[2] $E(\log f(Y_1; \theta))$. (An involved argument is necessary to establish this.) To maximise $E(\log f(Y_1; \theta))$, we consider its derivative

$$\frac{\partial}{\partial\theta} \int \log f(t;\theta) f(t;\theta_0)\, dt = \int \frac{\frac{\partial}{\partial\theta} f(t;\theta)}{f(t;\theta)} f(t;\theta_0) dt.$$

If $\theta = \theta_0$ this equation becomes

$$\int \frac{\partial}{\partial\theta} f(t;\theta_0) dt = \frac{\partial}{\partial\theta} \int f(t;\theta_0) dt = \frac{\partial}{\partial\theta}(1) = 0,$$

which shows that θ_0 is a stationary point and hopefully a maximum. Note that we have interchanged differentiation with integration and that the assumption of smoothness of f must be strong enough to justify this.

Lastly, property *(iv)* is related to *(v)*. To see why these properties are true we need some further definitions.

Definition 4.2 *For $Y_i \overset{\text{iid}}{\sim} F_\theta$ a random sample of size n, with p.d.f. $f(y;\theta)$, the score function is defined to be*

$$\begin{aligned} s(\theta; Y_1, \ldots, Y_n) &= \frac{d \log L(\theta; Y_1, \ldots, Y_n)}{d\theta} \\ &= \frac{d\ell(\theta; Y_1, \ldots, Y_n)}{d\theta} \\ &= \sum_{i=1}^{n} \frac{d \log f(Y_i;\theta)}{d\theta} \\ &= \sum_{i=1}^{n} s(\theta; Y_i). \end{aligned}$$

The Fisher information, also called expected information, is defined to be

$$\begin{aligned} I_n(\theta) &= \text{var}\left(s(\theta; Y_1, \ldots, Y_n)\right) \\ &= \text{var}\left(\sum_{i=1}^{n} s(\theta; Y_i)\right) \\ &= \sum_{i=1}^{n} \text{var}(s(\theta; Y_i)). \end{aligned}$$

For $n = 1$ we also write $I(\theta)$ instead of $I_1(\theta)$. Note that $I_n(\theta)$ here is exactly the same as that given earlier in Theorem 4.4, with the only exception being that now we give to $I_n(\theta)$ a proper name and show how it is related to the score function. Indeed, it can be shown that $E(s(Y_i;\theta)) = 0$ for any i. It then follows that $\text{var}(s(\theta; Y_1)) = E\left(s(\theta; Y_1)^2\right)$ (recall Footnote 2 above about the convention!). In fact we have the following result.

[2]Note that $E(\log f(Y_i;\theta))$ is the same for all i so by convention we use $i = 1$ and thus $E(\log f(Y_1;\theta))$.

Theorem 4.7 *For $Y_i \overset{\text{iid}}{\sim} F_\theta$ a random sample of size n, with p.d.f. $f(y;\theta)$*

$$I_n(\theta) = nI(\theta).$$

Furthermore,

$$\begin{aligned} I(\theta) &= -E\left(\tfrac{d^2 \log f(Y_1;\theta)}{d\theta^2}\right) \\ &= -\int \left(\tfrac{d^2 \log f(y;\theta)}{d\theta^2}\right) f(y;\theta)dy. \end{aligned}$$

Here is then what we wanted to see.

Theorem 4.8 (Asymptotic Normality of the MLE) *Under appropriate regularity conditions, the following hold:*

1. $\frac{\widehat{\theta}_n - \theta_0}{\sqrt{1/I_n(\theta_0)}} \xrightarrow{d} \mathrm{N}(0,1)$.
2. $\frac{\widehat{\theta}_n - \theta_0}{\sqrt{1/I_n(\widehat{\theta})}} \xrightarrow{d} \mathrm{N}(0,1)$.

The first statement says that $\widehat{\theta}_n$ is asymptotically distributed as $\mathrm{N}(\theta_0, I_n(\theta_0)^{-1})$. Thus, $\mathrm{var}(\widehat{\theta}_n)$ the variance of the MLE is asymptotically equal to $1/I_n(\theta_0)$, so the MLE is *asymptotically efficient*. The second statement says that the MLE still has the same normal distribution even though the unknown θ_0 is replaced by the MLE. It can be shown that the the MLE has a normal distribution even if we replace $I_n(\theta)$ by $J_n(\theta)$, the observed information (see Lecture 3).

A similar result holds for $\theta \in \mathbb{R}^d$, but this is outside the scope of this course. We will see further asymptotic results about the MLE and the likelihood function in the incoming lectures. For the time being, here is an example about the asymptotic distribution of the MLE.

Example 4.15 *Consider again Example 4.11. To determine the distribution of $\widehat{\lambda}_n = \overline{Y}$ note that $s(\lambda; Y_1) = -1 + Y_i/\lambda$, thus $\frac{ds(\lambda;Y_i)}{d\lambda} = \frac{d^2 \log f(\lambda;Y_i)}{d\lambda^2} = -Y_i/\lambda^2$. It follows that*

$$J_n(\lambda) = -\sum_{i=1}^n \tfrac{d^2 \log f(\lambda;Y_i)}{d\lambda^2} = \sum_{i=1}^n Y_i/\lambda^2$$

$$\begin{aligned} I_n(\lambda) &= nI(\lambda) \\ &= nE\left(s(\lambda;Y_1)^2\right) \\ &= -nE\left((Y_1/\lambda - 1)^2\right) = n/\lambda. \end{aligned}$$

Thus by Theorem 4.8 and letting λ_0 be the true parameter value, we have that

$$\widehat{\lambda}_n \xrightarrow{d} \mathrm{N}(\lambda_0, \lambda_0/n), \quad \text{and} \quad \widehat{\lambda}_n \xrightarrow{d} \mathrm{N}(\lambda_0, \overline{Y}/n),$$

as $n \to \infty$. Clearly, in this example the distribution of the MLE can be determined exactly. Indeed,

$$n\widehat{\lambda}_n = \sum_{i=1}^n Y_i \sim \mathrm{Poi}(n\lambda).$$

Remark 4.3

(i) In the above example we see that $E(J_n(\lambda)) = I_n(\lambda)$*, thus since we assume* $I_n(\lambda) < \infty$*, by the WLLN we have that* $J_n(\lambda) \xrightarrow{P} I_n(\lambda)$*; this result is not limited to the present example but holds more generally. Indeed, in all regular models for which the MLE is consistent, the Fisher information is equal to the expectation of the observed information.*

(ii) The result $\widehat{\lambda}_n \xrightarrow{d} \mathrm{N}(\lambda_0, \lambda_0/n)$*, where* $\widehat{\lambda}_n = \overline{Y}$ *can be obtained also by appealing to the CLT (do you see why?)*

(iii) In many practical applications the Fisher information is difficult to compute, if possible at all. Furthermore, θ_0 *is unknown. However, thanks to Theorem 4.8, using* $J_n(\widehat{\theta})$ *in place of* $I_n(\theta_0)$ *we still have the sample limiting distribution for the MLE. Thus, a practical advise is: if the Fisher information is computable, then use the standard error of the MLE given by* $\widehat{\mathrm{se}}(\widehat{\theta}) = \sqrt{1/I_n(\widehat{\theta})}$*; if not, use the standard error* $\widehat{\mathrm{se}}(\widehat{\theta}) = \sqrt{1/J_n(\widehat{\theta})}$*.*

The normality of the MLE is not limited to the scalar parameter case. Indeed, we have the following result.

Theorem 4.9 *Under appropriate regularity conditions on the model* F_θ*,* $\theta \subseteq \mathbb{R}^p$*, then:*

(i) $\widehat{\theta}_n \;\dot{\sim}\; \mathrm{N}_p\left(\theta, I_n(\widehat{\theta}_n)^{-1}\right)$*;*

(ii) $\widehat{\theta}_n \;\dot{\sim}\; \mathrm{N}_p\left(\theta, J_n(\widehat{\theta}_n)^{-1}\right)$*.*

The symbol "$\dot{\sim}$" means that the r.ve. on the left has a distribution which approaches the distribution on the right as $n \to \infty$. This result is extremely useful in practice since, it also says that each component of $\widehat{\theta}_n$ is also approximately normally distributed. In particular, if we let θ_i be ith component of the vector θ, $\widehat{\theta}_{n,i}$ the ith component of $\widehat{\theta}_n$ and if we let $I_n(\theta)^{ii}$ denote the cell (i,i) of the matrix $I_n(\theta)^{-1}$, then we have that

$$\widehat{\theta}_{n,i} \dot{\sim} \mathrm{N}\left(\theta_i, I_n(\theta)^{ii}\right).$$

A similar result holds with I_n replaced by J_n.

References

[LM18] Larsen, R. J. and Marx, M. L. (2018) *An Introduction to Mathematical Statistics and its Applications* (6th edition), Pearson Education, Inc..

Practice Lecture 4: Point estimation

Erlis Ruli (ruli@stat.unipd.it)

13 November 2020

1 Minimum least squares

We consier real data on bacterial counts taken from Ram et al. (2019)[1].

Let's first load the data in R. To do this we use the `read.table` command.

```
> bacteria <- read.table("bacteria.csv", sep=";", dec=",", header = T)
>
> # take a quick look of what's inside bacteria
> str(bacteria)
```

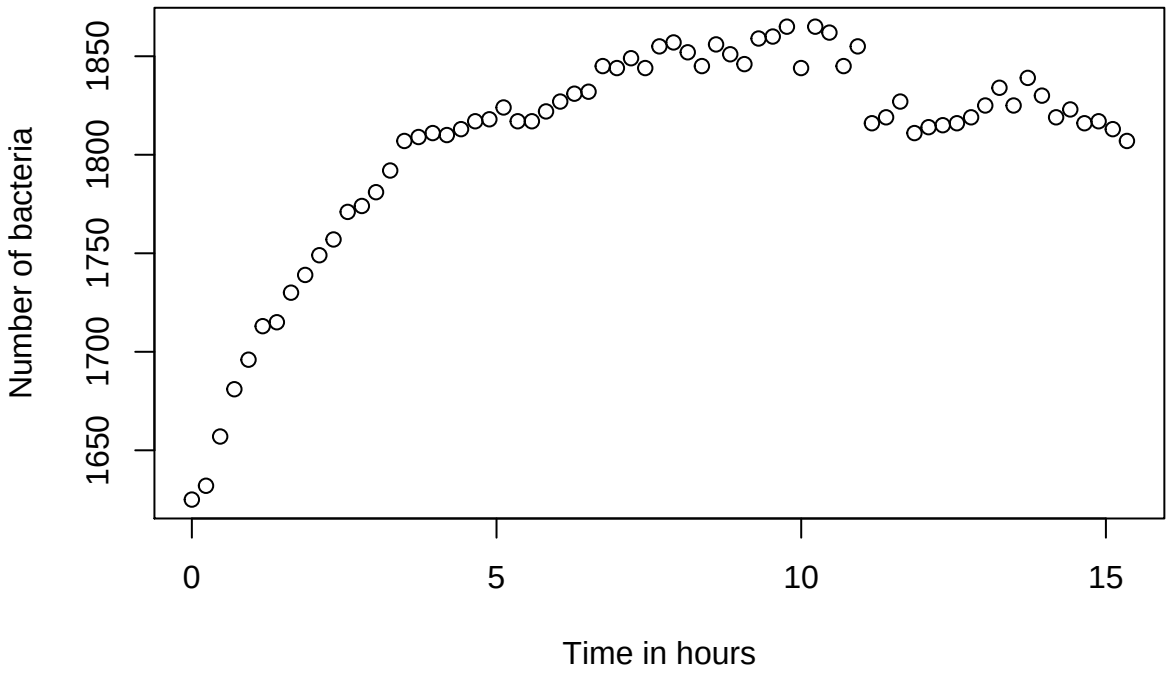
```
## 'data.frame':    67 obs. of  13 variables:
##  $ seconds: num  0 837 1674 2511 3348 ...
##  $ A1     : int  31906 30503 31362 30389 30886 31096 30445 30856 30571 30676 ...
##  $ A2     : int  33127 31315 31620 31923 32121 32158 32220 32315 32411 32357 ...
##  $ A3     : int  31601 31453 31833 32373 32430 32564 32575 32664 32589 32716 ...
##  $ A4     : int  32617 32689 33103 33269 33565 33431 33749 33639 33670 33760 ...
##  $ A5     : int  8329 8317 8395 8486 8405 8454 8458 8440 8528 8535 ...
##  $ A6     : int  8905 8877 8891 8975 9004 9006 9022 9037 9029 9014 ...
##  $ A7     : int  8603 8641 8646 8686 8781 8818 8786 8819 8829 8803 ...
##  $ A8     : int  9377 9285 9390 9476 9475 9515 9536 9475 9507 9462 ...
##  $ A9     : int  1620 1634 1658 1672 1701 1713 1713 1729 1733 1745 ...
##  $ A10    : int  1709 1711 1746 1764 1785 1798 1804 1822 1813 1836 ...
##  $ A11    : int  1679 1672 1696 1724 1744 1757 1769 1771 1785 1795 ...
##  $ A12    : int  1625 1632 1657 1681 1696 1713 1715 1730 1739 1749 ...
```

We see that we have 67 observations and 13 variables, i.e. columns. The first variable, `seconds`, is the time in seconds at which the measurements are taken. The other variables are the bacteria counts under 12 different conditions. We focus here on `A12`.

Before that, let's transform the time in hours.

```
> bacteria$hours <- bacteria$seconds/60^2
> with(bacteria,
+      plot(y=A12, hours, ylab="Number of bacteria",xlab = "Time in hours"))
```

[1] Yoav Ram, Eynat Dellus-Gur,Maayan Bibi, Kedar Karkare, Uri Obolski, Marcus W. Feldman,Tim F. Cooper, Judith Berman, and Lilach Hadany (2019). "Predicting microbial growth in a mixed culturefrom growth curve data", *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 14698–14707.



From the plot we see that the observed number of bacteria $y_1, \ldots, y_n$ is approximately quadratically related with time $t_1, \ldots, t_n$. Thus a model as the one of Example 4.10 seems reasonable. Let's fit this model by LS to the data.

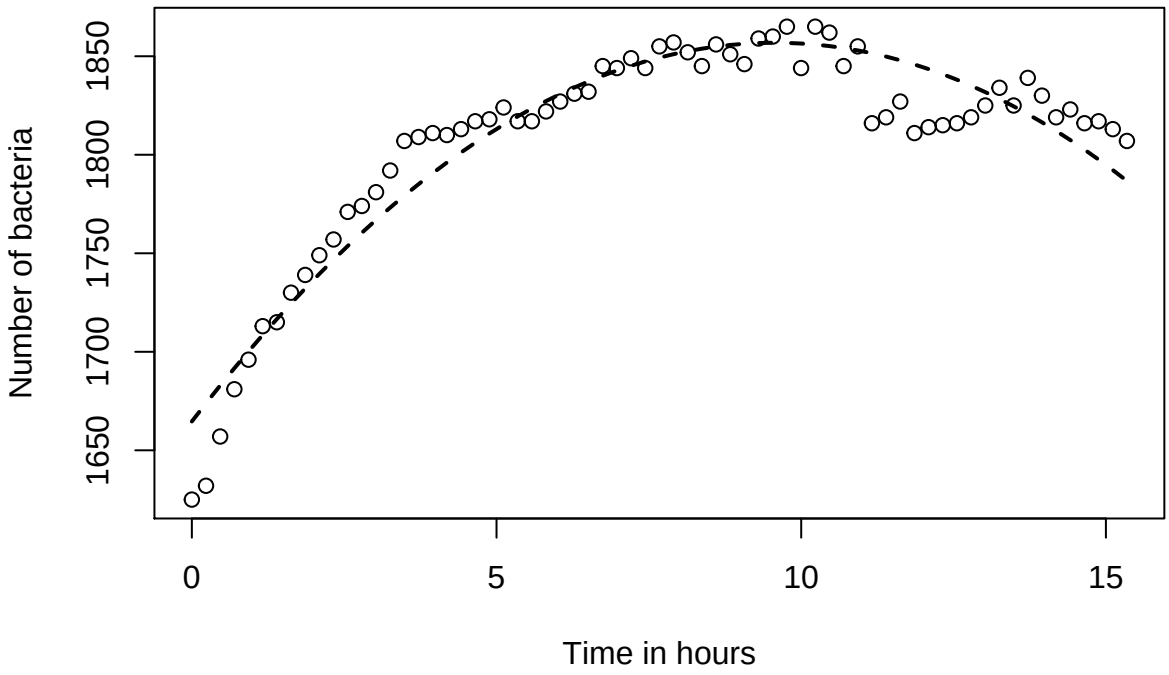
First we use the formula provided in Example 4.10.

```
> Y <- as.matrix(bacteria$A12)
> X <- as.matrix(cbind(1, bacteria$hours, bacteria$hours^2))
> (hat.theta.LS <- solve(t(X) %*% X) %*% t(X) %*% Y)
```

```
##             [,1]
## [1,] 1664.655325
## [2,]   40.181716
## [3,]   -2.101778
```

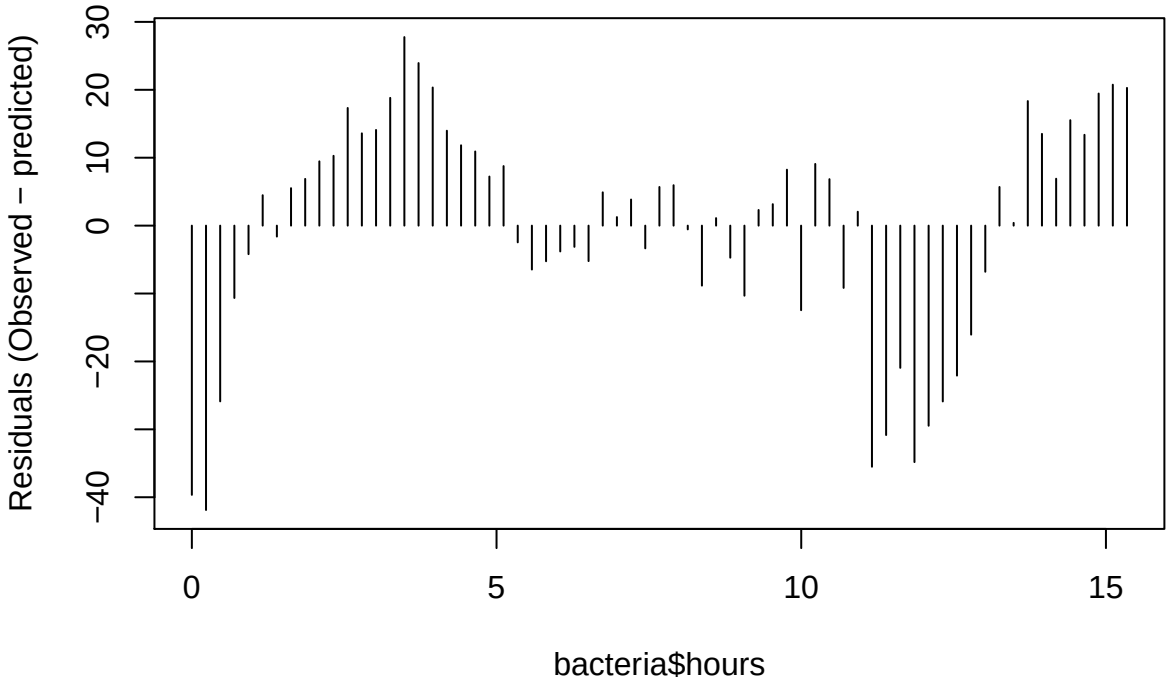
Thus the estimated model is $y_i = 1664.65 + 40.18t_i - 2.10t_i^2$, $i = 1, \ldots, 67$. We can also plot this function over the data

```
> with(bacteria,{
+   plot(y=A12, hours, ylab="Number of bacteria",xlab = "Time in hours");
+   points(hours, 1664.65 + 40.18*hours -2.10* hours^2, type="l", lwd=2,
+          lty=2)
+ })
```



Recall that by the property of LS, the quadratic function shown here is the one which minimises the squared vertical distances of y_i from the curve. These distances are

```
> res <- bacteria$A12 - (1664.65 + 40.18*bacteria$hours -2.10* bacteria$hours^2)
> plot(bacteria$hours,res,type="h", ylab="Residuals (Observed - predicted)")
```



Obviously, R has a much better way of fitting regression a model such as this: the command `lm`. Here it is how it can be done.

```
> ba.lm <- lm(A12~hours + I(hours^2), data=bacteria)
> summary(ba.lm)
```

```
##
## Call:
## lm(formula = A12 ~ hours + I(hours^2), data = bacteria)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.884  -6.476   3.298  10.613  27.772
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1664.6553     5.8238   285.8   <2e-16 ***
## hours         40.1817     1.7546    22.9   <2e-16 ***
## I(hours^2)    -2.1018     0.1106   -19.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.37 on 64 degrees of freedom
## Multiple R-squared:  0.9139, Adjusted R-squared:  0.9112
## F-statistic: 339.5 on 2 and 64 DF,  p-value: < 2.2e-16
```

The fromula "A12~hours + I(hours^2)" tells R that we want a model with A12 as response and `hours` and $\texttt{hours}^2$ as predictors. To make sure that the later is to be understood as the square of hours and not as the name of some other variable we use the function `I()`. This formula also includes the intercept θ_1 by default. The `lm` command thus fits the model, i.e. estimates the parameters θ. The output is saved in the object `ba.lm` and printed to the console by using the command `summary`, which gives an easy-to-read output.

