Appendix

T. Bartz-Beielstein et al. (eds.), *Experimental Methods for the Analysis of Optimization Algorithms*, DOI 10.1007/978-3-642-02538-9, © Springer-Verlag Berlin Heidelberg 2010

Appendix A
A Brief Introduction to Inferential Statistics

Dario Basso

Abstract This appendix introduces the elements of statistical theory that are used throughout the book. It starts by defining random variables and the theoretical models to describe them. It then briefly outlines the concepts underlying point estimation. The central part is dedicated to hypothesis testing and confidence intervals by means of which inference from sample statistics to population parameters is carried out. Subsequently, the treatment focuses on regression and modeling, which play a fundamental role in several chapters of this book. The presentation is necessarily limited to linear regression and to basic model fitting. For a broader treatment of statistical theory the reader is referred to any textbook of statistics and to Mood et al. (1974) and Davison (2008).

A.1 Introduction

Inferential statistics is a collection of techniques that allow us to deduce information from a set of observed data about a phenomenon under investigation in a *population* of interest.

The motivation for such techniques is the fact that surveying the phenomenon on the entire population (census) in order to gain complete knowledge of it is often unrealistic or too expensive. An alternative is then to extract a set of *statistical units* (i.e., a the phenomenon on this restricted set of units. We may do this because we expect the units in the sample to reproduce the phenomenon as it is in the population (with a certain degree of *variability* depending on the *size* of the sample).

A "well-chosen" sample should be representative of the population, therefore the statistical units to be included in the sample should be chosen independently of the

---

Dario Basso
Department of Statistics, University of Padua, Italy
e-mail: dario@stat.unipd.it

characteristics of the phenomenon in the population. This is usually known as a *random sampling* from a population.

There are several ways of sampling units from a population (e.g., with replacement or without replacement), and we may consider finite populations (e.g., the nonisomorphic graphs of a certain size for which we wish to determine the chromatic number) or nonfinite populations (such as the runtime of an optimization algorithm).

In inferential statistics, the observed data are usually considered as n independent realizations of a random experiment, whose possible outcomes are described by a *random variable* (r.v.).[1] The distribution of the random variable depends on some unknown parameters of the population and on how the sample has been extracted. We can also say that the random variable is a *model* of the phenomenon in the population, and that each datum is a realization of the same random variable. According to this model, the sample data are then considered as n realizations of *independent* and *identically distributed* (i.i.d.) random variables.

In *parametric* inference, the random variable describing the experiment is defined by a *probabilistic model*, which is usually a mathematical formula determining the probability of an event (or a set of events). The probabilistic model is identified by a parameter θ (or sometimes by a vector of parameters) that takes values in a *parameter space* Θ. The collection of probabilistic models identified by all possible values of θ in Θ is called a *parametric statistical model*. Here the inference on the phenomenon in the population is translated into an inference on the unknown parameter θ that identifies a specific distribution among those that belong to the statistical model. In the *likelihood* function approach, we seek the value for θ that is most in agreement (*likely*) with the observed data. There are other approaches to inference, such as nonparametric and Bayesian inference.

This probabilistic approach serves two purposes: (i) describing the variability of the sample outcomes and (ii) evaluating the uncertainty of the inference, e.g., by providing an interval of possible values for the true parameter θ, or by evaluating the risk of incorrectly answering the question "does the true parameter θ belong to a certain subset $\Theta_0 \in \Theta$?".

Let us end this paragraph with an explanatory example: suppose that we have a deterministic program, for example, a mixed integer programming solver, and that we want to determine its ability to solve a specific class of instances of a certain optimization problem, say the set covering problem.[2] Let θ be the true, unknown proportion of instances that can be solved within a runtime of, say $t_0 = 1{,}000$ s, that is, the amount of time we are prepared to accept before giving up. Suppose that a random sample of n instances is taken from the class of instances, and that the application of the solver to each instance of the sample is coded into two possible outcomes: 0 meaning "not solved within t_0", and 1 meaning "solved." We can thus describe each outcome with a random variable X_i assuming the values 0 and 1 with probability θ and $1-\theta$, $i = 1, \ldots, n$. We know from probability calculus that the random variable S, the sum of n independent and identically distributed dichotomous variables (such as X_i), has a binomial distribution whose probability function is

---

[1] A random variable is a function assigning a real number to each element of a probability space.

[2] In the optimization version of the set covering problem, we are given a universe $\mathcal{U}$ and a family $\mathcal{S}$ of subsets of $\mathcal{U}$, and we want to find a cover, that is, a subfamily $\mathcal{C} \subseteq \mathcal{S}$ of sets whose union is $\mathcal{U}$, that minimizes the number of selected sets. A class of instances can be determined, for example, by specifying a range for the number of objects in $\mathcal{U}$ and for the number of subsets in $\mathcal{S}$ and a certain structure in the elements covered by the subsets. Then, the class of instances, although large, is a finite set.

$$\Pr\{S = s\} = p_S(s;\theta) = \binom{n}{s}\theta^s(1-\theta)^{(n-s)} \qquad s \in \{0, 1, \ldots, n\}, \quad \theta \in (0,1).$$

Then the binomial distribution is the probabilistic model describing the outcome (sum) of the sample data, whereas the statistical model is the set $\mathcal{P} = \{p_S(s;\theta), \theta \in (0,1)\}$. The initial goal evaluating the proportion of solvable instances, is then translated into estimating the unknown parameter θ.

Once an estimate of θ has been obtained, it will be possible, through a specified probabilistic model, to answer questions such as: "is $\theta > 90\%$?" "Given that, if we change sample (i.e., if we repeat the experiment) we will have a different result, can we say something about the uncertainty of θ (i.e., can we give a set of reasonable values for θ)?" Of course, each of the previous questions cannot be answered with certainty. Inferential statistics can answer the previous questions while determining the probability of "incorrect" conclusions on θ.

A.1.1 Random Variables

A univariate quantitative random variable (r.v.) X is a variable taking values in a domain D_X with prespecified probabilities. There are two kinds of quantitative r.v.s: *discrete* and *continuous* (or absolutely continuous). In the former case the cardinality of the support is at most numerable (i.e., it has almost the same cardinality of $\mathbb{N}$), in the latter $D_X \subseteq \mathbb{R}$.

An r.v. X is therefore defined by the specification of the domain D_X and the probability associated with each possible outcome of X. Two very intuitive examples are the following. The outcome of a fair die is characterized by an r.v. X with $D_X = \{1, 2, 3, 4, 5, 6\}$ and the probability associated with each outcome (also called the *realization* of X) is 1/6. The launch of a fair coin can be described by the r.v. X assuming two values (or *modalities*): "head" and "tail", each with probability 1/2. This last example actually refers to a *categorical* variable (i.e., an r.v. whose possible outcomes are categories or adjectives), that can be recoded in order to obtain a discrete one (e.g., taking values in $D_X = \{0, 1\}$, with 0 being "head" and 1 being "tail").

As far as discrete r.v.s are concerned, the probability of the event $\{X = x\}$, $x \in D_X$, is given by the *probability function* $p_X(x)$ is summarized by a mathematical formula. For instance, if X is dichotomous and the probability of the event $\{X = 1\}$ is denoted by the parameter $\theta \in (0,1)$, then $p_X(x) = \theta^x(1-\theta)^{(1-x)}$. This distribution is known as the *Bernoulli* distribution. The probability function satisfies

$0 \leq p_X(x) \leq 1$ for all $x \in D_X$, where 0 denotes the probability of an impossible event, and 1 that of an (almost) sure event.

Another important function that is related to the r.v.s is the *cumulative distribution function* (cdf), or *distribution function*, which is defined as

$$F_X(x) = \Pr\{X \leq x\}, \qquad x \in \mathbb{R}.$$

Note that $F_X(x)$ is a continuous, nonnegative, nondecreasing function that satisfies $\lim\limits_{x\to-\infty} F_X(x) = 0$ and $\lim\limits_{x\to+\infty} F_X(x) = 1$. The probability of the event $X \in [a, b]$ can be computed as $F_X(b) - F_X(a)$.

In the continuous case, the event $\{X = x\}$ has zero probability for all $x \in D_X$, so the probability function does not apply here. The r.v. is then described by the *density function* $f_X(x)$, which is defined as

$$f_X(x) = \lim_{\delta\to 0} \frac{\Pr\{X \leq x+\delta\} - \Pr\{X \leq x\}}{\delta} = \frac{\partial F_X(x)}{\partial x}, \quad \delta > 0.$$

Therefore the relationship between $F_X(x)$ and $f_X(x)$ can be defined as

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t \qquad x \in \mathbb{R}.$$

Of course, a similar definition can be applied to the discrete case by letting

$$F_X(x) = \sum_{t\leq x} \Pr\{X = t\}, \qquad x \in \mathbb{R}.$$

Note that, when X is discrete, its cdf is typically a stepfunction.

The functions $p_X(x)$ and $f_X(x)$ are nonnegative and must, respectively, satisfy

$$\sum_{x\in D_X} p_X(x) = 1, \qquad \int_{x\in D_X} f_X(x)\mathrm{d}x = 1.$$

Usually, the cdf can also be specified by a closed mathematical formula.

The distribution of an r.v. is characterized by some indexes. One of them is the *expected value*, which is defined as

$$\mathrm{E}[X] = \sum_{x\in D_X} x p_X(x) \qquad \text{if } X \text{ is discrete}$$

$$\mathrm{E}[X] = \int_{x\in D_X} x f_X(x)\mathrm{d}x \quad \text{if } X \text{ is continuous.}$$

The expected value (some aliases are *expectation*, *mean of the distribution*, *first moment*) is a linear operator, i.e., $\mathrm{E}[a+bX] = a+b\mathrm{E}[X]$. For instance, the expected value of a fair die is equal to 3.5; the expected value of a Bernoulli variable X with $P\{X = 1\} = \theta$ is equal to $\mathrm{E}[X] = 1 \cdot \theta + 0 \cdot (1 - \theta) = \theta$. Note that $\mathrm{E}[X]$ is *not* an r.v., and it can sometimes be one of the parameters of the distribution.

