CHAPTER 6

COURSE IN STATISTICS

TABLE OF CONTENTS

Course in Statistics

The course in Statistics aims at providing an interpretation of the basic statistical concepts and methods, at an introductory level. Little previous knowledge is required, apart from the understanding of mathematical formulae.

The course is made up of five chapters, referred to by letters from A to E, from basic definitions to the estimation of discrete choice models.

Chapter A introduces the concepts of random variable, statistical population and units, samples and normal distribution. Chapter B is devoted to the binomial distribution, its asymptotic properties and the related multinomial distribution. Chapter C addresses the topic of data pooling, i.e. how to combine information from different sources. Chapter D deals with the statistical estimation of models, especially least squares estimation applied to linear regression models. Lastly, Chapter E addresses maximum likelihood estimation and its application to discrete choice models.

A Basic concepts

The lesson focuses on random variables and their application to statistical analysis. First we shall define a random variable (R.V.) and provide some traffic instances. Then we shall state the objectives and the basic methods of statistical analysis, from description of a population using statistical summaries, up to inference from random samples. Lastly we shall recall the definition and the elementary properties of the normal distribution.

A1 Random variable

A1a Intuitive definition

Let us consider a set Ω of elements ω, also called **disaggregate events**. An element ω may correspond to a given person, or object, or circumstance; then Ω corresponds to a population of persons or objects or circumstances.

We may also consider subsets A of elements: a subset is a group of disaggregate events, hence an aggregate event.

A **distribution of probability** P on Ω is a split of a unitary mass P(Ω) = 1 between all elements ω: then $P(\Omega) = \sum_{\omega \in \Omega} P(\omega)$. The number $P(\omega) \in [0,1]$ measures the relative importance of ω with respect to Ω and P. It is called the **relative frequency**, or **probability**, of ω.

When Ω is finite i.e. when its elements are in finite number N_Ω called the size of Ω, a well-known probability is the indifferent probability (also called equiprobability) such that each element has the same relative frequency $1/N_\Omega$.

The probability of an **aggregate event** A is defined as $P(A) = \sum_{\omega \in A} P(\omega)$. This measures the relative importance of A within Ω (and with respect to P).

In a human population with same relative frequency attributed to all individual persons, we may evaluate the proportion of male persons by the ratio of the number of them to the total size of the population: $P(A) = N_A / N_\Omega$.

A **random variable** X is a mapping of a set of elements Ω onto a set of values V. The noun "variable" means that alternative values $v_i \in V$ may be taken by $X(\omega)$ as ω ranges in Ω. Each value v_i may be thought of as a location, or site, in the space V.

When V is finite or denumerable, a random variable onto it is called a **discrete R.V.** When V is ordered, i.e. when we can compare any two values v_1 and v_2 and find which one is larger, an R.V. onto it is called an **ordered R.V.**

The adjective "random" means that we do not assume a perfect knowledge of X, because of either lack of information about Ω and the individual values $X(\omega)$, or of purposed omission.

We rather assume a limited amount of information, in the form of aggregate events $\{X \in V'\}$ for some subsets V' of V and the associated probabilities $P_X(V') = \Pr\{\omega; X(\omega) \in V'\}$.

The values taken by X in V are results or effects, as opposed to causes represented by the elements ω in Ω. A total, perfect knowledge of causes amounts to a deterministic model. A partial, limited knowedge of causes amounts to a random model. This enables us to model phenomena in which the exact causes are not known or difficult to describe in an accurate way.

Perhaps the simplest instance is the game of tossing a coin, with two possible outcomes Head and Tail. As it is difficult to explain for each outcome, the analyst prefers to describe the game by a random variable, characterized by the relative frequencies P(Head) and P(Tail).

A1b Traffic instances

In a population of persons ω, the number of trips performed on a given day by each of them.

In a population of households, the number of passenger cars owned (or used) by each of them.

In a given network link, the population of all the hours in a year, the R.V. "level of congestion" during one hour.

A1c Formal, axiomatic definition

A **measurable space** is a couple (Ω, τ_Ω) in which Ω is a set of elements denoted by ω, and τ_Ω is a set of subsets of Ω: each element A of τ_Ω is a group of elements ω of Ω.

The set τ_Ω is subject to the following requirements: $\Omega \in \tau_\Omega$; if $A \in \tau_\Omega$ then $\Omega \backslash A \in \tau_\Omega$; the union of any denumerable family I of elements A_i of τ_Ω belongs to τ_Ω. These requirements make the aggregation operation $A \cup B$ consistent.

A **probability space** is a triple (Ω, τ_Ω, P) in which (Ω, τ_Ω) is a measurable space and P is a mapping of τ_Ω on [0, 1] such that $P(\Omega) = 1$ and $P(\cup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$ for any finite or denumerable family I of disjoint elements A_i of τ_Ω.

Then P is called a probability on (Ω, τ_Ω). It associates a measure of relative mass to each subset A in τ_Ω, as compared to the largest subset Ω.

When $\Omega = \cup\{\omega_i\}$ is finite (or denumerable) the expression $P(\omega_i)$ makes sense and measures the relative importance of ω_i in Ω.

A **random variable** X is a mapping of a probability space (Ω, τ_Ω, P) called the set of events onto a measurable space (V, τ_V) called the set of values or sites, such that the reciprocal image $X^{-1}(V')$ of each element $V' \in \tau_V$ belongs to τ_Ω.

For $V' \in \tau_V$, the related probability $P(X^{-1}(V'))$ is also denoted by $P_X(V') = \Pr\{X(\omega) \in V'\}$. The mapping P_X of τ_V onto [0, 1] is a probability on (V, τ_V) and indicates the distribution of relative mass on V as induced by X.

A1d Comments on axiomatic formulation

The axiomatic theory of probability and random variables provides a consistent framework to derive the logical consequences from various combinations of relatively simple assumptions.

This framework is useful to analyze semantic models of physical or economic phenomena. Each semantic variable (eg. an electric intensity, a number of trips) may be interpreted as a

random variable. The axioms of probability enable one to study various combinations of random variables, including the sum, product, minimum of several basic R.V.s. They yield asymptotic properties for the results of large size random samples.

Of course, any application of the probability theory necessitates to check that the system under study satisfies the axiomatic assumptions.

A1e Partial, abridged description of R.V.

The precise amount of knowledge commonly assumed about a R.V. X is the following one: for each possible value $v \in V$, we know how many disaggregate events $\omega \in \Omega$ satisfy that $X(\omega) = v$, i.e. we know $P\{X = v\}$.

For a discrete R.V. this knowledge amounts to as many proportions p_i as there are values v_i in V. Then we can measure the probability of an aggregate event A as $P(A) = \sum_{\omega_i \in A} p_i$.

In the general case (discrete or continuous or mixed discrete-continuous), we can often avail ourselves of a **probability distribution function** (PDF), i.e. a mapping f_X of V onto $\Re_+$ such that $P(A) = \int_A f_X(v)\, dv$.

A simple instance is the uniform distribution on a real interval $[a, b]$: in this case $f_X(v) = \frac{1}{b-a}$ if $v \in [a, b]$ or 0 otherwise.

For an ordered R.V. X (eg. a real R.V.), we may also specify the probability of any interval $]-\infty, x]$: the associated function $F_X(x) = P\{X \in]-\infty, x]\} = P\{X \leq x\}$ is called the **cumulative distribution function** (CDF) of X.

The name of CDF derives from its relationship to the PDF: when a real R.V. X has a PDF f_X and a CDF F_X then $F_X(x) = \int_{]-\infty, x]} f_X(v)\, dv$.

In the case of a uniform distribution on $[a, b]$, $F_X(x) = \frac{x-a}{b-a}$ if $x \in [a, b]$ or 0 if $x \leq a$ or 1 if $x \geq b$.

A1f Conditional probability and independence

Let us consider the intersection $A \cap B$ of two aggregate events A and B. What is its relative importance as compared to A and B? A probabilistic answer to this question is to measure $P(A \cap B)$ and to compare it to P(A) and P(B).

When $P(B) > 0$ the conditional probability of event A with respect to B is

$$P(A|B) = P(A \cap B) / P(B)$$

also called the probability of A knowing B. The mapping $A \mapsto P(A|B)$ is a probability.

Two events A and B are independent when $P(A \cap B) = P(A)\, P(B)$, hence when $P(A|B) = P(A)$ if $P(B) > 0$.

Two R.V. X and Y are independent if for every couple of subsets (E, F) it holds that

$$P(X \in E \cap Y \in F) = P(X \in E)\, P(Y \in F).$$

A2 Statistical analysis

Statistics may be defined as "the science of collecting, classifying and interpreting information based on the numbers of things". Its two main objectives are the following:

(1) to provide an abbreviate description of statistical populations,

(2) to make use of partial measurement due to the observation of certain attributes only or certain individuals only.

A2a Definition and probabilistic setting

A statistical population is a collection of persons (original meaning) or objects of a same nature, called the statistical individuals or units. The number of individuals in the population is called the size of the population. Abstract populations may have infinite size.

In transportation a statistical individual may be a person, a household, a firm, a car, a truck, a trip, a unit of travelled distance, a traffic unit (eg. weight times distance).

Statistical individuals are described by their individual values of characters, also called attributes: each attribute takes one value for each individual, and by an individual's attribute it is meant the value of the attribute for the individual.

Instances of attributes include the age of a person, the income of a household, the origin point of a trip, its destination, the loading weight of a truck.

These statistical definitions are closely related to the probabilistic definitions, as demonstrated in the following table.

Tab. A. Correspondence between statistical and probabilistic concepts.

Statistical concept	Probabilistic concept
Population	Set of events Ω
Individual	Disaggregate event ω
Group	Event A
Character, attribute	Random variable
Value of attribute	Site of R.V.
Weight	Absolute frequency
Relative weight	Relative frequency, probability

A2b Statistical summaries

Statistical summaries provide an abridged, though incomplete, description of a statistical population. Let us consider a R.V. (attribute) X with value set V.

When V = $\{ v_1,..v_m \}$ is finite and has a small number of values v_k, we may describe the **individual probabilities** $p_k = \Pr\{X = v_k\}$.

Traffic instances include the proportions of cars and trucks in motorway traffic, of men and women in a population of passengers, of non-motorized households, of frequent users of public transport in a population of trip-makers.

When V is ordered, its distribution may be described by the fractiles: the **fractile** of X at level or order $\alpha \in [0, 1]$ is a value v_α in V such that $\Pr\{X < v_\alpha\} \leq \alpha$ and $\Pr\{X > v_\alpha\} \leq 1-\alpha$. A well-known fractile is the **median**, i.e. the fractile of order ½: it separates the distribution into two parts of equal weight.

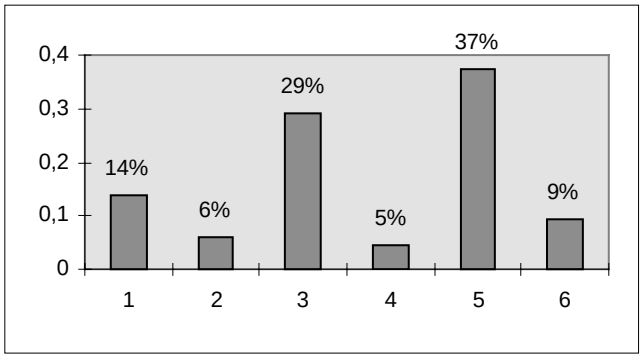
When V is imbedded in a vector space, we can define:

- The **mean** or average value, $\mathrm{E}[X] = \sum_{\omega \in \Omega} p(\omega) X(\omega)$: this measures the average location of X in V.
- The **variance** $\mathrm{var}[X] = \sum_{\omega \in \Omega} p(\omega)[X(\omega) - \mathrm{E}[X]]^2$ measures the average squared distance between X and the mean value $\mathrm{E}[X]$. It holds that $\mathrm{E}[X^2] = \mathrm{var}[X] + \mathrm{E}[X]^2$.
- The **standard deviation** $\sigma_X = \sqrt{\mathrm{var}\, X}$ square root of the variance, homogeneous to the mean.

All of these summaries involve real parameters. Some of them may be straightforwardly computed for compound real R.V.s:

- $\mathrm{E}[\lambda X + \mu Y] = \lambda \mathrm{E}[X] + \mu \mathrm{E}[Y]$.
- $\mathrm{Var}[\lambda X] = \lambda^2 \mathrm{var}[X]$.
- $\mathrm{var}[\lambda X + \mu Y] = \lambda^2 \mathrm{var}[X] + \mu^2 \mathrm{var}[Y] + 2\lambda\mu \,\mathrm{cov}[X, Y]$, where $\mathrm{cov}[X, Y] = \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]$ is the **covariance** of X and Y.

Numerical application. Compute the median, mean and variance of the distribution in the following figure. What is the fractile of order 0.8?



A2c Random samples and inference

A sample is a small separated part showing the quality of the whole (= a specimen). In statistics, a random sample S of size n consists in the observation of n individuals ω_i chosen from among a population in such a way that each individual ω has a non zero probability $p_i(\omega)$ of outcoming, i.e. to be the i-th result.

The random sample is simple if (1) the trials are independent and (2) for every trial i and event ω it holds that $p_i(\omega) = \mathrm{P}(\omega)$, i.e. the selection probability mirrors the relative frequencies in the population.

If we observe the values $X(\omega_i)$ of a real or vector-valued attribute X over the n sampled individuals, we may compute:

- the **empiric mean** $\hat{X}^{\text{emp}} = \frac{1}{n}\sum_{i=1}^{n} X(\omega_i)$,
- the **empiric variance** $\hat{S}^{2\text{emp}} = \frac{1}{n}\sum_{i=1}^{n} (X(\omega_i) - \hat{X}^{\text{emp}})^2$,

and use them to approximate the true mean and variance for the whole population.

We may also regard $\hat{X}^{\text{emp}}$ and $\hat{S}^{2\text{emp}}$ as random variables on the probability space Ω^n. Provided that the sample is random and simple, it holds that:

- $\mathrm{E}[\hat{X}^{\text{emp}}] = \mathrm{E}[X]$,
- $\mathrm{var}[\hat{X}^{\text{emp}}] = \frac{1}{n}\mathrm{var}[X]$,
- $\mathrm{E}[\hat{S}^{2\text{emp}}] = \frac{n-1}{n}\mathrm{var}[X]$.

These formulae make the foundation for **statistical inference**: given a random sample, $\hat{X}^{\text{emp}}$ will approximate E[*X*] whereas $\frac{n-1}{n}\hat{S}^{2\text{emp}}$ will approximate var[*X*]. Thus we may infer unknown parameters from partial observation, at the expense of some uncertainty which can be evaluated on an average basis.