The LS estimates of the three parameters are exactly those we obtained manually. In addition, `lm` also provides the standard errors of the estimates, which can be used for building a confidence interval (Lecture 5) or hypothesis testing (Lecture 6).

Note that, if we assume that the observed sample $y_1, \ldots, y_n$ is obtained from the random sample $Y_i \overset{\text{iid}}{\sim} \text{N}(g_i(\theta), \sigma^2)$, with $g_i(\theta) = \theta_0 + \theta_1 t_i + \theta_2 t_i^2$ and t_i fixed, then $\hat{\theta}_{\text{LS}}$ coincides with the MLE of θ.

2 Maximum likelihood for the Poisson model

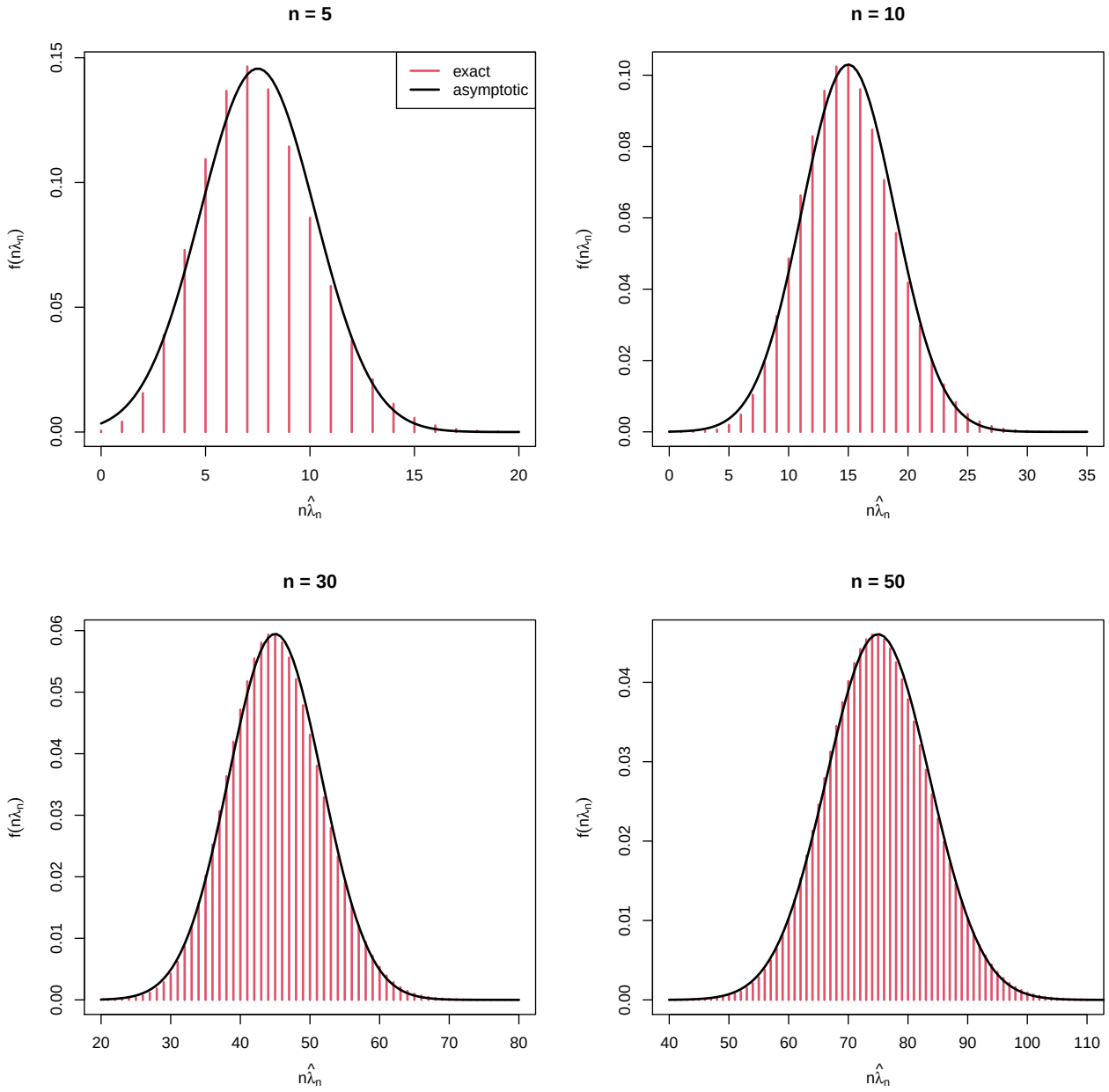
In Example 4.14 of L4 we derived two possible distributions for the MLE. Let's compare these two distributions and see what happens as n increases.

Let the true parameter value, be $\lambda_0 = 3/2$ and we consider samples of size $n \in \{5, 10, 30, 50\}$. By L4 we know that the distribution of the MLE $\hat{\theta}_n$ is approximately $\text{N}(3/2, 3/2n)$. Furthermore, $n\hat{\theta}_n$ follows a $\text{Poi}(3n/2)$. Note that $\hat{\theta}_n \sim \text{N}(3/2, 3/2n)$ implies $n\hat{\theta} \sim \text{N}(3n/2, 3n/2)$ by a property of the normal distribution.

The following figure compares the exact Poisson distribution with the asymptotic approximation (normal). We notice that, already for $n = 10$ the normal approximation is quite decent and as n increases the normal approximation and the exact distribution get closer and closer.

```
> par(mfrow=c(2,2))
>
> # sample size n=5
> n <- 5
> # exact distribution of n*hat.theta
> plot(x= 0:20,y = sapply(0:20, function(x) dpois(x,lambda=3*n/2)),
+      lwd=2,col=2, type="h",xlim=c(0,20), ylab=expression(f(n*hat(lambda)[n])),
+      xlab = expression(n*hat(lambda)[n]), main=paste("n =",n))
> # approximate dist. of n*hat.theta
> plot(function(x) dnorm(x, mean=n*3/2, sd=sqrt(3*n/2)),
```

```
+      lwd=2, add=TRUE, xlim = c(0,20))
> legend("topright", legend = c("exact", "asymptotic"), lwd=c(2,2), col=c(2,1),
+        lty=c(1,1))
>
> n <- 10
> plot(x= 0:35,y = sapply(0:35, function(x) dpois(x,lambda=3*n/2)),
+      lwd=2,col=2, type="h",xlim=c(0,35), ylab=expression(f(n*hat(lambda)[n])),
+      xlab = expression(n*hat(lambda)[n]),main=paste("n =",n))
> plot(function(x) dnorm(x, mean=n*3/2, sd=sqrt(3*n/2)),
+      lwd=2, add=TRUE, xlim = c(0,35))
>
>
> n <- 30
> plot(x= 20:80,y = sapply(20:80, function(x) dpois(x,lambda=3*n/2)),
+      lwd=2,col=2, type="h",xlim=c(20,80), ylab=expression(f(n*hat(lambda)[n])),
+      xlab = expression(n*hat(lambda)[n]),main=paste("n =",n))
> plot(function(x) dnorm(x, mean=n*3/2, sd=sqrt(3*n/2)),
+      lwd=2, add=TRUE, xlim = c(20,80))
>
> n <- 50
> plot(x= 40:110, y = sapply(40:110, function(x) dpois(x,lambda=3*n/2)),
+      lwd=2,col=2, type="h",xlim=c(40,110), ylab=expression(f(n*hat(lambda)[n])),
+      xlab = expression(n*hat(lambda)[n]),main=paste("n =",n))
> plot(function(x) dnorm(x, mean=n*3/2, sd=sqrt(3*n/2)),
+      lwd=2, add=TRUE, xlim = c(40,120))
```



In Lecture 3 we said that the observed information increases with the sample size. This holds also for the Fisher information and it can be immediately seen from the asymptotic distribution of the MLE. On the other hand, if the sample size increases, then the variance of the MLE decreases. In practice, this means that with increasing amounts of data, the likelihood function becomes more concentrated around the MLE and thus the distribution of the MLE becomes more concentrated around the true parameter value θ_0.

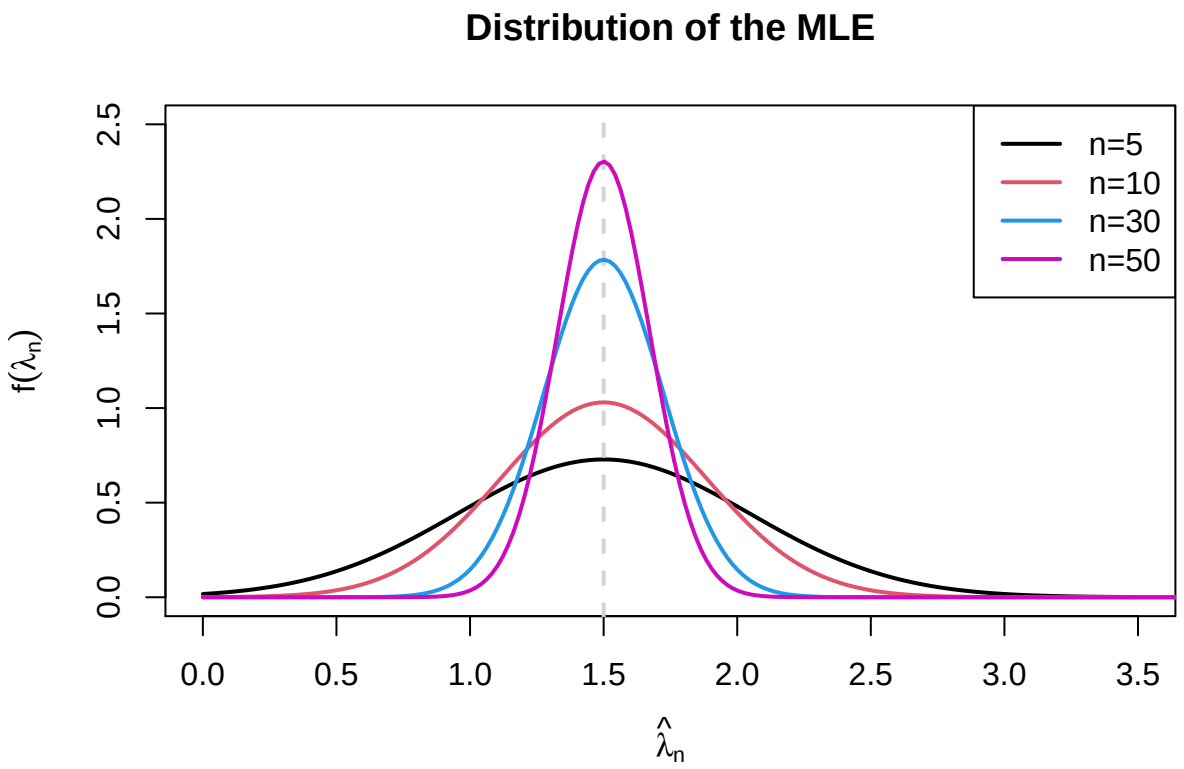
We illustrate this fact numerically. Below we plot the asymptotic distribution in the case of the Poisson model and under the assumption that samples of increasing sample size are drawn from the model with $\theta_0 = 3/2$.

```
> n1 = 5
> plot(function(x) dnorm(x, mean=3/2, sd=sqrt(3/(2*n1))),
+      lwd=2, xlim=c(0,3.5),n=500, ylim=c(0,2.5),
+      ylab=expression(f(hat(lambda)[n])),
+      xlab = expression(hat(lambda)[n]),
+      main = "Distribution of the MLE")
> abline(v=3/2, lwd=2, lty=2, col="lightgray")
```

```
> sapply(c(10, 30, 50),
+        function(y) plot(function(x) dnorm(x, mean=3/2, sd=sqrt(3/(2*y))),
+      lwd=2, xlim=c(0,10),col=round(1+y/10),n=500, add=TRUE))
```

```
##   [,1]        [,2]        [,3]
## x Numeric,500 Numeric,500 Numeric,500
## y Numeric,500 Numeric,500 Numeric,500
```

```
> legend("topright", legend = c("n=5", "n=10", "n=30", "n=50"),
+        lwd=c(2,2,2,2), col=c(1,2,4,6), lty=c(1,1,1,1))
```



Inferential Statistics **First semester, 2020**

Lecture 5: Confidence sets

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

Let us start again with a problem inspired by a real-life application.

For the problem of measuring energy consumption of WM's under the cotton 40°C washing program, by means of an estimator we can get an estimate for the mean of the population distribution. For instance, if we are willing to assume a parametric model, we can use $\widehat{\mu}$, the MLE of μ. However, announcing a single value as the estimate of a parameter calls for some auxiliary indication of how much that value can be trusted – a measure of reliability. Thus, we need to ask an additional question: how much variable is the estimate $\widehat{\mu}$? The notion of standard error we introduced in Lecture 4, is a commonly used measure, and sometimes estimates are given as point estimate $\pm$ a standard error. The standard error is not a maximum error, of course, so one could not expect the true value of the parameter to be within one standard deviation error of the estimate in every instance. Sometimes it will and sometimes it won't.

The essence of a confidence interval is thus to produce a lower limit and an upper limit which contain the true value θ_0 with a pre-specified probability. These lower and upper limit constitute a confidence interval and are the topic of the present lecture. An alternative name for confidence interval is interval estimation, since the aim of confidence interval theory is to estimate lower and upper limits for the true parameter value θ_0.

First we give some definitions.

Definition 5.1 *A random interval is a finite or infinite interval, where at least one of the end points is a r.v.*

Definition 5.2 *Given a random sample $Y_1, \ldots, Y_n$ from some distribution F_θ, let $L_n = L(Y_1, \ldots, Y_n)$ and $U_n = U(Y_1, \ldots, Y_n)$ be two statistics such that $L_n \leq U_n$. We say that the random interval $[L_n, U_n]$ is a confidence interval for θ with confidence level $1-\alpha$, $0 < \alpha < 1$ if*

$$P_\theta(L_n \leq \theta \leq U_n) \geq 1-\alpha, \quad \textit{for all } \theta \in \Theta.$$

The notation $P_\theta(\cdot)$ is to remind us that the probability has to be computed with respect to the distribution of the sample with the true parameter being θ, whatever is its value. Also we say that $[U_n, \infty)$ and $(-\infty, L_n]$ is an upper and a lower confidence limit for θ, respectively, with confidence level $1-\alpha$, if for all $\theta \in \Theta$

$$P_\theta(-\infty < \theta \leq U_n) \geq 1-\alpha$$

and

$$P_\theta(L_n \leq \theta < \infty) \geq 1-\alpha.$$

Thus $[L_n, U_n]$ traps θ with probability at least $1-\alpha$, no matter what $\theta \in \Theta$. We call $P_\theta(L_n \leq \theta \leq U_n)$ the *coverage* of the confidence interval and $1-\alpha$ the *confidence level*. In other words, the confidence level is the smallest coverage probability over all possible parameter values θ.

Be careful! $[L_n, U_n]$ is a random interval whereas θ is fixed.

Commonly, people use 95 or 99 percent confidence intervals, which correspond to choosing $\alpha = .05$ or $\alpha = .01$, respectively. If θ is a vector then we use a *confidence set*, such as an ellipse or an ellipsoid, instead of an interval.

Be careful! There is much confusion about how to interpret a confidence interval. A confidence interval is not a probability statement about θ since the latter is a fixed quantity (i.e. chosen by Nature and never revealed to us), not a random variable. A 95 percent confidence interval is interpreted as follows: if we repeat the experiment over and over again we will obtain each time a different confidence interval and the true parameter θ will be contained in 95 percent of those intervals.

5.1 Properties of confidence intervals

5.1.1 Expected length

It is quite possible that there exist more than one confidence interval for θ with the same confidence level $1 - \alpha$. In such a case, it is obvious that we would be interested in finding the shortest confidence interval on average and within a certain class of confidence intervals.

More formally, we define the *length* of the random confidence interval $[L_n, U_n]$ by

$$D_n = D(Y_1, \ldots, Y_n) = U_n - L_n,$$

and the expected length is defined by $\Delta = E(D_n)$. Thus the length of a random confidence interval, if it exists, is also a statistic. The *length criterion* then says that:

> Among all possible confidence intervals considered, all having the same confidence level, we should prefer the shortest on average.

This is because, the shorter is the confidence interval, the more precise is the inference about the true parameter θ.

5.2 Methods for confidence intervals

The general procedure for constructing confidence intervals is as follows: We start out with an r.v. $T_n(\theta) = T(Y_1, \ldots, Y_n; \theta)$ which depends on θ and on the Y_i's only through a hopefully sufficient statistic for θ, and whose distribution is completely determined. "Completely determined" means that the distribution does not depend on θ. Then L_n and U_n are some simple functions of $T_n(\theta)$ which are chosen in an obvious manner. The r.v. $T_n(\theta)$ whose distribution is completely determined is called *pivot* or *pivotal quantity*. If the r.v. $T_n(\theta)$ has a distribution which is completely determined as $n \to \infty$ then we say that $T_n(\theta)$ is an *asymptotic pivot*.

For every set B such that $P_\theta(T_n(\theta) \in B) \geq 1 - \alpha$, for all $\theta \in \Theta$, then the set

$$\{\theta \in \Theta : T_n(\theta) \in B\}$$

is a confidence region for θ of confidence level $1 - \alpha$.

Thus confidence intervals are constructed from a given pivotal quantity. Furthermore, the result below shows that when $g(\theta)$ is a strictly monotone function, a confidence interval for $g(\theta)$ can be readily derived from a confidence interval for θ.

Theorem 5.1 *If $[L_n, U_n]$ is a $(1-\alpha)$ confidence interval for θ and $g(\theta)$ is a monotone increasing function, then*

(i) $[g(L_n), g(U_n)]$ is a $(1-\alpha)$ confidence interval for $g(\theta)$ whenever $g(\cdot)$ is an increasing function.

(ii) $[g(U_n), g(L_n)]$ is a $(1-\alpha)$ confidence interval for $g(\theta)$ whenever $g(\cdot)$ is a decreasing function.

The following examples help at illustrating the point.

5.2.1 Exact pivots

Example 5.1 *Let $X_i \overset{\text{iid}}{\sim} \sim \text{Exp}(1/\theta)$ be a random sample of size n. We want to build a $1-\alpha$ level confidence interval for θ. To build a pivot for this problem first note that the MLE of θ is $\widehat{\theta} = \overline{X}$. Furthermore,*

$$\begin{aligned} n\overline{X} &= \sum_{i=1}^{n} X_i \sim \text{Ga}(n, 1/\theta) \iff 2n\overline{X} \sim \text{Ga}(n, 1/2\theta) \\ &\iff \tfrac{2n\overline{X}}{\theta} \sim \text{Ga}(n, 1/2). \end{aligned}$$

But $\text{Ga}(n, 1/2) = \text{Ga}(2n/2, 1/2) = \chi^2_{2n}$ (see L1), thus the quantity $2n\overline{X}/\theta$ is a pivot with distribution χ^2_{2n}. Let $0 < c_1 \leq c_2$ be two constants such that

$$P(c_1 \leq \chi^2_{2n} \leq c_2) = 1 - \alpha.$$

There are infinitely many possible choices for $c_1 \leq c_2$ that satisfy this inequality, but two of them are typically used. These are shown in Figure 5.1. In case (i), we restrict c_1, c_2 to have highest density. In this case c_1 and c_2 are found by cutting the density at the point which leads to c_1 and c_2 such that the area in the tails sums to α. In case (ii) we split α in two equal parts and choose c_1 and c_2 such that $P(\chi^2_{2n} \leq c_1) = P(\chi^2_{2n} \geq c_2) = \alpha/2$.

Option (i) typically leads to shorter intervals. Nevertheless, most practitioners prefer the equi-tailed option (ii) since it is much simpler to implement. Indeed, here $c_1 = \chi^2_{2n,1-\alpha/2}$ and $c_2 = \chi^2_{2n,\alpha/2}$, where $\chi^2_{2n,\alpha}$ is the upper αth quantile of the χ^2_{2n} distribution. In the case of the equi-tailed thresholds we have

$$\begin{aligned} 1 - \alpha &= P(\chi^2_{2n,1-\alpha/2} \leq \chi^2_{2n} \leq \chi^2_{2n,\alpha/2}) \\ &= P_\theta\left(\chi^2_{2n,1-\alpha/2} \leq \tfrac{2n\overline{X}}{\theta} \leq \chi^2_{2n,\alpha/2}\right) \\ &= P_\theta\left(\tfrac{1}{\chi^2_{2n,\alpha/2}} \leq \tfrac{\theta}{2n\overline{X}} \leq \tfrac{1}{\chi^2_{2n,1-\alpha/2}}\right) \\ &= P_\theta\left(\tfrac{2n\overline{X}}{\chi^2_{2n,\alpha/2}} \leq \theta \leq \tfrac{2n\overline{X}}{\chi^2_{2n,1-\alpha/2}}\right). \end{aligned}$$

So $\left[\tfrac{2n\overline{X}}{\chi^2_{2n,\alpha/2}}, \tfrac{2n\overline{X}}{\chi^2_{2n,1-\alpha/2}}\right]$ is confidence interval for θ with confidence level $1-\alpha$.