Another possible parameter of a probability distribution is the *variance*, which is defined as

$$\begin{aligned}\mathrm{V}[X] &= \mathrm{E}[X - \mathrm{E}[X]]^2 = \mathrm{E}[X^2 - 2X\mathrm{E}[X] + \mathrm{E}[X]^2] \\ &= \mathrm{E}[X^2] - 2\mathrm{E}[X]^2 + \mathrm{E}[X]^2 = \mathrm{E}[X^2] - \mathrm{E}[X]^2.\end{aligned}$$

Thus, the variance of X, when X is the outcome of a fair die is equal to $91/6 - 3.5^2 = 2.917$. If X is a Bernoulli variable with $P\{X = 1\} = \theta$, then $\mathrm{E}[X^2] = 1^2 \cdot \theta + 0^2 \cdot (1 - \theta) = \theta$, and therefore $\mathrm{V}[X] = \theta - \theta^2 = \theta \cdot (1 - \theta)$. Note that $\mathrm{V}[X] > 0$ (since $\mathrm{V}[X] = 0$ implies that X is actually a constant). The variance is also known as the *second central moment*. The expected value and variance may be nonfinite. For instance, the Cauchy distribution, whose (standard) density function is $f_X(x) = (1+x^2)^{-1}$, does not admit finite moments. There are other distributions that do not admit finite moments for some values of their parameters. One of them is the *Pareto distribution*, whose density function is

$$f_X(x; x_0, \theta) = \frac{\theta x_0^\theta}{x^{\theta+1}} \qquad x \geq x_0 > 0;\ \theta > 0.$$

For the Pareto distribution $\mathrm{E}[X] = \theta x_0/(1 - \theta)$, which exists when $\theta > 1$, and $\mathrm{E}[X^2] = x_0^2\theta/(2 - \theta)$, which exists when $\theta > 2$. In general, if $\mathrm{E}[X^r]$ is finite, then all the moments of order s with $s < r$ are also finite.

Other useful indicators of an r.v. are the *quantiles*. A quantile of order α, $\alpha \in (0, 1)$ is the *modality* x_α of an r.v. whose cdf satisfies $F_X(x_\alpha) = \alpha$. Note that, if X is continuous, there is a one-to-one relationship between $x_\alpha \in D_X$ and $\alpha \in (0, 1)$ (see uniform distribution). A very special quantile is the *median* of a distribution, defined as the quantile of order 1/2. Therefore, the *median* is the modality $x_{0.5}$ that satisfies $F_X(x_{0.5}) = 1/2$. Note the quantiles are always well defined only if X is continuous.

A.1.2 Examples of Statistical Models

In this section we review a few statistical models that are used in this book. The first two models are for discrete random variables while all the others are for continuous random variables. The description is necessarily concise; for extensive treatment see Johnson and Kotz (1970).

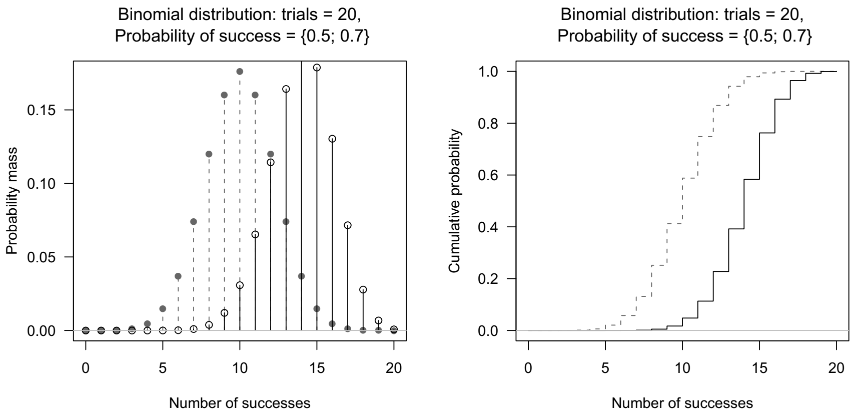

Fig. A.1: The binomial distribution for $n = 20$ and $\theta = 0.5$ (dashed line) or $\theta = 0.7$ (full line)

Binomial Random Variable

In the previous pages we encountered already the Bernoulli r.v.s. A *Binomial* r.v. is the sum of n i.i.d. *Bernoulli* r.v.s. We may indicate a random variable X with Bernoulli distribution using the notation $X \sim \mathrm{Bi}(1, \theta)$. Then, the notation for the Binomial is $X \sim \mathrm{Bi}(n, \theta)$. Its probability and distribution functions are, respectively,

$$p_X(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad F_X(x) = \Pr\{X \le x\} = \sum_{i=0}^{x}\binom{n}{i}\theta^i(1-\theta)^{n-i},$$

and are shown in Fig. A.1. The mean of the binomial distribution is $\mathrm{E}[X] = n\theta$. The variance of the distribution is $\mathrm{V}[X] = n\theta(1-\theta)$ (see next section).

Poisson Random Variable

The *Poisson* r.v. is used in modeling random arrivals. In this case we can see X as the number of arrivals in one unit of time and hence $D_X = \mathbb{N}$.

The probability function is

$$p_X(x) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}, \qquad \lambda \in \mathbb{R}^+,$$

and is identified by the parameter $\lambda > 0$, which is the mean of X. Figure A.2 shows an example with $\lambda = 1$. We denote this model by $X \sim Po(\lambda)$. To see that $\sum_{x\in D_X} p_X(x) = 1$, obtain Taylor's expansion of the function $\exp\{\lambda x\}$ in $x_0 = 0$, and let $x = 1$. The sum of n independent Poisson r.v.s with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$ is still a Poisson r.v. with parameter $\sum_{i=1}^{n} \lambda_i$.

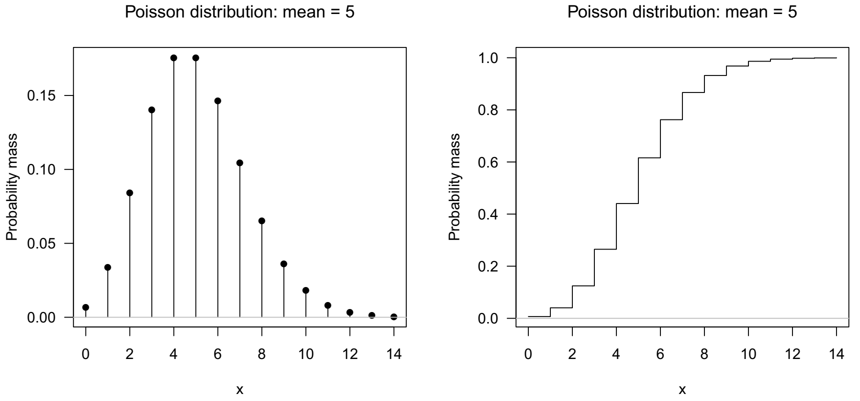

Fig. A.2: The Poisson distribution for $\lambda = 1$

Uniform Random Variable

This variable is defined in the interval $[a, b]$. We write it as $X \sim U[a, b]$. Its density and cumulative distribution functions are, respectively,

$$f_X(x) = \frac{I_{[a,b]}(x)}{b-a}, \qquad F_X(x) = \frac{1}{b-a} \int_{-\infty}^{x} I_{[a,b]}(t)\mathrm{d}t = \frac{x-a}{b-a},$$

where $I_{[a,b]}(\cdot)$ is the indicator function of the interval $[a, b]$. See Fig. A.3. Note that, if we set $a = 0$ and $b = 1$, we obtain $F_X(x) = x$, $x \in [0, 1]$. A typical example is the following: the cdf of a continuous r.v. is uniformly distributed in $[0, 1]$. The proof of this statement is as follows: For $u \in [0, 1]$, we have

$$\begin{aligned}\Pr\{F_X(X) \le u\} &= \Pr\{F_X^{-1}(F_X(X)) \le F_X^{-1}(u)\} = \Pr\{X \le F_X^{-1}(u)\} \\ &= F_X(F_X^{-1}(u)) = u.\end{aligned}$$

This means that, when X is continuous, there is a one-to-one relationship (given by the cdf) between $x \in D_X$ and $u \in [0, 1]$.

Normal (or Gaussian) Random Variable

This variable is defined on the support $D_X = \mathbb{R}$ and its density function is given by

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

The density function is identified by the pair of parameters (μ, σ^2), where $\mu \in \mathbb{R}$ is the mean (or location parameter) and $\sigma^2 > 0$ is the variance (or dispersion param-

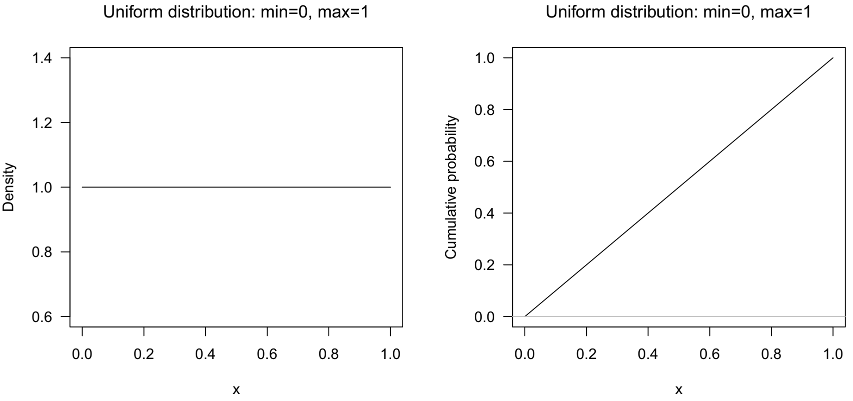

Fig. A.3: The uniform distribution in the interval $[0, 1]$

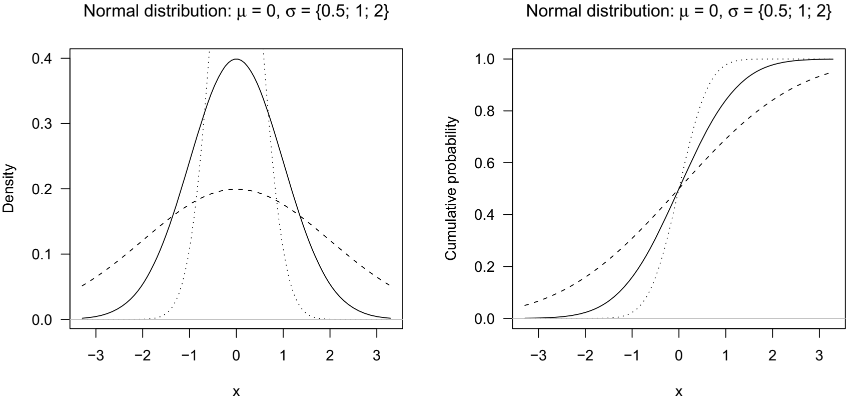

Fig. A.4: The normal distribution for $\sigma^2 = \{0.5; 1; 2\}$ (dotted, full, dashed line, respectively)

eter) of X. The density function is symmetric around μ. Some example of normal densities are given in Figure A.4 for different values of σ^2.