Numerical illustration. Compute $\hat{X}^{\text{emp}}$ and $\hat{S}^{2\text{emp}}$ and proxies of E[*X*], var[*X*] and var[$\hat{X}^{\text{emp}}$] for the following simple random sample of size 8: 12, 7, 3, 11, 2, 8, 9, 4.

A3 The normal distribution

A3a Definition

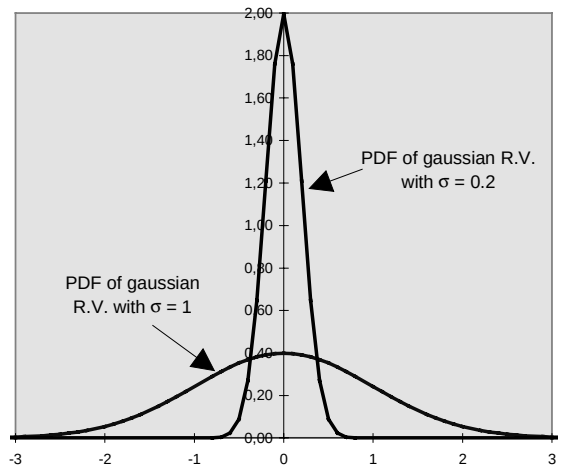
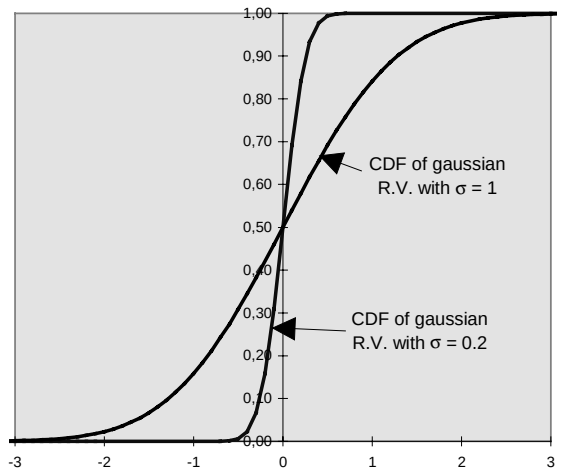
A real R.V. *X* has a gaussian, or normal, distribution with parameters μ and $\sigma > 0$ iff it has the following PDF:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

The graph of this function exhibits a characteristic bell shape, symmetric with respect to μ. The associated CDF has no closed-form analytic formula:

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \mathrm{d}t = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

where $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-t^2/2)\,\mathrm{d}t$.

Fig. a. Gaussian PDFs.	Fig. b. Gaussian CDFs.
	

A3b Elementary properties

The parameters μ and σ of a normal R.V. *X* correspond to its mean and standard deviation:

- E[*X*] = μ because of the symmetry,
- var[*X*] = σ^2 hence parameter σ determines the sharpness of the bell shape.

More precisely, a R.V. *X* is normal with parameters μ and σ iff the R.V. (*X* - μ)/σ is normal with parameters 0 and 1.

The sum of two **independent** normal R.V. *X* with parameters μ_X and σ_X and *Y* with parameters μ_Y and σ_Y is normal with mean $\mu_{X+Y} = \mu_X + \mu_Y$ and variance $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$.

A3c Instances

The normal distribution arises as the limit distribution for the sum or the mean of real R.V.s when their number increases.

It may also model a variable with known mean and variance.

A3d Computation

Standard statistical tables include the values of Pr(*U* > *z*) = 1-Φ(*z*) for a reduced gaussian variable *U* (with null mean and unit variance).

Standard statistical packages (including Excel) provide the PDF and CDF of a normal R.V. with any parameters μ and σ > 0.

A3e Distributions derived from the normal

Samples of gaussian R.V.s are important since they arise as limit distributions for large size samples (under some technical assumptions). Their study requires a knowledge of three "relatives" to the normal distribution, namely the chi-square distribution, the Fisher-Snedecor distribution and the Student distribution.

Let us consider a gaussian R.V. X with mean μ and variance σ^2. The reduced R.V. $U = (X - \mu)/\sigma$ is normal with null mean, unit variance and $E[U^4] = 3$.

A **chi-square** R.V. with ν degrees of freedom (DF) is the sum of ν independent R.V.s U_i^2, each of which is a squared reduced gaussian variable: $\chi^2_\nu = \sum_{i=1}^{\nu} U_i^2$.

Then it holds that $E[\chi^2_\nu] = \nu$ and $var[\chi^2_\nu] = 2\nu$. The PDF is

$$f_{\chi^2_\nu}(x) = \frac{1}{2}\frac{(x/2)^{\nu/2-1}\exp(-x/2)}{\Gamma(\nu/2)} \text{ if } x \geq 0 \text{ or 0 otherwise.}$$

The CDF is available in statistical tables or packages.

A **Fisher-Snedecor** R.V. $F(\nu_1, \nu_2)$ with two positive integer parameters ν_1 and ν_2 is a quotient $\frac{Y_1/\nu_1}{Y_2/\nu_2}$ in which Y_1 and Y_2 are independent chi-square variables with respectively ν_1 and ν_2 degrees of freedom. It holds that:

- $f_{\nu_1,\nu_2}(x) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B(\nu_1/2;\nu_2/2)} x^{\nu_1/2-1}(1+\frac{\nu_1}{\nu_2}x)^{-(\nu_1+\nu_2)/2}$ for $x > 0$ or 0 otherwise.
- $E[F(\nu_1,\nu_2)] = \frac{\nu_2}{\nu_2-2}$ if $\nu_2 > 2$.
- $var[F(\nu_1,\nu_2)] = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ if $\nu_2 > 4$.

A **Student R.V.** with ν degrees of freedom, denoted by T_ν, is a quotient $U/\sqrt{\chi^2_\nu/\nu}$ in which U and χ^2_ν are two independent R.V.s respectively reduced gaussian and chi-square with ν degrees of freedom. It holds that:

- $f_{T\nu}(x) = \frac{1}{\sqrt{\nu}B(1/2;\nu/2)}(1+\frac{1}{\nu}x^2)^{-(\nu+1)/2}$.
- $E[T_\nu] = 0$ if $\nu > 1$ (owing to symmetry).
- $var[T_\nu] = \nu/(\nu-2)$ if $\nu > 2$.

B The binomial and multinomial distributions

The binomial and multinomial distributions are useful to analyze the assignment of statistical individuals to a finite number of sites. Applications include quality control, opinion polls, marketing surveys and consumer analysis.

The lesson is made up of two parts. The first one deals with the binomial distribution, which arises when there are two sites only. The other one is devoted to the multinomial distribution, which extends the binomial to several sites. For both distributions we make a clear distinction between the probabilistic setting and the application to statistical inference.

B1 The binomial distribution

B1a Probabilistic setting: from Bernoulli to binomial

A Bernoulli R.V. X with parameter $p \in [0,1]$ is a R.V. with two sites of probability p and 1-p respectively. Typical couples of sites are 1/0, Success/Failure and so on. On considering the two sites 1 and 0, we obtain that

- $\mathrm{E}[X] = p$.
- $\mathrm{var}[X] = p(1-p)$.

A binomial R.V. $S_{n,p}$ with two parameters n a non-zero integer and $p \in [0, 1]$ is the sum of n independent Bernoulli R.V.s with sites 1 and 0 and same parameter p (i.e. independent, identically distributed R.V.s). Thus:

- $\mathrm{E}[S_{n,p}] = np$.
- $\mathrm{var}[S_{n,p}] = np(1-p)$.

The binomial distribution models the number of successes in a series of n independent trials. As the trials are independent, there are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ distinct trials which yield exactly k successes, and each of them has the individual probability $p^k(1-p)^{n-k}$. We also deduce that

- $\mathrm{P}\{S_{n,p} = k\} = \binom{n}{k} p^k(1-p)^{n-k}$ if $k \in \{0, 1, 2 \ldots n\}$ or 0 otherwise.

From the central limit theorem, when n is large the R.V. $\frac{S_{n,p} - \mathrm{E}[S_{n,p}]}{\sqrt{\mathrm{var}\, S_{n,p}}} = \frac{S_{n,p} - np}{\sqrt{np(1-p)}}$ is approximately reduced gaussian, hence the R.V. $S_{n,p}$ is approximately gaussian with mean np and variance $np(1-p)$.

B1b Statistical inference

The statistical inference problem associated with the binomial distribution is to recover an unknown proportion π from a simple random sample of n binary observations $y_i \in \{0, 1\}$.

The sample **empiric proportion** $\hat{p}^{\text{emp}} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n} S_{n,p}$ is an unbiased estimator of the true mean π. Its asymptotic distribution is gaussian with mean π and variance $\frac{1}{n}\pi(1-\pi)$. Under proviso that the conditions for normal approximation hold, we can also associate the following condidence interval to every confidence level 1-α:

$$\Pr(-u_{\alpha/2} < \frac{\hat{p}^{\text{emp}} - \pi}{\sqrt{\frac{1}{n}\pi(1-\pi)}} < u_{\alpha/2}) = 1\text{-}\alpha$$

in which $u_{\alpha/2}$ is the fractile of order 1-α/2 of a reduced normal variable *U*.

Several methods exist to overcome the dependency of the variance on π. The simplest one is to replace π by $\hat{p}^{\text{emp}}$ in the variance formula, yielding the approximate confidence interval

$$\Pr\{\, \pi \in \hat{p}^{\text{emp}} \pm u_{\alpha/2}\sqrt{\tfrac{1}{n}\hat{p}^{\text{emp}}(1-\hat{p}^{\text{emp}})}\,\} = 1\text{-}\alpha.$$

Another method is to solve the following quadratic inequality

$$\hat{p}^{\text{emp}} - \pi \le u_{\alpha/2}^2 \frac{1}{n}\pi(1-\pi), \Leftrightarrow \pi^2(1+\frac{u^2}{n}) - \pi(\frac{u^2}{n} + 2\hat{p}^{\text{emp}}) + \hat{p}^{\text{emp}2} \le 0$$

$$\Leftrightarrow \pi \in \frac{\hat{p}^{\text{emp}} + \frac{u^2}{2n} \pm \sqrt{\frac{u^4}{4n^2} + \hat{p}^{\text{emp}}(1-\hat{p}^{\text{emp}})\frac{u^2}{n}}}{1+u^2/n}.$$

A prior consistency condition is $n \ge 50$. Additional, posterior conditions are that $np > 5$ and $n(1\text{-}p) > 5$ in which $p = \frac{\hat{p}^{\text{emp}} + u^2/(2n)}{1+u^2/n}$.

B1c Illustration

A roadside interview survey was performed on a given road to reveal the proportion of drivers with a certain destination. 500 drivers were surveyed, among whom 35 had the desired destination. Estimate the true proportion of the destination in the total link traffic and compute the confidence intervals at confidence levels 1-α = 80% ; 90% ; 95% and 99%.

B2 The multinomial distribution

B2a Probabilistic setting

Let us consider a discrete R.V. *X* with *m* sites v_k with respective probabilities p_k. Let S_k be the number of occurrences of site v_k in a series of *n* independent trials of variable *X*. Thus S_k is a binomial R.V. with parameters *n* and p_k. The multinomial distribution with parameters *n* and $(p_k)_{k=1..m}$ is the distribution of the *m*-tuple $(S_k)_{k=1..m}$.

It can be shown that:

$$\text{P}(\textstyle\bigcap_{k=1}^{m}\{S_k = n_k\}) = n!\prod_{k=1}^{m}\frac{(p_k)^{n_k}}{n_k!} \text{ if } \sum_{k=1}^{m} n_k = n \text{ or 0 otherwise.}$$

From this formula it is straightforward to compute $E[S_k S_\ell] = n(n\text{-}1)\, p_k\, p_\ell$ if $k \neq \ell$, hence $\mathrm{cov}[S_k, S_\ell] = -n\, p_k\, p_\ell$ if $k \neq \ell$.

From the central limit theorem, when n is large the random vector $(S_k)_{k=1..m}$ is approximately m-dimensional gaussian with mean $(p_k)_{k=1..m}$ and covariance matrix C as follows: diagonal terms $C_{kk} = \mathrm{var}[S_k / n] = \frac{1}{n} p_k (1 - p_k)$ and off-diagonal terms $C_{k\ell} = \mathrm{cov}[S_k / n; S_\ell / n] = -p_k\, p_\ell / n$ for $k \neq \ell$.

However this distribution is degenerate (it has no PDF) because $\sum_{k=1}^{m} S_k / n = 1$. By subtracting the last component S_m we obtain an (m-1)-tuple $(S_k)_{k=1..m-1}$ with nondegenerate asymptotic distribution. The inverse reduced covariance matrix is $\overline{C}^{-1}$ with diagonal terms $n(\frac{1}{p_k} + \frac{1}{p_m})$ and off-diagonal terms $1/\, p_m$.

A theorem on multivariate normal variables implies that a j-dimensional gaussian variable Y with mean μ and covariance matrix Σ gives rise to a R.V. $D^2 = (Y - \mu)^t \Sigma^{-1} (Y - \mu)$ which is distributed chi-square with j degrees of freedom.

Here we apply the theorem to the gaussian R.V. Y which approximates the reduced multinomial R.V.: in this case $D^2 = \sum_{k=1}^{m} \frac{(n_k - n p_k)^2}{n p_k}$.

B2b Statistical inference

A simple random sample of size n yields m empiric absolute frequencies $\hat{S}_k^{\mathrm{emp}} = n_k$ and m empiric relative frequencies $\hat{p}_k^{\mathrm{emp}} = n_k / n$. The vector of empiric proportions $(\hat{p}_k^{\mathrm{emp}})_{k=1..m}$ is an unbiased estimator of the true proportions $(\pi_k)_{k=1..m}$.

The asymptotic result involving a chi-square variable may be used to test a null hypothesis $H_0 = \{ \tilde{p}_k = \hat{p}_k^{\mathrm{emp}} \ \forall k \in 1..m \}$ against the alternative $H_1 = \{ \exists k \in 1..m,\ \tilde{p}_k \neq \hat{p}_k^{\mathrm{emp}} \}$.

The test consists in evaluating $d^2 = \sum_{k=1}^{m} \frac{(n_k - n\tilde{p}_k)^2}{n\tilde{p}_k}$ and computing the critical probability $\alpha = \Pr\{ \chi^2_{m-1} \geq d^2 \}$. A very low value of the critical probability α (1% or less) should push the analyst into rejecting the null hypothesis H_0.