As a numerical example, suppose X_i's are time in seconds taken by a HPC server to reboot in $n = 10$ reboots, and let the observed sample be (48.386, 65.418, 26.510, 15.830, 22.381, 21.882, 30.121, 18.714,

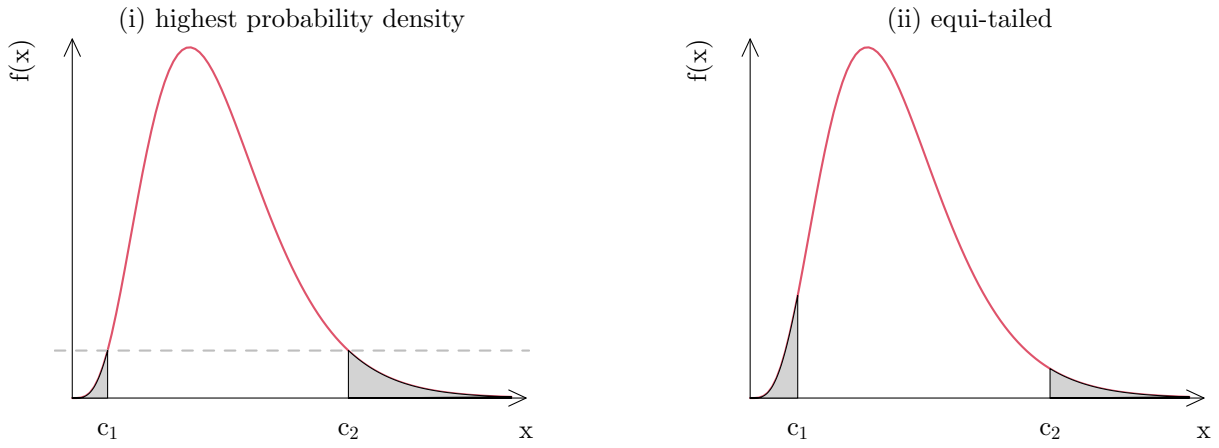


Figure 5.1: Two different approaches for defining thresholds c_1, c_2 which leave in the tails an overall probability equal to α. In (i) the thresholds have highest density but lead to tails of different probabilities. In case (ii), thresholds have different density values but they lead to equal-probability tails, i.e. tails of probability $\alpha/2$.

10.874, 1.759). Then $\overline{x} = 26.1875$. *With* $\alpha = .10$ *we have that* $\chi^2_{20,.05} = 31.41$ *and* $\chi^2_{20,.95} = 10.85$. *Thus the 90 percent confidence interval for* θ *is* $[16.67, 48.27]$.

We interpret this result by saying that we are 90 percent confident that the interval $[16.67, 48.27]$ *will contain the true parameter value* θ. *"90 percent confident" means the following. If we could draw a large number of samples of size* $n = 10$ *and compute, for each of them, a 0.90 confidence interval, then we expect that 90 percent of these intervals will contain the true parameter value under which the data are actually generated.*

Example 5.2 *Let* $Y_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ *be a random sample of size n. First, suppose that* σ *is known, so that* μ *is the only unknown parameter. Consider the r.v.* $\sqrt{n}(\overline{Y} - \mu)/\sigma$. *This depends on the sample only through the sufficient statistic* $\overline{Y}$ *of* μ *and its distribution is the same as the r.v.* $Z \sim N(0, 1)$, *for all* μ.

Next determine two numbers $c_1 \leq c_2$ *such that*

$$\begin{aligned} 1 - \alpha &= P(c_1 \leq Z \leq c_2) \\ &= P_\mu\left(c_1 \leq \tfrac{\sqrt{n}(\overline{Y}-\mu)}{\sigma} \leq c_2\right) \\ &= P_\mu\left(\overline{Y} - c_2 \tfrac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} - c_1 \tfrac{\sigma}{\sqrt{n}}\right), \end{aligned}$$

is a confidence interval for μ *with coverage* $1 - \alpha$. *Its length is equal to* $(c_2 - c_1)\sigma/\sqrt{n}$ *which is also equal to the expected length. Form this it follows that, among all confidence intervals with coverage* $1 - \alpha$ *which have the above form, the shortest one is that for which* $c_2 - c_1$ *is smallest. This happens if* $c_2 = z_{\alpha/2}$ *and* $c_1 = -c_2$, $z_{\alpha/2}$ *is the upper* $(\alpha/2)$*th quantile of the* $\mathrm{N}(0, 1)$ *distribution. Therefore, the shortest confidence interval for* μ *with coverage* $1 - \alpha$ *(and which is of the form above) is given by*

$$\left[\overline{Y} - z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}, \overline{Y} + z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}\right].$$

Example 5.3 *Let $Y_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ be a random sample of size n. Now suppose that μ is known, so that σ^2 is the unknown parameter and consider the r.v.*

$$\frac{n\widehat{\sigma^2}_\mu}{\sigma^2}, \quad \textit{where} \quad \widehat{\sigma^2}_\mu = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2.$$

This r.v. depends on the sample only through the sufficient statistic $\widehat{\sigma^2_\mu}$ for σ^2 and its distribution is χ^2_n for all σ^2.

Now determine two numbers c_1 and c_2, $0 < c_1 \leq c_2$ such that

$$\begin{aligned} 1-\alpha &= P(c_1 \leq \chi^2_n < c_2) \\ &= P_{\sigma^2}\left(c_1 \leq \frac{n\widehat{\sigma^2}_\mu}{\sigma^2} \leq c_2\right) \\ &= P_{\sigma^2}\left(\frac{n\widehat{\sigma^2}_\mu}{c_2} \leq \sigma^2 \leq \frac{n\widehat{\sigma^2}_\mu}{c_1}\right), \end{aligned}$$

Thus

$$\left[\frac{n\widehat{\sigma^2}_\mu}{c_2}, \frac{n\widehat{\sigma^2}_\mu}{c_1}\right]$$

is a confidence interval for σ^2 with coverage $(1-\alpha)$ and length equal to $(1/c_1 - 1/c_2)n\widehat{\sigma^2}_\mu$. The expected length is equal to $(1/a - 1/b)n\sigma^2$. As in Example 5.1, it is possible to determine c_1 and c_2 such that the resulting confidence interval of the form above has the shortest length. This can be achieved by thresholds which have highest density (option (i) in Figure 5.1).

However, in practice c_1 and c_2 are often chosen by assigning probability $\alpha/2$ to each one of the tails of the χ^2_n distribution. The resulting equi-tailed confidence interval is then

$$\left[\frac{n\widehat{\sigma^2}_\mu}{\chi^2_{n,\alpha/2}}, \frac{n\widehat{\sigma^2}_\mu}{\chi_{n,1-\alpha/2}}\right].$$

Note that this is not the best choice because then the corresponding interval is not the shortest one.

Furthermore, thanks to Theorem 5.1, a $(1-\alpha)$ confidence interval for $\sigma = \sqrt{\sigma^2}$ is

$$\left[\sqrt{\frac{n\widehat{\sigma^2}_\mu}{\chi^2_{n,\alpha/2}}}, \sqrt{\frac{n\widehat{\sigma^2}_\mu}{\chi_{n,1-\alpha/2}}}\right],$$

Example 5.4 *Let $Y_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ be a random sample of size n, where both μ and σ^2 are unknown. We consider confidence intervals for μ and σ^2 separately. Consider the two pivots*

$$\frac{\sqrt{n}(\overline{Y}-\mu)}{\sqrt{S^2}} \sim t_{n-1}, \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1};$$

see Lecture 4. Thus a confidence interval of level $1-\alpha$ for μ and a confidence interval of level $1-\alpha$ for σ^2 are

$$\left[\overline{Y} - t_{n-1,\alpha/2}\sqrt{\frac{S^2}{n}}, \overline{Y} + t_{n-1,\alpha/2}\sqrt{\frac{S^2}{n}}\right]$$

and

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}}\right],$$

respectively; here $t_{n-1,\alpha}$ denotes the upper αth quantile of the t-Student distribution with $n-1$ degrees of freedom.

Here are two two-sample problems. In the first one we wish to make inference about the difference between the means of the two normal populations, whereas in the second we wish to make inference about the ratio of their variances.

Example 5.5 *Let $Y_i \overset{\text{iid}}{\sim} N(\mu_1, \sigma_1^2)$ be a random sample of size n and $X_j \overset{\text{iid}}{\sim} N(\mu_2, \sigma_2^2)$, be an i.i.d. random sample of size m where we assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and μ_1, μ_2, σ^2 are unknown. We wish to build a confidence interval for $\mu_1 - \mu_2$. Consider the r.v.*

$$\frac{(\overline{Y}-\overline{X})-(\mu_1-\mu_2)}{\sqrt{S^2_{pool}\left(\frac{1}{m}+\frac{1}{n}\right)}}, \tag{5.1}$$

where

$$S^2_{pool} = \frac{(n-1)S_Y^2+(m-1)S_X^2}{n+m-2}$$

is sometimes referred to as the pooled variance estimate. It can be shown that the r.v. in (5.1) *has distribution t_{n+m-2}, thus it is a pivotal quantity. A confidence interval of level $1-\alpha$ for $\mu_1 - \mu_2$ is then*

$$\left[(\overline{Y} - \overline{X}) - t_{n+m-2,\alpha/2}\sqrt{S^2_{pool}\left(\tfrac{1}{m} + \tfrac{1}{n}\right)}, (\overline{Y} - \overline{X}) + t_{n+m-2,\alpha/2}\sqrt{S^2_{pool}\left(\tfrac{1}{m} + \tfrac{1}{n}\right)}\right].$$

Example 5.6 *Let $Y_i \overset{\text{iid}}{\sim} N(\mu_1, \sigma_1^2)$ be a random sample of size n and $X_j \overset{\text{iid}}{\sim} N(\mu_2, \sigma_2^2)$, be an i.i.d. random sample of size m and $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ are all unknown. We wish to build a confidence interval for σ_2^2/σ_1^2. The r.v.*

$$\frac{S_Y^2/\sigma_1^2}{S_X^2/\sigma_2} \sim F_{n-1,m-1}$$

provides a pivotal quantity for σ_1^2/σ_2^2. Thus a $(1-\alpha)$ confidence interval for σ_2^2/σ_1^2 can be obtained by setting the quantiles for the F distribution with degrees of freedom $n-1$ and $m-1$ such that

$$1-\alpha \quad = \quad P_\theta\left(F_{n-1,m-1,1-\alpha/2} \le \frac{S_Y^2/\sigma_1^2}{S_X^2/\sigma_2^2} \le F_{n-1,m-1,\alpha/2}\right)$$

and thus the confidence interval is

$$\left[\frac{S_X^2}{S_Y^2}F_{n-1,m-1,1-\alpha/2}, \frac{S_X^2}{S_Y^2}F_{n-1,m-1,\alpha/2}\right],$$

where $F_{n,m,\alpha}$ denotes the upper αth quantile of the $F_{n,m}$ distribution.

Example 5.7 *(Paired samples). In all examples above we assumed independent samples. In some cases, such as test-retest experiments, dependent samples are appropriate. For example, to measure the effect of new type of motor for WM and compare it with the old motor, we would select n WM's at random from the production line and measure their energy consumption both with the old motor and with the new motor. The observations would be independent between pairs, but the observations within a pair would not be independent because they were taken on the same WM. Other examples are: effectiveness of a diet plan on weight loss for a sample of people, in which the weight is measure before and after the diet, measuring the performance of two different prediction algorithms on the same validation set, etc.*

We have a random sample made of n pairs (X_i, Y_i) *and we assume that the differences* $D_i = Y_i - X_i$*, for* $i = 1, \ldots, n$ *are normally distributed with mean* $\delta = \mu_1 - \mu_2$ *and variance* $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$*, or*

$$D_i \sim \mathrm{N}(\delta, \sigma_D^2).$$

All parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12}$ *are unknown. Now let* $\overline{D} = \sum_{i=1}^n D_i/n = \overline{Y} - \overline{X}$*, and*

$$S_D^2 = \tfrac{\sum_{i=1}^n (D_i - \overline{D})^2}{n-1}.$$

It follows that $\overline{D} \sim \mathrm{N}(\delta, \sigma_D^2/n)$ *and thus*

$$\tfrac{\overline{D} - \delta}{\sqrt{S_D^2/n}} \sim t_{n-1},$$

is a pivotal quantity. A $(1-\alpha)$ *confidence interval for* $\delta = \mu_1 - \mu_2$ *is given by*

$$\left[\overline{D} - t_{n-1,\alpha/2}\sqrt{\tfrac{S_D^2}{n}}, \overline{D} + t_{n-1,\alpha/2}\sqrt{\tfrac{S_D^2}{n}}\right].$$

5.2.2 Asymptotic pivots

All pivots considered in the previous section have exact distributions. When a pivotal quantity is not available, it may still be possible to determine a confidence region for a parameter θ if a statistic exists with an asymptotic distribution that depends on θ but not on any other unknown nuisance parameters. Specifically, let $Y_1, \ldots, Y_n$ have joint p.d.f. $f(y_1, \ldots, y_n; \theta)$ and let $T_n = T(Y_1, \ldots, Y_n) \sim g(t; \theta)$ as n diverges; thus $g(\cdot; \theta)$ is some distribution that depends only on θ. T_n may be a sufficient statistic for θ, or possibly some reasonable estimator such as an MLE.

We saw in Lecture 4 that under suitable regularity condition the MLE is asymptotically normally distributed, i.e.

$$\tfrac{\widehat{\theta}_n - \theta}{\mathrm{se}(\widehat{\theta}_n)} \dot{\sim} \mathrm{N}(0,1).$$

The quantity on the right hand side is thus an asymptotic pivot for the parameter θ, and

$$\left[\widehat{\theta} - z_{\alpha/2}\mathrm{se}(\widehat{\theta}_n), \widehat{\theta} + z_{\alpha/2}\mathrm{se}(\widehat{\theta}_n)\right],$$

is an approximate confidence interval for θ with coverage level approximately $(1-\alpha)$; here "approximately" means that

$$P_\theta\left[\widehat{\theta} - z_{\alpha/2}\text{se}(\widehat{\theta}_n) \leq \theta \leq \widehat{\theta} + z_{\alpha/2}\text{se}(\widehat{\theta}_n)\right] \to 1-\alpha, \quad \text{as} \quad n \to \infty.$$

These type of confidence intervals are called Wald intervals, are always symmetric and are easy to compute since only the MLE and its standard error are required. However, in problems of small sample size in which the actual distribution of the MLE could be far from the normal, these type of intervals may be inaccurate. That is, their actual coverage probability may be far from the confidence level $1-\alpha$. Another issue with Wald confidence intervals is that they do not take care of the constraints on the parameter space Θ. For instance, we may get $[-1.2, 1.6]$ as the confidence interval for the mean λ of a Poisson distribution. Since negative values for λ are not allowed, in such a case, it is customary to discard the negative values and set the lower bound of the interval to the lowest value of the parameter space, i.e. zero in this case.

Under the maximum likelihood framework, it is possible to build likelihood-based confidence intervals which are more accurate than Wald intervals and solve many of the shortcomings of the latters. We will see this in the next lecture.

For the time being here are some examples.

Example 5.8 *Let $Y_1, \ldots, Y_n$ be an i.i.d. sample from the $\text{Poi}(\lambda)$ distribution and we wish to build a confidence interval for λ. Consider the sufficient statistic $T_n = \sum_{i=1}^n Y_i$, for which we know that $T_n \sim \text{Poi}(n\lambda)$. Note that this time we do not have a pivot since T_n does not depend on λ.*

However, we know that the MLE of λ is $\widehat{\lambda} = \overline{Y}$ and thus by the consistency of the MLE we know that

$$\frac{\sqrt{n}(\overline{Y}-\lambda)}{\sqrt{\overline{Y}}} \xrightarrow{d} \text{N}(0,1).$$

Thus an approximate $(1-\alpha)$ Wald confidence interval for λ is

$$\left[\overline{Y} - z_{\alpha/2}\sqrt{\frac{\overline{Y}}{n}}, \overline{Y} + z_{\alpha/2}\sqrt{\frac{\overline{Y}}{n}}\right].$$

Example 5.9 *Let $Y_1, \ldots, Y_n$ be an i.i.d. sample from the $\text{Ber}(\theta)$ distribution and we wish to build a confidence interval for θ. Note again that the sufficient statistic $T_n = \sum_{i=1}^n Y_i$, for which we know that $T_n \sim \text{Bin}(n, \theta)$, is not a pivot since T_n does not depend on θ.*

However, we know that the MLE of θ is $\widehat{\theta} = \overline{Y}$ and thus by the consistency of the MLE we know that

$$\frac{\sqrt{n}(\overline{Y}-\theta)}{\sqrt{\overline{Y}(1-\overline{Y})}} \xrightarrow{d} \text{N}(0,1).$$

Thus an approximate $(1-\alpha)$ Wald confidence interval for θ is

$$\left[\overline{Y} - z_{\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}, \overline{Y} + z_{\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}\right].$$

Example 5.10 *Let $Y_1, \ldots, Y_n$ be an i.i.d. sample from $\text{Ber}(\theta_1)$ and let $X_1, \ldots, X_m$ be an i.i.d. sample from $\text{Ber}(\theta_2)$. We wish to build a confidence interval for $\theta_2 - \theta_1$, the difference between the two success probabilities.*

Let $\widehat{\theta}_1 = \overline{Y}$ and the $\widehat{\theta}_2 = \overline{X}$ be the MLE of θ_1 and θ_2, respectively. By the consistency of the MLE we know that

$$\frac{\sqrt{n}(\overline{Y}-\theta_1)}{\sqrt{\overline{Y}(1-\overline{Y})}} \xrightarrow{d} \mathrm{N}(0,1) \quad \textit{and} \quad \frac{\sqrt{m}(\overline{X}-\theta_2)}{\sqrt{\overline{X}(1-\overline{X})}} \xrightarrow{d} \mathrm{N}(0,1)$$

Furthermore, since the X_i's and Y_i's are independent, then also $\widehat{\theta}_1$ is independent from $\widehat{\theta}_2$. By the properties of the normal distribution we have that

$$\widehat{\theta}_2 - \widehat{\theta}_1 \sim \mathrm{N}\left(\theta_2 - \theta_1, \frac{\overline{Y}(1-\overline{Y})}{n} + \frac{\overline{X}(1-\overline{X})}{m}\right)$$

Thus an approximate $(1-\alpha)$ Wald confidence interval for $\theta_2 - \theta_1$ is

$$\left[\overline{Y} - \overline{X} \pm z_{\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n} + \frac{\overline{X}(1-\overline{X})}{m}}\right].$$

Example 5.11 *Let $Y_1, \ldots, Y_n$ and $X_1, \ldots, X_m$ be as in Example 5.5 but without assuming variance equality. When variances are not equal it is not easy to eliminate them to obtain a pivotal quantity for $\mu_1 - \mu_2$. One possible approach would be by noticing that, as n and m both diverge to ∞,*

$$\frac{\overline{Y}-\overline{X}-(\mu_1-\mu_2)}{\sqrt{S_Y^2/n + S_X^2/m}} \xrightarrow{d} \mathrm{N}(0,1).$$

Thus for large sample sizes, approximate confidence limits for $\mu_1 - \mu_2$ may be easily obtained from this expression. Note that the above limiting result also holds if the samples are not from normal distributions, so this provides a general large-sample result for differences of means. How good the limiting approximation will be in a particular sample depend somewhat on the form of the densities. A better approximation can be obtained by using the t-Student distribution as limiting distribution. This is

$$\frac{\overline{Y}-\overline{X}-(\mu_1-\mu_2)}{\sqrt{S_Y^2/n + S_X^2/m}} \sim t_\nu,$$

where the degrees of freedom are estimated by the formula

$$\nu = \frac{(S_Y^2/n + S_X^2/m)^2}{\frac{(S_Y^2/n)^2}{n-1} + \frac{(S_X^2/m)^2}{m-1}},$$

where this time ν may not necessarily be an integer. The general problem of making inferences about $\mu_1 - \mu_2$ with unequal variances is known as the Behrens-Fisher problem.

References

[LM18] Larsen, R. J. and Marx, M. L. (2018) *An Introduction to Mathematical Statistics and its Applications* (6th edition), Pearson Education, Inc..

Practice Lecture 5: Confidence intervals

Erlis Ruli (ruli@stat.unipd.it)

30 November 2020

In this handouts we illustrate examples considered in Lecture 5 in practice using `R`. In some examples we generate fictitious data from the model and in others we consider real-life data.

1 Exact pivots

1.1 Example 5.2 (Normal population with known variance)

Let $Y_1, \ldots, Y_n$ be a random sample of size n from $N(\mu, 2)$ and our aim is to build a confidence interval for μ. In particular, suppose that the sample size is $n = 10$ and it is obtained from the above distribution with true parameter $\mu_0 = 0$.