The normal distribution belongs to the location-scale family distributions. This means that, if $Z \sim N(0, 1)$ (read, Z has a standard normal distribution; i.e., with $\mu = 0$ and $\sigma^2 = 1$), and we consider the linear transformation $X = \mu + \sigma Z$, then $X \sim N(\mu, \sigma^2)$ (read, X has a normal distribution with mean μ and variance σ^2). This means that one can obtain the probability of any interval $(-\infty, x]$, $x \in \mathbb{R}$ for any normal distribution (i.e., for any pair of the parameters μ and σ) once the quantiles of the standard normal distribution are known. Indeed

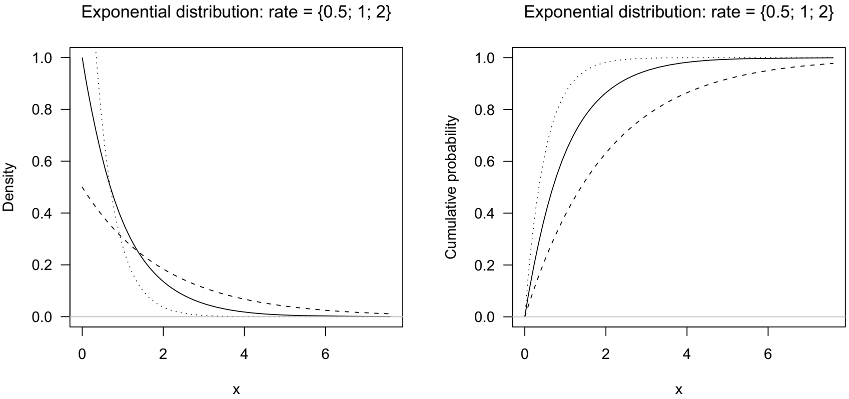

Fig. A.5: The exponential distribution for $\lambda = \{0.5, 1, 2\}$ (dashed, full, dotted line, respectively)

$$
\begin{aligned}
F_X(x) &= \Pr\{X \le x\} = \Pr\left\{\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right\} \\
&= \Pr\left\{Z \le \frac{x-\mu}{\sigma}\right\} = F_Z\left(\frac{x-\mu}{\sigma}\right) \qquad x \in \mathbb{R}.
\end{aligned}
$$

The quantiles of the standard normal distribution are available in any statistical program. The density and cumulative distribution function of the standard normal r.v. at point x are usually denoted by the symbols $\phi(x)$ and $\Phi(x)$.

Exponential Random Variable

This is defined on the support $(0, +\infty)$. We write it as $X \sim \mathrm{Exp}(\lambda)$. The density and distribution functions are:

$$
f_X(x) = \lambda \mathrm{e}^{-\lambda x}, \qquad F_X(x) = 1 - \mathrm{e}^{-\lambda x} \qquad \lambda > 0
$$

and are shown in Fig. A.5. The mean of this distribution is equal to $1/\lambda$; the variance is equal to $1/\lambda^2$. There is a useful reparameterization of this density function which is called *reparameterization with the mean* and can be obtained by letting $\lambda = 1/\theta$; we write this $X \sim \mathrm{Exp}(1/\theta)$. It is easy to prove that the mean and variance of the distribution, according to this reparameterization, are $\mathrm{E}[X] = \theta$ and $\mathrm{V}[X] = \theta^2$. The exponential distribution is used to describe the times at which random arrivals occur. Relevant in this context is the memoryless property, that is, $\Pr\{X > s + x \mid X > s\} = \Pr\{X > x\}$ for all $s, x \ge 0$. Hence, for exponentially distributed arrivals, the probability that we have to wait x seconds for a new arrival, after we had waited s seconds, is not different from the probability that we wait x seconds. The similarity between random arrivals and runtime of stochastic algorithms led to attempts to use this model and its theoretical consequences also in this latter field.

Gamma Random Variable

The exponential distribution is a special case of the *Gamma* distribution. Random variables with this distribution have density function:

$$f_X(x) = \frac{\lambda^\nu x^{\nu-1}\mathrm{e}^{-\lambda x}}{\Gamma(\nu)} \qquad \nu, \lambda > 0, \qquad \Gamma(\nu) = \int_0^{+\infty} \lambda^\nu t^{\nu-1}\mathrm{e}^{-\lambda t}\mathrm{d}t.$$

Here λ is the *scale* parameter and ν the *shape* parameter. We write it as $X \sim \mathrm{Ga}(\nu, \lambda)$. The *Gamma function* $\Gamma(\nu)$ is a standardizing constant, as it satisfies $\lim_{x\to+\infty} F_X(x) = 1$. Moreover, it satisfies the property $\Gamma(\nu+1) = \nu\Gamma(\nu)$, therefore if ν is integer, $\Gamma(\nu) = (\nu-1)!$.

It can be shown that the mean of the Gamma r.v. is ν/λ and the variance is ν/λ^2. Another important property of the Gamma distribution is that the sum of n independent Gamma r.v.s with the same scale parameter λ and shape parameters ν_i, $i = 1, \ldots, n$ is still distributed as a Gamma r.v. with scale parameter λ and shape parameter $\sum_{i=1}^n \nu_i$

Weibull Random Variable

This is defined on $(0, +\infty)$; its density and distribution functions are

$$f_X(x) = \lambda\nu(\lambda x)^{\nu-1}\exp\{-(\lambda x)^\nu\}, \qquad F_X(x) = 1 - \exp\{-(\lambda x)^\nu\}.$$

and are shown in Fig. A.6. We write it as $X \sim \mathrm{We}(\nu, \lambda)$ and it can be shown that $\mathrm{E}[X] = 1/\lambda\Gamma(1 + 1/\nu)$ and $\mathrm{V}[X] = \lambda^{-2}[\Gamma(1 + 2/\nu) - \Gamma(1 + 1/\nu)^2]$.

The exponential distribution can be seen as a special case of the Weibull distribution when $\nu = 1$.

Note that it is always possible to translate the distribution of a r.v. by applying the transformation $Y = X - \mu$. This transformation does not affect the variance and the shape parameters, but affects the support D_Y. By applying this transformation to the examples of Gamma and Weibull r.v.s the support becomes $D_Y = (\mu, +\infty)$.

A.2 Point Estimation

In the previous section, we have seen that the core point of the inferential process is the choice of an adequate statistical model, which is identified by some parameters. In order to model the observed data, the estimation of the parameters of the probability distribution is required. There are several ways to obtain the estimate of the parameters of interest. For instance, moment estimation consists of estimating the parameter(s) of interest with the equivalent sample quantity. Let us assume that the phenomenon under study can be modeled with an r.v. X with unknown param-

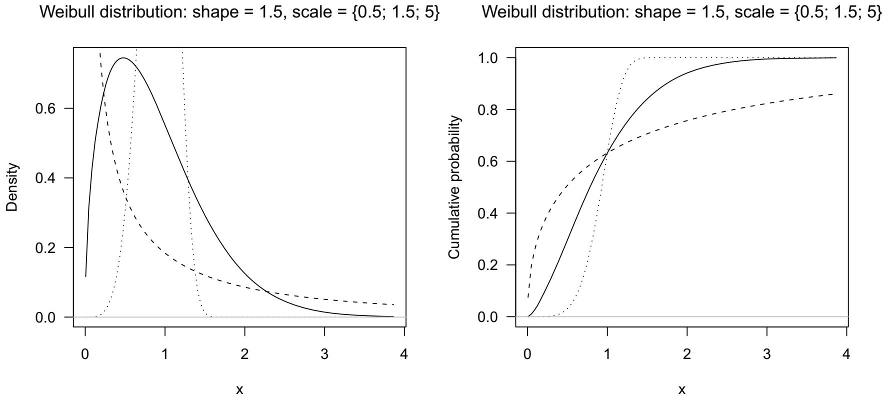

Fig. A.6: The Weibull distribution for $\lambda = 1$ and $\nu = \{0.5; 1.5; 5\}$ (dashed, full, dotted line, respectively)

eters μ and σ^2, where μ and σ^2 are, respectively, the mean and the variance of the population. From the previous sections, we know that the following relationships hold:

$$\mu = \mathrm{E}[X], \qquad \sigma^2 = \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

Now the *sample estimators* of $\mathrm{E}[X]$ and $\mathrm{E}[X^2]$ are, respectively,

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2;$$

therefore the moment estimators of μ and σ^2 are

$$\hat{\mu} = \bar{X} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.$$

Now suppose we extract an i.i.d. sample of size n from a population to investigate the phenomenon, say the vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]$. Then, the *(point) estimates* of μ and σ^2 obtained with the moment estimation are:

$$\hat{\mu} = \frac{1}{n} x_i \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

Note that the estimator is itself an r.v., whereas the estimate is a realization of an r.v., although we will use the same symbols, the context being sufficiently clear.

Other useful sample indicators are the *minimum* and the *maximum*, respectively denoted by the symbols $X_{(1)} = \min_i X_i$ and $X_{(n)} = \max_i X_i$.

There are other ways to obtain a point estimate, to cite but two, the *maximum likelihood* and the *least squares estimation*, which will be discussed later in this appendix.

Distribution of the Most Common Sample Estimators When Observations Are i.i.d.

The estimator of the parameter of interest is an r.v. because its realization depends on the sample given. The estimator of the mean of the distribution is usually the sample mean. This estimator often has the same distribution as the observations in the sample. If the sample is made of independent and identically distributed (i.i.d.) observations, it is easy to obtain the expected value and variance of $\bar{X}$. For instance, if $[X_1, \dots, X_n]$ is a vector of i.i.d. r.v.s from a distribution with mean μ and variance σ^2, then

$$\begin{aligned}
\mathrm{E}[\bar{X}] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu;\\
\mathrm{V}[\bar{X}] &= \mathrm{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathrm{V}[X_i] + 2\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{COV}(X_i, X_j)\right]\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n},
\end{aligned}$$

where $\mathrm{COV}(X_i, X_j)$ is the covariance between X_i and X_j. Note that the last result is due to the assumption of independence among observations, which implies $\mathrm{COV}(X_i, X_j) = 0, i \neq j$.