C Pools and panels

It is seldom the case that the information on a complex system emanates from a single source. On the contrary there are numerous instances in which information from several sources must be combined in a consistent way: eg. O-D trip information obtained from several en-route surveys. This lesson focuses on systematic methods to aggregate and mix data from several sources, known as data pooling. It also deals with panel surveys, in which a given sample is observed twice, before and after a change is made.

The lesson is made up of five parts. Part 1 recalls the properties of the sum of several random variables and their application to sampling. Part 2 introduces stratified sampling, in which subpopulations are sampled independently to improve accuracy. Part 3 is devoted to the product of independent random variables, with application to the temporal expansion of en-route survey. Part 4 deals with the combination of several en-route surveys to estimate an O-D trip matrix. Lastly, Part 5 introduces panel surveys.

C1 The additive combination of random variables

C1a A reminder

For any two random variables X and Y and real constants λ and μ the following properties hold:

- $\mathrm{E}[\lambda X + \mu Y] = \lambda\,\mathrm{E}[X] + \mu\,\mathrm{E}[Y]$
- $\mathrm{var}[\lambda X + \mu Y] = \lambda^2\,\mathrm{var}[X] + \mu^2\,\mathrm{var}[Y] + 2\lambda\mu\,\mathrm{cov}[X,Y]$.

Furthermore, when X and Y are independent the last formula reduces to

- $\mathrm{var}[\lambda X + \mu Y] = \lambda^2\,\mathrm{var}[X] + \mu^2\,\mathrm{var}[Y]$.

C1b Application to sampling

Assuming two random samples $(X_i)_{i=1..n}$ and $(X_{n+i})_{i=1..m}$ of independent, identically distributed RVs X_i, their empiric means $\hat{X}^{\mathrm{emp}} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\hat{Y}^{\mathrm{emp}} = \frac{1}{m}\sum_{i=1}^{m} X_{n+i}$ are independent random variables. Both may be used to estimate the mean of X_i. However the best solution is to make use of all the available information by defining a combined estimator $Z_\lambda = \lambda\hat{X}^{\mathrm{emp}} + (1-\lambda)\hat{Y}^{\mathrm{emp}}$ for $\lambda \in [0,1]$.

By construct $\mathrm{E}[Z_\lambda] = \mathrm{E}[X_i]$. The variance $\mathrm{var}[Z_\lambda] = \lambda^2\,\frac{\mathrm{var}[X_i]}{n} + (1-\lambda)^2\,\frac{\mathrm{var}[X_i]}{m}$ is minimal iff λ achieves the minimum value of $f(\lambda) = \frac{\lambda^2}{n} + \frac{(1-\lambda)^2}{m}$. By equating to zero the first order derivative $f'(\lambda) = 2(\frac{\lambda}{n} - \frac{1-\lambda}{m})$ we obtain $\frac{\lambda}{1-\lambda} = \frac{n}{m}$ i.e. $\lambda = n/(n+m)$ and $1-\lambda = m/(n+m)$.

The pooled estimator $Z_{n/(n+m)} = \frac{1}{n+m}\sum_{i=1}^{n+m} X_i$ has mean $\mathrm{E}[X_i]$ and variance $\mathrm{var}[X_i]/(n+m)$; it is the most efficient convex combination of $\hat{X}^{\mathrm{emp}}$ and $\hat{Y}^{\mathrm{emp}}$.

C1c The case of en-route surveys

Let us consider two en-route surveys performed on a same network link. Each observation X_i is a Bernoulli random variable with parameter π_j, the probability that a trip belongs to O-D pair *j*.

Each survey $k \in \{1,2\}$ yields an estimator $\hat{p}_k^{\mathrm{emp}} = \frac{1}{n_k}\sum_{i=1}^{n_k} X_i$. The pooled estimator is $\hat{p}^{\mathrm{emp}} = \frac{n_1}{n_1+n_2}\hat{p}_1^{\mathrm{emp}} + \frac{n_2}{n_1+n_2}\hat{p}_2^{\mathrm{emp}} = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2} X_i$, with mean π_j and variance $\frac{\pi_j(1-\pi_j)}{n_1+n_2}$.

C2 Stratified samples

C2a Assumptions and properties

Let us divide a finite statistical population into *S* strata (subpopulations) indexed by *s*, each one with finite size N_s. The ratio $w_s = \mathrm{N}_s/\mathrm{N}$ where $\mathrm{N} = \sum_{s\in S}\mathrm{N}_s$ is called the relative weight of stratum *s*.

The mean μ of a R.V. *X* on the population may be estimated by the empiric mean $\hat{X}^{\mathrm{emp}}$ of a simple random sample. An alternative is to estimate μ using a stratified estimator defined in the following way.

Assume that in each stratum *s* there is a simple random sample of size n_s, yielding the empiric mean estimator $\hat{X}_s^{\mathrm{emp}}$ in the stratum. The stratified estimator for the whole population is

$$\hat{X}^{\mathrm{str}} \equiv \sum_s w_s \hat{X}_s^{\mathrm{emp}}.$$

It is unbiased since $\mathrm{E}[\hat{X}^{\mathrm{str}}] = \sum_s w_s \mathrm{E}[\hat{X}_s^{\mathrm{emp}}] = \sum_s w_s \mathrm{E}[X_s] = \mathrm{E}[X]$.

Assuming independence between the subsamples, the variance amounts to

$$\mathrm{var}[\hat{X}^{\mathrm{str}}] = \sum_s w_s^2 \mathrm{var}[\hat{X}_s^{\mathrm{emp}}].$$

Furthermore, if each subsample is chosen without replacement then $\mathrm{var}[\hat{X}^{\mathrm{str}}] = \sum_s w_s^2 \frac{\sigma_s^2}{n_s}\frac{E_s - n_s}{E_s - 1}$, in which σ_s^2 is the true variance of *X* restricted to stratum *s*.

C2b Application

A stratified estimator is useful when each stratum is relatively homogeneous with respect to the R.V.: then, for a given sample size n_s, the variance $\mathrm{var}[\hat{X}_s^{\mathrm{emp}}]$ is likely to be much smaller than $\mathrm{var}[X]$, leading to a variance $\mathrm{var}[\hat{X}^{\mathrm{str}}]$ smaller than $\mathrm{var}[\hat{X}^{\mathrm{emp}}]$.

Another application is to mix samples collected with different sampling rates n_s /N$_s$.

C2c An instance of daily traffic

Let us apply a stratified estimator to the annual average daily traffic on a network link. We can divide the population of the days in the year into three periods: Summer, Weekday out of Summer and Weekend out of Summer. Then $E = 365.25$, $E_1 = 3 \times 30.5 = 91.5$, $E_2 = 9 \times 21.7 = 195.3$, E_3 = 9 x 8.7 = 78.3 and the respective weights are $w_1 = 0.25$, w_2 = 0.54 and w_3 = 0.21.

The overall and strata summaries of the daily traffic are as follows (in veh/day):

E[X]	= 17,925	SD[X]	= 8,280
E[X_1]	= 23,570	SD[X_1]	= 8,340
E[X_2]	= 16,810	SD[X_2]	= 5,650
E[X_3]	= 14,070	SD[X_3]	= 10,270

Compute the variance of the empiric mean estimator in a simple random sample of size 15. Compute the variance of the stratified estimator, assuming that each stratum subsample is of size 5.

C3 The product of two independent random variables

C3a Basic properties

For any two real RVs X and Y which are independent, it holds that

- $\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$
- $\mathrm{var}[XY] = \mathrm{var}[X]\,\mathrm{var}[Y] + \mathrm{E}[X]^2\,\mathrm{var}[Y] + \mathrm{E}[Y]^2\,\mathrm{var}[X]$.

(as $\mathrm{var}[XY] = \mathrm{E}[X^2Y^2] - \mathrm{E}[XY]^2 = \mathrm{E}[X^2]\,\mathrm{E}[Y^2] - \mathrm{E}[X]^2\,\mathrm{E}[Y]^2 = \mathrm{E}[X^2](\mathrm{E}[Y^2] - \mathrm{E}[Y]^2) + \mathrm{E}[Y]^2(\mathrm{E}[X^2] - \mathrm{E}[X]^2)$ which can be averaged with the same expression for var[YX]).

C3b Application to a sampled category link count

The problem is to derive the mean and variance of the number of trips Q_{aj} of category j traversing link a. A possible estimator is $\hat{Q}_{aj} = \hat{X}_a \hat{p}_{ja}$, in which $\hat{X}_a$ is an estimator of the mean link traffic and $\hat{p}_{ja}$ is an estimator of the proportion of trips on link a which belong to category j.

Assuming that $\hat{X}_a$ and $\hat{p}_{ja}$ are independent, we can derive from the previous properties that

- $\mathrm{E}[\hat{Q}_{aj}] = \mathrm{E}[\hat{X}_a]\,\mathrm{E}[\hat{p}_{ja}]$
- $\mathrm{var}[\hat{Q}_{aj}] = \mathrm{E}[\hat{p}_{ja}^2]\,\mathrm{var}[\hat{X}_a] + \mathrm{E}[\hat{X}_a]^2\,\mathrm{var}[\hat{p}_{ja}] = \mathrm{E}[\hat{X}_a^2]\,\mathrm{var}[\hat{p}_{ja}] + \mathrm{E}[\hat{p}_{ja}]^2\,\mathrm{var}[\hat{X}_a]$.

When $\hat{X}_a$ and $\hat{p}_{ja}$ are unbiased, so is $\hat{Q}_{aj}$. When $\hat{X}_a$ is an empiric mean estimator with sample size N_a and $\hat{p}_{ja}$ an empiric mean estimator with sample size n_a, the variance of $\hat{Q}_{aj}$ amounts to

- $\text{var}[\hat{Q}_{aj}] = (\pi_{ja}^2 + \frac{\pi_{ja}(1-\pi_{ja})}{n_a})\frac{\text{var}[X_a]}{N_a} + \text{E}[X_a]^2\frac{\pi_{ja}(1-\pi_{ja})}{n_a}$
  $= \pi_{ja}^2\frac{\text{var}[X_a]}{N_a} + \frac{\pi_{ja}(1-\pi_{ja})}{n_a}(\text{E}[X_a]^2 + \frac{\text{var}[X_a]}{N_a})$.

C3c Numerical illustration

Using the empiric mean of daily link traffic considered in the previous part, and n_a = 400 and $p_{aj} = 3\%$, evaluate the estimators of $\text{E}[\hat{Q}_{aj}]$ and $\text{var}[\hat{Q}_{aj}]$.

C4 The pooling of en-route surveys

C4a Problem setting

Let us consider a set of en-route surveys, each of which is referred to by the network link *a* along which it was conducted. It is assumed that any two or more surveys along a same link are pooled as indicated in the first part. Also associated with each surveyed link is a sample of period flows, yielding an estimator $\hat{X}_a$ of the mean period link flow. By period we refer to a common unit of time (often one day or one hour).

We tackle the problem of estimating mean O-D flows for O-D pairs *i* which are sufficiently surveyed, i.e. for each of them there is a subset C_i of surveyed links such that the estimator $\hat{Q}_i = \sum_{a\in C_i}\hat{Q}_{ai}$ is unbiased. As previously $\hat{Q}_{ai}$ denotes an estimator of the average period flow of O-D pair *i* through link *a*.

Two questions are in order here. Firstly, is there any such cut C_i for each O-D pair? If not the case we must turn to much more sophisticated estimation principles. Secondly, what about alternative cuts C_i and C_i' for a given O-D pair? A selection rule could be to choose the cut with minimum variance of $\hat{Q}_i(C_i)$. However this may be largely improved by more sophisticated analysis.

We rather focus on the computation of $\text{var}[\hat{Q}_i]$ and $\text{cov}[\hat{Q}_i, \hat{Q}_j]$.

C4b Probabilistic analysis

The first step is to develop

$\text{cov}[\hat{Q}_i, \hat{Q}_j] = \text{cov}[\sum_{a\in C_i}\hat{Q}_{ai}, \sum_{b\in C_j}\hat{Q}_{bj}]$

$= (\sum_{a\in C_i, b\in C_j}\text{E}[\hat{Q}_{ai}, \hat{Q}_{bj}]) - (\sum_{a\in C_i}\text{E}[\hat{Q}_{ai}])(\sum_{b\in C_j}\text{E}[\hat{Q}_{bj}])$

$= \sum_{a\in C_i, b\in C_j}\text{cov}[\hat{Q}_{ai}, \hat{Q}_{bj}]$.

The second step is to substitute $\hat{X}_a\hat{p}_{ia}$ for $\hat{Q}_{ai}$, leading to

$\mathrm{cov}[\hat{Q}_i, \hat{Q}_j] = (\sum_{a\in C_i, b\in C_j} \mathrm{E}[\hat{X}_a \hat{p}_{ia} \hat{X}_b \hat{p}_{jb}]) - (\sum_{a\in C_i} \mathrm{E}[\hat{X}_a \hat{p}_{ia}])(\sum_{b\in C_j} \mathrm{E}[\hat{X}_b \hat{p}_{jb}])$

$= \sum_{a\in C_i, b\in C_j} \mathrm{E}[\hat{X}_a \hat{X}_b]\mathrm{E}[\hat{p}_{ia} \hat{p}_{jb}] - \mathrm{E}[\hat{X}_a]\mathrm{E}[\hat{p}_{ia}]\mathrm{E}[\hat{X}_b]\mathrm{E}[\hat{p}_{jb}]$

by using the independence assumption between $\hat{X}_a$ and $\hat{p}_{ia}$ (resp. $\hat{X}_b$ and $\hat{p}_{jb}$).

The third step is to separate two cases, depending on whether $a = b$ or not. If $a \neq b$ then $\mathrm{E}[\hat{X}_a \hat{X}_b] = \mathrm{E}[\hat{X}_a]\mathrm{E}[\hat{X}_b]$ yielding that $\mathrm{cov}[\hat{Q}_{ai}, \hat{Q}_{bj}] = \mathrm{E}[\hat{X}_a]\mathrm{E}[\hat{X}_b]\ \mathrm{cov}[\hat{p}_{ia}, \hat{p}_{jb}]$, in which a further reasonable assumption that $\hat{p}_{ia}$ and $\hat{p}_{jb}$ are independent may be introduced, leading to a null contribution.