So Nature takes this model and generates the following data (or if you prefer to think more practically: suppose you are measuring the difference in diameter of bearing spheres with respect to a target value, and you have n such measurements)

```
> set.seed(5) # fix the random seed
> n <- 10 # set the sample size
> mu0 <- 0
> sigma2.0 <- 2
> yobs <- rnorm(n, mu0, sqrt(sigma2.0))
> yobs # measurements we are given by Nature
```

```
##  [1] -1.18914922  1.95777976 -1.77553362  0.09919685  2.42034289 -0.85264064
##  [7] -0.66774411 -0.89855073 -0.40414495  0.19531452
```

We saw in Lecture 5 the procedure for building a confidence interval for this problem, i.e. $\bar{Y} \pm z_{\alpha/2}\sigma/\sqrt{n}$. Let us consider a 0.95 confidence interval, i.e. $\alpha = 0.05$. Thus for the sample at hand we replace the random quantities with their observed counterparts to obtain

```
> alpha = 0.05
> (bar.y <- mean(yobs)) # sample mean
```

```
## [1] -0.1115129
```

```
> (se <- sqrt(sigma2.0/n)) # this is the standard error of the estimate
```

```
## [1] 0.4472136
```

```
> bar.y + c(-1,1)*qnorm(p = alpha/2, lower.tail =FALSE)*se
```

```
## [1] -0.9880355  0.7650096
```

```
> # note: we used lower.tail =FALSE in the quantile function of the
> # normal distribution in order to have the upper quantile as desire.
```

Thus our interval is $[-0.99, 0.77]$. At this point it is good to pause a bit and think about it. This interval is the observed (i.e. numerical) version of the random interval $\bar{Y} \pm z_{\alpha/2}\sigma/\sqrt{n}$. Thus, the probability that

the interval $[-0.99, 0.77]$ contains μ_0 can only be either 1 or 0 depending on whether μ_0 is inside or not, respectively. By the way, in this particular case, the observed interval does contain μ_0 so the probability of coverage is 1.

Recall that, in practice, we do not know μ_0, so we will never know if μ_0 is inside the particular observed interval at hand or not. However, we if we could compute in this way a large number of intervals, then 95% of them will contain μ_0. So we can only say that we are 95% **confident** that our observed interval $[-0.99, 0.77]$ contains μ_0.

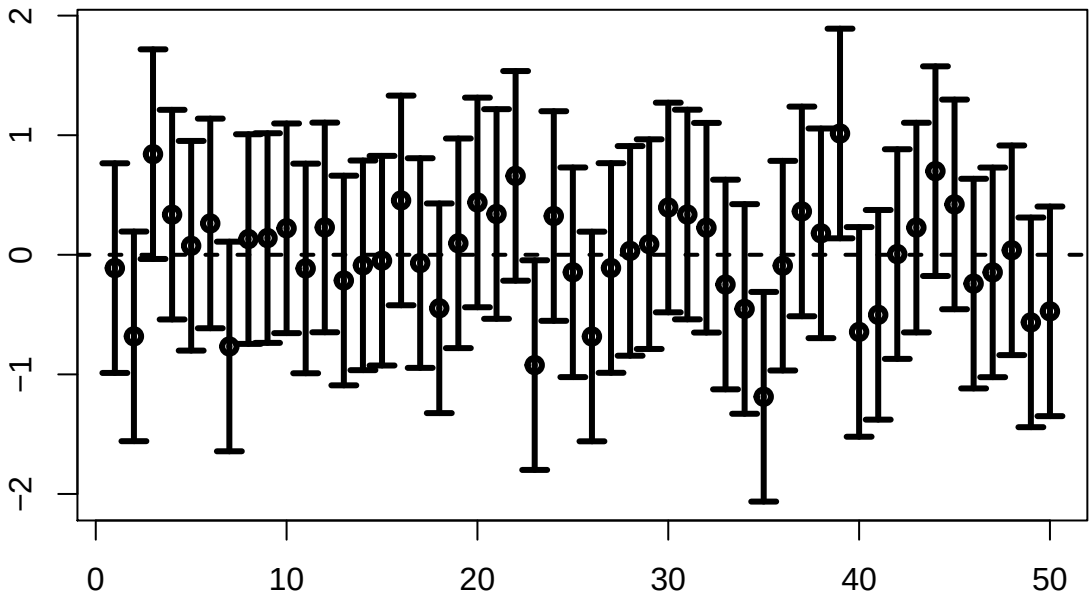
To see this better, we will perform a *simulation study* in which we compute a large number of such intervals, say $N = 10^6$ (i.e. we ask Nature to give us N datasets from the same model it used to generate the first sample). The idea is the following. We generate $N = 10^6$ observed samples from the same model as above and for each observed sample, we compute the associated observed 95 percent interval. Here is the `R` code for this.

```
> set.seed(5)
> N = 1e+6 # the number of intervals we want to calculate
> my.N.CI <- matrix(NA, nrow = N, ncol = 2)# put the obs.intervals here
> for(i in 1:N){
+   yi <- rnorm(n, mu0, sqrt(sigma2.0))
+   bar.yi <- mean(yi)
+   my.N.CI[i,] <- bar.yi + c(-1,1)*qnorm(p = alpha/2, lower.tail =FALSE)*sqrt(sigma2.0/n)
+ }
> # here are the first 6 observed intervals out of N
> head(my.N.CI)
```

```
##             [,1]      [,2]
## [1,] -0.98803546 0.7650096
## [2,] -1.55854687 0.1944982
## [3,] -0.03504896 1.7179961
## [4,] -0.54076425 1.2122808
## [5,] -0.80106181 0.9519833
## [6,] -0.61434304 1.1387020
```

We plot below the first 50 observed confidence intervals along with the true parameter value

```
> # we use the plotrix library for doing this plot
> # if not already installed use
> # install.packages(plotrix)
> library(plotrix)
> plotCI(x= 1:50, y = apply(my.N.CI[1:50,], 1, mean),
+        li = my.N.CI[1:50,1],ui = my.N.CI[1:50,2],
+        xlab="Observed confidence intervals for mu",ylab=NA, lwd=3,)
> abline(h = 0, lwd=2, lty=2)
```



Observed confidence intervals for mu

Inspecting the first 50 observed intervals for μ, we note that there are three intervals which do not contain the true value (here denoted by the horizontal dashed line). How may of the N intervals do contain the true value μ_0?

```
> mu0.inside <- apply(my.N.CI, MARGIN = 1,
+                     function(x) ifelse(mu0 >= x[1]  & mu0 <= x[2],1,0))
> # m0.inside is a vector of 1 and 0
> head(mu0.inside)
```

```
## [1] 1 1 1 1 1 1
```

```
> # how many ones are there relative to N?
> mean(mu0.inside)
```

```
## [1] 0.949681
```

Thus we conclude that the fraction of the intervals that contain μ_0 is essentially 0.95.

Remark: Obviously it is not exactly 0.95 because 0.949681 is only a sample average which targets the confidence level $(1-\alpha) = 0.95$. By the LLN, this sample average converges to the target as $N \to \infty$, but our lazy choice was $N = 10^6 < \infty$. Also, do not confuse n, the sample size with N the number of simulations we performed. The larger N the closer will be the sample average to the target value $(1-\alpha)$. On the other hand, the pivot we used to build the confidence intervals has an exact distribution, no matter what is n, thus the latter plays no role in this particular case.

Remark: The higher the confidence level, i.e. the lower α the wider is the confidence interval. Thus for $\alpha = 0$, our interval is simply $\mathbb{R}$, thus we are obviously certain that our interval will contain μ_0.

1.2 Example 5.3 (Normal population with known mean)

This time we have μ known and σ^2 is the unknown parameter of interest. Let us set $\mu = 0$, thus the model we are considering is $N(0, \sigma^2)$.

Assume that the true variance is $\sigma_0^2 = 1$ and let $n = 10$. Under this assumption Nature generates the following data (this time we are interested in the variability of the differences in diameter of our bearing spheres).

```
> set.seed(5) # fix the random seed
> n <- 10 # set the sample size
> mu0 <-  0
> sigma2.0 <- 1
> yobs <- rnorm(n, mu0, sqrt(sigma2.0))
```

We saw that the random interval $\left[\frac{n\hat{\sigma}^2_\mu}{\chi^2_{n,\alpha/2}}, \frac{n\hat{\sigma}^2_\mu}{\chi_{n,1-\alpha/2}}\right]$ is a $(1-\alpha)$ confidence interval for σ^2. Let us consider a 0.95 confidence interval. Thus for the sample at hand we replace the random quantities with their observed counterparts to obtain

```
> (hat.sig2.mu <- sum((yobs-mu0)^2)/n)
```

```
## [1] 0.8224575
```

```
> # CI is
> c(n*hat.sig2.mu/qchisq(p=alpha/2, df=n, lower.tail = FALSE),
+   n*hat.sig2.mu/qchisq(p=1-alpha/2, df=n, lower.tail = FALSE))
```

```
## [1] 0.4015283 2.5329977
```

Again, this is an observed interval and may or may not contain the true value σ^2_0. All we can say is that we are 95% confident that it will.

Exercise. Compute a 0.95 confidence interval for $\log \sigma^2$.

1.3 Example 5.4 (Normal population)

This time both μ and σ^2 are unknown parameters, thus the model we are considering is $N(\mu, \sigma^2)$. We want to compute a confidence interval for each of the parameters.

This time we have two different pivotal quantities (see Lecture 5). Assume that the true mean is $\mu_0 = 0$ and the true variance is $\sigma^2_0 = 1$ and let $n = 10$. Under this assumption Nature generates the following data (this time we are interested both in the average and in the variability of the differences in diameter of our bearing spheres).

```
> set.seed(5) # fix the random seed
> n <- 10 # set the sample size
> mu0 <-  0
> sigma2.0 <- 1
> yobs <- rnorm(n, mu0, sqrt(sigma2.0))
```

Let us consider 0.95 confidence intervals. Thus for the sample at hand we replace the random quantities with their observed counterparts to obtain

```
> # recall that var divides by n-1!
> (hat.sig2 <- var(yobs))
```

```
## [1] 0.9069332
```

```
> (hat.mu <- mean(yobs))
```

```
## [1] -0.07885155
```

```
> # the standard error of hat.mu
> se <- sqrt(hat.sig2/n)
>
> # CI for mu
> c(hat.mu + c(-1,1)*qt(alpha/2, df=n-1, low=F)*se)
```

```
## [1] -0.7601077  0.6024046
```

```
> # CI for mu
> c(n*hat.sig2/qchisq(p=alpha/2, df=n-1, lower.tail = FALSE),
+   n*hat.sig2/qchisq(p=1-alpha/2, df=n-1, lower.tail = FALSE))
```

```
## [1] 0.476762 3.358527
```

Again, these are observed intervals, which may or may not contain their respective values. All we can say is that we are 95% confident that each interval will contain its true value.

Since R is a statistical software we do not need to do everything "by hand" as above. In this example we can instead use the built-in R command `t.test` which, besides other things, computes also a confidence interval for μ.

```
> # for the CI for mu
> t.test(yobs)
```

```
##
##  One Sample t-test
##
## data:  yobs
## t = -0.26183, df = 9, p-value = 0.7993
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.7601077  0.6024046
## sample estimates:
##   mean of x
## -0.07885155
```

Do not worry about the meaning of rest of the output, we will see it in the incoming lecture. Notice how much wider are the confidence intervals obtained in this example as compared to those of the previous two examples. The higher uncertainty due to wider intervals is the price we have to pay when the parameters are unknown and thus have to be estimated from data.

1.4 Example 5.5 (Difference of means for two normal samples)

To illustrate this example we consider real data on energy consumption of two type of WM's, which differ only on the type of motor they are equipped with. The data are not paired, in the sense that the WM's in the two groups have different ID's.

In this problem we are interested in the difference in of the two population means $\mu_1 - \mu_2$. We first load the data in R by

```
> # read the file, specifying header=TRUE since the first row contains the names of the variables.
> motors <- read.table("wm_motors.txt", header = TRUE)
> head(motors)
```

```
##      energy motor
## 1 0.7423580  GEN1
## 2 0.6971112  GEN1
## 3 0.7268793  GEN1
## 4 0.7188058  GEN1
## 5 0.7218520  GEN1
## 6 0.7777559  GEN1
```

```
> # another quick view for data frames is
> str(motors)
```

```
## 'data.frame':    30 obs. of  2 variables:
```

```
##  $ energy: num  0.742 0.697 0.727 0.719 0.722 ...
##  $ motor : chr  "GEN1" "GEN1" "GEN1" "GEN1" ...
```

The command `read.table` reads the external file and loads the data into an `R` object, here called `motors`, which is of type: `data.frame`. The latter generalises matrices, as created by the command `matrix`, since they are able to store numbers as well as strings, i.e. in a matrix you can only store things of the same type.

First we perform some reshaping of the data in order to have the two variables separated.

```
> # energy consumption with motor GEN1
> y <- motors$energy[motors$motor == "GEN1"]
>
> # energy consumption with motor GEN2
> x <- motors$energy[motors$motor == "GEN2"]
```

Now we compute the confidence interval for $\mu_1 - \mu_2$ by assuming that the two samples come from two normal distributions, resp. $Y_i \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ and $X_i \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$, assuming equal variances, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma$.

```
> # sample averages of energy consumption under GEN1 and GEN2
> bar.y <- mean(y)
> bar.x <- mean(x)
>
> # compute the size of the two samples
> n <- length(y)
> m <- length(x)
>
> # sample variances
> s2.y <- var(y)
> s2.x <- var(x)
>
> # pooled variance
> pooled.s2 <- ((n-1)*s2.y + (m-1)*s2.x)/(n+m-2)
>
> # the standard error is thus
> se = sqrt(pooled.s2*(1/n+1/m))
>
> # the confidence interval is then
> (bar.y-bar.x) + c(-1,1)*qt(alpha/2, df=n+m-2, low=F)*se
```

```
## [1] -0.044354504 -0.008654067
```

We see that both limits of the observed confidence interval are negative. To interpret this result, suppose that the true value of μ_1, say $\mu_{1,0}$ and the true value of μ_2, $\mu_{2,0}$, are equal, i.e. $\mu_{1,0} = \mu_{2,0}$ so then $\mu_{1,0} - \mu_{2,0} = 0$. Actually, this was what the engineers wanted to check in the study. Indeed, they suspected that the two motors do not consume on average the same energy, other things held equal. We see that zero is not inside the confidence interval. Thus, what we can conclude from this study is that we are 95% confident that the means of energy consumption in the two groups of WM's are not equal, i.e. with 95% confidence the two motors consume on average different amounts of energy.

Again, there is a quicker way to compute this interval: the `t.test` command.

```
> t.test(y,x, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  y and x
## t = -3.0415, df = 28, p-value = 0.005068
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.044354504 -0.008654067
## sample estimates:
## mean of x mean of y
## 0.7139261 0.7404304
```

1.5 Example 5.6 (Ratio of variances for two normal samples)

We consider the same data of the example above but now we assume different variances between the samples and we are interested on the ratio of the variances. We compute the 0.95 confidence interval for the ratio σ_1^2/σ_2^2, both manually, by applying its mathematical definition, and by using the `R` command `var.test`.

```
> c(s2.y/s2.x * qf(alpha/2, df1=n-1, df2=m-1),
+   s2.y/s2.x * qf(1- alpha/2, df1=n-1, df2=m-1))
```

```
## [1]  8.632077 76.583642
```

```
> # the same done with the built-in R command
> var.test(y,x)
```

```
##
##  F test to compare two variances
##
## data:  y and x
## F = 25.711, num df = 14, denom df = 14, p-value = 2.897e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   8.632077 76.583642
## sample estimates:
## ratio of variances
##            25.7114
```

Thus we see that with 95% of confidence the ratio of the variances σ_1^2/σ_2^2 varies from 8.6 to 76.6.

Exercise If the engineers claim that the true variances $\sigma_{1,0}^2$ and $\sigma_{2,0}^2$ are equal. Do the data support this claim?

2 Asymptotic pivots

In the case of exact pivots (as in Section 1), the only way the sample size n affects the results is by leading to narrower confidence intervals.

In the case of asymptotic pivots, not only n does affect the width of the interval, in the same way as above, but it does affect also the coverage probability of the interval itself. Thus, the coverage of confidence intervals from asymptotic pivots will not be exactly equal to $1-\alpha$ but it will converge to $(1-\alpha)$ as $n \to \infty$.

In practice our dataset has a fixed sample size n and this is result is not very useful since it cannot tell us how accurate our intervals will be in that particular sample size. However, we can check the coverage of our intervals for a given sample size by simulation. We will do this in the next example.

2.1 Example 5.8 (Poisson population)

Consider the same setting as in Lecture 5 and the aim is to check the coverage probability of the approximate $(1-\alpha)$ Wald confidence intervals, i.e. we want to compute

$$P_\lambda\left(\bar{Y} + z_{\alpha/2}\sqrt{\frac{\bar{Y}}{n}} \leq \lambda \leq \bar{Y} + z_{\alpha/2}\sqrt{\frac{\bar{Y}}{n}}\right),$$

for different values of n. The aim is thus to see if the coverage probability goes towards $(1-\alpha)$ as n increases.

This probability can be computed analytically, but here we will approximate it by simulation, i.e. we will use the Monte Carlo method (see P1). The Monte Carlo method can be used to approximate any feature of a probability distribution that is analytically difficult.

We set the true parameter $\lambda_0 = 3/2$ and take sample sizes $n = 5, 50$. We will approximate such the coverage probability by its empirical average in a large Monte Carlo sample N. By the LLN we know that as $N \to \infty$, the sample average will converge to the true probability coverage of the interval.

```
> lambda0 <- 3/2
> N <- 1e+6
>
> # we first build a function which takes as input an observed sample and
> # outputs a confidence interval
> ci.poisson <- function(ysamp, alpha){
+   n = length(ysamp)
+   bar.y = mean(ysamp)
+   se = sqrt(bar.y/n)
+   oo = c(bar.y + c(-1,1)*qnorm(alpha/2, low=F)*se)
+   return(oo)
+ }
```

Now we generate datasets from the Poison model and compute the confidence intervals by our newly defined function `ci.poisson`.

With sample of size $n = 5$ we have

```
> set.seed(5)
> n <- 5
> CI.5 <- matrix(NA, nrow = N, ncol = 2)# put the obs.intervals here
> for(i in 1:N) {
+   y5 <- rpois(n, lambda0)
+   CI.5[i,] <- ci.poisson(y5,alpha = 0.05)
+ }
> lambda0.inside5 <- apply(CI.5, MARGIN = 1,
+                      function(x) ifelse(lambda0 >= x[1]  & lambda0 <= x[2],1,0))
> # how many ones are there?
> mean(lambda0.inside5)
```

```
## [1] 0.936302
```

With sample of size $n = 50$ we have

```
> set.seed(5)
> n <- 50
> CI.50 <- matrix(NA, nrow = N, ncol = 2)# put the obs.intervals here
> for(i in 1:N) {
+   y50 <- rpois(n, lambda0)
+   CI.50[i,] <- ci.poisson(y50,alpha = 0.05)
+ }
> lambda0.inside50 <- apply(CI.50, MARGIN = 1,
+                      function(x) ifelse(lambda0 >= x[1]  & lambda0 <= x[2],1,0))
```

```
> # how many ones are there?
> mean(lambda0.inside50)
```

```
## [1] 0.952314
```

Thus we see that with $n = 5$ and when $\lambda_0 = 3/2$ the probability coverage of the interval is approximately equal to 0.9363 whereas with $n = 50$ such a probability is 0.9523, much closer to the nominal value 0.95.

Exercise Use the fact that $n\bar{X} \sim \text{Poi}(n\lambda_0)$, to compute the coverage of the above intervals exactly.

2.2 Example 5.9 (Is Mendel's theory supported?)

We illustrate this problem by means of a practical example.

In one of his experiments with pea-plants Mendel crossed a certain number of green pod plants with yellow pod plants. The first generation (F1) he got only green plants (given green was the dominant trait colour). Successively plants of the F1 were let to self-pollinate leading thus to second generation (F2) plants. The F2 plants Mendel got where 39 green and 17 were yellow. Mendel wanted to know the proportion of the green plants.

Let us formalise this problem from a statistical point of view. Let Y_i denote a binary r.v. which takes value 1 if the ith F2 plant is green and takes 0 otherwise, let $\theta = P(X_i = 1)$, i.e. the probability of having a green plant among the F2 plants. It is reasonable to assume that the plants of the F2 are between them independent, thus we have that $Y_1, \ldots, Y_n$, with $Y_i \sim \text{Ber}(\theta)$, and in this case $n = 56$. The aim is then to compute a confidence interval for θ.

With the data above we have that

```
> # 36 green out of 56
> bar.y = 39/(39+17)
> se = sqrt((bar.y)*(1-bar.y)/56)
>
> # the 0.95 CI is
> bar.y + c(-1,1)*qnorm(alpha/2, low=F)*se
```

```
## [1] 0.5760019 0.8168553
```

According to Mendel's theory, since the green colour is dominant, it must appear in 75% of the plants. In other words, $\theta_0 = 0.75$. With 0.95 confidence we can say that Mendel's theory is supported.