The above results are valid whenever the sample is made of i.i.d. observations and when the common distribution admits finite expected value and variance. As regards the distribution of $\bar{X}$, it really depends on the distribution of X_i. For instance, if $[X_1, \dots, X_n]$ is a vector of independent r.v.s and $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \dots, n$, then

$$\sum_{i=1}^{n}(a_i + b_i X_i) \sim N\left(\sum_{i=1}^{n}(a_i + b_i\mu_i), \sum_{i=1}^{n} b_i^2\sigma_i^2\right) \qquad a_i, b_i \in I\!R.$$

As a consequence, if we let $\mu_i = \mu$, $\sigma_i^2 = \sigma^2$ (hence the X_i's are i.i.d.), $a_i = 0$, and $b_i = n^{-1}$ for all i, then:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Another common example is given by considering a sample of i.i.d. observations from a Bernoulli distribution; that is, when $X_i = 1$ has a Bernoulli distribution with parameter θ. Recall from Sect. A.1.2 that $S = \sum_{i=1}^{n} X_i \sim \mathrm{Bi}(n, \theta)$; since $\bar{X} = S/n$, $\bar{X}$ has the same probability function of S, but a different domain. In particular $D_{\bar{X}} = \{0, 1/n, 2/n, \ldots, 1\}$ and $\Pr\{\bar{X} = s/n\} = \Pr\{S = s\}$. This means that, in this case, it is possible to obtain exact inference on the parameter θ (see Sect. A.3).

Given that $\mathrm{E}[X_i] = \theta$ and $\mathrm{V}[X_i] = \theta(1-\theta)$, we have that, from the general results on $\bar{X}$ when observations are i.i.d.,

$$\mathrm{E}[\bar{X}] = \theta \qquad \text{and} \qquad \mathrm{V}[\bar{X}] = \frac{\theta(1-\theta)}{n}.$$

The distribution of $X_{(n)}$, the *maximum* of n i.i.d. random variables distributed as $F_X(x)$, is obtained by realizing that the event $\{X_{(n)} \leq x\}$, $x \in D_X$ implies the event $\cap_{i=1}^{n}\{X_i \leq x\}$, and by the definition of independent r.v.s. Thus

$$F_{X_{(n)}}(x) = \Pr\{X_{(n)} \leq x\} = \prod_{i=1}^{n} \Pr\{X_i \leq x\} = F_X(x)^n.$$

The distribution of $X_{(1)}$, the *minimum* of n i.i.d. random variables distributed as $F_X(x)$, is obtained by realizing that the event $\{X_{(1)} > x\}$ implies the event $\cap_{i=1}^{n}\{X_i > x\}$, and by the definition of independent r.v.s. Thus

$$F_{X_{(1)}}(x) = 1 - \Pr\{X_{(1)} > x\} = 1 - \prod_{i=1}^{n} \Pr\{X_i > x\} = 1 - [1 - F_X(x)]^n.$$

The density functions of $X_{(1)}$ and $X_{(n)}$ are, respectively,

$$f_{X_{(1)}}(x) = n f_X(x)[1 - F_X(x)]^{(n-1)} \qquad \text{and} \qquad f_{X_{(n)}}(x) = n f_X(x) F_X(x)^{(n-1)}.$$

Their distributions can be written explicitly only in some cases, for instance, when $X \sim U[a,b]$ or when $X \sim \mathrm{Exp}(\lambda)$.

Properties of "Good" Estimators

Since an estimator is an r.v., it is possible to obtain the expected value and variance. These quantities are very useful when comparing different estimators. A first requirement of an estimator is that, on average, it yields the true value of the parameter of interest. This property is called *unbiasedness*. Formally, if $\hat{\theta}$ is an estimator of the parameter θ, the requirement can be written as

$$\mathrm{E}[\hat{\theta}] = \theta \qquad\qquad \forall\, \theta \in \Theta.$$

This computation can usually be done because of the assumption that the sample observations are i.i.d. from a specified statistical model. For instance, let $X_1, \ldots, X_n$ be a random sample from a normal distribution with parameters μ and σ^2. Then, because the r.v.s are identically distributed, we have $\mathrm{E}[X_i] = \mu$ and $\mathrm{V}[X_i] = \sigma^2$ for all i. For instance, the expected value of the moment estimator of μ is

$$\mathrm{E}[\hat{\mu}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu,$$

and therefore $\hat{\mu}$ is an unbiased estimator of μ. This is not true for the moment estimator of the variance, since

$$\begin{aligned}\mathrm{E}[\hat{\sigma}^2] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i^2] - \mathrm{E}[\bar{X}^2] \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mu^2 + \sigma^2) - \left(\mu^2 + \frac{\sigma^2}{n}\right) = \sigma^2\left(\frac{n-1}{n}\right).\end{aligned}$$

In the last equation, we have used the relationships

$$\begin{aligned}\mathrm{V}[X_i] = \mathrm{E}[X_i^2] - \mathrm{E}[X_i]^2 \quad &\Rightarrow \quad \mathrm{E}[X_i^2] = \mathrm{V}[X_i] + \mathrm{E}[X_i]^2 = \sigma^2 + \mu^2; \\ \mathrm{V}[\bar{X}] = \mathrm{E}[\bar{X}^2] - \mathrm{E}[\bar{X}]^2 \quad &\Rightarrow \quad \mathrm{E}[\bar{X}^2] = \mathrm{V}[\bar{X}] + \mathrm{E}[\bar{X}]^2 = \frac{\sigma^2}{n} + \mu^2.\end{aligned}$$

The expectation of $\hat{\sigma}^2$ tells us that, on average, the estimator $\hat{\sigma}^2$ underestimates the true value of the parameter σ^2, and therefore $\hat{\sigma}^2$ is said to be *biased*. However, note that

$$\lim_{n \to +\infty} \mathrm{E}[\hat{\sigma}^2] = \sigma^2,$$

and $\hat{\sigma}^2$ is then asymptotically unbiased. Usually, statisticians prefer to consider the following unbiased estimator of σ^2:

$$s^2 = \left(\frac{n}{n-1}\right)\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}[X_i - \bar{X}]^2.$$

Another important property of an estimator is *consistency*. Formally, the (weak) consistency of an estimator requires that the estimator converges in probability to the true value of the parameter of interest. That is, given an estimator $\hat{\theta}_n$ (that depends on the sample size n),

$$\lim_{n \to +\infty} \Pr\left\{ |\hat{\theta}_n - \theta| > \epsilon \right\} = 0 \qquad \forall\, \epsilon > 0.$$

Roughly speaking, if the amount of information increases, then we expect the estimator to be distributed around the true value of the parameter, and the accuracy should increase with n (i.e., the variance of the distribution of $\hat{\theta}$ should decrease with n). Two sufficient conditions to ensure that an estimator is (weakly) consistent are

$$\mathrm{E}[\hat{\theta}] = \theta \qquad \text{and} \qquad \lim_{n \to +\infty} \mathrm{V}[\hat{\theta}] = 0.$$

For instance, the estimator of the mean of an i.i.d. sample $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$ is weakly consistent since $\mathrm{E}[\hat{\mu}] = \mu$, and $\mathrm{V}[\hat{\mu}] = \sigma^2/n$.

Central Limit Theorem

A sequence $Z_1, \ldots, Z_n$ of r.v.s with distribution functions $F_{Z_1}(t), \ldots, F_{Z_n}(t)$ converges in distribution to an r.v. Y with distribution function $F_Y(y)$, written $Z_n \xrightarrow{d} Y$, if

$$\lim_{n \to +\infty} F_{Z_n}(t) = F_Y(t) \qquad\qquad t \in I\!R.$$

We can then state the central limit theorem as follows.

Let $[X_1, X_2, \ldots, X_n]$ be n i.i.d. r.v.s from a common distribution $F_X(x)$, with finite first and second moment, and let $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then, as n increases, we have

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mathrm{E}[X_i])}{\sqrt{\mathrm{V}(X_i)}} \quad \xrightarrow{d} \quad N(0,1).$$

This theorem is important because it allows us to obtain the distribution of some test statistics depending on the mean of n i.i.d. r.v.s.

A.3 Hypothesis Testing

Let us go back to the explanatory example of Sect. A.1. Suppose that we are interested in evaluating whether the probability of an algorithm to find a solution within a certain time limit t_0 is more than or equal to 90%. To do that, suppose that we run the algorithm $1,000$ times, and find out that the runtime is less than t_0 874 times. A point estimation of the parameter θ gives $\hat{\theta} = 0.874$. This value is less than 0.9, but is it sufficiently far from 0.9 to be sure enough that our conjecture cannot be realistic? What would change if we had run the algorithm $n = 100$ times and found that the runtime is less than t_0 87 times? Would our conclusion be the same?

Null and Alternative Hypotheses

The question "is the probability of the algorithm to have a runtime less than t_0 more than or equal to 90%?" is called the *null hypothesis*. There is an *alternative* hypothesis which can be true, i.e., that the probability of the algorithm having a runtime less than t_0 is less than 90%.

Since our decision must be made on the available information (that of the sample), it is impossible to answer the question with no margin of error. The theory of hypothesis testing was born in order to answer these questions, bounding and quantifying the probability of incorrectly rejecting the null hypothesis.

In the previous paragraphs we saw that the estimator of a parameter is an r.v. and that its distribution is described by a probability law that depends on some parameters of the population. How would this probability distribution look if the null hypothesis were true? Is the probability of observing the estimate of θ given by the sample "too small" to trust that the null hypothesis is true? Or in our previous example, is $\hat{\theta}$ "too far" from 0.9 to decide that the null hypothesis should be rejected?

The null and alternative hypotheses can be formulated as follows:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases},$$

where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$. The subset Θ_0 specifies the values of the parameter which are in agreement with the null hypothesis. In our previous example $\Theta_0 = [0.9, 1]$, and $\Theta_1 = [0, 0.9)$. Note that in the null hypothesis there is always a well-specified value of the parameter θ (e.g., the value 0.9 belongs to Θ_0; this will be clearer in what follows).

Statistical Test and Acceptance/Rejection Region in the Sample Space

A *statistical test* is a partition of the *sample space* $\mathcal{X}$, where the sample space is the set of all possible values of the random vector of sample data $\mathbf{X}$. In other words, there are some points of the sample space $\mathbf{x} \in \mathcal{X}_0 \subset \mathcal{X}$ which are in agreement with the null hypothesis, and others $\mathbf{x} \in \mathcal{X}_1 \subset \mathcal{X}$ which are too unlikely to assume that the null hypothesis holds. Thus, the observed data may lead to the *rejection* of the null hypothesis or not. The information of the data is summarized by the *test statistic* $T = T(\mathbf{X})$, which is a function of the random vector of data $\mathbf{X}$ whose probability distribution is known if the null hypothesis is true. Given that the distribution of the test statistic is known under the null hypothesis, we may take a decision based on $T(\mathbf{x})$, the value of the test statistic computed with the vector of observed data $\mathbf{x}$.

The domain of $T(\mathbf{X})$ can be partitioned into an *acceptance region* $A(\mathbf{X})$ and a *rejection region* $R(\mathbf{X})$ of the null hypothesis.