If $a = b$, i.e. $a \in C_i \cap C_j$, the corresponding term amounts to

$\mathrm{cov}[\hat{Q}_{ai}, \hat{Q}_{aj}] = \mathrm{E}[\hat{X}_a^2]\mathrm{E}[\hat{p}_{ia} \hat{p}_{jb}] - \mathrm{E}[\hat{X}_a]^2\ \mathrm{E}[\hat{p}_{ia}]\mathrm{E}[\hat{p}_{jb}]$

$= \mathrm{var}[\hat{X}_a]\mathrm{E}[\hat{p}_{ia}]\mathrm{E}[\hat{p}_{jb}] + \mathrm{E}[\hat{X}_a^2]\mathrm{cov}[\hat{p}_{ia}, \hat{p}_{jb}]$.

The conclusion is

$\mathrm{cov}[\hat{Q}_i, \hat{Q}_j] = \sum_{a\in C_i \cap C_j} \mathrm{var}[\hat{X}_a]\mathrm{E}[\hat{p}_{ia}]\mathrm{E}[\hat{p}_{ja}] + \mathrm{E}[\hat{X}_a^2]\mathrm{cov}[\hat{p}_{ia}, \hat{p}_{ja}]$.

Fourthly, let us assume that $\hat{X}_a$ is an empiric mean estimator with sample size N_a and that $\hat{p}_{ia}$ and $\hat{p}_{ja}$ are empiric proportion estimators from a multinomial sample of size n_a. Then

$\mathrm{cov}[\hat{p}_{ia}, \hat{p}_{ja}] = \frac{1}{n_a}(\delta_{ij}\pi_{ia} - \pi_{ia}\pi_{ja})$ in which $\delta_{ij} = 1$ if $i = j$ or 0 otherwise.

When $i = j$ we obtain

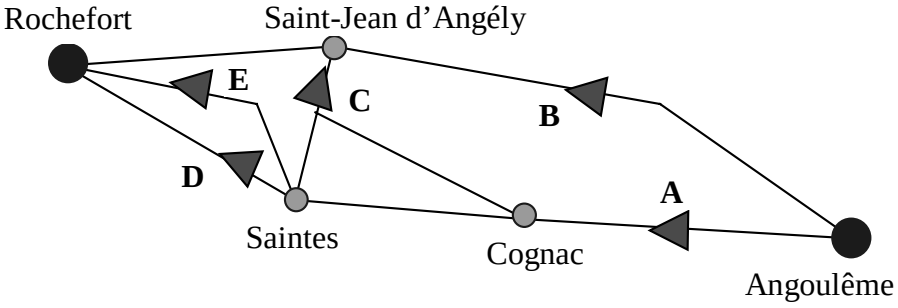
$\mathrm{var}[\hat{Q}_i] = \sum_{a\in C_i} \pi_{ia}^2\ \mathrm{var}[\hat{X}_a] + \frac{\pi_{ia}(1-\pi_{ia})}{n_a}\mathrm{E}[\hat{X}_a^2]$

$= \sum_{a\in C_i} \pi_{ia}^2 \frac{\mathrm{var}[X_a]}{N_a} + \frac{\pi_{ia}(1-\pi_{ia})}{n_a}(\mathrm{E}[X_a]^2 + \frac{\mathrm{var}[X_a]}{N_a})$

$= \sum_{a\in C_i} \mathrm{var}[\hat{Q}_{ia}]$ of course.

C4c Instance

The following picture is extracted from the French roadway network. Red arrows indicate en-route surveys and are designated hereafter by letters from A to E.



Each surveyed link *a* is characterized by two samples, one sample of daily counts and the other one of roadside interviews yielding the origin and destination of the intercepted trips. The first sample has size N_a and results in empiric mean count $\hat{X}_a^{\text{emp}}$ and empiric variance $\hat{S}_a^{2\text{emp}}$. The other sample has size n_a and results in absolute frequencies n_{ia} for O-D pairs *i*.

Link *a*	N_a	$\hat{X}_a^{\text{emp}}$	$\sqrt{\hat{S}_a^{2\text{emp}}}$	n_a	n_{ia}
A	5	5,000	2,100	800	31
B	8	3,000	1,260	500	15
C	5	1,500	750	400	6
D	10	9,500	5,200	1,000	8
E	20	18,000	8,400	2,000	22

The absolute frequencies in the table correspond to the O-D pair from Angoulême to Rochefort. Find a traffic cut appropriate to that O-D pair, estimate the O-D flow and predict its accuracy.

C5 On panel surveys

Panel surveys are used to measure the variation in an empiric variable between two different epochs, eg. before and after a change is made.

C5a Justification

Two methods are available to estimate the average evolution of an empiric variable. The first one can be called the naive approach: it consists in computing the difference between the estimated means of each epoch, based on two different samples. The other one is the panel approach, which amounts to average the individual differences associated with a constant sample.

The panel method is more efficient with respect to sample size. However it also has limits when the two epochs are separated by a long time, because in this case the initial sample may not be adapted to the final population.

C5b Probabilistic setting

Let us denote by $\mathrm{Y_I}$ and $\mathrm{Y_F}$ the R.V. at the initial and final epochs respectively, and assume that the sample size $T = T_\mathrm{I} = T_\mathrm{F}$ is the same in all cases.

In the naive method, the initial and final empiric means $\hat{Y}_\mathrm{I} = \frac{1}{T}\sum_t Y_{\mathrm{I}t}$ and $\hat{Y}_\mathrm{F} = \frac{1}{T}\sum_{t'} Y_{\mathrm{F}t'}$ are independent RVs with respective means of $\overline{\mathrm{Y}}_\mathrm{I}$ and $\overline{\mathrm{Y}}_\mathrm{F}$ and variances of $\mathrm{var}(\mathrm{Y_I})/T$ and $\mathrm{var}(\mathrm{Y_F})/T$. Hence the difference $\hat{Y}_\mathrm{F} - \hat{Y}_\mathrm{I}$ is a RV with mean $\overline{\mathrm{Y}}_\mathrm{F} - \overline{\mathrm{Y}}_\mathrm{I}$ and variance $V_{\text{naive}} = (\mathrm{var}(\mathrm{Y_I}) + \mathrm{var}(\mathrm{Y_F}))/T$.

In the panel method, the empiric mean $\hat{D} = \frac{1}{T}\sum_t (Y_{\mathrm{F}t} - Y_{\mathrm{I}t})$ is a RV with similar mean of $\overline{\mathrm{Y}}_\mathrm{F} - \overline{\mathrm{Y}}_\mathrm{I}$ but variance of $V_{\text{panel}} = (\mathrm{var}(\mathrm{Y_F} - \mathrm{Y_I}))/T$.

Then the ratio of the two variances $V_{\text{panel}} / V_{\text{naive}} = \text{var}(Y_F - Y_I) / (\text{var}(Y_I) + \text{var}(Y_F))$ is inferior to 1. It is all the more reduced as the evolution $\text{var}(Y_F - Y_I)$ is smaller between the two epochs.

In agriculture, a common statistical practice is to compare paired samples, i.e. twin fields submitted to different treatments (eg. doses of fertilizer). Thus regression analysis may be performed on several panel samples.

C5c Instance: the French nationwide automobile panel

In transportation panel surveys are not used to measure spatial variables such as demand zones or O-D pairs. Their main use is to estimate distances travelled by transport mode or type of vehicle.

The French nationwide automobile panel aims at measuring the motorization rate and car usage in France. The panel sample comprises 10,000 households surveyed on every year. One third of the sample is renewed each year to maintain representativity. About 80% of the sampled households respond to the questionnaire, which enables dynamic analysis on about 4,000 households for three consecutive years A-1, A and A+1.

The statistical unit is the household. In each sampled household a person is asked about the composition of the household, the available cars (their age, fuel type and power) and associated annual travelled distance (decomposed into urban, interurban and freeways).

Let us quote some 1996 results: 25,6 millions of cars (including vehicles up to 3.5 tons). Average annual distance amounts to 13,960 km (+/- 120) against 11,200 in 1985. Diesel cars are in proportion of 30% and constitute 44% of travelled distances. The average year is 6.5 years, and average total distance is 84,000 km. More than one car out of two is used on every day, and four out of five drivers drive their car almost every day.

Specific panel investigation yielded the following conclusions: the transition from gas to diesel is especially frequent when the previous annual travelled distance is high (> 18,000 km) and it is associated to a strong increase in travelled distance (+ 4,000 km). More generally, a new car is more intensively used than the previous one (+ 3,000 km a year).

C5d Numerical illustration

The following table, derived from the French automobile panel, indicates the mean and standard deviation of household travelled distance (by car). The sample size is 1,400. Compute the mean change and the associated variance using first the naive method, second the panel method.

Tab. B. Exercise on panel method.

Year	A-1	A+1	Ecart
Average household annual travelled distance (km)	19 740	21 310	+ 1 570
Standard deviation (km)	13 954	14 142	12 820

D Statistical estimation and linear regression

Rational decision-making is based on information and judgment. As information is costly, only limited, uncertain information is available. Statistical estimation aims at using limited information in an efficient way. It provides point estimators (eg. empiric mean, maximum likelihood) and also confidence intervals that contain the true unknown value of a parameter with a given level of confidence. Such intervals may be used to select one hypothesis out of two, which is referred to as hypothesis testing.

The lesson deals with statistical estimation, tests, linear regression models and transportation instances. The linear regression model is the simplest, yet most useful, statistical model to analyse the dependencies between two quantitative random variables. We shall indicate briefly its estimation and then depict its application to three transportation models: a zone-based generation model, a binary logit choice model and a varying parameter binary choice model.

D1 Statistical estimation

D1a Definition and objectives

To estimate a parameter Θ of a R.V. X is to apply a function τ to the result of a random sample $\mathbf{x}(\omega) = (x_i)_{i=1..n}$, in order to recover an approximate value of Θ. The R.V. $\tau(\mathbf{x}(\omega))$ is called an **estimator** of Θ.

Instances include the empiric sample mean to recover the mean of a R.V., the empiric sample variance to recover its variance.

The theory of statistical estimation deals with:

- The definition of point estimators,
- The definition of desirable properties for estimators,
- The assessment of whether an estimator possesses certain properties,
- The identification of performant, efficient point estimators.

Its primary application is to recover information about parameters useful to decision-making.

D1b Performance of point estimation

The estimation error of an estimator T of parameter Θ splits into

$$T - \Theta = E[T] - \Theta + T - E[T]$$

in which $E[T]-\Theta$ is the estimator **bias** and $T-E[T]$ the estimator **dispersion**.

From the definition or E[T], $E[(T-\Theta)^2] = var[T] + (E[T]-\Theta)^2$, i.e. the mean square error is the sum of the estimator variance and the squared bias.

When several estimators are available to recover the value of a given parameter, a reasonable selection principle for evaluation and comparison is to select the estimator with minimum square error.

Ex. The empiric variance $\hat{S}^{2emp} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{x}^{emp})^2$ has mean $\frac{n-1}{n}\sigma^2$ hence it is biased. The corrected empiric variance $\hat{S}^{2*emp} = \frac{n}{n-1}\hat{S}^{2emp}$ is an unbiased estimator of σ^2 but has larger variance than $\hat{S}^{2emp}$. When the true mean μ is known, the estimator $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ is unbiased with variance $\frac{1}{n}(E[(X-\mu)^4] - \sigma^4)$ inferior to that of the corrected empiric variance estimator, $var[\hat{S}^{2*emp}] = \frac{1}{n}(E[(X-\mu)^4] - \frac{n-3}{n-1}\sigma^4)$, hence T performs better than $\hat{S}^{2*emp}$.

A slightly different notion of performance is useful to analyse estimators more complex than the empiric mean and variance. It involves a generic estimator $\mathbf{T} = (T_n)_{n\in N}$ in which T_n corresponds to a sample size of *n*.

A generic estimator **T** is asymptotically convergent iff its distribution concentrates on Θ when the sample size tends to infinity. This implies that $E[T_n] \to \Theta$ and $var[T_n] \to 0$ as $n \to \infty$.

D1c Interval estimation

Point estimation contains no information on the accuracy of its result, i.e. about how close the result is to the true value. The purpose of interval estimation is to provide confidence intervals which contain the true value with a given probability.

Precisely, a confidence interval I_α for Θ at confidence level 1-α satisfies $Pr(\Theta \in I_\alpha) = 1-\alpha$. The larger 1-$\alpha$, the wider I_α. The probability α measures the risk that the parameter should not belong to I_α: it is also called the significance level.

A confidence interval at confidence level 1-α may be built in the following way, based on a point estimator T of parameter Θ. From the distribution of T knowing Θ, we can choose two probabilities α_1 and α_2 that add up to α and determine two numbers h_1 and h_2 such that $Pr(T < \Theta - h_1) = \alpha_1$ and $Pr(T > \Theta + h_2) = \alpha_2$. Thus

$$Pr(\Theta - h_1 \le T \le \Theta + h_2) = 1 - \alpha_1 - \alpha_2 = 1-\alpha.$$

If h_1 and h_2 do not depend on Θ, we can equivalently state that with probability 1-α it holds that $T - h_2 \le \Theta \le T + h_1$, or equivalently $Pr(T - h_2 \le \Theta \le T + h_1) = 1-\alpha$, which constitutes the confidence interval.

Given a random sample and the resulting value t of estimator T, there is a probability 1-α that $\Theta \in [t - h_2; t + h_1]$. As $t - h_2$ and $t + h_1$ depend on the random sample, $[t - h_2; t + h_1]$ is a random interval.

D1d Confidence intervals for the parameters of a normal distribution

Let us consider a gaussian R.V. *X* with mean μ and variance σ^2. Assuming that the variance is known, the empiric mean satisfies $Pr(\mu - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \hat{x}^{emp} < \mu - u_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$ in which $u_{\alpha/2}$ is defined by $Pr(U > u_{\alpha/2}) = \alpha/2$ for a reduced gaussian R.V. *U*. The resulting confidence interval at level 1-α for the true mean is $\mu \in \hat{x}^{emp} \pm u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

When the variance is unknown, we consider the corrected empiric variance $\hat{S}^{*2emp} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{x}^{emp})^2$ because the quotient $Y = \frac{\hat{x}^{emp} - \mu}{\hat{S}^{*emp}/\sqrt{n}} = \frac{(\hat{x}^{emp} - \mu)}{\sigma/\sqrt{n}} / \sqrt{\frac{(n-1)\hat{S}^{*2emp}/\sigma^2}{n-1}}$ is a Student R.V. with *n*-1 degrees of freedom (as the quotient of a reduced gaussian R.V. by an independent R.V. which is the square root of the χ^2_{n-1} R.V. $(n\text{-}1)\hat{S}^{*2emp}/\sigma^2$ divided by its number of DF).