A confidence interval for the probability of success θ can also be obtained by the `R` command `prop.test` as follows

```
> prop.test(x=39, n=39+17, correct = F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  39 out of 39 + 17, null probability 0.5
## X-squared = 8.6429, df = 1, p-value = 0.003283
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5666413 0.8009967
## sample estimates:
##         p
## 0.6964286
```

Note the 0.95 confidence interval here does not exactly coincide with the Wald-type confidence interval we computed above. This is because `prop.test` uses another approximate pivot which is based on the score function. We will not illustrate it here but we suffices to say that in large samples the two pivot will give

approximately equal results. By this command we can build also a confidence interval for a two sample problem as in Example 5.10.

2.3 Example 5.11 (Two normal populations with unequal variances)

This example can be handled similarly as in Example 5.3, where this time the option `var.equal = FALSE` in the command `t.test` must be specified.

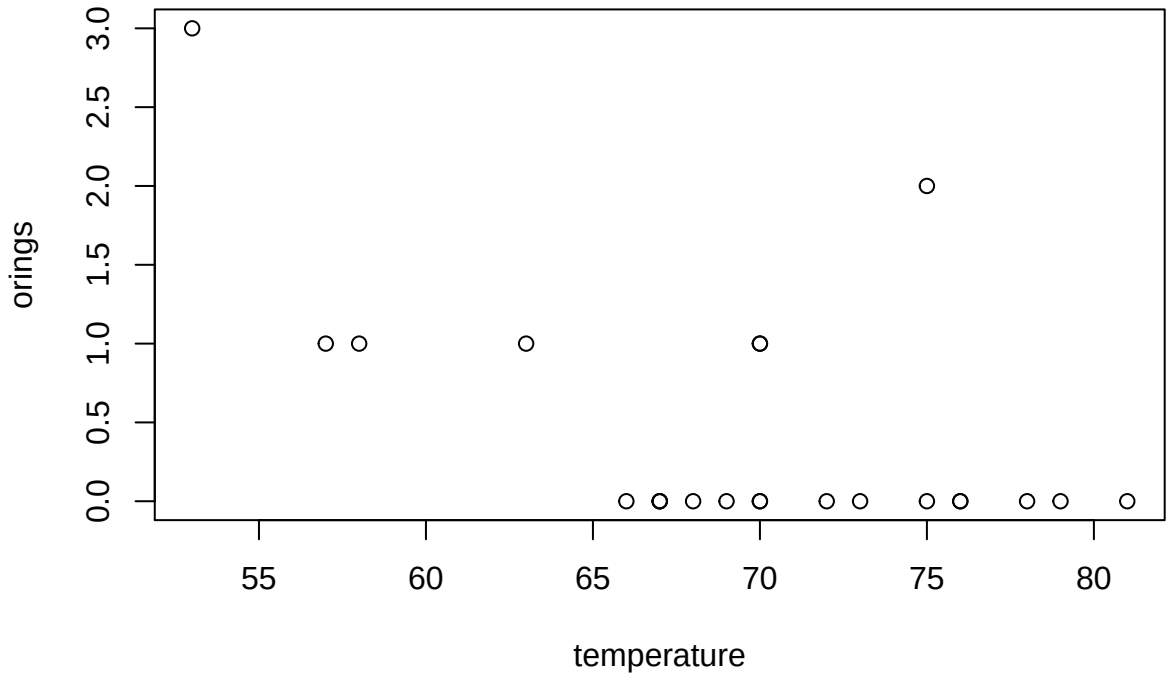
2.4 Was the Challenger disaster predictable?

Lastly, consider Example 2.8 (Lecture 2) but applied to a different problem, the space shuttle Challenger[1].

On January 28, 1986, a routine launch was anticipated for the space shuttle named Challenger. Seventy-three seconds into the flight, a disaster happened: the shuttle broke apart, killing all seven crew members on board. An investigation into the cause of the disaster focused on a critical seal of the solid rocket boosters, called O-ring, and it is believed that damage to these O-rings during a shuttle launch may be related to the ambient temperature during the launch. Below we summarize observational data on O-rings from 23 pre-Challenger shuttle missions, where the mission order is based on the temperature at the time of the launch. Also reported are the number of O-rings damaged after the launch (`orings`).

First we load and plot the data

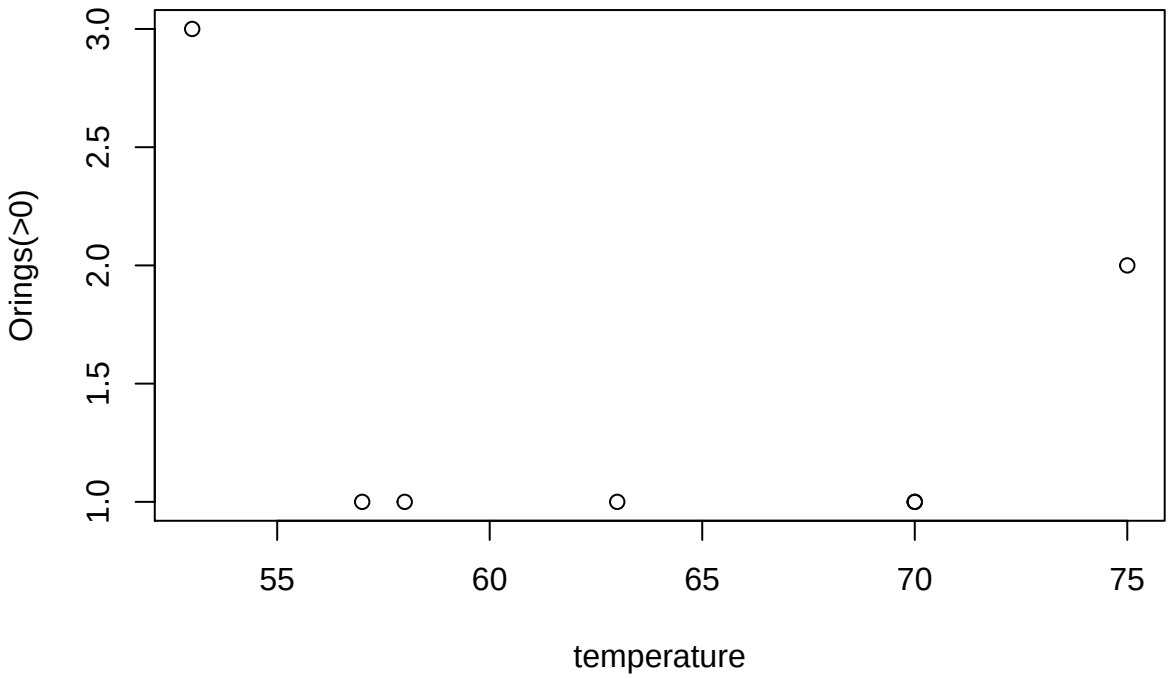
```
> challenger <- read.table("challenger2.dat", header = TRUE)
> plot(challenger)
> points(x=53, y=5, col=2, pch="*", cex=2)
```



The data we see were analysed by NASA's engineers the day before the Challenger disaster. There was some perplexity about launching the shuttle on that day, due to very low temperatures. Some of the engineers plotted a truncated version of the data, discarding all the zeros, as in the figure below.

[1]Dalal, Fowlkes and Hoadley (1989) Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure. Journal of the American Statistical Association Vol. 84, No. 408, pp.945-957. See also here https://www.space.com/18084-space-shuttle-challenger.html

```
> plot(challenger[challenger$orings>0,],
+      ylab = "Orings(>0)")
> points(x=53, y=5, col=2, pch="*", cex=2)
```



The conclusion was that there is no clear relationship between the number of broken o-rings and temperature, so they decided to launch the shuttle...

To formulate a statistical model, we assume that the launches are independent. Furthermore, note that each shuttle was equipped with 6 o-rings. A reasonable model for these data is the same as that of Example 2.8, where now Y_i is a $\text{Bin}(6, \theta_i)$ r.v. that reports the number of damaged o-rings; the probability of "success" θ_i depends on t_i, a fixed variable that gives the temperature at the ith shuttle launch. In particular, the assumed model is

$$\begin{aligned} \text{orings}_i &\sim \text{Bin}(6, \theta_i), \quad \text{with oring}_i \text{ independent from oring}_j, \text{ for all } i \neq j = 1, \ldots, 23, \\ \text{logit}(\theta_i) &= \alpha + \beta \text{temperature}_i, \quad i = 1, \ldots, 23, \end{aligned}$$

The parameter β concerns the relationship between the number of broken o-rings and the temperature; $\beta < 0$ implies that there is a negative relation.

Obviously the parameters are unknown and we will estimate them from the data. Since there is no exact pivotal quantity for this problem we will appeal to Theorem 4.9 for building a confidence interval for β.

First let us code the log-likelihood ourselves and then compute the MLE.

```
> logLBin <- function(theta, data){
+
+   n  = nrow(data)
+   alpha. = theta[1]
+   beta. = theta[2]
+
+   # compute the success prob. for each launch (observation)
+   thetai = plogis(alpha. + beta.*data$temperature)
+
```

```
+   # log-likelihood for each launch (observation)
+   oo = dbinom(x = data$orings, size = 6, prob = thetai, log=TRUE)
+
+   # sum all the log-likelihood contributions
+   return(sum(oo))
+ }
```

To maximise the likelihood function we take as starting value $(\alpha = 10, \beta = 0)$.

```
> start0 <- c(10, 0)
> (oo <- optim(par = start0, fn = function(x) -logLBin(x, data = challenger),
+              hessian = TRUE))
```

```
## $par
## [1]  6.7601709 -0.1398558
##
## $value
## [1] 16.37869
##
## $counts
## function gradient
##       77       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
##
## $hessian
##             [,1]      [,2]
## [1,]   8.402383   535.328
## [2,] 535.328034 34572.101
```

We see that the numerical optimisation routine converged and yelded:

- the MLE of $\theta = (\alpha, \beta)$ given by $\hat{\theta} \doteq (6.76, -0.14$ and $-\log L(\hat{\theta}) \doteq 16.38$
- the observed information matrix

$$J_n \doteq \begin{pmatrix} 8.4 & 535.33 \\ 535.33 & 3.45721 \times 10^4 \end{pmatrix}$$

Here it is an approximate (Wald-type) 0.95 confidence interval for β

```
> var.theta <- solve(oo$hessian)
> (se.beta <- sqrt(var.theta[2,2]))
```

```
## [1] 0.04634533
```

```
> oo$par[2] + c(-1,1)*qnorm(alpha/2, low=F)*se.beta
```

```
## [1] -0.2306910 -0.0490206
```

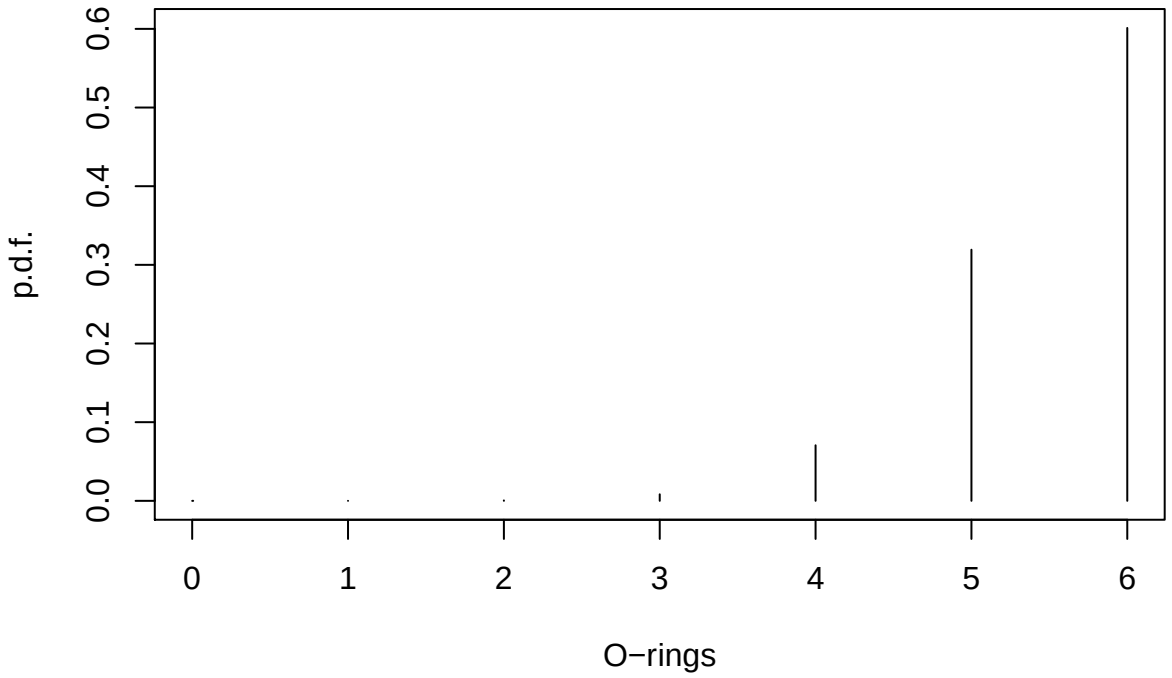
The night before the launch, NASA's engineers supposed there was relation between o-rings and temperature, that is, in terms of our model $\beta_0 = 0$. However, the data say that with an approximate 95% confidence, there is a negative relation between o-rings and ambient temperature, i.e. $\beta_0 < 0$; thus the lower the temperature the higher is on average the number broken o-rings.

For a given temperature value t, the estimated model is $\text{Bin}(6, \hat{\theta}(\text{t}))$, where $\hat{\theta}(t) = \frac{e^{\hat{\alpha}+\hat{\beta}t}}{1+e^{\hat{\alpha}+\hat{\beta}t}}$. By the way, the temperature at the launch of the Challenger shuttle was 31°F. So at the day of the lunch, the estimated distribution of the number of failures of o-rings is $\text{Bin}\left(6, \frac{e^{\hat{\alpha}+\hat{\beta}\cdot 31}}{1+e^{\hat{\alpha}+\hat{\beta}\cdot 31}}\right)$ and is plotted below

```
> # estimated probability of "success"
> (theta.t <- plogis(oo$par[1] + oo$par[2]*31))
```

```
## [1] 0.9186872
```

```
> plot(x = 0:6, dbinom(0:6, 6, theta.t), type="h",
+      xlab="O-rings", ylab="p.d.f.")
```



We can also calculate an estimated probability for the failure of the Challenger space shuttle at $t = 31$, this is just the sum of the probabilities that more than 3 o-rings fail, i.e.

```
> sum(dbinom(4:6,6,theta.t))
```

```
## [1] 0.9910897
```

```
> # or also by
> 1-pbinom(3,6, prob = theta.t,)
```

```
## [1] 0.9910897
```

which is sadly close to 1.

The model we fitted so far is called logistic regression and this can also be performed using built-in `R` command, `glm` as follows.

```
> oo2 <- glm(cbind(orings, 6-orings)~temperature, data=challenger,
+            family = binomial())
> summary(oo2)
```

```
##
## Call:
## glm(formula = cbind(orings, 6 - orings) ~ temperature, family = binomial(),
##     data = challenger)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9876  -0.7798  -0.4987  -0.2975   2.7483
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.75183    2.97989   2.266  0.02346 *
## temperature -0.13971    0.04647  -3.007  0.00264 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.761  on 22  degrees of freedom
## Residual deviance: 19.093  on 21  degrees of freedom
## AIC: 36.757
##
## Number of Fisher Scoring iterations: 5
```

The description of this command is lengthy and outside the scope of this course, but just notice how close are the estimates and the standard errors provided by `glm` with ours. The rest of the output need not concern us for the moment. We will see what the columns Z and Pr(>|Z|) are useful for in Lecture 6.

Inferential Statistics **First semester, 2020**

Lecture 6: Hypothesis testing

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

It is known that the population of WM's has a finite average energy consumption μ. A manufacturer of motors for WM's claims that his next generation motors (called NGM1) are energetically more efficient than, while achieving the same performance as, the currently available motors. We take a group of WM's with the old motor and a group of WM's with NGM1 and measure their energy consumptions. Consider the following two hypothesis:

Null Hypothesis: the average energy consumption is the same in the two groups.
Alternative hypothesis: the average energy consumption is not the same in the two groups.

If the NGM1 group has much lower (or higher) energy consumption we will reject the null hypothesis and conclude that the evidence favours the alternative hypothesis. This is an example of hypothesis testing.

Other similar questions in scientific activity are: is a new drug effective? Does a lot of manufactured items contain an excessive number of defectives? Is the mean lifetime of a component at least some specified amount? Ordinarily, information about such phenomena can be obtained only by performing experiments whose outcomes have some bearing on the hypotheses of interest.

More formally, suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta_1. \tag{6.1}$$

We call H_0 the null hypothesis and H_1 the alternative hypothesis. Let X be an r.v. with range $\mathcal{X}$. We test a hypothesis by finding an appropriate subset of outcomes $R \subset \mathcal{X}$, called

the *rejection region* or *critical region*. If the outcome of the experiment is in R, i.e. $X \in R$, we reject the null hypothesis, otherwise, we do not reject the null hypothesis:

$$X \in R \Longrightarrow \text{reject } H_0$$
$$X \notin R \Longrightarrow \text{retain (do not reject) } H_0.$$

In most practical cases the rejection region R is of the form

$$R = \{X : T(X) \geq c\},$$

where $T(X)$ is a *test statistic* and c is a *critical value*. The problem in hypothesis testing is to find an appropriate test statistic $T(X)$ and an appropriate critical value c.

Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggest that he is guilty. Similarly, we retain H_0 unless there is strong evidence to reject H_0. There are two types of errors we can make. Rejecting H_0 when H_0 is true is *called type I error*. Retaining (or accepting) H_0 when H_1 is true is called *type II error*. The possible outcomes for hypothesis testing are

(C1) accept H_0 when H_0 is true

(C2) reject H_0 when H_1 is true

(W1) reject H_0 when H_0 is true

(W2) accept H_0 when H_1 is true.

(C1) and (C2) are correct decisions, whereas (W1) and (W2) are wrong or incorrect decisions. The probability of each of the above incorrect decisions is called *size of the error*, which we state in the definition below. The size of the type I error is typically denoted by α and the size of the type II error is denoted by β. More formally, we have the following definition.

Definition 6.1 *The power function of a test with rejection region R is defined by*

$$\gamma(\theta) = P_\theta(X \in R).$$

The size of a test is defined to be

$$\alpha = \sup_{\theta \in \Theta_0} \gamma(\theta).$$

A test is said to have level α_0 *if* $\alpha_0 \leq \alpha$. *Furthermore, if* $\alpha_0 = \alpha$, *then we say that the test has size* α_0. *We also define the size of type II error by* $\beta(\theta) = 1 - \gamma(\theta)$ *for all* $\theta \in \Theta_1$.

A hypothesis of the form $\theta = \theta_0$ is called *simple hypothesis* because the underlying distribution is completely determined. A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a *composite hypothesis*. A test for hypothesis of the form

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0,$$

is called *two-sided* test. A test for

$$H_0 : \theta \leq \theta_0 \quad \text{against} \quad H_1 : \theta > \theta_0,$$

or

$$H_0 : \theta \geq \theta_0 \quad \text{against} \quad H_1 : \theta < \theta_0,$$

is called a *one-sided test*. The most common tests are two-sided.

The problem of hypothesis testing reduces to that of finding an *optimal* region R such that the size of the errors are minimised. In principle, an optimal R is the one for which $\alpha = \beta = 0$, that is, a rejection region which leads to zero errors. However this is impossible to achieve. Intuitively, we if wanted $\alpha = 0$ then we need to take R large enough in order to contain *all* the possible sample points, so that if H_0 is true we will never be able to reject H_0. But this would lead us to *always* accept H_0, even though H_1 is true, which implies $\beta = 1$. Thus we cannot have an optimal rejection region without placing further restrictions.

The restriction typically adopted is to fix α in advance. The basic problem of testing a hypothesis H_0 is then to find a critical region of size α that will minimise β. If there exists a critical region of size α that minimises β among all critical regions whose size does not exceed α, it is called *best critical region of size* α. The value of α is often chosen taking into account practical considerations, and only critical regions of this size or less are permitted in the competition. A test that is based on such a best critical region is called a *best test of size* α. Since the size of type II error rate is one minus the power of the test, then a best test of size α is also also a test with highest power.

We would like then to be able to construct tests with highest power under H_1, among all size α tests and for every θ. Such a test, when it exists, is called *uniformly most powerful test* (UMP).

To illustrate these concepts, consider the following example about testing a simple null hypothesis H_0 against a simple alternative hypothesis H_1.

Example 6.1 *Let X be a discrete r.v. whose p.d.f. depends upon a parameter θ and assume that we wish to test $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$, on the basis of a single observed value x. Hence, $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. Let the p.d.f. of X for $\theta = \theta_0$ and $\theta = \theta_1$ be as in the following table.*

X	*0*	*1*	*2*	*3*	*4*	*5*
$f(x;\theta_0)$	*.02*	*.03*	*.05*	*.05*	*.35*	*.50*
$f(x;\theta_1)$	*.04*	*.05*	*.08*	*.12*	*.41*	*.30*

If we choose $\alpha = .05$, the possible critical regions of this size are the following three regions

(a) $R = \{0, 1\}$

(b) $R = \{2\}$

(c) $R = \{3\}$.