Once the sample data $\mathbf{x}$ have been observed, we may conclude that

$$\begin{cases} \text{we cannot reject } H_0 & \text{if } T(\mathbf{x}) \in A(\mathbf{X}) \\ \text{we reject } H_0 & \text{if } T(\mathbf{x}) \in R(\mathbf{X}). \end{cases}$$

Type I and Type II Errors

The acceptance or rejection of H_0 is therefore induced by the sample data $\mathbf{x}$ through the test statistic T. Hence, there are two kinds of errors that may arise, which are known as type I (α) and type II (β) errors:

$$\alpha = \Pr_{\theta \in \Theta_0} \{T(X) \in R(\mathbf{X})\}$$

$$\beta = \Pr_{\theta \in \Theta_1} \{T(X) \in A(\mathbf{X})\}$$

That is, α is the probability of incorrectly rejecting H_0 when H_0 is true; β is the probability of not rejecting H_0 when the alternative hypothesis H_1 is true. A theoretical "perfect" test should satisfy $\alpha = \beta = 0$. In practice, the variability of the sample data vector $\mathbf{X}$ cannot ensure this ideal condition. Operatively, it is impossible to control both kinds of error, because the true value of θ is unknown (especially under the alternative hypothesis).

The role of inferential statistics is thus to try to control at least one of these errors. The error that we can control is α, as there is always a well-specified value of θ in Θ_0. Indeed, the well-specified value of θ in Θ_0 typically maximizes the probability of a type I error. Thus, the rejection region $R(\mathbf{X})$ is determined for an a priori chosen α that satisfies the condition

$$\alpha = \sup_{\theta \in \Theta_0} \Pr_{\theta}\{T(X) \in R(\mathbf{X})\}, \tag{A.1}$$

which is also known as the *significance level* of the test.[3] We will better explain this last sentence by referring to the introductory example of this section.

In Sect. A.2, we have seen that $n\hat{\theta} = \sum_{i=1}^{n} X_i \sim \mathrm{Bi}(n, \theta)$ when the X_i's are i.i.d. r.v.s with a Bernoulli distribution and $\Pr\{X_i = 1\} = \theta$. Now recall the null hypothesis $H_0 : \theta \geq 0.9$ of the example. This means that, if H_0 is true, θ belongs to the (closed) interval $[0.9, 1]$. There are infinitely many points in this interval, each specifying a probability distribution of the r.v. $n\hat{\theta}$ under H_0. Thus, in order to compute the probability of making a type I error, we have to specify the rejection region of the test $R(\mathbf{X})$ first. This region will be made of the points $\mathbf{x} \in \mathcal{X}$ for which $\hat{\theta}$ is "too small" with respect to the border value $\theta = 0.9$ of Θ_0; that is, $R(\mathbf{X})$ will be an interval that satisfies

$$R(\mathbf{X}) = \{\mathbf{x} \in \mathcal{X} : n\hat{\theta} \leq c(\alpha)\},$$

for a constant $c(\alpha)$ to be specified, known as the *critical value* of the test. Now let $T(X) = n\hat{\theta}$ be the test statistic, whose distribution is $\mathrm{Bi}(n, \theta)$, $\theta \in \Theta_0$, when the null hypothesis is true. For any value of $\theta \in \Theta_0$ we can obtain a constant $c(\alpha)$ satisfying the above condition on $R(\mathbf{X})$. Given that the rejection region has the form $[0, c(\alpha)]$, and given that if $\theta_1 \leq \theta_2$ are two points of Θ_0

$$\Pr_{\theta_1}\{n\hat{\theta} \leq c\} \geq \Pr_{\theta_2}\{n\hat{\theta} \leq c\},$$

the value of $\theta \in \Theta_0$ that maximizes the type I error is the boundary point $\theta = 0.9$; that is

$$\alpha = \sup_{\theta \in \Theta_0} \Pr_{\theta}\{T(X) \in R(\mathbf{X})\} = \Pr_{\theta}\{T(X) \in R(\mathbf{X})\}_{|\theta=0.9}.$$

The above probability is known as the significance level of the test, and the rejection region $R(\mathbf{X})$ will then be specified by choosing a desired α-level $\in (0, 1)$ and by focusing attention on the case when $n\hat{\theta} \sim \mathrm{Bi}(n, 0.9)$ and $n = 1{,}000$. In other words, we will base the inference on θ on the Binomial model with parameters $n = 1{,}000$

---

[3] Note that we have used the same symbol α to specify both the type I error and the significance level of the test. This is because, when the null hypothesis is of the kind $H_0 : \theta = \theta_0$ (i.e., the null parameter space consists in only one point), the two definitions coincide.

and $\theta = 0.9$, because the other elements of Θ_0 would lead to a smaller type I error. The probabilistic model maximizing the type I error is known as the *null distribution* of the test statistic.

The significance level α indicates how much we are willing to risk (in terms of probability) an incorrect rejection of H_0 when the latter is true. Some typical choices of α are 1%, 5%, and 10%, although the significance level of the test is a subjective choice (and the p-value approach described in next paragraph will reduce the role of α). To fix the ideas, suppose that we choose $\alpha = 5\%$: The discrete nature of the binomial distribution does not allows us to find a quantile which satisfies $\Pr_{\theta=0.9}\{n\hat{\theta} < c(\alpha)\} = \alpha$ exactly. We can thus choose the quantile of the null distribution whose cdf is closest to the chosen α-level of the test (or not bigger than α). By looking at the quantiles of the binomial distribution with parameters $n = 1,000$ and $\theta = 0.9$, we find that

$$\Pr_{\theta=0.9}\{1000 \cdot \hat{\theta} \leq 883\} = 0.0433, \qquad \text{and} \qquad \Pr_{\theta=0.9}\{1000 \cdot \hat{\theta} \leq 884\} = 0.0534.$$

Thus, we can set the critical value equal to $c(\alpha') = 883$, and perform the test with a significance level which is actually equal to $\alpha' = 4.33\%$. Therefore, the rejection and acceptance regions of the test will be

$$R(\mathbf{X}) = \{\mathbf{x} \in \mathcal{X} : 1000 \cdot \hat{\theta} \leq 883\} \qquad \text{and} \qquad A(\mathbf{X}) = \{\mathbf{x} \in \mathcal{X} : 1000 \cdot \hat{\theta} \geq 884\}.$$

In our example, $1,000 \cdot \hat{\theta} = 874$, which belongs to $R(\mathbf{X})$; then, we will reject the null hypothesis at a significance level $\alpha' = 4.33\%$.[4]

Note that the null sample space $\mathcal{X}_0$ is identified by the points of $A(\mathbf{X})$ and vice versa; that is:

$$\mathbf{x} \in \mathcal{X}_0 \qquad \Longleftrightarrow \qquad T(\mathbf{x}) \in A(\mathbf{X}).$$

There is another way to solve the testing problem above. Given that in our example n is very large, we could have applied the central limit theorem in order to specify the null distribution of a different test statistic $T_n(\mathbf{X})$. If X_i $i = 1, \ldots, n$ are i.i.d. dichotomous variables with $\theta = \Pr\{X_i = 1\}$, then we known that $\mathrm{E}[X_i] = \theta$ and $\mathrm{V}[X_i] = \theta(1-\theta)$. Thus, if we let

$$T_n(\mathbf{X}; \theta) = \sqrt{n}\frac{(\hat{\theta} - \theta)}{\sqrt{\theta(1-\theta)}},$$

we know by the central limit theorem that $T_n(\mathbf{X}; \theta)$ is approximately distributed as a standard normal r.v., if θ is the true value of the parameter. Now under the null hypothesis, $\theta \geq 0.9$ and note that $\{T_n(\mathbf{X}; \theta) \geq 0\}$ implies $\{\hat{\theta} \geq \theta\}$. This means that, if we set $\theta = 0.9$, positive values of $T_n(\mathbf{X}; \theta)$ are in accordance with the null hypothesis, whereas negative values of $T_n(\mathbf{X};\theta)$ are in disagreement with the null hypothesis. Hence, the rejection region $R(\mathbf{X})$ will be of the form $(-\infty, c(\alpha)]$, with $c(\alpha)$ being the critical value of the test. It is easy to see that, given two points of Θ_0, say $\theta_1 \leq \theta_2$, $\Pr_{\theta_1}\{T_n(\mathbf{X};\theta_1) \leq c(\alpha)\} \geq \Pr_{\theta_2}\{T_n(\mathbf{X};\theta_1) \leq c(\alpha)\}$, again the boundary point $\theta = 0.9$ maximizes the type I error for any given $c(\alpha)$. Therefore we have that

---

[4] It is worth noting that, if the observed value of the test statistic belongs to $A(\mathbf{X})$, this does not mean that we have a further knowledge about the true value of the parameter θ in the population. We can only conclude that the observed data do not disagree "enough" with H_0 to reject it at the specified significance level.

$$T_{1000}(\mathbf{x};\theta = 0.9) = \sqrt{1000}\frac{.874 - 0.9}{\sqrt{0.9 \cdot 0.1}} = -2.740641$$

is now the observed value of the test statistic, and that this value can be compared with the quantiles of the limiting distribution of $T_n(\mathbf{X})$ (that is, the standard normal) in order to determine whether $T(\mathbf{x};\theta = 0.9)$ falls into the rejection region of the null hypothesis or not. Let $\alpha = 5\%$; the quantile $z_{0.05}$ of the standard normal distribution is equal to -1.6448, therefore

$$\begin{aligned} R(\mathbf{X}) &= \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x};\theta = 0.9) \leq -1.6448\}, \\ A(\mathbf{X}) &= \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x};\theta = 0.9) > -1.6448\}. \end{aligned}$$

Since the observed value of the test statistic falls into the rejection region, we will reject the null hypothesis at a 5% significance level.

The two testing approaches introduced in this paragraph lead to the same conclusion. It is worth noting that they are indeed slightly different: in the first approach we have compared the observed value of the test statistics $n\hat{\theta}$ with the quantiles of its theoretical distribution evaluated the boundary point $\theta = 0.9$, whereas in the second approach we have compared the value $T_n(\mathbf{x})$ with its asymptotic distribution, which is a standard normal as the sample size n tends to infinity. The first test is said to be an *exact* test, whereas the second is said to be *asymptotic*. The difference between the two approaches tends to vanish as n tends to infinity, but may not be negligible for small n, as we will show in the next paragraph.

As a final remark, the conclusions of hypothesis testing are conditioned by the amount of information available. To see this, suppose $n = 100$ and $\hat{\theta} = 0.87$; by applying both the exact and asymptotic tests one would not reject the null hypothesis $H_0 : \theta \geq 0.9$ at a significance level of $\alpha = 5\%$. This happens because, when the available amount of information increases, a "good" test should better discriminate between the null and alternative hypotheses. This property is known as the *consistency* of a test, and it may not hold for some tests.