We may also evaluate $t_{\alpha/2}^{(n-1)}$ such that $\Pr(Y > t_{\alpha/2}^{(n-1)}) = \alpha/2$ and obtain

$$\Pr(-t_{\alpha/2}^{(n-1)} \le \frac{\hat{x}^{emp} - \mu}{\hat{S}^{*emp}/\sqrt{n}} \le t_{\alpha/2}^{(n-1)}) = 1 - \alpha$$

or equivalently $\Pr(\mu \in \hat{x}^{emp} \pm t_{\alpha/2}^{(n-1)} \frac{\hat{S}^{*emp}}{\sqrt{n}}) = 1 - \alpha$.

As regards confidence intervals on the variance σ^2 when the mean is unknown, the R.V. $n\hat{S}^{2emp}/\sigma^2$ is chi-square with *n*-1 degrees of freedom. For a given confidence level of 1-α we can choose two numbers h_1 and h_2 such that $\Pr(h_1 \le \chi^2_{n-1} \le h_2) = 1 - \alpha$, yielding

$$\Pr(\frac{n}{h_2}\hat{S}^{2emp} \le \sigma^2 \le \frac{n}{h_1}\hat{S}^{2emp}) = 1 - \alpha.$$

D2 Hypothesis testing

The theory of statistical tests consists in the systematic formulation and judgment of hypotheses about parameters or shape of R.V., on the basis of sample information and also of the acceptance of some risk in decision-making. We shall restrict ourselves to tests on numerical values of parameters.

D2a Formulation of null and alternative hypotheses

A hypothesis is a quantitative assertion on the parameters or shape of a distribution, eg. that the mean of a R.V. *X* is equal to m_0. It may arise from prior knowledge. Its judgment consists in comparing the value m_0 to a sample estimator and deciding whether or not to reject the hypothesis.

In a statistical test under standard form, there are two hypotheses, namely a null hypothesis denoted by H_0 and a contradictory, alternative hypothesis denoted by H_1 and defined by contrast to H_0, eg. that the mean of R.V. *X* is different from m_0, or equal to $m_1 \neq m_0$ etc.

D2b The two types of risk in judging the hypotheses

Decision-making on which one out of two alternative states-of-nature is true involves two types of risk, as depicted in the following decision table.

State of nature	Judgment	Quality of decision
H_0 true, H_1 false	Reject H_0	Wrong: Type I Error
	Accept H_0	Right
H_0 false, H_1 true	Reject H_0	Right
	Accept H_0	Wrong: Type II Error

Type I error is measured by the Type I risk, defined as $\alpha = \Pr\{\text{reject } H_0 \mid H_0 \text{ true}\}$ also called the **significance level** of the test. Type II error is measured by the Type II risk, defined as $\beta = \Pr\{\text{accept } H_0 \mid H_1 \text{ true}\}$ also called the **power of the test**.

The decision rule to accept or reject H_0 depends on a sample-based **decision variable** T. Assuming that H_0 holds, to the decision variable T is associated an **acceptance region** A which depends on H_0 and α, in such a way that $\Pr\{T \in A \mid H_0 \text{ true}\} = 1-\alpha$.

For instance if $H_0 = \{E[X] = m_0\}$ the acceptance region A may be a confidence interval around m_0 at confidence level $1-\alpha$. The complementary set of A is $\overline{A}$, called the **critical region** of the test since $\Pr\{T \in \overline{A} \mid H_0 \text{ true}\} = \alpha$ is the Type I risk.

The Type II error β depends on both the alternative hypothesis H_1 and the acceptance region A, since $\beta = \Pr\{\overline{A} \mid H_1\}$.

Under a given sample size *n*, decreasing α will increase β. Under a given α, increasing *n* will decrease β hence the test will be more powerful. Under given *n* and α, narrowing the gap between H_0 and H_1 will decrease β.

D2c Testing the mean of a normal distribution

Let *X* be a gaussian R.V. with unknown mean μ and m_0 be a reference value for the following null hypothesis $H_0 = \{\mu = m_0\}$.

The alternative hypothesis may be $H_1 = \{\mu = m_1 > m_0\}$ (or $\{\mu = m_1 < m_0\}$).

The decision variable is the sample empiric mean $\hat{x}^{\text{emp}}$, which under H_0 is distributed normal with mean m_0 and variance σ^2 / n.

The acceptance region A is a one-sided interval $]-\infty; m_\alpha]$ such that $\Pr\{m_0 + \frac{\sigma}{\sqrt{n}} U \le m_\alpha\} = 1-\alpha$ i.e. $m_\alpha = m_0 + \sigma u_\alpha / \sqrt{n}$. Then the critical region is $\overline{A} =]m_\alpha; +\infty[$.

The decision rule is to accept H_0 when $\hat{x}^{\text{emp}} \le m_\alpha$ and to reject H_0 when $\hat{x}^{\text{emp}} > m_\alpha$.

The Type II error is $\beta = \Pr\{\hat{x}^{\text{emp}} \le m_\alpha \mid H_1\} = \Pr\{m_1 + \frac{\sigma}{\sqrt{n}} U \le m_\alpha\} = \Pr\{U \le \frac{m_\alpha - m_1}{\sigma / \sqrt{n}}\}$.

D3 The linear regression model

Statistical regression aims at describing in a generic way the dependency of an endogenous, "explained" R.V. *Y* on an exogenous, "explanatory" variable *X* which may be random or not. It results in a formula $E[Y \mid X = x] = F(\Theta, X)$ in which Θ is a parameter and F a mathematical function.

The usual form of a regression problem is to specify function F and to estimate parameter Θ based on a random sample of joint observations $(X_i, Y_i)_{i=1..n}$.

D3a Assumptions in the linear regression model

A linear regression model of R.V. *Y* on a variable *X* is formulated as

$$Y = \alpha + \beta X + \varepsilon \text{ (or } Y = \alpha + \beta g(X) + \varepsilon),$$

in which ε is an additional R.V. called the error and assumed independent from *X*. Thus

$$E[Y \mid X = x] = \alpha + \beta X + E[\varepsilon].$$

The variable *X* may be a vector with *m* components X_r, in which case β is also a vector of *m* components β_r and $\beta.X = \beta^t X = \sum_{r=1}^{m} \beta_r X_r$.

The observations consist in a random sample of size *n* of the couple (*X*, *Y*), the result of which is denoted by (X_i, Y_i). Was the parameter $\Theta = (\alpha, \beta)$ known, we would interpret the *n* values $Y_i - \alpha - \beta X_i$ as independent observations of ε. We would also derive the empiric mean $\hat{\varepsilon}^{emp} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \alpha - \beta X_i$ and empiric variance $\hat{S}_\varepsilon^{2\,emp} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i - \hat{\varepsilon}^{emp})^2$.

We can take account of $E[\varepsilon]$ within α and therefore assume that $E[\varepsilon] = 0$. In this case an improved estimator of the variance is provided by $T_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2$.

However we assume no prior knowledge of α and β. The estimation problem is to recover α and β, and maybe also σ_ε^2.

When *X* is a deterministic variable with known value X_i for each observation *i*, the interpretation of ε is somewhat different: associated to the *i*-th observation is a model $Y_i = \alpha + \beta X_i + \varepsilon_i$ in which ε_i is the *i*-th random error. Here a standard assumption is that the variables $(\varepsilon_i)_{i=1..n}$ are independent, identically distributed R.V. with null mean.

D3b Least squares estimation

The principle of least squares estimation is to select the value of the parameter which minimizes the squared distance between the observed Y_i and the associated predictions $\tilde{Y}_i = \alpha + \beta X_i$. The objective function to minimize is also

$$F(\alpha, \beta, \sigma_\varepsilon^2) = \sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2.$$

The derivatives are $\frac{\partial F}{\partial \alpha} = -2\sum_{i=1}^{n} Y_i - \alpha - \beta X_i$ and $\frac{\partial F}{\partial \beta_r} = -2\sum_{i=1}^{n} (X_i)_r (Y_i - \alpha - \beta X_i)_r$. The first-order necessary conditions for optimality are $\frac{\partial F}{\partial \alpha} = 0$ hence $n\alpha = \sum_{i=1}^{n} Y_i - \beta X_i$, and $\frac{\partial F}{\partial \beta_r} = 0$ hence $\alpha \sum_{i=1}^{n} X_{ir} = (\sum_{i=1}^{n} X_{ir} Y_i) - \sum_{i=1}^{n} X_{ir} (\beta . X_i)_r$.

When $m = 1$ this reduces to $\hat{\alpha}^{LS}(\sum_{i=1}^{n} X_i) + \hat{\beta}^{LS}(\sum_{i=1}^{n} X_i^2) = \sum_{i=1}^{n} Y_i X_i$. On replacing α by its expression derived from the first condition, we obtain the following LS estimators

$\hat{\beta}^{LS} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}^{emp})(X_i - \hat{X}^{emp})}{\sum_{i=1}^{n} (X_i - \hat{X}^{emp})^2}$ (ratio of empiric covariance to empiric variance of X)

$\hat{\alpha}^{LS} = \hat{Y}^{emp} - \hat{\beta}^{LS} \hat{X}^{emp}$.

In the general case, let $\mathbf{Y}$ denote the column matrix $(Y_i)_{i=1..n}$ and $\mathbf{X}$ the matrix of n line vectors $(1, X_i)$ of dimension $1+m$. Then $\hat{\Theta}^{LS} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}$.

We may even associate an individual weight w_i to each observation i, yielding a diagonal matrix of weights $\mathbf{W} = \text{diag}[w_i]_{i=1..n}$. Then the weighted LS estimator of Θ is

$$\hat{\Theta}^{WLS} = (\mathbf{X}^t \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{W} \mathbf{Y}.$$

D3c Properties of the LS estimators

Conditional on the values of the X_i, which is denoted by a superscript **x**, the following results hold (expectation is taken with respect to ε):

(1) $\hat{\alpha}^{LS}$ and $\hat{\beta}^{LS}$ are unbiased, and so is a prediction $\tilde{Y}(x) = \hat{\alpha}^{LS} + \hat{\beta}^{LS} x$.

(2) $\text{cov}[\hat{\beta}^{LS}, \hat{Y}^{emp}] = 0$.

(3) An unbiased estimator of σ_ε^2 is $\hat{\sigma}_\varepsilon^{2LS} = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \tilde{Y}_i)^2$.

(4) When $m = 1$, $\text{var}^{\mathbf{x}}[\hat{\beta}^{LS}] = \sigma_\varepsilon^2 / \sum_{i=1}^{n} (X_i - \hat{X}^{emp})^2$.

(5) $\text{var}^{\mathbf{x}}[\hat{\alpha}^{LS}] = \sigma_\varepsilon^2 (\frac{1}{n} + \frac{(\hat{X}^{emp})^2}{\sum_{i=1}^{n} (X_i - \hat{X}^{emp})^2})$.

(6) $\text{cov}^{\mathbf{x}}[\hat{\alpha}^{LS}; \hat{\beta}^{LS}] = -\hat{X}^{emp} \text{var}^{\mathbf{x}}[\hat{\beta}^{LS}]$.

The results (1), (2) and the unbiasedness of $\hat{\sigma}_\varepsilon^{2LS}$ extend to the case when X is random.

As regards any sample, the residual errors $e_i = Y_i - \tilde{Y}(X_i)$ have null empiric mean (from the definition of $\hat{\beta}^{LS}$) and empiric variance $\hat{S}_{Y/X}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2 = (1 - \rho^2) \hat{S}_Y^2$, in which $\rho = \text{cov}^{\mathbf{x}}[X, Y] / (\hat{S}_X \hat{S}_Y)$ is the empiric **coefficient of correlation** between X and Y. The coefficient of correlation takes its value in [-1, 1]: values close to 1 or –1 indicate that the linear model of Y with respect to X performs well.

D3d Confidence intervals and tests in the gaussian case

From now it is assumed that the residual variable ε is gaussian $N(0,\sigma_\varepsilon^2)$. Then, conditional on $X = x$, the R.V. Y is gaussian $N(\alpha+\beta x,\sigma_\varepsilon^2)$. The R.V. $\hat{\alpha}^{LS}$, $\hat{\beta}^{LS}$ and $\tilde{Y}(x)$ are also gaussian with mean and variance as established in the general case.

The R.V. $\frac{(n-2)\hat{\sigma}_\varepsilon^{2\,LS}}{\sigma_\varepsilon^2}$ is independent from $\hat{\alpha}^{LS}$, $\hat{\beta}^{LS}$ and $\hat{Y}^{emp}$ and it is distributed χ^2_{n-2}.

To obtain confidence intervals on $\hat{\alpha}^{LS}$, $\hat{\beta}^{LS}$ and $\tilde{Y}(x)$ when σ_ε^2 is unknown, we can combine the gaussian assumption about each of them to the chi-square assumption on $(n-2)\hat{\sigma}_\varepsilon^{2\,LS}/\sigma_\varepsilon^2$. Thus $(\hat{\alpha}^{LS}-\alpha)[\hat{\sigma}_\varepsilon^{2\,LS}(\frac{1}{n}+\frac{(\hat{X}^{emp})^2}{\sum_{i=1}^n (X_i-\hat{X}^{emp})^2})]^{-1/2}$ is a Student R.V. with n-2 degrees of freedom, so is $(\hat{\beta}^{LS}-\beta)[\hat{\sigma}_\varepsilon^{2\,LS}/\sum_{i=1}^n (X_i-\hat{X}^{emp})^2]^{-1/2}$.

The relevance of the linear model may be evaluated by testing hypothesis $H_0 = \{\hat{\beta}^{LS} \neq 0\}$ against $H_1 = \{\hat{\beta}^{LS} = 0\}$. This amounts to computing the critical probability $\alpha = 2\,Pr\{T^{(n-2)} > |\hat{\beta}^{LS}|\}$: the assumption H_0 cannot be rejected at the confidence level 1-α.

D3e Prediction with a linear model

We now focus on the R.V. Y conditional on a (possibly unobserved) value x_0 of X. The predicted value of Y is $\tilde{Y}(x_0)=\hat{\alpha}^{LS}+\hat{\beta}^{LS}x_0$ which is distributed normal with mean $\hat{\alpha}^{LS}+\hat{\beta}^{LS}x_0$ and variance $\sigma_\varepsilon^2(\frac{1}{n}+\frac{(\hat{X}^{emp})^2}{\sum_{i=1}^n (X_i-\hat{X}^{emp})^2})$.