The value of β corresponding to these critical regions is (a) .91, (b) .92 and (c) .88. Among these critical regions, the region $R = \{3\}$ is therefore to be preferred because it leads to the lowest type II error rate, i.e it leads to highest power among the three critical regions of the same size. Before we can claim that it is the best critical region of size $\alpha = .05$, i.e. it is a UMP test, it is necessary to make certain that there is no better critical region of size less than .05. The non trivial critical regions are (d) $R = \{0\}$ and (e) $R = \{1\}$. Since the corresponding values of β are (d) .96 and (e) .95, it follows that the test based on choosing $x = 3$ as critical region is a most powerful test of size .05.

A UMP test, when it exists, can be constructed through the following famous result.

Theorem 6.1 (Neyman-Pearson Lemma) *Let $x_1, \ldots, x_n$ be an observed sample of size n from the random sample $X_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$ with p.d.f. $f(x;\theta)$. If there exists a critical region C of size α and a nonnegative constant k such that*

$$\frac{\prod_{i=1}^n f(x_i;\theta_1)}{\prod_{i=1}^n f(x_i;\theta_0)} \geq k \quad \textit{for points in } C \tag{6.2}$$

and

$$\frac{\prod_{i=1}^n f(x_i;\theta_1)}{\prod_{i=1}^n f(x_i;\theta_0)} < k \quad \textit{for points not in } C, \tag{6.3}$$

then C is a best critical region of size α.

Proof: To simplify notation let $x = (x_1, \ldots, x_n)$ and $dx = dx_1 \cdots dx_n$. Furthermore, we write $L_j(x) = \prod_{i=1}^n f(x_i;\theta_j)$, for $j = 1, 2$. Let C^* be any other critical region of size less then or equal to α. The two critical regions C and C^* may be represented by the sets of points labeled C and C^* in Figure 6.1. Their intersection is denoted by e and their non-intersecting parts by a and b, respectively.

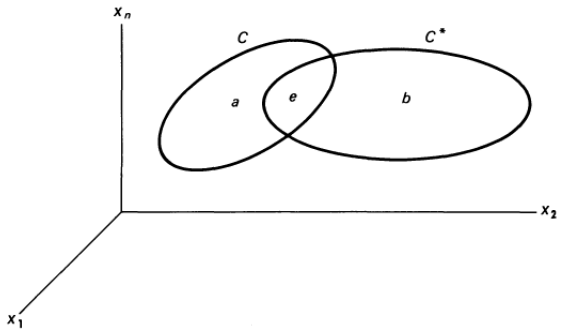


Figure 6.1: A graphical representation of the best critical region C compared to another critical region C^*.

Since C and C^* are critical regions of sizes α and $\leq \alpha$, respectively, it follows by the definition of the size of a critical region that

$$\int_C \cdots \int \prod_{i=1}^n f(x_i;\theta_0)\, dx_1 \cdots dx_n = \int_C L_0(x)\, dx = \alpha \tag{6.4}$$

and that

$$\int_{C^*} \cdots \int \prod_{i=1}^{n} f(x_i;\theta_0)\, dx_1 \cdots dx_n = \int_{C^*} L_0(x)\, dx \leq \alpha. \tag{6.5}$$

Hence

$$\int_{C} L_0(x)\, dx \geq \int_{C^*} L_0(x)\, dx. \tag{6.6}$$

Writing $C = a + e$ and $C^* = b + e$, we may cancel the integral over e from both sides of (6.6) to reduce it to

$$\int_{a} L_0(x)\, dx \geq \int_{b} L_0(x)\, dx. \tag{6.7}$$

Let β and β^* denote the sizes of the type II error for the critical regions C and C^*, respectively. Since the size of a type II error is the probability that the sample point will fall outside the critical region when H_1 is true, which in turn is equal to one minus the probability that will fall inside the critical region when H_1 is true, we may write

$$\beta = 1 - \int_{C} L_1(x)\, dx, \quad \text{and} \quad \beta^* = 1 - \int_{C^*} L_1(x)\, dx.$$

Hence

$$\begin{aligned} \beta^* - \beta &= \int_{C} L_1(x)\, dx - \int_{C^*} L_1(x)\, dx \\ &= \int_{a} L_1(x)\, dx - \int_{b} L_1(x)\, dx, \end{aligned} \tag{6.8}$$

where in the last equality we have cancelled the integral over the common part of C and C^*.

From the definition of C given in (6.2), it follows that $L_1(x) \geq kL_0(x)$ for all points in C, and hence for all points in a, and therefore that

$$\int_{a} L_1(x)\, dx \geq k \int_{a} L_0(x)\, dx.$$

Similarly, since b lies outside C, every point of b satisfies (6.3), namely $L_1(x) < kL_0(x)$; consequently

$$\int_{b} L_1(x)\, dx < k \int_{b} L_0(x)\, dx.$$

Applying these two results to (6.8) will yield the inequality

$$\beta^* - \beta \geq k \int_a L_0(x)\,dx - k \int_b L_0(x)\,dx. \tag{6.9}$$

But from (6.7) the right side must be nonnegative; therefore we arrive at the conclusion that

$$\beta^* \geq \beta.$$

Since β^* is the size of the type II error for any critical region, other than C, of size less then or equal to α, this proves that C is a best critical region of size α. ■

Here is an an illustration of how this theorem enables us to find a UMP test.

Example 6.2 *Let* $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$*, for* $i = 1, \ldots, n$ *and consider the problem of testing the hypothesis* $H_0 : \mu = \mu_0$ *against* $H_1 : \mu = \mu_1 > \mu_0$*, with* σ^2 *being known. Here*

$$\begin{aligned} \tfrac{L_1}{L_0} &= \frac{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu_1)^2\right]}{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu_0)^2\right]} \\ &= \exp\left[\tfrac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^n X_i + \tfrac{n(\mu_0^2-\mu_1^2)}{2\sigma^2}\right]. \end{aligned}$$

From (6.2) *it follows that the critical region* C *will be determined by the inequality*

$$\exp\left[\tfrac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^n X_i + \tfrac{n(\mu_1^2-\mu_0^2)}{2\sigma^2}\right] \geq k$$

for some constant $k > 0$*. Taking logarithms will yield the equivalent inequality*

$$\tfrac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^n X_i \geq \log k + \tfrac{n(\mu_1^2-\mu_0^2)}{2\sigma^2}.$$

It was assumed under H_1 *that* $\mu_1 > \mu_0$*, so the previous inequality reduces to*

$$\sum_{i=1}^n X_i \geq \tfrac{\sigma^2 \log k}{\mu_1-\mu_0} + \tfrac{n(\mu_1+\mu_0)}{2}.$$

Since k may be chosen to be any nonnegative number, as it ranges over values from 0 to ∞, the right side of this inequality will assume values from $-\infty$ to $+\infty$; therefore this inequality is equivalent to the inequality

$$\sum_{i=1}^{n} X_i \geq a,$$

where a may be chosen to be any real number. The equation $\sum_i X_i = a$ is that of a plane in n dimensional sample space. In Figure 6.2 the part of this plane with positive coordinates is sketched.

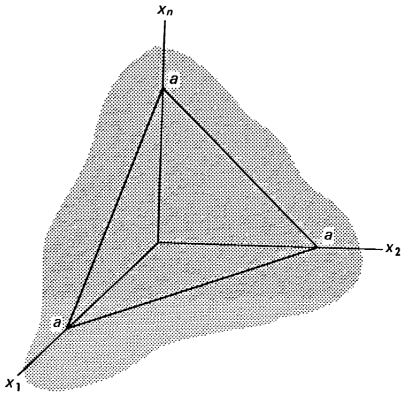


Figure 6.2: Graphical representation of the rejection region in Example 6.2.

The critical region C is therefore that part of the sample space which lies above this plane. As a assumes increasingly large numerical values, the region C includes increasingly less of the sample space, and as a goes from $-\infty$ to ∞ C shifts from including all of the sample space to none of it. Thus it is clear that by a proper choice of a, which is equivalent to a proper choice of k we can choose C to be a region of any desired probability size α, other than 0 or 1.

The selected value of α affects only the value of the constant a and not the shape of the region C; therefore a test based on this type of critical region must be a best test of its size, whatever is the size. Furthermore, since $\sum_i X_i \geq a$ is equivalent to $\overline{X} > a/n$, the best test here is equivalent to one that is based on the statistic $\overline{X}$ and which chooses as critical region the interval $\overline{X} \geq b$, where b is a constant selected to satisfy $P_\theta(\overline{X} \geq b|H_0) = \alpha$. The striking feature of this test is its simplicity, in that its critical region can be made to depend only upon the statistic $\overline{X}$ rather than upon the n dimensional r.ve. $X_1, \ldots, X_n$.

As a numerical illustration, consider the problem of testing for the mean energy consumption of a population of WM's being either 8 or 10. Suppose that the experience has shown that energy consumption may be treated as a normal variable with $\sigma = 2$ *and suppose that a random sample of size* $n = 16$ *yielded* $\overline{x} = 9$*. The problem then is to test the hypothesis* $H_0 : \mu = 8$ *against* $H_1 : \mu = 10$ *by means of the sample information. We shall choose* $\alpha = .05$*. Since the best test here is based on the critical region* $\overline{X} \geq b$*, where* b *is chosen to satisfy* $P_\theta(\overline{X} \geq b|H_0) = .05$*. Now when* $\mu = \mu_0 = 8$*,* $\overline{X} \sim \mathrm{N}(8, 2/\sqrt{16})$*, so* $Z = \frac{\overline{X}-8}{.5} \sim \mathrm{N}(0, 1)$*, when* H_0 *is true.*

Hence

$$\begin{aligned} P(\overline{X} \geq b|\mu = 8) &= P\left(\tfrac{\overline{X}-8}{.5} \geq \tfrac{b-8}{.5}\big|\mu = 8\right) \\ &= P(Z \geq \tfrac{b-8}{.5}). \end{aligned}$$

By the properties of the standard normal distribution we have that $P(Z \geq 1.645) = .05$*; consequently* b *must be chosen to satisfy the equation* $\frac{b-8}{.5} = 1.645$*, which is equivalent to* $b = 8.823$*. Our critical region of size* .05 *therefore consists of those sample points for which* $\overline{X} \geq 8.823$ *and the UMP test is thus*

$$\textit{Reject } H_0 \textit{ if } \overline{X} \geq 8.823.$$

The observed sample value $\overline{x} = 9$ *falls in this critical region, hence* H_0 *is rejected in favour of* H_1*.*

Example 6.3 (Example 6.2 cont.: Calculation of β**)** *Now let us evaluate* β*. For the example concerned with the testing a normal mean, assume again that* $\mu_0 = 8$*,* $\mu_1 = 10$*,* $\sigma = 2$*,* $n = 16$ *and* $\alpha = .05$*. The critical region for that problem was found to be* $\overline{X} \geq 8.823$*; therefore*

$$\beta = P_\mu(\overline{X} \leq 8.823|H_1).$$

Under H_1 $\overline{X}$ *is a normal r.v. with mean* 10 *and standard deviation* .5*, hence* $Z = \frac{\overline{X}-10}{.5} \sim \mathrm{N}(0, 1)$*. Using properties of the normal distribution we obtain*

$$\begin{aligned} \beta = P_\mu(\overline{X} \leq 8.823|H_1) &= P\left(\tfrac{\overline{X}-10}{.5} \leq \tfrac{8.823-10}{.5}|\mu = 10\right) \\ &= P(Z \leq -2.36) = .009. \end{aligned}$$

The geometrical meaning of α and β for this problem are displayed in Figure 6.3. Note that the power of the test γ is given by $1 - \beta$.

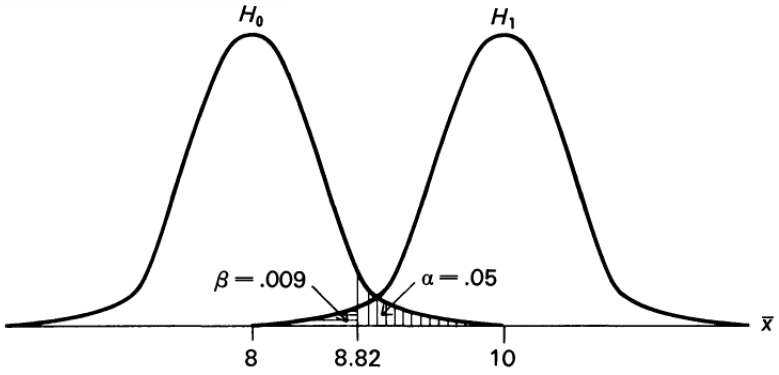


Figure 6.3: Geometrical representation of type I error and type II error when testing a normal mean.

The Neyman-Pearson Lemma is designed for testing simple null hypothesis versus a simple alternatives, i.e. $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. However, it can also be used to test one-sided hypothesis $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$, provided the model at hand satisfies the *monotone likelihood ratio* property.

Definition 6.2 *A family of distributions indexed by the real parameter θ is said to have a monotone likelihood ratio if there is a statistic T_n such that for each pair (θ, θ'), where $\theta > \theta'$, the likelihood ratio $L(\theta)/L(\theta')$ is a non decreasing function of T_n.*

Example 6.4 *Let $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$ where σ^2 is known. We wish to test $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$. Hence in this case we have $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, \infty)$, a composite null versus a composite alternative.*

To see if a UMP test exists we have to check if the model satisfies the monotone likelihood ratio property. For consider the pair (μ, μ'), where $\mu > \mu'$, then by the previous example we have that

$$\frac{L(\mu)}{L(\mu')} = \exp\left[\frac{\mu-\mu'}{\sigma^2}\sum_{i=1}^{n} X_i + \frac{n(\mu^2-(\mu')^2)}{2\sigma^2}\right].$$

With $T_n = \overline{X}$ we see that the monotone likelihood ratio property is satisfied since, $\mu - \mu' > 0$,

thus the NP Lemma tells us that the test:

$$\textit{Reject } H_0 \textit{ if } \overline{X} \geq c,$$

is UMP for $H_0 : \theta \leq \theta_0$ *against* $H_1 : \theta > \theta_0$*. To define the rejection region we apply the same reasoning as in Example 6.2. The rejection region is thus*

$$R = \{(X_1 \ldots, X_n) : \overline{X} \geq c\}.$$

Since H_1 *is composite, the power* γ *is now a function of* μ*. Thus we have*

$$\begin{aligned}
\gamma(\mu) &= P_\mu\left(\overline{X} \geq c\right) \\
&= P_\mu\left(\tfrac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \geq \tfrac{\sqrt{n}(c-\mu)}{\sigma}\right) \\
&= P\left(Z \geq \tfrac{\sqrt{n}(c-\mu)}{\sigma}\right) \\
&= 1 - \Phi\left(\tfrac{\sqrt{n}(c-\mu)}{\sigma}\right).
\end{aligned}$$

This function is increasing in μ *as it can be seen from Figure 6.4. Hence the size of the test is*

$$\textit{size } = \sup_{\mu \leq 0} \gamma(\mu) = \gamma(0) = 1 - \Phi\left(\tfrac{\sqrt{n}c}{\sigma}\right).$$

If we want a test of level α_0*, we set the size equal to* α_0 *and solve for* c *to get*

$$c = \tfrac{\sigma\Phi^{-1}(1-\alpha_0)}{\sqrt{n}}.$$

We reject when $\overline{X} \geq \sigma\Phi^{-1}(1-\alpha_0)/\sqrt{n}$ *or, equivalently, we reject when* $\frac{\sqrt{n}(\overline{X}-0)}{\sigma} \geq z_{\alpha_0}$*; we have used the fact that* $\Phi^{-1}(1-\alpha_0)$ *is the quantile function of the standard normal distribution evaluated at* $1-\alpha_0$*, which is also equal to the* z_{α_0}*, the upper* α_0*th quantile of* $Z \sim N(0,1)$*.*

Most powerful test do not always exist. For instance, in the above problem for the hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ there is no UMP test. Thus instead of going deeper into UPM tests we'll just consider three widely used tests: the Wald test, the χ^2 test, and the likelihood ratio test.

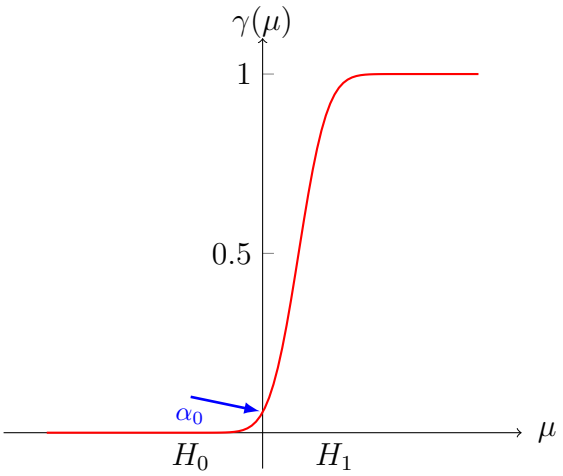


Figure 6.4: The power function for Example 6.4. The size of the test is the largest probability of rejecting H_0 when H_0 is true. This occurs at μ_0, hence the size is $\gamma(0)$. We choose the critical value c so that $\gamma(0) = \alpha_0$.

6.1 Wald test

Let θ be a scalar parameter, let $\widehat{\theta}$ be an estimate of θ and let $\widehat{\text{se}} = \widehat{\text{se}}(\widehat{\theta})$, be the estimated standard error of $\widehat{\theta}$. Consider testing

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0,$$

and assume that $\widehat{\theta}$ is asymptotically normal, i.e. for large n,

$$\widehat{\theta} \;\dot{\sim}\; \text{N}(\theta_0, \widehat{\text{se}}^2).$$

The Wald test of size α is to reject H_0 when $|W_n| > z_{\alpha/2}$, where $W_n = \frac{\widehat{\theta}-\theta_0}{\widehat{\text{se}}}$ is called the *Wald statistic*.

The Wald test has size α asymptotically, as $n \to \infty$. Indeed, under $H_0 : \theta = \theta_0$ we have that

$$\begin{aligned} P_{\theta_0}(|W_n| > z_{\alpha/2}) &= P_{\theta_0}\left(\tfrac{|\widehat{\theta}-\theta_0|}{\widehat{\text{se}}} > z_{\alpha/2}\right) \\ &\to P(|Z| > z_{\alpha/2}) \\ &= \alpha. \end{aligned}$$

where $Z \sim \text{N}(0,1)$.

Remark 6.1 *An alternative version of the Wald test statistic is $W_{n,0} = (\widehat{\theta} - \theta_0)/\text{se}_0$, where se_0 is the standard error, i.e. the standard deviation of $\widehat{\theta}$, computed at θ_0. Both versions of the test are valid and are asymptotically equivalent.*

Let us consider the power of the Wald test when the null hypothesis is false. Suppose that the true value is $\theta_\star \neq \theta_0$. The power $\gamma(\theta_\star)$, which is the probability of correctly rejecting the null hypothesis, is approximately equal to

$$1 - \Phi\left(\tfrac{\theta_0-\theta_\star}{\widehat{\text{se}}} + z_{\alpha/2}\right) + \Phi\left(\tfrac{\theta_0-\theta_\star}{\widehat{\text{se}}} - z_{\alpha/2}\right).$$

With all other things held equal, as $n \to \infty$ we see that the power goes to 1; recall that $\widehat{\text{se}} \to 0$ as $n \to \infty$ since $\widehat{\theta}$ is consistent. Furthermore, from the power function of the Wald test we also note that the power is large if $\theta_\star$ is far from θ_0.