The p-Value Approach

The acceptance–rejection method of testing hypotheses that we have just introduced is unable to capture all the latent information in the data. For instance, if we had repeated the experiment and found that in 870 cases the runtime did not exceed t_0, we would have rejected the null hypothesis $H_0 : \theta \geq 0.9$ as well. However, clearly an estimate of θ equal to 0.870 is slightly smaller than the previous one, which was

$\hat{\theta} = 0.874$. It would be better to have an idea of how far the observed data are from the boundary point of the null parameter space Θ_0, for instance. This can be done by computing the *observed significance level* (or p-value), which is defined as *the minimum significance level for which the null hypothesis would be rejected.* Formally, the p-value is defined as

$$p\text{-value} = \min_{\alpha} \sup_{\theta \in \Theta_0} \Pr_{\theta}\{\mathbf{X} \in R(\mathbf{X}; \alpha)\},$$

where the emphasis is on the fact that the rejection region is specified by α. Going back to our previous example (the exact testing approach), we have that

$$\Pr_{\theta=.9}\{T(\mathbf{X}) \leq 874\} = 0.0045,$$

so, with the observed data, we should have chosen a significance level $\alpha = 0.45\%$ in order to reject the null hypothesis.

The p-value is much more informative about the rejection of the null hypothesis than the acceptance–rejection approach because, ceteris paribus, we could have chosen a significance level about ten times smaller than 4.33% and rejected the null hypothesis as well.

According to the asymptotic approach, the p-value is equal to $\Phi(-2.740641) = 0.0031$, so now the difference between the exact and asymptotic approaches becomes more evident. In both cases there is strong evidence against the null hypothesis. Of course, for fixed α, there is the equivalence

$$p\text{-value} < \alpha \qquad \Longleftrightarrow \qquad T(\mathbf{X}) \in R(\mathbf{X}; \alpha).$$

The example above is a test with one-sided alternative (i.e., when H_1 is of the form $\theta < \theta_0$ or $\theta > \theta_0$). There are also tests with two-sided alternatives, when $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$. In this kind of testing problem, the null hypothesis should be rejected whenever $|\hat{\theta} - \theta_0|$ is "too big." In this case, the acceptance and rejection regions are of the form

$$\begin{aligned} A(\mathbf{X}) &= \{\mathbf{x} \in \mathcal{X} : c_1(\alpha) \leq T(\mathbf{X}) \leq c_2(\alpha)\}, \\ R(\mathbf{X}) &= \{\mathbf{x} \in \mathcal{X} : T(\mathbf{X}) < c_1(\alpha) \ \cup \ T(\mathbf{X}) > c_2(\alpha)\}; \end{aligned}$$

with α being the significance level of the test. Usually, $c_1(\alpha)$ and $c_2(\alpha)$ are determined in order to be $\alpha/2 = \Pr_{\theta_0}\{T(\mathbf{X}) < c_1(\alpha)\} = \Pr_{\theta_0}\{T(\mathbf{X}) > c_2(\alpha)\}$.

The p-value computation, in this case, is

$$p\text{-value} = 2 \min \left\{ \Pr_{\theta_0}\{T(\mathbf{X}) \leq T(\mathbf{x})\}, \Pr_{\theta_0}\{T(\mathbf{X}) \geq T(\mathbf{x}))\} \right\}.$$

For instance, if we apply an exact test assessing $H_0 : \theta = 0.9$, then $\Pr_{\theta_0}\{T(\mathbf{X}) \leq 874) = 0.0045$, $\Pr_{\theta_0}\{T(\mathbf{X}) \geq 874) = 0.9966$, therefore the p-value is equal to $2 \times 0.0045 = 0.009$.

Brief Introduction to Permutation Tests

In the previous paragraphs, we have introduced two examples of *parametric* statistical tests. The parametric approach requires that the statistical model generating the observed data be known; moreover, the null distribution of the test statistic is often approximated, and there are several examples where the conditions (*regularity conditions*) that ensure the properties of the *likelihood ratio test* are not met. Permutation tests are a further approach that removes some of these issues. It has been becoming more popular in the recent years thanks to the advent of fast computers, although the theory behind it can be traced back to Fisher in the 1930s.

Permutation tests are exact testing procedures that do not require assumptions on the probability distribution of data. They are based on the notion of *exchangeability* of data. The r.v.s $X_1, X_2, \ldots, X_n$ are said to be exchangeable if their joint distribution is equal to the joint distribution of a permutation of $X_1, X_2, \ldots, X_n$. For instance, n i.i.d. r.v.s are always exchangeable, because their joint distribution can be written as the product of their densities/probability functions (because of the commutative property of the product operator). Let $\mathbf{X} = [X_1, X_2, \ldots, X_n]$ be an n-dimensional vector, then $X_1, X_2, \ldots, X_n$ are exchangeable if

$$\Pr\{\mathbf{X}\} = \Pr\{X_1, X_2, \ldots, X_n\} = \Pr\{X_{\pi_1}, X_{\pi_2}, \ldots, X_{\pi_n}\} = \Pr\{\mathbf{X}^*\},$$

where $\pi_1, \pi_2, \ldots, \pi_n$ is a random permutation of the first n integers.

Suppose that we want to compare two algorithms on n instances. At each run we record the dichotomous variable X_i $i = 1, \ldots, n$, where $X_i = 1$ means "algorithm A has smaller runtime than algorithm B," and $X_i = 0$ meaning the opposite. We want to test the null hypothesis that the two algorithms have the same performance against the alternative that algorithm A is better. This can be translated into $H_0 : \theta = 0.5$ versus $H_1 : \theta > 0.5$. The sample space of the experiment is made of 2^n points. Since the test is a partition of the sample space $\mathcal{X}$, one can obtain the exact distribution of the test statistics by computing the value of a test statistic at all points of $\mathcal{X}$. This is equivalent to the exact parametric testing approach to this problem, i.e., if the test statistic is $n\hat{\theta} = \sum_{i=1}^{n} X_i$, the resulting probability distribution is again binomial with parameters n and $\theta = 0.5$.

To see this note that, if $n = 5$ and $\mathbf{x} = [0, 0, 1, 1, 0]$, there are $\binom{5}{2} = 10$ permutations of the vector $\mathbf{x}$ that lead to the same estimate of $\hat{\theta} = 0.2$ over $2^5 = 32$ possible permutations of $\mathbf{x}$. Therefore, the probability of observing $\hat{\theta} = 0.2$ in the sample space is $10/32 = 0.3125$. The binomial model introduced in the previous paragraphs would give us the same result: the probability of the event $X = 2$, when $X \sim \text{Bi}(5, 0.5)$ is

$$\binom{5}{2} 0.5^2 (1 - 0.5)^3 = 10 \cdot 0.03125 = 0.3125.$$

If, in addition, the runtimes of algorithms A and B are also recorded (see Table A.1), then we are dealing with a continuous variable which can be modeled as

	1	2	3	4	5
X_A	0.591	1.587	0.210	0.158	0.797
X_B	0.490	0.315	0.641	1.413	0.401
X	0	0	1	1	0

Table A.1: Runtime results (in seconds) of algorithms A (X_A) and B (X_B) in $n = 5$ instances. The event $X = 1$ means "A is faster than B"

$X_{iA} = \mu + \delta + \varepsilon_{iA}$ and $X_{iB} = \mu + \varepsilon_{iB}$. Let us assume that the r.v.s ε_{iA} and ε_{iB} are i.i.d. Under this assumption, the random variable X_{iA} satisfies $X_{iA} \overset{d}{<} X_{iB}$ (i.e., algorithm A is faster than B) only if $\delta < 0$. Note that we have only assumed the r.v.s ε_{iA} and ε_{iB} to be i.i.d. (it would be sufficient that they are exchangeable, not necessarily i.i.d.). The null hypothesis of equal runtime performances can be expressed by $H_0 : \delta = 0$; note that under H_0 we have $X_{iA} \overset{d}{=} X_{iB}$, i.e., the random variables X_{iA} and X_{iB} have identical distribution. This means that, if H_0 is true, the observed data $\mathbf{x} = [x_{1A}, \ldots, x_{nA}, x_{1B}, \ldots, x_{nB}]$ are independent realizations of the same r.v. If this is true, the probability of observing $\mathbf{x}$ is the same as that of observing $\mathbf{x}^*$, where $\mathbf{x}^*$ is a random permutation of $\mathbf{x}$. In other words, the data of algorithm A could have been generated from X_{iB} and vice versa. Thus, in order to perform a permutation test to assess $H_0 : \delta = 0$ against $H_1 : \delta < 0$ we consider all $\binom{2n}{n}$ possible permutations of $\mathbf{x}$, choose a suitable test statistics (for instance, the difference of the means $T^* = T(\mathbf{x}^*) = \bar{x}^*_A - \bar{x}^*_B$, and obtain its null distribution by computing the value of T^* for any (distinct) random permutation of $\mathbf{x}$. The observed value of the test statistic $T = T(\mathbf{x}) = \bar{x}_A - \bar{x}_B$ (i.e., the value of the test statistic obtained from the observed data) will then be compared with the null (permutation) distribution of T^* in order to compute a p-value. Note that, in this example, small (negative) values of T are significant against the null hypothesis.

Formally, let $T^*_{(1)} \leq T^*_{(2)} \leq \ldots, \leq T^*_{(M)}$ be the values of T^* computed at each point $\mathbf{x}^* : \mathbf{x}^* = \pi(\mathbf{x}), \pi \in \Pi$, where Π is the set of all permutations of the first $2n$ natural integers and N is its cardinality. Let $T^*_{[N\alpha]}$ be the α-quantile of the permutation distribution. Then $H_0 : \delta = 0$ will be rejected in favor of the alternative $H_1 : \delta < 0$ at a significance level α if $T \leq T^*_{[N\alpha]}$. Note that this is equivalent to obtaining a p-value from the null distribution and comparing it with the nominal significance level α. We can define the p-value of this example as

$$p = \hat{F}_{T^*}(T) = \frac{1}{N} \sum_{b=1}^{N} I(T^*_{(b)} \leq T).$$

Thus, the exact p-value in case that the runtime results are recorded as Bernoulli variables (last row of Table A.1), will be equal to: $p = \Pr_{\theta=0.5}\{5 \cdot \hat{\theta} \leq 2\} = 0.5$.