From the basic modeling assumption, we also know that the distribution of Y conditional on $X = x_0$ is gaussian with mean $\alpha+\beta x_0$ and variance σ_ε^2. As $\tilde{Y}(x_0)$ depends on the sample observation only, $Y_{X=x_0}$ and $\tilde{Y}(x_0)$ are independent. Their difference is distributed gaussian with null mean and variance $\sigma_\varepsilon^2(1+\frac{1}{n}+\frac{(\hat{X}^{emp})^2}{\sum_{i=1}^n (X_i-\hat{X}^{emp})^2})$, which implies that $\{Y_{X=x_0}-\tilde{Y}(x_0)\}[\hat{\sigma}(1+\frac{1}{n}+\frac{(\hat{X}^{emp})^2}{\sum_{i=1}^n (X_i-\hat{X}^{emp})^2})]^{-1/2}$ is a Student R.V. with n-2 degrees of freedom.

D4 Linear models in transportation

We shall now provide three instances of linear models in transport demand analysis.

D4a Linear generation model

Let us consider the following estimation model

$$P_o = a + bX_o + e_o$$

in which the index *o* identifies an observation, P_o is an observed zonal production, X_o is an observed zonal attribute and e_o is a statistical error that includes design error as well as measurement error.

Let us denote by *O* the set of observations. We further assume that the errors e_o are independent samples of identically distributed random variables ε_o, with null mean and variance of σ^2. The least squares estimates of *a* and *b* are respectively

$\hat{a} = \overline{P} = \dfrac{\sum_{o \in O} P_o}{|O|}$ where $|O|$ denotes the number of observations in *O*, and

$\hat{b} = \dfrac{\sum_{o \in O} P_o (X_o - \overline{X})}{\sum_{o \in O} (X_o - \overline{X})^2}$ where $\overline{X} = \dfrac{\sum_{o \in O} X_o}{|O|}$.

These estimators are unbiased (i.e. the means of $\hat{a}$ and $\hat{b}$ over all possible samples are the true values of *a* and *b* respectively). Their variances follow the general case.

The model readily extends to the multidimensional case, in which both *b* and X_o are vectors of several numbers.

D4b Binary choice model

Let us consider a model of binary choice between two sites 1 and 0 which may represent transport modes or network routes. We assume that, given attributes *X* of the decision-maker and of the two sites, the probability of choosing site 1 is

$$p_X = \mathrm{F}(\Theta.X),$$

in which Θ is a vector of parameters and F is a monotonic mathematical function with values in [0, 1]. This can be inverted into

$$\Theta.X = \mathrm{F}^{-1}(p_X).$$

Each observation *i* is the outcome of a binary R.V. $y_i \in \{0,1\}$. If there is a large number n_v of observations *i* for the value *v* of *X*, the sample frequency of site 1 is $\hat{p}_v = \frac{1}{n_v} \sum_{i=1}^{n_v} y_i$ expectedly close to p_v. Then a Taylor expansion formula yields that

$$\mathrm{F}^{-1}(\hat{p}_v) \approx \Theta.v + (\hat{p}_v - p_v) / \mathrm{f}(\mathrm{F}^{-1}(p_v))$$

where f is the derivative of F. The error term $\eta_v = (\hat{p}_v - p_v)/\mathrm{f}(\mathrm{F}^{-1}(p_v))$ has null mean and variance $\mathrm{var}[\eta_v] = \frac{p_v(1-p_v)}{n_v\,\mathrm{f}(\mathrm{F}^{-1}(p_v))^2}$.

The so-called minimum chi-square estimator $\hat{\Theta}^{\mathrm{chi}}$ is defined as the weighted least squares estimator of the linear model $\mathrm{F}^{-1}(\hat{p}_v) = \Theta.v + \eta_v$, with $\mathrm{var}[\eta_v]$ evaluated at $\hat{p}_v$. Thus

$$\hat{\Theta}^{\mathrm{chi}} = [\sum_v \frac{vv^t}{\mathrm{var}(\eta_v)}]^{-1}[\sum_v \frac{\mathrm{F}^{-1}(\hat{p}_v)}{\mathrm{var}(\eta_v)} v],$$

which differs from the unweighted LS estimator only by the weights $\mathrm{var}[\eta_v]$. It can be shown that $\hat{\Theta}^{\mathrm{chi}}$ is consistent and approximately normal with asymptotic covariance matrix $[\sum_v vv^t / \mathrm{var}(\eta_v)]^{-1}$, which is identical to that of the corresponding maximum likelihood estimator.

D4c Application to binary logit

In the logit model, function $\mathrm{F}(z) = 1/[1+\exp(-z)]$ hence $\mathrm{F}^{-1}(u) = \ln\frac{u}{1-u}$ and $\mathrm{f}(z) = \mathrm{F}'(z) = \mathrm{F}(z)[1-\mathrm{F}(z)]$. Thus $\mathrm{f}(\mathrm{F}^{-1}(p)) = p(1-p)$ and $\mathrm{var}[\eta_v] = 1/n_v$.

D4d Binary model with a log-normal coefficient

Let us consider a model of competition between a free route 1 and a toll route 0. The probability of choosing the free route is equal to the probability of a log-normal value-of-time being inferior to the cut-off value-of-time $v^* = \frac{P_1 - P_0}{T_0 - T_1}$, i.e. $p = \Pr(v_{\mu,\sigma} \le v^*) = \mathrm{H}_{\mu,\sigma}(v^*) = \Phi(\frac{1}{\sigma}[\ln(v^*) - \mu])$ with μ and σ the mean and standard deviation of the normal R.V. $\ln v_{\mu,\sigma}$, $\mathrm{H}_{\mu,\sigma}$ the CDF of $v_{\mu,\sigma}$ and Φ the CDF of a reduced gaussian variable.

Here we have a two-component vector $X = (\ln(\frac{P_1 - P_0}{T_0 - T_1}); -1)$, a two-component parameter $\Theta = (\frac{1}{\sigma}; \frac{\mu}{\sigma})$ and $\mathrm{f}(z) = \mathrm{F}'(z) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z^2)$.

E Maximum likelihood estimation and DCM

Maximum likelihood (ML) is one of the two classical estimation methods in Statistics, the other one being least squares. ML is particularly well suited to analyze discrete dependent variables, which is the justification for introducing it together with the estimation of DCM.

The lesson comprises three parts. Part 1 defines ML estimation and states its main properties. Part 2 describes its application to DCM. Lastly, Part 3 demonstrates the ML estimation of several binary choice models in the case of the Prado-Carénage study.

E1 On ML estimation

E1a Defintion

Let $f(x|\Theta)$ denote the probability density function of a R.V. *X* at point *x* with respect to parameter Θ. It may be used to derive the PDF $g(\mathbf{x}|\Theta)$ of a random sample $\mathbf{x} = (x_i)_{i=1..n}$. When the sample is simple $g(\mathbf{x}|\Theta) = \prod_{i=1}^{n} f(x_i|\Theta)$.

Given a random sample **x**, the **likelihood function** of parameter Θ is

$$L(\Theta|\mathbf{x}) = g(\mathbf{x}|\Theta).$$

The **maximum likelihood (ML) estimator** of Θ is a point $\hat{\Theta}^{ML}$ which maximizes the likelihood function $L(\Theta|\mathbf{x})$. As it depends on the sample **x**, $\hat{\Theta}^{ML}$ is a random variable.

The principle of ML estimation amounts to choosing the value $\hat{\Theta}^{ML}$ at which point the probability of observing the result **x** is maximum.

E1b Elementary properties

The ML estimator possesses **functional invariance**: for every continuous function ϕ it holds that $\hat{\phi}(\Theta)^{ML} = \phi(\hat{\Theta}^{ML})$.

To compute the ML estimator, the usual way (assuming there is no additional constraint on Θ) is to solve the first-order necessary condition for maximality: $\frac{\partial}{\partial\theta_i} L(\Theta|\mathbf{x}) = 0$, or equivalently $\frac{\partial}{\partial\theta_i} \Lambda(\Theta|\mathbf{x}) = 0$ in which $\Lambda(\Theta|\mathbf{x}) = \ln L(\Theta|\mathbf{x})$ is called the log likelihood function. When there are additional constraints on Θ, Kuhn-Tucker conditions should rather be used.

E1c ML estimation of a gaussian distribution

We can compute the ML estimators of the parameters of a gaussian R.V. *X* distributed $N(\mu, \sigma^2)$ in the following way, based on a simple random sample $\mathbf{x} = (x_i)_{i=1..n}$.

$$L(\mu,\sigma^2 \mid \mathbf{x}) = g(\mathbf{x} \mid \mu,\sigma^2) = \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2})$$

hence $\Lambda(\mu,\sigma^2 \mid \mathbf{x}) = -\frac{1}{2}\sum_{i=1}^{n}\ln(2\pi\sigma^2) + \frac{(x_i-\mu)^2}{\sigma^2} = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}$.

The first order derivatives are $\frac{\partial\Lambda}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$ and $\frac{\partial\Lambda}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^4}$.

From $\frac{\partial\Lambda}{\partial\mu} = 0$ we get $\hat{\mu}^{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}x_i$ whereas $\frac{\partial\Lambda}{\partial(\sigma^2)} = 0$ yields $\hat{\sigma}^{2\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$.

In the case of a gaussian R.V., the ML estimators are identical to the empiric estimators.

E1d ML estimation of a multinomial distribution

The estimation problem is to recover the parameters $\boldsymbol{\pi} = (\pi_k)_{k=1..m}$ of a multinomial distribution, from a random sample of size *n* yielding absolute frequencies $(n_k)_{k=1..m}$. Thus $n_k \in \{0,1..n\}$ and $\sum_{k=1}^{m} n_k = n$.

The PDF of a single trial *i* is $\pi_{k(i)}$, in which *k*(*i*) is the site obtained in the trial. Thus the random sample has PDF $\prod_{k=1}^{m}(\pi_k)^{n_k}$, which defines the likelihood of parameter $\boldsymbol{\pi}$ and also the log likelihood function $\Lambda(\boldsymbol{\pi}) = \sum_{k=1}^{m} n_k \ln\pi_k$.

The ML estimator is obtained by maximizing $\Lambda(\boldsymbol{\pi})$ subject to the constraints $\pi_k \geq 0$ and $\sum_{k=1}^{m}\pi_k = 1$. Let us associate a multiplier *z* to the equality constraint. The lagrangean function of the optimization problem is

$$L(\boldsymbol{\pi}, z) = \Lambda(\boldsymbol{\pi}) - z(\sum_{k=1}^{m}\pi_k - 1) = [\sum_{k=1}^{m} n_k \ln(\pi_k) - z\pi_k] - z$$

with first-order derivatives $\frac{\partial L}{\partial\pi_k} = \frac{n_k}{\pi_k} - z$ and $\frac{\partial L}{\partial z} = 1 - \sum_{k=1}^{m}\pi_k$. The first-order necessary conditions for optimality are also sufficient since Λ is concave in $\boldsymbol{\pi}$; these are stated as

$$\frac{\partial L}{\partial\pi_k} \leq 0,\ \pi_k\frac{\partial L}{\partial\pi_k} = 0 \text{ and } \frac{\partial L}{\partial z} = 0.$$

By summing the complementary slackness conditions we obtain that $\sum_{k=1}^{m} n_k - z = 0$ hence *z* = *n*. For each category *k*, either $n_k = 0$ hence $\hat{\pi}_k^{\mathrm{ML}} = 0$, or $n_k > 0$ hence $\hat{\pi}_k^{\mathrm{ML}} = n_k / n$. Whatever the case, the ML estimator $\hat{\pi}_k^{\mathrm{ML}} = n_k / n$ is identical to the empiric relative frequency.

E1e Asymptotic properties

Asymptotically (for very large sample size) the ML estimator is convergent, i.e. it concentrates on the true value. Thus in the limit it has no bias and its deviation tends to zero.

Let us define the information matrix $I_n(\Theta) = [\mathrm{cov}(\frac{\partial \Lambda}{\partial \theta_i}, \frac{\partial \Lambda}{\partial \theta_j})]_{i,j}$ where expectation is conducted over the random samples **x** of size *n*. It can be shown that $I_n(\Theta) = -\mathrm{E}[\frac{\partial^2 \Lambda}{\partial \theta_i \partial \theta_j}]_{i,j}$ and is related to the asymptotic distribution of $\hat{\Theta}^{\mathrm{ML}}$.

Precisely, $\sqrt{n}(\hat{\Theta}^{\mathrm{ML}} - \Theta)$ is asymptotically distributed multivariate normal with mean zero and covariance matrix $\lim_{n\to\infty}(\frac{1}{n} I_n(\Theta))^{-1}$. This asymptotic normality of the ML estimator enables us to state confidence intervals and to perform tests of significance on the estimated parameters.

We may test whether the *i*-th component $\hat{\theta}_i^{\mathrm{ML}}$ is different from 0 at the 1-α confidence level by constructing the ratio $t = \hat{\theta}_i^{\mathrm{ML}} / \sqrt{\mathrm{var}(\hat{\theta}_i^{\mathrm{ML}})}$ and computing the probability that a reduced gaussian variable has larger absolute value than *t*.

E1f Comparison of models

Another application of ML is to evaluate the likelihood of a restricted model, as compared to a reference model. Let us assume that in the restricted model the vector Θ' of *N'* parameters is a subvector of Θ and corresponds to a value of Θ with the *N*-*N'* components out of Θ' set at zero value.

Then the log of the squared ML ratio, $\ln([\mathrm{L}(\hat{\Theta}^{\mathrm{ML}}) / \mathrm{L}(\hat{\Theta}'^{\mathrm{ML}})]^2) = 2[\Lambda(\hat{\Theta}^{\mathrm{ML}}) - \Lambda(\hat{\Theta}'^{\mathrm{ML}})]$, is distributed as a chi-square random variable with *N*-*N'* degrees of freedom. We can also evaluate the critical probability of the restriction hypothesis $\Theta = \Theta' \cup \{0\}_{N-N'}$. This is the probability that a chi-square variable with *N*-*N'* degrees of freedom has value larger than the observed ratio, say *x*: it can be written as $\Pr(\chi^2_{N-N'} \geq x)$.

E2 Likelihood function of DCM

The ML estimation of a DCM model involves the following three steps: (1) the formation of the likelihood function, which depends on the model formulation and the sampling scheme; (2) the maximization of the likelihood function, which amounts to solving a mathematical program; (3) the design of confidence intervals and tests.

All steps will be illustrated in the next Section. We shall concentrate here on the first one.