Example 6.5 *(Comparing Two Prediction Algorithms). We test a prediction algorithm on a test set of size n_1 and we test a second prediction algorithm on a second test set of size n_2. Let X be the number of incorrect predictions for algorithm 1 and let Y be the number of incorrect predictions of algorithm 2. Then $X \sim \text{Bin}(n_1, \theta_1)$ and $Y \sim \text{Bin}(n_2, \theta_2)$. We wish to verify if the two algorithms give the same number of incorrect predictions, thus the null hypothesis is $H_0 : \theta_1 = \theta_2$ against the alternative $H_1 : \theta_1 \neq \theta_2$. Letting $\delta = \theta_1 - \theta_2$, these two hypotheses can also be stated as follows*

$$H_0 : \delta = 0 \quad \textit{against} \quad H_1 : \delta \neq 0.$$

Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be the MLE of θ_1 and θ_2, respectively. By the equivariance principle of the MLE, we have that the MLE of δ is $\widehat{\delta} = \widehat{\theta}_1 - \widehat{\theta}_2$. The standard error of $\widehat{\delta}$ can be found by

noting that $\widehat{\theta}_1$ *and* $\widehat{\theta}_2$ *are asymptotically independent normals, thus*

$$\widehat{\text{se}}(\widehat{\delta}) = \sqrt{\frac{\widehat{\theta}_1(1-\widehat{\theta}_1)}{n_1} + \frac{\widehat{\theta}_2(1-\widehat{\theta}_2)}{n_2}}.$$

The size α *Wald test is to reject* H_0 *when* $|W_n| > z_{\alpha/2}$ *where*

$$W_n = \frac{\widehat{\delta}-0}{\widehat{\text{se}}(\widehat{\delta})} = \frac{\widehat{\theta}_1-\widehat{\theta}_2}{\sqrt{\frac{\widehat{\theta}_1(1-\widehat{\theta}_1)}{n_1}+\frac{\widehat{\theta}_2(1-\widehat{\theta}_2)}{n_2}}}.$$

The power function of this test will be largest when θ_1 *is far from* θ_2 *and when* n_1 *and* n_2 *are large.*

What if we tested both algorithms on the same test set? The two samples are no longer independent. Instead we have to resort to the following strategy. Let $X_i = 1$ *if algorithm 1 is correct on test case* i *and* $X_i = 0$ *otherwise. Let* $Y_i = 1$ *if algorithm 2 is correct on test case* i*, and* $Y_i = 0$ *otherwise. Define* $D_i = X_i - Y_i$*. A typical dataset will be something like this:*

Test case	x_i	y_i	$d_i = x_i - y_i$
1	*1*	*0*	*1*
2	*1*	*1*	*0*
3	*1*	*1*	*0*
4	*0*	*1*	*-1*
5	*0*	*0*	*0*
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	*0*	*1*	*-1*

Let $\delta = E(D_i) = E(X_i) - E(Y_i) = P(X_i = 1) - P(Y_i = 1)$*. We can estimate* δ *by the sample average of* $d_1, \ldots, d_n$*, thus let* $\widehat{\delta} = \overline{d} = n^{-1}\sum_{i=1}^n d_i$*. Furthermore, we can estimate the sampling variance of* $\widehat{\delta}$ *by* s_d^2/n *where* $s_d^2 = (n-1)^{-1}\sum(d_i - \overline{d})^2$ *and thus set* $\widehat{\text{se}}(\widehat{\delta}) = \sqrt{s_d^2/n}$*. To test* $H_0 : \delta = 0$ *against* $H_1 : \delta \neq 0$ *we use the test statistic* $W_n = \widehat{\delta}/\widehat{\text{se}}(\widehat{\delta})$ *and reject* H_0 *if* $|W_n| > z_{\alpha/2}$*. This last is called a test for paired samples.*

Example 6.6 *(Nonparametric Comparison of Two Means). Let* $X_1, \ldots, X_m$ *and* $Y_1, \ldots, Y_n$ *be two independent random samples from populations with means* μ_1 *and* μ_2*, respectively,*

with both populations having finite variance. We are interested in testing the null hypothesis that $\mu_1 = \mu_2$. *Write this as* $H_0 : \delta = 0$ *against* $H_1 : \delta \neq 0$, *where* $\delta = \mu_1 - \mu_2$. *Let* $\widehat{\delta} = \widehat{\mu}_1 - \widehat{\mu}_2$, *where* $\widehat{\mu}_1 = \overline{X}$ *and* $\widehat{\mu}_2 = \overline{Y}$. *For large values of* m *and* n *the standard error of* $\widehat{\delta}$ *is*

$$\widehat{\text{se}} = \sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}.$$

The size α *Wald test rejects* H_0 *when* $|W_n| > z_{\alpha/2}$, *where*

$$W_n = \frac{\widehat{\delta} - 0}{\widehat{\text{se}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}.$$

There is close connection between Wald tests and Wald confidence intervals, which permits us to perform hypothesis testing through confidence intervals. Indeed, the size α Wald test rejects $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin \text{IC}_{1-\alpha}$, where $\text{IC}_{1-\alpha}$ is the $1 - \alpha$ confidence interval

$$\text{IC}_{1-\alpha} = [\widehat{\theta} \pm z_{\alpha/2}\, \widehat{\text{se}}].$$

Remark 6.2 *When we reject* H_0 *we often say that the result is statistically significant. A result might be statistically significant and yet the size of the effect might be practically or scientifically negligible. Thus a result could be statistically significant but not scientifically significant. The difference between statistical significance and scientific significance can be better understood in light the above connection between hypothesis testing and confidence intervals. Any interval that excludes* θ_0 *corresponds to a test which rejects* $H_0 : \theta = \theta_0$. *But the values in the interval could be close to* θ_0 *(not scientifically significant) or far from* θ_0 *(scientifically significant); see Figure 6.5. The message from this figure is that statistical significance does not imply that the finding is of scientific importance. Furthermore, confidence intervals are often more informative than tests.*

6.2 p-values

Reporting "reject H_0" or "accept H_0" is not very informative. Instead, we could try to see, for every α, whether the test rejects at that level. Generally, if the test rejects at level α it will also reject at level $\alpha' > \alpha$. Thus, there is smallest α at which the test rejects and we call this number the p-value.

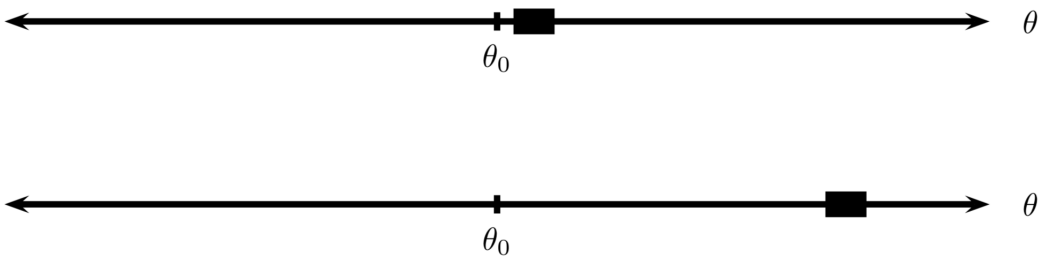


Figure 6.5: Scientific significance against statistical significance. A level α test rejects $H_0 : \theta = \theta_0$ if and only if the $1 - \alpha$ confidence interval does not include θ_0. Here are shown two different confidence intervals, both excluding θ_0, so in both cases the test would both reject H_0. But in case on top, the estimated value of θ is close to θ_0 so the finding is probably of little scientific or practical value. In the bottom case, the estimated value of θ is far from θ_0 so the finding is of scientific value.

Definition 6.3 *Suppose that for every $\alpha \in (0, 1)$ we have a size α test with rejection region R_α. Then,*

$$p\text{-value} = \inf\{\alpha : T(X_1, \ldots, X_n) \in R_\alpha\}.$$

That is, the p-value is the smallest size at which we can reject H_0.

Informally, the p-value is a measure of evidence against H_0: the smallest the p-value, the stronger is the evidence against H_0. Typically, researchers use the following evidence scale:

- p-value $< .01 \Rightarrow$ very strong evidence against H_0.
- p-value $\in [.01, .05) \Rightarrow$ strong evidence against H_0.
- p-value $\in [.05, .1) \Rightarrow$ weak evidence against H_0.
- p-value $> .1 \Rightarrow$ little or no evidence against H_0.

Be aware that a large p-value is not strong evidence in favour of H_0. A large p-value can occur because (i) H_0 is true or (ii) H_0 is false but the test has low power. Another common confusion about the p-value is that it is sometimes interpreted as the probability of H_0 being true given the data. This is clearly false.

The following result shows how the p-value is computed.

Theorem 6.2 *Suppose that the size α test is of the form:*

$$\textit{reject } H_0 \textit{ if and only if } T(X_1, \ldots, X_n) \geq c.$$

Then,

$$p\text{-value} = \sup_{\theta \in \Theta_0} P_\theta \left(T(X_1, \ldots, X_n) \geq t(x_1, \ldots, x_n) \right).$$

If $\Theta_0 = \{\theta_0\}$ *then*

$$p\text{-value} = P_{\theta_0}(T(X_1, \ldots, X_n) \geq t(x_1, \ldots, x_n)).$$

In words, Theorem 6.2 says that the p-value is the probability under H_0 of observing a value of the test statistic the same as or more extreme than what was actually observed.

In the case of the Wald test, if we let $w_n = (\widehat{\theta} - \theta_0)/\widehat{\text{se}}$ denote the observed value of the Wald statistic W_n, the p-value is given by

$$p\text{-value} = P_{\theta_0}(|W_n| \geq |w_n|) \doteq P(|Z| \geq |w_n|) = 2\Phi(-|w_n|).$$

This is further illustrated in Figure 6.6.

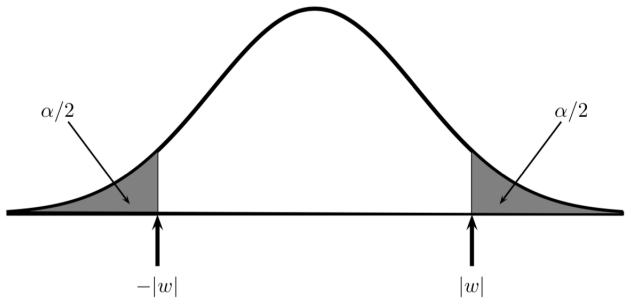


Figure 6.6: The p-value is the smallest α at which you would reject H_0. To find the p-value for the Wald test, we find α such that $|w_n|$ and $-|w_n|$ are just at the boundary of the rejection region. Here w_n is the observed value of the Wald statistic. Thus the p-value is the tail area $P(|Z| \geq |w_n|)$.

Note that the p-value depends on the observed data through w_n, thus it is an r.v. We state this more formally by the following theorem.

Theorem 6.3 *If the test statistic has a continuous distribution, then under* $H_0 : \theta = \theta_0$*, the* p-value *has distribution* $\text{Unif}(0, 1)$*. Therefore, if we reject* H_0 *when the* p-value *is less then* α*, the probability of a type I error is* α*.*

In other words, if H_0 is true, the p-value is like a random draw from a $\text{Unif}(0, 1)$ distribution. If H_1 is true, the distribution of the p-value will tend to concentrate closer to 0.

Example 6.7 *Regarding our motivating application about energy consumption of WM's, suppose the average energy consumption measured from* $m = 10$ *WM's with the old motor be* $\overline{x} = 216$ *and the let the sample variance be* $s_1^2 = 5$*. Furthermore, suppose that from a sample of* $n = 15$ *WM's with NGM1 we got* $\overline{y} = 213$ *and* $s_2^2 = 2.5$*. To verify that the average energy consumption of the population of WM's with the old motor* μ_1 *is equal to* μ_2*, the average energy consumption of the population of WM's with NGM1, consider* $H_0 : \delta = 0$ *against* $H_1 : \delta \neq 0$*, where* $\delta = \mu_1 - \mu_2$*. The observed Wald test statistic is*

$$w_n = \frac{\widehat{\delta}-0}{\widehat{\text{se}}} = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}} = \frac{216-195}{\sqrt{0.5+0.167}} = 3.43.$$

The p-value is given by

$$P(|Z| \geq 3.43) = 2P(Z \leq -3.43) = .0006$$

which is very strong evidence against the null hypothesis.

6.3 Pearson's χ^2 test for multinomial data

Recall that if the r.ve. $(X_1, \ldots, X_k) \sim \text{Mn}(k, \theta_1, \ldots, \theta_k)$, then the MLE of $\theta = (\theta_1, \ldots, \theta_k)$ is $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$, where $\widehat{\theta}_i = X_i/n$, with $n = \sum_{i=1}^k X_i$.

Let $\theta = (\theta_1, \ldots, \theta_k)$ and let $\theta_0 = (\theta_{01}, \ldots, \theta_{0k})$ be some fixed vector and suppose we want to test

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0.$$

Consider the statistic

$$T_n = \sum_{i=1}^k \frac{(X_j - n\theta_{0j})^2}{n\theta_{0j}} = \sum_{i=1}^k \frac{(X_j - E_j)^2}{E_j},$$

where $E_j = E(X_j) = n\theta_{0j}$ is the expected value of X_j under H_0. It can be shown that under H_0,

$$T_n \xrightarrow{d} \chi^2_{k-1} \quad \text{as } n \to \infty.$$

Hence the test: reject H_0 if $T_n \geq \chi^2_{k-1,\alpha}$, has asymptotic level α; here $\chi^2_{k-1,\alpha}$. The p-value is $P(\chi^2_{k-1} \geq t_n)$, where t_n is the observed value of test statistic T_n.

Example 6.8 *Consider again Mendel's experiment on peas, where round yellow seeds are breed with wrinkled green seeds. There are four type of progeny: round yellow, wrinkled yellow, round green, wrinkled green. The number of each type is a multinomial with probability* $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$*. His theory of inheritance predicts that* θ *is equal to*

$$\theta_0 = \left(\tfrac{9}{16}, \tfrac{3}{16}, \tfrac{3}{16}, \tfrac{1}{16}\right).$$

In $n = 556$ *trials he observed* $x = (315, 101, 108, 32)$*. We will test* $H_0 : \theta = \theta_0$ *against* $H_1 : \theta \neq \theta_0$*. Since* $n\theta_{01} = 312.75$*,* $n\theta_{02} = n\theta_{02} = 104.25$ *and* $n\theta_{04} = 34.75$*, the observed test statistic is*

$$t_n = \tfrac{(315-312.75)^2}{312.75} + \tfrac{(101-104.25)^2}{104.25} + \tfrac{(108-104.25)^2}{104.25} + \tfrac{(32-34.75)^2}{34.75} = 0.47.$$

With $\alpha = .05$ *the threshold is* $\chi^2_{3,.05} = 7.815$*. Since 0.47 is not larger than 7.815 we do not reject* H_0*. In addition, the p-value is equal to* $P(\chi^2_3 \geq 0.47) = .93$*, which is not evidence against* H_0*. Hence the data do not contradict Mendel's theory.*

6.4 The likelihood ratio test

The Wald test is mostly useful for testing a scalar parameter, although vector parameters can also be tested. The likelihood ratio test is more general and can be used for testing scalar and vector-valued parameters.

Consider testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \notin \Theta_0$. The *likelihood ratio statistic* is defined as

$$\lambda_n = 2 \log \left(\frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)} \right) = 2 \log \left(\frac{L(\widehat{\theta})}{L(\widehat{\theta}_0)} \right), \tag{6.10}$$

where $\widehat{\theta}$ is the MLE and $\widehat{\theta}_0$ is the restricted MLE in the space Θ_0. The *likelihood ratio test* is then

$$\text{if the observed } \lambda_n \text{ is greater or equal to } c \Longrightarrow \text{ reject } H_0.$$

Here c is a positive critical value such that $\sup_{\theta \in \Theta_0} P_\theta(\lambda_n \geq c) = \alpha$.

The rationale behind the likelihood ratio test is that $\sup_{\theta \in \Theta} L(\theta)$ will be greater than

$\sup_{\theta\in\Theta_0} L(\theta)$ when H_1 is true. Thus large values of

$$\Lambda_n = L(\widehat{\theta})/L(\theta_0),$$

or of the likelihood ratio statistic $\lambda_n = 2\log\Lambda_n$ indicate that the hypothesis H_1 is more likely to be true than H_0.

To determine the critical value c we need to derive the distribution of λ_n. In many cases this is not possible, but under mild regularity conditions on the model F_θ, it can be shown that λ_n has an asymptotic χ^2 distribution.

To see this formally, let's consider a more general case in which θ is a vector. In particular, let Θ_0 consists of all parameter values θ such that some coordinates of θ are fixed at some particular values and the rest are left free.

Theorem 6.4 *Suppose that* $\theta = (\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots, \theta_r)$. *Let*

$$\Theta_0 = \{\theta : (\theta_{q+1}, \ldots, \theta_r) = (\theta_{0,q+1}, \ldots, \theta_{0,r})\},$$

and let λ_n *be the likelihood ratio test statistic* (6.10). *Under* $H_0 : \theta \in \Theta_0$,

$$\lambda_n \xrightarrow{d} \chi^2_{r-q} \quad \text{as } n \to \infty,$$

where the degrees of freedom $r - q$ *are given by* $dim(\Theta) - dim(\Theta_0)$; $dim(S)$ *denotes the dimension of the space* S.

For example, if $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ and we want to test the null hypothesis that $\theta_4 = \theta_5 = 0$, then the limiting distribution of the likelihood ratio test has $5 - 3 = 2$ degrees of freedom. Or equivalently, the degrees of freedom in the χ^2 distribution of the likelihood ratio test are equal to the number of parameters being fixed under H_0.

The critical value for a size α likelihood ratio test is $c = \chi^2_{r-q,\alpha}$ and the p-value for this test is $P(\chi^2_{r-q} \geq \lambda_n^{obs})$, where λ_n^{obs} is the observed value of likelihood ratio statistic.

Example 6.9 *(Example 6.8 revisited) The observed value of the likelihood ratio test in the*

case of Example 6.8 is

$$\begin{aligned}\lambda_n &= 2\log\left(\tfrac{L(\widehat{\theta})}{L(\theta_0)}\right)\\ &= 2\sum_{j=1}^{k} x_j \log\left(\tfrac{\widehat{\theta}_j}{\theta_{0j}}\right)\\ &= 2\left(315\log\left(\tfrac{315/556}{9/16}\right) + 101\log\left(\tfrac{101/556}{3/16}\right) + 108\log\left(\tfrac{108/556}{3/16}\right) + 32\log\left(\tfrac{32/556}{1/16}\right)\right)\\ &= 0.48.\end{aligned}$$

Under H_1 there are four parameters. However, the parameters must sum to one, so the dimension of the parameter space is three. Under H_0 there are no free parameters so the dimension of the restricted parameter space is zero. The difference of these two dimension is three, so the limiting distribution of Λ_n under H_0 is χ^2_3 and the p-value is $P(\chi^2_3 \geq 0.48) = .92$. The conclusion is the same as with the Pearson's χ^2 test.

6.4.1 Likelihood confidence sets

Under suitable regularity conditions the likelihood ratio test λ_n is thus an asymptotic pivot. As seen in Lecture 5, this asymptotic pivot can be inverted to construct asymptotic confidence sets. We call these confidence sets likelihood confidence sets. If the parameter θ has p elements, we define a $1-\alpha$ likelihood confidence set for θ by the set

$$\{\theta : 2(\ell(\widehat{\theta}) - \ell(\theta)) \leq \chi^2_{p,\alpha}\}.$$

With respect to Wald confidence intervals, likelihood confidence sets tend to have probability coverage closer to the prescribed confidence level. Furthermore, typically they are to be found numerically since the λ_n is typically not invertible analytically.

6.5 Further examples of tests

6.5.1 Tests with exact null distributions

Example 6.10 (Example 5.4 revised) *As in Example 5.4, let $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ be a random*

sample of size n from a normal distribution with both parameters being unknown; thus $\theta = (\mu, \sigma^2)$. *We wish to test the hypothesis* $H_0 : \mu = \mu_0$ *against* $H_1 : \mu \neq \mu_0$. *Under* H_0, *the test statistic*

$$T_n = \frac{\sqrt{n}(\overline{X}-\mu_0)}{\sqrt{S^2}} \sim t_{n-1},$$

and the test which rejects H_0 *if* $|T_n| \geq t_{n-1,\alpha/2}$ *is called the t-test or Student t-test and has size* α. *The p-value is computed as* $P(|T_{n-1}| \geq t_n^{obs}) = 2P(t_{n-1} \geq |t_n^{obs}|)$; *here we use* t_n^{obs} *to denote the observed value of* T_n *in order to not confuse it with* t_{n-1}, *the t-Student distribution with* $n-1$ *degrees of freedom.*

Equivalently, the test accepts H_0 *if*

$$\begin{aligned} |T_n| \leq t_{n-1,\alpha/2} &\iff |\overline{X} - \mu_0| \leq t_{n-1,\alpha/2}\sqrt{S^2/n} \\ &\iff -t_{n-1,\alpha/2}\sqrt{S^2/n} \leq \overline{X} - \mu_0 \leq t_{n-1,\alpha/2}\sqrt{S^2/n} \\ &\iff \overline{X} - t_{n-1,\alpha/2}\sqrt{S^2/n} \leq \mu_0 \leq \overline{X} + t_{n-1,\alpha/2}\sqrt{S^2/n}, \end{aligned}$$

i.e. if $\mu_0 \in [\overline{X} \pm t_{n-1,\alpha/2}\sqrt{S^2/n}]$, *the* $1-\alpha$ *confidence interval for* μ *includes* μ_0.

It can be shown that the Student t-test is also a likelihood ratio test for the same null and alternative hypothesis as above. Let's work this out.