If the observed runtimes are as in the first two rows of Table A.1, the observed value of the test statistic is $T = 0.0166$; Now there are $N' = 10!$ possible distinct permutations of the whole vector of data. If we compute the test statistics for each permutation, we will realize that there are many repetitions. For instance, if we separately permute the elements of $\mathbf{x}_A$ and $\mathbf{x}_B$ in all possible ways, we will obtain the same value of the test statistic $(5!)^2$ times. Thus the number of really informative permutations is in fact $N = \binom{10}{5}$, i.e., all the possible distinct combinations of ten elements in groups of five. The null distribution of T^* can be obtained as follows. Let $\mathbf{x}^*$ be a random permutation of $\mathbf{x}$, consider the first $n_A = 5$ elements of $\mathbf{x}^*$ to be the results of algorithm A, and the last $n_B = 5$ elements of $\mathbf{x}^*$ to be the results of algorithm B, and compute the value of the test statistic $T^* = \bar{x}_A^* - \bar{x}_B^*$, where $\bar{x}_A^* = \sum_{i=1}^{n} x_{iA}^*/n_A$ and $\bar{x}_B^* = \sum_{i=1}^{n} x_{iB}^*/n_B$. Repeat this procedure for all, N informative permutations. From the null distribution we can then obtain the p-value, which is equal to 0.5238. Note that this p-value is not necessarily equal to the previous one (obtained with data coded as 0/1), since the support of the test statistic now is made of $N = 252$ points instead of 32. Nevertheless, the conclusion is the same: there is no evidence in the observed data that the null hypothesis should be rejected. What has changed is the amount of information available (the runtimes).

Permutation tests are *conditional* procedures, where conditioning is with respect to the *permutation sample space* $\mathcal{X}^* = \{\mathbf{x}^* : \mathbf{x}^* = \pi(\mathbf{x}), \pi \in \Pi\}$. That is, the sample space is built on the observed vector of data $\mathbf{x}$ and it is induced by the null hypothesis. The distribution function $\hat{F}_{T^*}(t)$, $t \in \mathbb{R}$, is the exact conditional distribution of T^* on $\mathcal{X}^*$. When n is large, it might be impossible to perform all possible N informative permutations: in such a case $\hat{F}_{T^*}(t)$ can be approximated by considering a large number $B < N$ of random permutations of $\mathbf{x}$.

Note that the test statistic we have applied is *permutationally equivalent* to $T^{*\prime} = \bar{x}_A^*$, in the sense that $T^{*\prime}$ leads to the same inferential conclusions (i.e., to the same p-value). This is because, conditionally on $\mathbf{x}$, the mean of the whole vector of data is a constant and, for any permutation $\pi \in \Pi$, there is the relationship $2n\bar{x} = (n_A\bar{x}_A^* + n_B\bar{x}_B^*)$, so $T(\mathbf{x}^*) = \bar{x}_A^*(1 + n_A/n_B) - n\bar{x}/n_B$. Since n_A, n_B, and $\bar{x}$ are constants, $T^* \overset{\pi}{\sim} T^{*\prime}$, where the symbol $\overset{\pi}{\sim}$ means "*is permutationally equivalent to*". It can also be shown that there is no need to standardize the test statistic as we usually do in the parametric framework.

Finally, given the data collected, the minimum possible significance level is the inverse of the cardinality of informative permutations, i.e., $\min_{\mathbf{x}^* \in \mathcal{X}^*}(p\text{-value}) = 1/N$. In our example, with the Bernoulli variable X, $\min_{\mathbf{x}^* \in \mathcal{X}^*}(p\text{-value}) = 1/32$, and if we consider the runtimes, $\min_{\mathbf{x}^* \in \mathcal{X}^*}(p\text{-value}) = 1/252$.

Finally, the p-value of the parametric two-sample t-test is 0.5197. However, the t-test assumes that data are normally distributed, and this is not the case here since there cannot be negative runtimes.

A.4 Confidence Intervals

The result of point estimation, discussed in Sect. A.2, depends on the observed sample. If we repeat the experiment, the resulting estimate of the parameter will differ, because the data will be different (if X is continuous, the probability of observing the same data set is zero). Therefore, it is better to provide an interval of possible values for the unknown parameter θ, rather than a single value of the estimate. The construction of the confidence intervals is based on the *pivotal quantity*, i.e., a statistic that depends on the observed data and on the unknown parameter, and whose probability distribution does not depend on θ. For instance, recall that, by the central limit theorem,

$$T(\mathbf{X};\theta) = \sqrt{n}\frac{\hat{\theta}-\theta}{\sqrt{\theta(1-\theta)}} \qquad \xrightarrow{d} \qquad N(0,1),$$

where θ is the true value of the parameter as n increases. Then $T(\mathbf{X};\theta)$ is a pivotal quantity since its (asymptotic) distribution is standard normal, not depending on θ. Thus, we can define a *random interval* $C(\mathbf{X}) = [c_1(\alpha), c_2(\alpha)]$ such that

$$\Pr\{T(\mathbf{X};\theta) \in C(\mathbf{X})\} = 1-\alpha,$$

which only depends on α and not on θ. Now, by the applying the inverse function $T^{-1}(\cdot)$ with respect to θ, we may write the probability above as

$$\Pr\{B(\mathbf{X}) \ni \theta\} = 1-\alpha,$$

where $B(\mathbf{X})$ is equal to $T^{-1}(C(\mathbf{X}))$. The probability $1-\alpha$ is called the *confidence level*, and it represents how much we trust that the true value of the parameter is contained in the interval $B(\mathbf{X})$ before the experiment takes place. Note that this is an a priori probability, concerning the random interval $C(\mathbf{X})$. Once the data have been collected we obtain the realization of the r.v. $C(\mathbf{x}) = [c_1(\mathbf{x};\alpha), c_2(\mathbf{x};\alpha)]$. Now $C(\mathbf{x})$ is no longer a random interval, so it does not make sense to write "the probability of θ being included in $B(\mathbf{x})$ is $1-\alpha$." But if we could repeat the same experiment (i.e., sampling data from the same population) a large number of times, then we would find that $(1-\alpha)\%$ of the times the interval $B(\mathbf{x})$ contains the true value of the parameter θ. Thus $B(\mathbf{x})$ will contain or not the true value of θ with probability 1 (and we do not know whether this is happening or not), but we have a *confidence* of $(1-\alpha)\%$ that θ is included in $B(\mathbf{X})$.

From the formulas above we derive that

$$\Pr\left\{c_1(\alpha) \leq \sqrt{n}\frac{\hat{\theta}-\theta}{\sqrt{\theta(1-\theta)}} \leq c_2(\alpha)\right\} = 1-\alpha.$$

This means that, a priori (and if n is large enough for the CLT to take effect)

$$\Pr\left\{\hat{\theta} - c_2(\alpha)\sqrt{\frac{\theta(1-\theta)}{n}} \leq \theta \leq \hat{\theta} - c_1(\alpha)\sqrt{\frac{\theta(1-\theta)}{n}}\right\} = 1 - \alpha.$$

Since $B(\mathbf{x})$ must not depend on the unknown parameter θ, it will be evaluated by plugging in the estimate of θ. Thus

$$B(\mathbf{x}) = \left[\hat{\theta} - c_2(\alpha)\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}, \hat{\theta} - c_1(\alpha)\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right]$$

is an approximate (i.e., asymptotic) confidence interval of level $1-\alpha$. We only need to fix the constants $c_1(\alpha)$ and $c_2(\alpha)$ in order to obtain the confidence interval for θ. In practice, they are determined by letting

$$\Pr\{T(\mathbf{X};\theta) \leq c_1(\alpha)\} = \Pr\{T(\mathbf{X};\theta) \geq c_2(\alpha)\} = \alpha/2.$$

For instance, if we choose a confidence level of 95%, then $c_2(\mathbf{X};\alpha) = -c_1(\mathbf{X};\alpha) = 1.96$.

Going back to our example in the previous section, where we were trying to establish the probability θ for an algorithm to solve an instance within a time limit of t_0, the (approximate) confidence interval for θ with confidence level equal to 95% is given by:

$$\left[0.874 - 1.96\sqrt{\frac{0.874\cdot(1-0.874)}{1000}}, 0.874 + 1.96\sqrt{\frac{0.874\cdot(1-0.874)}{1000}}\right] = [0.853, 0.894]$$

Note that the confidence intervals are always "centered" on the estimate of the parameter of interest.

It can be shown that there is a one-to-one correspondence between (two-sided) statistical tests and confidence intervals.

Therefore, a quick way to test for two-sided alternative hypotheses at a significance level α is choosing the desired confidence level $1-\alpha$, obtaining the corresponding confidence interval, and checking whether the null value of the parameter θ_0 is included in the interval or not. Of course, one can construct "one-sided" intervals, in order to perform a one-sided test.

Returning to our example, we could build a confidence interval on the parameter θ in order to test for $H_0 : \theta \geq 0.9$ in the following form:

$$1 - \alpha = \Pr\{T(\mathbf{X};\theta) \geq c(\alpha)\}.$$

where $c(\alpha)$ should be the α-quantile of a standard normal distribution. Then

$$1-\alpha = \Pr\left\{\sqrt{n}\frac{\hat{\theta}-\theta}{\sqrt{\theta(1-\theta)}} \geq c(\alpha)\right\} = \Pr\left\{\theta \leq \hat{\theta} - c(\alpha)\sqrt{\frac{\theta(1-\theta)}{n}}\right\},$$

If we choose $1-\alpha = 95\%$, then $c(\alpha) = -1.6448$. As before, we plug in the estimate of θ and obtain the confidence interval

$$\text{C.I.} = \left[0, \hat{\theta} - c(\alpha)\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right].$$

An approximate C.I. for θ of this kind with a confidence level of 95% is $[0, 0.8912]$. Since $\theta_0 = 0.9$ does not belong to the interval, the null hypothesis $H_0 : \theta \geq 0.9$ is rejected at a significance level of 5%.

A.5 Regression and Modeling

In statistics, wide use is made of *linear regression* and *model fitting*. The linear (regression) model is the simplest model trying to describe the (linear) linkage between one *response* variable Y and (one or) some *explicative* variables X_j, $j = 1, \ldots, p$, $p \geq 1$. We talk about *model fitting* when we are trying to investigate the probabilistic model beneath the r.v. X generating the data.

Let us consider again the algorithmic example introduced in Sect. A.1: an algorithm solving a set of instances of an optimization problem within a certain time limit. Let us assume now that the time limit is large enough that the algorithm always terminates with a solution found. Linear models can then be used to find a relationship between some parameters of the algorithm and its run time performance, while model fitting can be used to determine that the runtimes are, for example, exponentially distributed.

A.5.1 Linear Regression

The simplest model is the one linking two variables. The goal is to describe how Y varies on average as a function of X. To do that, we collect n independent observations of the bivariate variable $[X_i, Y_i]$, $i = 1, \ldots, n$ and fit a linear model that models the observed data "better." Once the linear model has been fitted, we can use it to determine whether there is a significant correlation between the response and the explicative variable(s), or try to predict the (average) value of $\mathrm{E}[Y]$ corresponding to a given a new observation x_0.