E2a Random sample

When *n* decision-makers *i* are sampled at random and independently, the observation reveals the chosen variants $(k(i))_{i=1..n}$ and yields absolute frequencies $(n_k)_{k=1..m}$. Assuming that the true relative frequency is a function $\pi_k(\Theta)$ of parameter Θ, the probability of the observation is also $\prod_{k=1}^{m} [\pi_k(\Theta)]^{n_k}$.

As a likelihood function, it depends on parameter Θ (which differs from the likelihood function of a multinomial distribution). The associated log likelihood function is

$$\Lambda(\Theta) = \sum_{k=1}^{m} n_k \ln \pi_k(\Theta) .$$

E2b Random sample of a category model

Let us define a category model as a set of models *i* which share the same parameter Θ. Instances include route- or mode-choice model analyzed at the level of the O-D pair *i*. We assume that in model *i* the true relative frequency of variant *k* is a function $\pi_{ki}(\Theta)$ of parameter Θ.

A random sample of *n* decision-makers decomposes into category subsamples with sizes n_i and variant absolute frequencies n_{ki}. Thus the observation has probability

$$\prod_{i\in I}\prod_{k=1}^{m}[\pi_{ki}(\Theta)]^{n_{ki}} \ .$$

The associated log likelihood function is

$$\Lambda(\Theta) = \sum_{i\in I}\sum_{k=1}^{m} n_{ki} \ln \pi_{ki}(\Theta) \ .$$

E2c Choice-based sample

A choice-based sample consists in as many subsamples as there are choice options *k*. Assuming that the observed attributes of an individual *t* coincide with the definition of a category *i* (eg. when only the origin and destination of a trip are surveyed), the probability of observing an individual *t* is $\Pr(k_t \cap i_t)$. This depends on the conditional probability of choosing *k* for the *i*-th category, $\Pr(k|i)$, on the basis of

$$\Pr(k_t \cap i_t) \ = \ \Pr(k|i) \ . \ p_i$$

in which p_i denotes the (marginal) probability of category *i*.

The marginal probability p_i is itself estimated from the variant subsamples. Let n_k be the size of the *k*-th subsample and N_k the population size of variant *k*. Then the sample rate of variant *k* is n_k / N_k.

The size of category *i* is estimated by

$$\hat{N}_i = \sum_k n_{ki} \frac{N_k}{n_k} \ .$$

Thus we obtain the following estimator of the marginal probability:

$$\hat{p}_i = \frac{\hat{N}_i}{\sum_{j\in I}\hat{N}_j} = \frac{\hat{N}_i}{\sum_k N_k} \ .$$

The overall sample is a stratified sample with *m* strata. The probability to belong to stratum *k* is n_k / n. Given the stratum *k*, the probability to belong to category *i* is $\Pr(i|k) = \Pr(k \cap i)/\Pr(k)$ $= \Pr(k|i).\Pr(i)/\Pr(k) \ = \ \Pr(k|i) \ \dfrac{p_i}{\sum_{j\in I}\Pr(k|j)p_j}$ since $\Pr(k) = \sum_{j\in I}\Pr(k|j)p_j$.

The overall likelihood function is also

$$\mathrm{L}(\Theta) = \prod_{k=1}^{m}\prod_{t\in k}\frac{n_k}{n}\frac{\Pr(k|i_t)p_{i_t}}{\sum_{j\in I}\Pr(k|j)p_j} \ = \prod_{k=1}^{m}\prod_{i\in I}[\frac{n_k}{n}\frac{\Pr(k|i)p_i}{\sum_{j\in I}\Pr(k|j)p_j}]^{n_{ki}} \ .$$

The associated log likelihood function is

$$\Lambda(\Theta) = \sum_{k=1}^{m} \sum_{i \in I} n_{ki} \ln\left(\frac{n_k}{n} \frac{\Pr(k|i)\, p_i}{\sum_{j \in I} \Pr(k|j)\, p_j}\right)$$

$$= \left(\sum_{k,i} n_{ki} \ln \Pr(k|i)\right) - \left(\sum_k n_k \ln\left[\frac{N_k}{N n_k} \sum_i \Pr(k|i)\, p_i\right]\right).$$

We replace the unknown p_i by the values $\hat{p}_i$ estimated from the sample, to obtain the following estimated log likelihood function, which differs from its simple counterpart by its second term, which depends on $\Pr(k|i)$:

$$\hat{\Lambda}(\Theta) = \left(\sum_{k,i} n_{ki} \ln \Pr(k|i)\right) - \left(\sum_k n_k \ln\left[\frac{N_k}{N n_k} \sum_i \Pr(k|i)\, \hat{p}_i\right]\right).$$

E3 Case study of the Prado-Carénage tunnel

The Prado-Carénage tunnel in Marseilles was opened to car traffic in September 1993. A major survey was implemented in March 1995, which yielded a considerable amount of valuable information on the route choices made by drivers in Marseilles.

We shall first describe briefly the survey results. Then we shall explain the econometric processing of the base model, i.e. a binary logit model with a varying parameter. Lastly we shall indicate the main results of the estimation of the base and simplified models.

E3a Survey results

Measurements of journey times and objective knowledge about paths. Travel times were surveyed on a set of road routes which pass through the tunnel or its main competing roads. Each was subjected to several types of periodic measurement (morning peak, evening peak or the rest of the day), so that the mean journey time on these routes be known to within two minutes at the 95% confidence level. Remark that the routes in the survey are the "main legs" of the true routes, and *they ignore the end legs of these*, which are specific to each trip. The distance and price of paths can also be found. The distance is measured in the field or on a map. The price is estimated from the toll if the route goes through the tunnel and by estimating the costs of car travel on the basis of the product of the distance and a ratio per unit distance. Note that in 1995 the toll fare was 11 F, for a tunnel 2.5 km long.

The origin-destination (O-D) surveys and the revelation of preferences. The O-D surveys intercepted the vehicles in the tunnel or on the competing roads and revealed not only certain characteristics of the drivers (age, sex) but also some of the circumstances surrounding the trip (purpose, location of departure and arrival and, of course, the time of passage through the survey station). The O-D surveys also *reveal the route choices* made by drivers because a driver who is intercepted at a station has chosen the route he or she is on in preference to alternative routes. This provides indirect information on the VoT distribution.

Additional information from a network assignment model. As data about the routes which are not chosen is also required, a network assignment model was used to compute the "shortest" alternative path (i.e. that with the shortest journey time) on the basis of the modelled journey times.

E3b Econometric processing

Let us recall the economic formula for the difference in random utility of individual *i* between the free route and the toll route:

$$\Delta U(i) = -(P_{free} - P_{toll}) - v_i(T_{free} - T_{toll}) - \theta_D(D_{free} - D_{toll}) + \Delta\varepsilon \qquad (A)$$

We shall replace (A) by the following:

$$\Delta U'(i) = P - v_i \Delta T' + \Delta\varepsilon' \qquad (B)$$

where $P = P_{toll} - P_{free}$, $\Delta T'$ denotes the difference between the estimated mean main leg journey times between the surface, free route and the competing tunnel route while $\Delta\varepsilon'$ is a new disruption which includes both the disruption $\Delta\varepsilon$ in the economic model and the difference $\Delta T - \Delta T'$ between the true time saving, ΔT, and what is substituted for it, $\Delta T'$.

Thus the disruption $\Delta\varepsilon'$ represents (i) the uncertainty in the economic behaviour model, (ii) the fluctuation affecting the sampling of observations, (iii) the uncertainty affecting main leg journey times and end leg journey times,, (iv) the uncertainty affecting computation of the alternative route, (v) the uncertainty affecting perception of times, (vi) the uncertainty affecting the aggregation of O-D pairs. The combined effect of these sources of variability and uncertainty is to interfere with the exploitation of observations with a view to finding important unknown quantities such as the distribution parameters for VoT. The ultimate aim of the econometric model is to free us as far as possible from uncertainties because it frees us from the variabilities which we are able to make explicit. Here the purpose is to separate the variability which is associated with VoTs from the residual uncertainty $\Delta\varepsilon'$.

Modal market share. Given P, $\Delta T'$ and a VoT of *v*, the individual toll route market share is

$$\Pr(\text{Toll} \mid v) = \Pr\{\Delta U'(v) \le 0 \mid v\} = \Pr\{\Delta\varepsilon' \le v\Delta T' - P \mid v\} = Z(v\Delta T' - P) \qquad (C)$$

In which we let Z denote the cumulative density function of $\Delta\varepsilon'$. This is the *individual toll route market share*, specific to a given VoT *v*. Under the additional assumption that Z does not depend on *v*, we obtain the following *aggregate market share* of the toll route

$$\begin{aligned}\Pr(\text{toll}) = \Pr\left(\left\{\Delta U'(i) \le 0\right\}\right) &= \int_v \Pr\left(\left\{\Delta U'(i) \le 0 \mid v\right\}\right) d\Pr(v) \\ &= \int_v \Pr\left(\left\{\Delta\varepsilon' \le v\Delta T' - P \mid v\right\}\right) dH(v) \qquad (D)\end{aligned}$$

in which function H is the cumulative distribution function of the VoT, so $dH(v) = d\Pr(v)$. By changing variables $\alpha = H(v)$ and letting H^{-1} denote the reciprocal of H, this is transformed into

$$\Pr(\text{Toll}) = \int_0^1 Z\left(H^{-1}(\alpha)\Delta T' - P\right) d\alpha \,. \qquad (E)$$

Parametric estimation problem. Formula (E) gives the toll route market share as a function of the distributions H and Z. We shall formulate an estimation problem in which H and Z are the unknowns; more precisely, we assume that H and Z belong to certain functional classes within which they are identified by the numeric values of certain parameters. We assume that $Z(x) = Z_0((x - \Delta\bar{\varepsilon}')/\sigma_{\Delta\varepsilon'})$ where the two parameters $\Delta\bar{\varepsilon}'$ and $\sigma_{\Delta\varepsilon'}$ are respectively the mean and standard deviation of $\Delta\varepsilon'$, and Z_0 is the (known) CDF of a reduced variate. For example, we can assume that Z_0 is the distribution function Φ of a reduced normal variate,

which gives a probit model; if in addition we specify that H is a log-normal distribution we construct a probit model with a log-normal varying coefficient.

The issue of identifiability. As P does not vary in the observations, the parameters are related to each other. Letting μ and σ be the mean and standard deviation of the log VoT, only the scalar terms $(\Delta\overline{\varepsilon}' + P)/\sigma_{\Delta\ \varepsilon'}$, $\mu - \ln(\sigma_{\Delta\varepsilon'})$ and σ can be identified. Thus the mean VoT cannot be found, as we can only know $\mu - \ln(\sigma_{\Delta\varepsilon'})$. On the other hand σ can be identified. If we assume that the VoT are log-normally distributed, the ratio between the standard deviation and the mean depends only on σ on the basis of $(\exp(\sigma^2) - 1)^{1/2}$.

E3c Results

Estimated parameters. We estimated a logit model with log-normal VoT by the maximum likelihood method, yielding that $(\Delta\overline{\varepsilon}' + P)/\sigma_{\Delta\varepsilon'}$ = 1.535, $\mu - \ln(\sigma_{\Delta\varepsilon'})$ = -2.03, σ = 0.662. We applied the likelihood ratio test to the hypothesis "There is only one VoT", i.e. "σ = 0": as compared to a logit model with a single VoT, there is a 12 point log-likelihood gain, which means that the probability of a unique VoT is less than 10^{-4}!

Alternative specifications. Estimation of fixed coefficient stochastic models, distributed coefficient models with no residual random variate, logit models with a log-logistic or rectangular coefficient, confirms firstly the dispersion of values-of-time and the clear statistical superiority of the models which make it explicit, and secondly the importance of the role played by the residual random variable.

Tab. C. Results of alternative models.

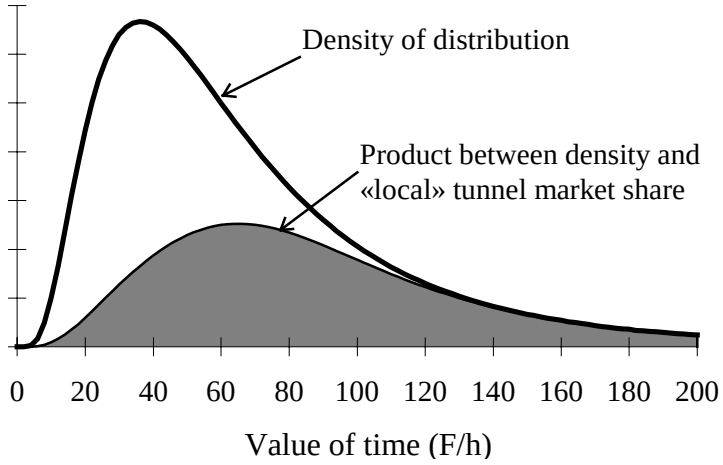
Model specification	Value of Time (F/h)			$\sigma_{\Delta\varepsilon'}$	Log Likelihood
	Mean	Median	SD		
Logit with single VoT	63.0	63.0	0	8.37	-1 599.27
Probit with single VoT	60.2	60.2	0	7.60	-1 596.65
VoT Log-Nor, $\Delta\varepsilon = 0$	253	58.2	1070	0	-1 471.94
VoT Log-Log, $\Delta\varepsilon = 0$	381	59.2	∞	0	-1 453.49
Logit with Log-Nor VoT	70.3	56.5	52.2	7.17	-1 587.80
Logit with Log-Log VoT	73.0	56.9	76.7	7.12	-1 587.76
Logit with uniform VoT	63.6	63.6	51.5	7.05	-1 587.39

Inference of the mean VoT. Assuming that the mean random disruption is zero, we obtain that $\sigma_{\Delta\varepsilon'} = 11/1.535 \approx 7.17$ F, from which we can compute a median VoT of 56.5 F/h, a mean of 70.3 F/h and a standard deviation of 52.2 F/h. The uncertainty relating to the VoT parameters is of the order of 2 or 3 F/h.

Confidence intervals on the mean VoT, the median VoT and the standard deviation of VoT were obtained by propagating the estimated variances $\text{var}(\hat{\mu})$, $\text{var}(\hat{\sigma}^2)$ and covariance $\text{cov}(\hat{\mu}, \hat{\sigma}^2)$ through the definitional relationships $\text{Mean} = \exp(\mu + \sigma^2/2)$, $\text{Median} = \exp(\mu)$ and $\text{SD} = \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)^{1/2}$.