Under H_0 *we have that*

$$\sup_{\theta\in\Theta_0} L(\theta) = \frac{\exp\left[-\frac{1}{2\widehat{\sigma^2_{\mu_0}}}\sum_{i=1}^n (X_i-\mu_0)^2\right]}{(2\pi)^{n/2}\widehat{\sigma^2_{\mu_0}}^{n/2}}, \tag{6.11}$$

where $\widehat{\sigma^2_{\mu_0}} = \sum_{i=1}^n (X_i - \mu_0)^2/n$. *Under* H_1 *we have*

$$\sup_{\theta\in\Theta} L(\theta) = \frac{e^{-n/2}}{(2\pi)^{n/2}}\left[\frac{\sum_i (X_i-\overline{X})^2}{n}\right]^{-n/2}. \tag{6.12}$$

Dividing (6.12) *by* (6.11) *gives the ratio of likelihoods*

$$\Lambda_n = \left[\frac{\sum_i (X_i-\overline{X})^2}{\sum_i (X_i-\mu_0)^2}\right]^{-n/2},$$

which has critical region $\Lambda_n \geq a$. *But this critical region is also equivalent to the critical region*

$$\begin{aligned}\frac{\sum_i(X_i-\mu_0)^2}{\sum_i(X_i-\overline{X})^2} \geq b &\iff \frac{n(\overline{X}-\mu_0)^2}{\sum_i(X_i-\overline{X})^2} \geq b-1 \\ &\iff \frac{n(\overline{X}-\mu_0)^2(n-1)}{\sum_i(X_i-\overline{X})^2} \geq (b-1)(n-1) \\ &\iff \frac{n(\overline{X}-\mu_0)^2}{S^2} \geq (b-1)(n-1) \\ &\iff T_n^2 \geq (b-1)(n-1).\end{aligned}$$

We thus see that the critical region of the likelihood ratio test is of the type $T_n^2 \geq d$ *or* $|T_n| \geq \sqrt{d} = c$*. In order to define a size* α *test it is sufficient then to find* c *such that* $P(t_{n-1} \geq c) = \alpha/2$*. This is clearly given by* $t_{n-1,\alpha/2}$*.*

As a numerical example suppose $x_1, \ldots, x_n$ *is a sample of energy consumption of* $n = 10$ *WM's for which* $\overline{x} = 201$ *and* $s^2 = 5^2$ *and suppose we wish to test* $H_0 : \mu = 200$ *against* $H_1 : \mu \neq 200$ *at the level* $\alpha = .05$*. Then*

$$t_n^{obs} = \frac{\sqrt{10}(201-200)}{5} = 2.$$

Since $t_{9,0.025} = 2.26$ *and 2 is not greater than 2.26 we do not reject* H_0 *at level* $\alpha = .05$*. Furthermore, the p-value for this observed statistic is* $2P(t_{n-1} > 2) = 0.077$*, which is only a mild evidence against* H_0*.*

Example 6.11 (Example 5.4 revised II) *As in Example 5.4, let again* $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ *be a random sample of size n from a normal distribution with both parameters being unknown. We wish to test the hypothesis* $H_0 : \sigma^2 = \sigma_0^2$ *against* $H_1 : \sigma^2 \neq \sigma_0^2$*. Under* H_0*, the test statistic*

$$T_n = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1},$$

rejects H_0 *at the level* α *if* $T_n > \chi_{n-1,\alpha/2}$ *or if* $T_n < \chi_{n-1,1-\alpha/2}$*; The p-value is computed as* $2\min\{P(\chi^2_{n-1} > t_n^{obs}), 1 - P(\chi^2_{n-1} > t_n^{obs})\}$*. The rationale for this p-value is that we may reject* H_0 *either because the sample variance is greater than* σ_0^2*, in this case* $P(\chi^2_{n-1} > t_n^{obs})$ *is low or because the sample variance is much lower than* σ_0^2*. Alternatively, we reject* H_0 *if* σ_0^2 *is outside the equi-tailed* $1-\alpha$ *confidence interval obtained in Example 5.4.*

It can be shown that this test is also a likelihood ratio test for the same null and alternative hypothesis as above.

Let's consider a numerical example again. Suppose we wish to test $H_0 : \sigma^2 = 1$ *against* $H_1 : \sigma \neq 1$ *at the level* $\alpha = .05$. *From the observed with* $n = 10$ *suppose we got* $s^2 = 2^2$. *Then the observed value of the statistic* T_n *is* $t_n^{obs} = \frac{9\times 4}{1} = 36$. *Since* $\chi_{9,.025} = 19.02277$, *and* $t_n^{obs} > 19.02277$ *we reject* H_0. *On the other hand the p-value is*

$$p\text{-}value = \min[P(\chi_9^2 > t_n^{obs}), P(\chi_9^2 > t_n^{obs})] = 2\min(1 - 0.99996, 0.99996) = 2(1 - 0.99996).$$

Example 6.12 (Example 5.5. revised) *As in Example 5.5, let* $Y_i \overset{\text{iid}}{\sim} N(\mu_1, \sigma_1^2)$ *be a random sample of size* n *and* $X_j \overset{\text{iid}}{\sim} N(\mu_2, \sigma_2^2)$, *be an i.i.d. random sample of size* m *where we assume* $\sigma_1^2 = \sigma_2^2 = \sigma^2$ *and* μ_1, μ_2, σ^2 *are unknown. Now we wish to test* $H_0 : \mu_1 = \mu_2$ *against* $H_1 : \mu_1 \neq \mu_2$. *Considering the pivot given in Example 5.5, under* H_0 *we have that*

$$T_n = \frac{\overline{Y} - \overline{X}}{\sqrt{S_{pool}^2\left(\frac{1}{m} + \frac{1}{n}\right)}},$$

where S_{pool}^2 *is the pooled variance (see L5, p.8). Thus the test rejects* H_0 *for values* $|T_n| \geq t_{n-1,\alpha/2}$. *This test is called the two-sample t-test and has level* α. *The p-value is computed by* $P(|t_{n-1}| \geq t_n^{obs})$, *where* t_n^{obs} *is the observed value of* T_n.

By a reasoning similar to that applied in Example 6.10 it is easy to see that H_0 *will be rejected if the* $1 - \alpha$ *confidence interval*

$$\left[\overline{Y} - \overline{X} \pm t_{n-1,\alpha/2}\sqrt{S_{pool}^2\left(\tfrac{1}{m} + \tfrac{1}{n}\right)}\right]$$

does not contain 0.

It is also possible to show that this test is also a likelihood ratio test. This sheds some light on where did the pivot introduced in L5 come from.

In applied work it often happens that one has measurements under k different experimental conditions and interest is on testing if the means of the k experimental conditions are equal. This problem is known as the Analysis of Variance (ANOVA).

Example 6.13 (The ANOVA test) *Following Example 2.6, and changing slightly notation in order deal with a more general situation, let* y_{ij} *be the observed measurements across the* $j = 1, \ldots, k$ *experimental conditions and the* $i = 1, \ldots, n_j$ *replications or sample units at each experimental condition. A typical dataset for this problem looks like the table below.*

Conditions				
1	*2*	*3*	$\cdots$	k
y_{11}	y_{12}	y_{13}	$\cdots$	y_{1k}
y_{21}	y_{22}	y_{23}	$\cdots$	y_{2k}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$y_{n_1 1}$	$y_{n_2 2}$	$y_{n_3 3}$	$\cdots$	$y_{n_k k}$

In Example 2.6 we assumed each variable to have normal distribution with mean μ_j and variance σ^2. In this case we have k variables (or treatments, in the ANOVA jargon), thus we have

$$Y_{ij} = \mu_j + \epsilon_{ij}\,, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, k$$
$$\epsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2),$$

where μ_j are often referred to as <u>treatment means</u>.

Let $\mu = \frac{1}{n}\sum_{j=1}^{k} n_j \mu_j$ be the overall mean, that is, the common mean of all treatments pulled in a single variable, where $n = \sum_{j=1}^{n} n_j$.

The ANOVA test is a test for the null hypothesis that all treatment means are equal against the alternative that at least two treatments have different means, i.e.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \textit{against} \quad H_1 : \mu_j \neq \mu_l, \ \textit{for some } j, l = 1, \ldots, k.$$

To find a suitable test statistic for these hypothesis let's first estimate the μ_j by the usual sample average estimator $\overline{Y}_{\bullet j} = n_j^{-1}(Y_{1j} + \ldots, Y_{n_j j})$; the filled dot is to remind us that we are summing over i's. Similarly, we can estimate μ by its sample average estimator $\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} Y_{ij} = \frac{1}{n}\sum_{j=1}^{k} n_j \overline{Y}_{\bullet j} = \overline{Y}$. Now consider the overall variability of the data around the estimator of the overall mean, which we call <u>total sum of squares</u> (SST), given

by

$$\begin{aligned} SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y})^2 &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}[(\overline{Y}_{\bullet j}-\overline{Y})+(Y_{ij}-\overline{Y}_{\bullet j})]^2 \\ &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{Y}_{\bullet j}-\overline{Y})^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{\bullet j})^2 \\ &= SSR+SSE, \end{aligned}$$

Thus the overall variability around the estimator of the overall mean μ is equal to the variability of treatments' means estimators around the estimator of the overall mean (SSR) plus the variability of the data around the estimators of the treatment means (SSE). Now, if H_0 is true, then treatments' means estimators $\overline{Y}_{\bullet j}$ would all be approximately equal to the overall mean estimate $\overline{Y}$, thus SSR would be approximately zero. On the other hand if $Y_{\bullet j}$ are all different, SSR would be high, dominating over SSE. Hence a test statistic based on the ratio SSR/SSE seems a reasonable choice.

Let S_j^2 be the sample variance of the variable for the j treatment, thus

$$S_j^2 = \tfrac{1}{n_j-1}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{\bullet j})^2,$$

and note that

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{\bullet j})^2 = \sum_{j=1}^{k}(n_j-1)S_j^2.$$

But since $(n_j-1)S_j^2/\sigma^2 \sim \chi^2_{n_j-1}$ and S_j^2's are independent, then by the closure with respect to addition property of the gamma distribution it follows that

$$SSE/\sigma^2 \sim \chi^2_{n-k}.$$

Furthermore, it can be shown that under H_0, $SSR/\sigma^2 \sim \chi^2_{k-1}$ and that SSE and SSR are independent. Therefore the statistic

$$F = \frac{\frac{SSR}{\sigma^2(k-1)}}{\frac{SSE}{\sigma^2(n-k)}} = \frac{SSR/(k-1)}{SSE/(n-k)},$$

has distribution $F_{k-1,n-k}$ and an α level test for $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ rejects the null hypothesis for values of $F \geq F_{k-1,n-k,\alpha}$. The p-value is computed as $P(F_{k-1,n-k} \geq F^{obs})$, where F^{obs} is the value of the F statistic at the observed data.

The quantities involved in the ANOVA test are often reported in a table such as the following

	d.f.	*SS*	*Mean Square*	F	*p-value*
Treatment	$k-1$	*SSR*	*SSR/(k-1)*	$\frac{SSR/(k-1)}{SSE/(n-k)}$	$P(F_{k-1,n-k} \geq F^{obs})$
Residual	$n-k$	*SSE*	*SSE/(n-k)*		
Total	$n-1$	*SST*			

As a numerical illustration consider the following problem. Gas mileages are recorded during a series of road tests with four new models of Japanese luxury sedans. We wish to test the null hypothesis that all four models, on the average, give the same mileage.

	model A	*model B*	*Model C*	*Model D*
	22	*28*	*29*	*23*
	26	*24*	*32*	*24*
		29	*28*	
$Y_{\bullet j}$	*24*	*27*	*29.67*	*23.5*
$(n_j - 1)S_j^2$	*8*	*14*	*8.66*	*.5*

The overall sample average is $\overline{Y} = 26.5$. The ANOVA for this data is thus

	d.f.	*SS*	*Mean Square*	F	*p-value*
Treatment	3	*61.34*	*20.45*	3.94	$P(F_{k-1,n-k} \geq 3.94) = 0.072$
Residual	6	*31.16*	*5.19*		
Total	9	*92.5*			

Since the p-value is higher than $\alpha = .05$ we do not reject H_0. On the other hand, $F_{3,6,.05} = 4.757$ and since the observed value of the F statistic is 3.94, again we do not reject H_0.

Remark 6.3 *If the ANOVA test rejects H_0, there are many different ways in which the means of the k variables could be different, i.e $\mu_1 \neq \mu_2$ or $\mu_1 \neq \mu_3$ or both, etc. Typically a practitioner is interested in these differences and not on H_0 per se, thus after H_0 has been*

rejected, it is often of interest to learn which of the pair of means is significantly different from zero. This can be done either through t-tests or confidence intervals for the difference of means $\mu_i - \mu_j$, for $i \neq j = 1, \ldots, k$. In ANOVA jargon this is known as post-hoc analysis.

Given that there are k means, we could compute at most $k(k-1)/2$ confidence intervals (or tests). But while the level of each confidence intervals is the nominal value $1-\alpha$, when we conduct simultaneously $k(k-1)/2$ confidence intervals, the joint confidence level is much lower. To see this let $\mathrm{IC}_{i,j}^{\alpha}$ denote the $1-\alpha$ confidence interval for $\delta_{i,j} = \mu_i - \mu_j$ for all $i \neq j = 1, \ldots, k$. Then, the probability that all confidence intervals contain their true parameter value is

$$
\begin{aligned}
P\left(\bigcap_{\substack{i,j=1\\ i<j}}^{k} \left\{\delta_{i,j} \in \mathrm{IC}_{i,j}^{\alpha}\right\}\right) &= P\left(\left\{\delta_{1,2} \in \mathrm{IC}_{1,2}^{\alpha}\right\} \cap \cdots \cap \left\{\delta_{k-1,k} \in \mathrm{IC}_{k-1,k}^{\alpha}\right\}\right) \\
&= 1 - P\left(\bigcup_{\substack{i,j=1\\ i<j}}^{k} \left\{\delta_{i,j} \notin \mathrm{IC}_{i,j}^{\alpha}\right\}\right) \\
&\geq 1 - \sum_{\substack{i,j=1\\ i<j}}^{k} \alpha \\
&= 1 - k(k-1)\alpha/2.
\end{aligned}
$$

Thus the probability that all the confidence intervals contain their true parameter is far from the nominal value. What's even worse is that, for large k the joint confidence level decreases to zero. The message from this is that when conducting a large number of confidence intervals (or test) simultaneously we could end up by finding a significant results just by luck. This is known as the simultaneous inference or the multiple comparison problem.

There are many solutions to the simultaneous inference problem in the statistical literature, but Bonferroni is the easiest. In post-hoc analysis and when the desired confidence level is, say 95%, the Bonferroni's method is to fix $\alpha < 5\%$, depending on the number of confidence intervals (or hypothesis tests) performed. For instance if there are k variables, and after rejecting H_0 one is interested in the confidence intervals for all $\delta_{i,j}$, then Bonferroni's methods is to use as α the value $\alpha_B = 2\alpha/k(k-1)$. Indeed, with α_B in place of α we see that the joint confidence level of the intervals is no less than $1-\alpha$. The price paid for correcting

for simultaneous inference is: wider confidence intervals for $\delta_{i,j}$. For instance, if $k = 4$ and we wanted simultaneous inference with confidence 95%, then the confidence intervals for $\delta_{i,j}$ should be built using $\alpha = \alpha_B \doteq 0.0083$.

6.5.2 Tests based on the likelihood ratio

Let us now see some examples of hypothesis testing performed via the likelihood ratio test.

Example 6.14 (Poisson population). *Consider again Example 5.8. Let $Y_i \overset{\text{iid}}{\sim} \text{Poi}(\theta)$, $i = 1, \ldots, n$, and suppose we wish to test $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$. Let $\widehat{\theta} = \overline{Y}$ and consider the likelihood ratio*

$$\Lambda_n = \frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)} = \frac{e^{-n\widehat{\theta}} \widehat{\theta}^{\sum_{i=1}^n Y_i}}{e^{-n\theta_0} \theta_0^{\sum_{i=1}^n Y_i}} = e^{n(\theta_0 - \overline{Y})} \left(\frac{\overline{Y}}{\theta_0} \right)^{n\overline{Y}}.$$

By Theorem 6.4 we know that under H_0,

$$\lambda_n = 2 \log \Lambda_n = 2n(\theta_0 - \overline{Y}) + 2n\overline{Y}[\log \overline{Y} - \log \theta_0] \dot{\sim} \chi^2_1, \quad \text{as } n \to \infty.$$

Thus an asymptotic α level likelihood ratio test is to reject H_0 if the statistic $2n(\theta_0 - \overline{Y}) + 2n\overline{Y}[\log \overline{Y} - \log \theta_0]$ computed at the observed sample is larger than $\chi^2_{1,\alpha}$. The p-value for this test is given by $P(\chi^2_1 \geq 2 \log \Lambda_n^{obs})$, where Λ_n^{obs} is the observed value of Λ_n.

As a numerical example consider the following number of bugs observed during the operation of a certain software installed on a server:

(5, 4, 1, 0, 0, 1, 1, 2, 1, 1)

Suppose we want to test $H_0 : \theta = 3$ vs $H_1 : \theta \neq 3$. Since the value of the statistic at this observed sample (7.884) is greater than $\chi^2_{1,.05}$ (3.841) we reject H_0. In addition, the p-value is .005, much lower than $\alpha = .05$ thus again we reject the null hypothesis.

We can also compute a likelihood confidence set for θ using Section 6.4.1. We show this set pictorially in Figure 6.14 in which a 95% confidence set is illustrated. In this example, the confidence set is an interval.

To obtain this confidence set we consider λ_n as a function of θ. We cut this function horizontally at the level $\chi^2_{1,\alpha}$, with $\alpha = .05$ and find the two points of intersection. The 95% likelihood confidence interval is then given by all values of θ which are within the two intersection points. The 95% confidence interval found is thus [0.93,2.52] and since 3, the value under H_0, is not included in this interval we reject H_0 and conclude again that the population mean is statistically different from 3.

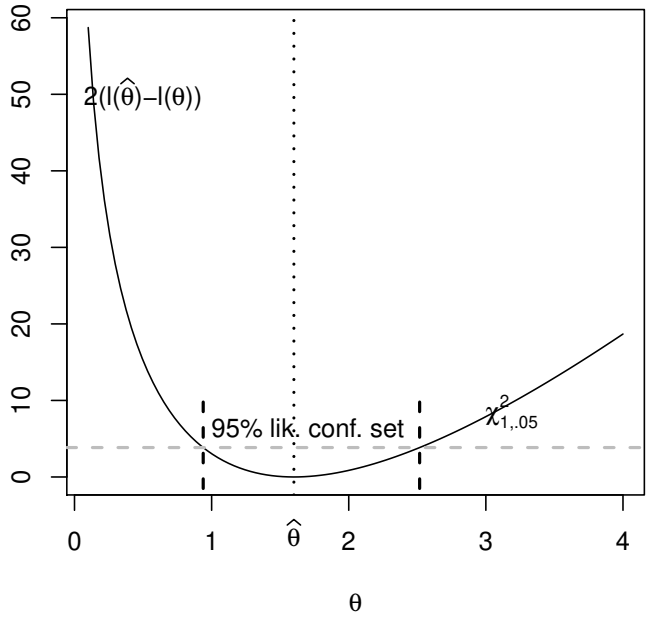


Figure 6.7: Graphical representation of a 95% likelihood confidence set for the mean θ of a Poisson population.

References

[LM18] Larsen, R. J. and Marx, M. L. (2018) *An Introduction to Mathematical Statistics and its Applications* (6th edition), Pearson Education, Inc..

Some exam-type exercises

Erlis Ruli (ruli@stat.unipd.it)

20 December 2020

The following are some problems that you may find on a final exam session of Inferential Statistics. The solution provided here are kept short for the sake of brevity. During the exam session you must be as detailed as possible, by justifying and explaining the reasoning you are applying in orderd to solve the problem.

Exercise 1

Let $X \sim F_\theta$ be a discrete r.v. with $\theta \in \Theta$, where the parameter space is $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$. The four possible distributions are as follows.

X	1	2	3
$f(x;\theta_1)$	.4	.1	.5
$f(x;\theta_2)$	.2	.1	.7
$f(x;\theta_3)$	.2	.4	.4
$f(x;\theta_4)$	.6	.3	.1

Let (X_1, X_2) be an i.i.d. random sample from F_θ.

(a) Compute the maximum likelihood estimator

(b) Compute the most powerful test of level $\alpha = .09$ for testing the hypthesis $H_0 : \theta = \theta_2$ against $H_1 : \theta = \theta_4$.

(c) Compute the power of the test under the hypothesis in (b).

(d) Perform a test of level $\alpha = .09$ for testing the hypthesis $H_0 : \theta = \theta_2$ against $H_1 : \theta \neq \theta_2$.

Solution

Exercise 2

Let $X_i \overset{iid}{\sim} \text{Wei}(\alpha, 1/\lambda)$, for $i = 1, \ldots, n$, $\alpha > 0, \beta > 0$. Note that $\text{Wei}(\alpha, 1/\lambda)$ has p.d.f.

$$f(x;\alpha,\lambda) = \alpha\lambda x^{\alpha-1} e^{-\lambda x^\alpha},\ x > 0.$$

Note:

$$E(X) = \frac{\lambda^{-1/\alpha}}{\alpha}\Gamma(1/\alpha),\ E(X^2) = \lambda^{-2/\alpha}\Gamma[(2+a)/a].$$

Assume $\alpha = 1$.

(a) Find $\hat{\lambda}_{MM}$, the method of moments estimator for λ.

(b) Compute the bias and the variance of $\hat{\lambda}_{MM}$.

(c) Is $\hat{\lambda}_{MM}$ consistent?

(d) Is $\hat{\lambda}_{MM}$ efficient ?

(e) Compute $\hat{\lambda}$, the maximum likelihood estimator for λ.

(f) Compute the exact and an approximate distribution of $\hat{\lambda}$.

(g) If possible find a UMP test for $H_0 : \lambda \leq \lambda_0$ aginst $H_1 : \lambda > \lambda_0$ with size α.

(h) Does there exists an UMP test for $H_0 : \lambda = \lambda_0$ vs $H_0 : \lambda \neq \lambda_0$? Why ?

(i) Compute an approximate confidence interval for λ.

(j) In a study about the lifetime of washing machines (measured in years), the observed sample of size $n = 20$ led to $\sum_{i=1}^{20} x_i = 9.849$. Get the p-value of the hypothesis $H_0 : \lambda = 1$ vs $H_0 : \lambda \neq \lambda_0$ using an exact test of size $\alpha = .05$ and compare it by the p-value obtained by an approximate test of the same size.

Solution