The share of variability. We can compare the variability caused by the dispersion of VoT to the residual variability var[$\Delta\varepsilon'$]. For a fixed time saving of ΔT' and no distance saving, the variance of ΔU' can be broken down into $\Delta T'^2 \, \mathrm{var}(v) + \mathrm{var}(\Delta\varepsilon')$. If we consider a median time saving ΔT' = 11 mn and a lognormal VoT, we obtain $\Delta T'^2 \, \mathrm{var}(v) \approx 91 F^2$ in comparison to $\mathrm{var}(\Delta\varepsilon') \approx 52 F^2$. Thus the variability due to the VoT is almost twice the residual variability, which means that it is thoroughly justified to make it explicit!

Fig. c. Modal share of the tunnel according to the value-of-time (for $\Delta T'$ = 11 mn).



E3d Comments

The results we shall conserve are as follows: firstly, the dispersion of VoT for car journeys in Marseilles. The importance of residual causes other than the mean time saving or the distance saving should also be noted - these are responsible for one third of the variability of choices (through the residual random variate). We should also not forget the supplementary hypothesis which we have used to determine a mean VoT: the absence of a mean difference (other than that relating to time, distances or the toll) between the tunnel and competing routes. To free ourselves from this hypothesis the problem would need to be examined with several levels of tolls.

F Statistical tables

F1 Reduced normal distribution

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Area under the reduced normal distribution

Cell contains $\Phi(x+y)$-0.5 in which x is the line value and y the column value

F2 Student distributions

DF	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.005$
1	6,314	12,706	63,656
2	2,920	4,303	9,925
3	2,353	3,182	5,841
4	2,132	2,776	4,604
5	2,015	2,571	4,032
6	1,943	2,447	3,707
7	1,895	2,365	3,499
8	1,860	2,306	3,355
9	1,833	2,262	3,250
10	1,812	2,228	3,169
11	1,796	2,201	3,106
12	1,782	2,179	3,055
13	1,771	2,160	3,012
14	1,761	2,145	2,977
15	1,753	2,131	2,947
16	1,746	2,120	2,921
17	1,740	2,110	2,898
18	1,734	2,101	2,878
19	1,729	2,093	2,861
20	1,725	2,086	2,845
21	1,721	2,080	2,831
22	1,717	2,074	2,819
23	1,714	2,069	2,807
24	1,711	2,064	2,797
25	1,708	2,060	2,787
26	1,706	2,056	2,779
27	1,703	2,052	2,771
28	1,701	2,048	2,763
29	1,699	2,045	2,756
30	1,697	2,042	2,750
40	1,684	2,021	2,704
50	1,676	2,009	2,678
60	1,671	2,000	2,660
120	1,658	1,980	2,617
180	1,653	1,973	2,603
240	1,651	1,970	2,596
300	1,650	1,968	2,592
360	1,649	1,967	2,590
420	1,648	1,966	2,588
480	1,648	1,965	2,586
540	1,648	1,964	2,585
600	1,647	1,964	2,584

Inverse CDF of the Student distribution

Cell contains the value *t* yielding CDF of 1-α (under the nb of DF)

F3 Chi-square distributions

DF	0,995	0,975	0,950	0,500	0,100	0,050	0,025	0,010	0,005
1	3,9E-05	0,00098	0,00393	0,455	2,706	3,841	5,024	6,635	7,879
2	0,010	0,051	0,103	1,386	4,605	5,991	7,378	9,210	10,597
3	0,072	0,216	0,352	2,366	6,251	7,815	9,348	11,345	12,838
4	0,207	0,484	0,711	3,357	7,779	9,488	11,143	13,277	14,860
5	0,412	0,831	1,145	4,351	9,236	11,070	12,832	15,086	16,750
6	0,676	1,237	1,635	5,348	10,645	12,592	14,449	16,812	18,548
7	0,989	1,690	2,167	6,346	12,017	14,067	16,013	18,475	20,278
8	1,344	2,180	2,733	7,344	13,362	15,507	17,535	20,090	21,955
9	1,735	2,700	3,325	8,343	14,684	16,919	19,023	21,666	23,589
10	2,156	3,247	3,940	9,342	15,987	18,307	20,483	23,209	25,188
11	2,603	3,816	4,575	10,341	17,275	19,675	21,920	24,725	26,757
12	3,074	4,404	5,226	11,340	18,549	21,026	23,337	26,217	28,300
13	3,565	5,009	5,892	12,340	19,812	22,362	24,736	27,688	29,819
14	4,075	5,629	6,571	13,339	21,064	23,685	26,119	29,141	31,319
15	4,601	6,262	7,261	14,339	22,307	24,996	27,488	30,578	32,801
16	5,142	6,908	7,962	15,338	23,542	26,296	28,845	32,000	34,267
17	5,697	7,564	8,672	16,338	24,769	27,587	30,191	33,409	35,718
18	6,265	8,231	9,390	17,338	25,989	28,869	31,526	34,805	37,156
19	6,844	8,907	10,117	18,338	27,204	30,144	32,852	36,191	38,582
20	7,434	9,591	10,851	19,337	28,412	31,410	34,170	37,566	39,997
21	8,034	10,283	11,591	20,337	29,615	32,671	35,479	38,932	41,401
22	8,643	10,982	12,338	21,337	30,813	33,924	36,781	40,289	42,796
23	9,260	11,689	13,091	22,337	32,007	35,172	38,076	41,638	44,181
24	9,886	12,401	13,848	23,337	33,196	36,415	39,364	42,980	45,558
25	10,520	13,120	14,611	24,337	34,382	37,652	40,646	44,314	46,928
26	11,160	13,844	15,379	25,336	35,563	38,885	41,923	45,642	48,290
27	11,808	14,573	16,151	26,336	36,741	40,113	43,195	46,963	49,645
28	12,461	15,308	16,928	27,336	37,916	41,337	44,461	48,278	50,994
29	13,121	16,047	17,708	28,336	39,087	42,557	45,722	49,588	52,335
30	13,787	16,791	18,493	29,336	40,256	43,773	46,979	50,892	53,672
40	20,707	24,433	26,509	39,335	51,805	55,758	59,342	63,691	66,766
50	27,991	32,357	34,764	49,335	63,167	67,505	71,420	76,154	79,490
60	35,534	40,482	43,188	59,335	74,397	79,082	83,298	88,379	91,952
70	43,275	48,758	51,739	69,334	85,527	90,531	95,023	100,425	104,215
80	51,172	57,153	60,391	79,334	96,578	101,879	106,629	112,329	116,321
90	59,196	65,647	69,126	89,334	107,565	113,145	118,136	124,116	128,299
100	67,328	74,222	77,929	99,334	118,498	124,342	129,561	135,807	140,170

Inverse CDF of the chi-square distribution

Cell contains the value x yielding CDF of 1-α (under the nb of DF)

F4 Fisher-Snedecor distribution

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	4052,2	98,502	34,116	21,198	16,258	13,745	12,246	11,259	10,562	10,044	9,3303	8,6832	8,096	7,8229	7,5624	7,3142	7,0771	6,8509
2	4999,3	99	30,816	18	13,274	10,925	9,5465	8,6491	8,0215	7,5595	6,9266	6,3588	5,849	5,6136	5,3903	5,1785	4,9774	4,7865
3	5403,5	99,164	29,457	16,694	12,06	9,7796	8,4513	7,591	6,992	6,5523	5,9525	5,417	4,9382	4,7181	4,5097	4,3126	4,1259	3,9491
4	5624,3	99,251	28,71	15,977	11,392	9,1484	7,8467	7,0061	6,4221	5,9944	5,4119	4,8932	4,4307	4,2185	4,0179	3,8283	3,6491	3,4795
5	5764	99,302	28,237	15,522	10,967	8,7459	7,4604	6,6318	6,0569	5,6364	5,0644	4,5556	4,1027	3,8951	3,699	3,5138	3,3389	3,1735
6	5859	99,331	27,911	15,207	10,672	8,466	7,1914	6,3707	5,8018	5,3858	4,8205	4,3183	3,8714	3,6667	3,4735	3,291	3,1187	2,9559
7	5928,3	99,357	27,671	14,976	10,456	8,26	6,9929	6,1776	5,6128	5,2001	4,6395	4,1416	3,6987	3,4959	3,3045	3,1238	2,953	2,7918
8	5981	99,375	27,489	14,799	10,289	8,1017	6,8401	6,0288	5,4671	5,0567	4,4994	4,0044	3,5644	3,3629	3,1726	2,993	2,8233	2,6629
9	6022,4	99,39	27,345	14,659	10,158	7,976	6,7188	5,9106	5,3511	4,9424	4,3875	3,8948	3,4567	3,256	3,0665	2,8876	2,7185	2,5586
10	6055,9	99,397	27,228	14,546	10,051	7,8742	6,6201	5,8143	5,2565	4,8491	4,2961	3,8049	3,3682	3,1681	2,9791	2,8005	2,6318	2,4721
12	6106,7	99,419	27,052	14,374	9,8883	7,7183	6,4691	5,6667	5,1115	4,7058	4,1553	3,6662	3,2311	3,0316	2,8431	2,6648	2,4961	2,3363
15	6157	99,433	26,872	14,198	9,7223	7,559	6,3144	5,5152	4,9621	4,5582	4,0096	3,5222	3,088	2,8887	2,7002	2,5216	2,3523	2,1915
20	6208,7	99,448	26,69	14,019	9,5527	7,3958	6,1555	5,3591	4,808	4,4054	3,8584	3,3719	2,9377	2,738	2,5487	2,3689	2,1978	2,0346
24	6234,3	99,455	26,597	13,929	9,4665	7,3128	6,0743	5,2793	4,729	4,3269	3,7805	3,294	2,8594	2,6591	2,4689	2,288	2,1154	1,95
30	6260,4	99,466	26,504	13,838	9,3794	7,2286	5,992	5,1981	4,6486	4,2469	3,7008	3,2141	2,7785	2,5773	2,386	2,2034	2,0285	1,86
40	6286,4	99,477	26,411	13,745	9,2912	7,1432	5,9084	5,1156	4,5667	4,1653	3,6192	3,1319	2,6947	2,4923	2,2992	2,1142	1,936	1,7628
60	6313	99,484	26,316	13,652	9,202	7,0568	5,8236	5,0316	4,4831	4,0819	3,5355	3,0471	2,6077	2,4035	2,2079	2,0194	1,8363	1,6557
120	6339,5	99,491	26,221	13,558	9,1118	6,969	5,7373	4,9461	4,3977	3,9965	3,4494	2,9594	2,5168	2,3099	2,1108	1,9172	1,7263	1,533

Inverse function of CDF of Fisher-Snedecor distribution with parameters Line, Column and argument 1%

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	161,45	18,513	10,128	7,7086	6,6079	5,9874	5,5915	5,3176	5,1174	4,9646	4,7472	4,5431	4,3513	4,2597	4,1709	4,0847	4,0012	3,9201
2	199,5	19	9,5521	6,9443	5,7861	5,1432	4,7374	4,459	4,2565	4,1028	3,8853	3,6823	3,4928	3,4028	3,3158	3,2317	3,1504	3,0718
3	215,71	19,164	9,2766	6,5914	5,4094	4,7571	4,3468	4,0662	3,8625	3,7083	3,4903	3,2874	3,0984	3,0088	2,9223	2,8387	2,7581	2,6802
4	224,58	19,247	9,1172	6,3882	5,1922	4,5337	4,1203	3,8379	3,6331	3,478	3,2592	3,0556	2,8661	2,7763	2,6896	2,606	2,5252	2,4472
5	230,16	19,296	9,0134	6,2561	5,0503	4,3874	3,9715	3,6875	3,4817	3,3258	3,1059	2,9013	2,7109	2,6207	2,5336	2,4495	2,3683	2,2899
6	233,99	19,329	8,9407	6,1631	4,9503	4,2839	3,866	3,5806	3,3738	3,2172	2,9961	2,7905	2,599	2,5082	2,4205	2,3359	2,2541	2,175
7	236,77	19,353	8,8867	6,0942	4,8759	4,2067	3,7871	3,5005	3,2927	3,1355	2,9134	2,7066	2,514	2,4226	2,3343	2,249	2,1665	2,0868
8	238,88	19,371	8,8452	6,041	4,8183	4,1468	3,7257	3,4381	3,2296	3,0717	2,8486	2,6408	2,4471	2,3551	2,2662	2,1802	2,097	2,0164
9	240,54	19,385	8,8123	5,9988	4,7725	4,099	3,6767	3,3881	3,1789	3,0204	2,7964	2,5876	2,3928	2,3002	2,2107	2,124	2,0401	1,9588
10	241,88	19,396	8,7855	5,9644	4,7351	4,06	3,6365	3,3472	3,1373	2,9782	2,7534	2,5437	2,3479	2,2547	2,1646	2,0773	1,9926	1,9105
12	243,9	19,412	8,7447	5,9117	4,6777	3,9999	3,5747	3,2839	3,0729	2,913	2,6866	2,4753	2,2776	2,1834	2,0921	2,0035	1,9174	1,8337
15	245,95	19,429	8,7028	5,8578	4,6188	3,9381	3,5107	3,2184	3,0061	2,845	2,6169	2,4034	2,2033	2,1077	2,0148	1,9245	1,8364	1,7505
20	248,02	19,446	8,6602	5,8025	4,5581	3,8742	3,4445	3,1503	2,9365	2,774	2,5436	2,3275	2,1242	2,0267	1,9317	1,8389	1,748	1,6587
24	249,05	19,454	8,6385	5,7744	4,5272	3,8414	3,4105	3,1152	2,9005	2,7373	2,5055	2,2878	2,0825	1,9838	1,8874	1,7929	1,7001	1,6084
30	250,1	19,463	8,6166	5,7459	4,4957	3,8082	3,3758	3,0794	2,8637	2,6996	2,4663	2,2468	2,0391	1,939	1,8409	1,7444	1,6491	1,5543
40	251,14	19,471	8,5944	5,717	4,4638	3,7743	3,3404	3,0428	2,8259	2,6609	2,4259	2,2043	1,9938	1,892	1,7918	1,6928	1,5943	1,4952
60	252,2	19,479	8,572	5,6878	4,4314	3,7398	3,3043	3,0053	2,7872	2,6211	2,3842	2,1601	1,9464	1,8424	1,7396	1,6373	1,5343	1,429
120	253,25	19,487	8,5494	5,6581	4,3985	3,7047	3,2674	2,9669	2,7475	2,5801	2,341	2,1141	1,8963	1,7896	1,6835	1,5766	1,4673	1,3519

Inverse function of CDF of Fisher-Snedecor distribution with parameters Line, Column and argument 5%