Thus, the model assumes that there is a linear relationship of the kind

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where β_0 and β_1 are, respectively, the *intercept* and the *slope* of the line fitting the response data as a linear function of the explicative data, and ε_i is an r.v. (often known as *experimental error*) describing the variability of Y that does not depend on X. Note here that x_i is assumed to be the realization of the r.v. X_i. The errors are assumed to be an i.i.d. r.v. satisfying:

$$\mathrm{E}[\varepsilon_i] = 0, \qquad \mathrm{V}[\varepsilon_i] = \sigma^2, \qquad i = 1, \ldots, n.$$

These are the (minimal) assumptions on the ε_i's, i.e., we still have not specified the probability distribution of ε_i. An equivalent way to write the model is $\mathrm{E}[Y] = \beta_0 + \beta_1 X$, where the emphasis is on the fact that the fitting line models the expected value of Y as a function of X, rather than Y itself. In order to determine the model we require the estimates of β_0 and β_1. There are several ways to obtain these estimates, and here we will only refer to the *least squares error* estimation. Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction of $\mathrm{E}[Y]$ corresponding to the point x_i, $i = 1, \ldots, n$. The least squares estimates minimize the objective function

$$\mathrm{SSR}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

In other words, $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the squared (Euclidean) distance among the observed and the predicted sets of points, so the fitted line "goes through" the observed data, which are points in $\mathbb{R}^2$. By differentiating $\mathrm{SSR}(\hat{\beta}_0, \hat{\beta}_1)$ with respect to $\hat{\beta}_0$ we obtain

$$\frac{\partial \mathrm{SSR}(\hat{\beta}_0, \hat{\beta}_1)}{\hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{x}$ and $\bar{y}$ are the sample means of, respectively, X and Y. By plugging in the estimate of β_0 and differentiating with respect to $\hat{\beta}_1$ we obtain

$$\frac{\partial \mathrm{SSR}(\hat{\beta}_1)}{\hat{\beta}_1} = -2 \sum_{i=1}^{n} [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})](x_i - \bar{x}) = 0,$$

and therefore

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})/n}{\sum_{i=1}^{n} (x_i - \bar{x})^2/n} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{V}(X)},$$

where $\mathrm{V}(X)$ is the sample variance of X and $\mathrm{Cov}(X, Y)$ is the sample *covariance* between X and Y. The equation $\hat{y}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$ gives the line fitting $E(Y)$ as a function of X.

The estimators of the parameters are unbiased:

$$\mathrm{E}[\hat{\beta}_1] = E\left[\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right] = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \mathrm{E}[Y_i - \bar{Y}]}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1;$$
$$\mathrm{E}[\hat{\beta}_0] = \mathrm{E}[\bar{Y} - \hat{\beta}_1 \bar{x}] = \mathrm{E}[\bar{Y}] - \mathrm{E}[\hat{\beta}_1]\bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0.$$

Note that in the above equations we have stressed that X is not an r.v. (whereas Y is an r.v. since it depends on ε), and applied one property of the expected value for i.i.d. observation: $\mathrm{E}[\bar{Y}] = \mathrm{E}[\bar{Y}_i]/n = \beta_0 + \beta_1 \bar{x}$. It can be proved that the variance of $\hat{\beta}_1$ is

$$\mathrm{V}[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Accuracy of the Regression Model

Once the linear model has been fitted, we may ask something about the goodness of the fit. Clearly, a perfect fit should satisfy the condition $y_i = \hat{y}_i$ for all i; on the other hand, the worst model ever is such that the predicted values do not depend on X (hence there is no relationship between X and Y), e.g., $\hat{y}_i = k$ for all i, with k being a constant. The accuracy of the model is then evaluated by looking at the proportion of the variability of Y that is “explained” by X. The total sample *deviance* (the deviance of an r.v. is its variance multiplied by a constant) of Y can be decomposed as follows:

$$\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i \pm \hat{y}_i - \bar{y})^2$$
$$= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 - 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \bar{y}),$$

where the double product is equal to zero because it is equivalent to $\partial \mathrm{SSR}(\hat{\beta}_1)/\partial \hat{\beta}_1$. The total deviance can thus be written

$$\mathrm{SST} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathrm{SSE} + \mathrm{SSR},$$

where SSE is the *explained* deviance (i.e., the variability of Y due to the model) and SSR is the residual deviance (i.e. the variability of Y not depending on X).

We have $\mathrm{SST} = \mathrm{SSE}$ when the observed points are already on a line (i.e., there is a perfect linear relationship between Y and X), and $\mathrm{SST} = \mathrm{SSR}$ when $\hat{\beta}_1 = 0$, so the fitted model is actually of the kind $\hat{y}_i = \bar{y}$. Thus, an index of the accuracy of the model is given by the ratio

$$R^2 = \rho(X,Y)^2 = \frac{\mathrm{SSE}}{\mathrm{SST}} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\mathrm{COV}(X,Y)^2}{\mathrm{V}[X]\mathrm{V}[Y]}.$$

Clearly, $0 \leq \rho(X,Y)^2 \leq 1$, and high values of $\rho(X,Y)^2$ indicate the presence of a strong *linear* relationship between X and Y. The index $\rho(X,Y)^2$ is the square of the *correlation coefficient*, which is a standardized measure of the covariance between X and Y and satisfies $-1 \leq \rho(X,Y) \leq 1$. The index R^2 is also known as the *coefficient of determination*.

Testing the Slope Coefficient

We have just seen that when $\hat{\beta}_1 = 0$ there is no (observed) linear relationship between Y and X, i.e., the variables are *uncorrelated*, but $\hat{\beta}_1$ is an estimate of the true parameter determining the (true) linear relationship β_1. Therefore it is important to evaluate whether β_1 is significantly different from 0 or not. That is, we want to perform a statistical test assessing the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$.

We require some further assumptions on the error distribution in order to perform a parametric test. Thus, the errors ε_i are assumed to be i.i.d. r.v.s with normal distributions. If $\varepsilon_i \sim N(0,\sigma^2)$, then also Y is normally distributed, specifically $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Being a linear combination of normal r.v.s, we have that $\hat{\beta}_1 \sim N(\beta_1, \sigma^2[\sum_{i=1}^{n}(x_i - \bar{x})^2]^{-1})$. If σ^2 is known, then the r.v.

$$T(\hat{\beta}_1, \beta_1) = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\frac{\hat{\beta}_1 - \beta_1}{\sigma} \quad \sim \quad N(0,1)$$

is a pivotal quantity, and therefore we can specify the acceptance/rejection regions of the test or obtain a two-sided p-value by considering the standard normal distribution as the null distribution of the test statistic. In practice, σ^2 is unknown and needs to be estimated from the data. A natural estimate of σ^2 is given by the (unbiased) variance estimator of the residuals

$$\hat{\sigma}^2 = \mathrm{V}(\hat{\varepsilon}) = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2,$$

which is equal to the residual deviance divided by its degrees of freedom.[5]

Then if we replace σ^2 by its estimate, we have that the test statistic

$$T(\hat{\beta}, \beta_1) = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\frac{\hat{\beta} - \beta_1}{\hat{\sigma}} \quad \sim \quad t_2,$$

that is, $T(\hat{\beta}, \beta_1)$ has a Student t distribution with $n-2$ degrees of freedom (the Student t distribution with k degrees of freedom is defined as the ratio between

---

[5] It can be shown that the expected value of SSR is equal to $\sigma^2(n-2)$. It can also be proved that SSR $\sim \sigma^2\chi^2_{n-2}$, i.e., SSR has a χ^2 distribution with $n-2$ degrees of freedom.

Z and $\sqrt{\chi_k^2/k}$, where $Z \sim N(0,1)$ and χ_k^2 is a chi-square r.v. with k degrees of freedom).

Thus we will reject the null hypothesis on β_1 for large values of $|T(\hat{\beta}, \beta_1)_{|\beta_1=0}|$.

Multiple Regression

The theory of simple linear regression can be easily extended to the more general case where $\mathrm{E}[Y]$ is modeled as a linear function of p explicative variables X_j, $j = 1, \ldots, p$. That is, if we consider the linear model

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i.$$

With multiple linear regression we want to describe the variability of Y as a linear function of p explicative variables that may jointly influence the response. The assumptions on errors are as in the simple linear regression model, and it is possible to perform a statistical test on each slope coefficient β_j, $j = 1, \ldots, p$. What changes here is that the estimates of the parameters are now functions of the *partial correlations* between Y and X_j, i.e., the correlation between Y and X_j computed after removing the correlations between the other explicative variables and Y. The decomposition of the total deviance still holds, but now the degrees of freedom of the residual deviance are $n-p-1$ (in simple linear regression $p = 1$). The adequacy of the model can be evaluated with an index of determination which accounts for the presence of p explicative variables. Indeed, if the number of variables increases the residual deviance decreases, even if none of the explicative variables is correlated with the response. This fact can be explained, for instance, by considering that the estimation of the parameters requires the solution of a system of $p+1$ equations. Therefore if $n = p+1$ there is only one solution that jointly satisfies all $p+1$ equations. Another intuitive example is given by the polynomial regression that has the form

$$Y_i = a + bx_i + cx_i^2 + dx_i^3 + \cdots + \varepsilon_i,$$

where the data (points of $\mathbb{R}^2$) are modeled by a polynomial function of X. It is known that there is only one line passing through two points, only one parabola passing through three points, etc. Thus, we must modify the simple coefficient of determination in order to take into account this geometric property. Define the *adjusted coefficient of determination*

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}.$$

An R^2_{adj} equal to 1 indicates perfect matching between the set of the responses and the set of predicted values (this happens when the nonadjusted R^2 is equal to 1). Note that, in some cases, R^2_{adj} could also be negative (e.g., when $R^2 = 0$): a similar result does not make sense in terms of the proportion of the variability explained by the model, and it must simply be interpreted as an index of a "terrible" model, i.e., a model where there is absolutely no correlation between the response and the explicative variables. The first thing to do when fitting a multiple regression model is to evaluate if at least one of the explicative variables has a significant correlation with the response. If this does not happens, the model is completely useless. We can express this situation with the null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots, \beta_p = 0$ against the alternative $H_1 : \exists \, \beta_i \neq 0$. Note that, if the null hypothesis is true, then we are considering a model with intercept only (such a model always satisfies $\mathrm{SST} = \mathrm{SSR}$ and $R^2 = 0$). We can evaluate if our model is not significantly different from the intercept-only model by looking at their related explicative deviances (or residual deviances). Thus let SSR_0 and SSR_p be the residual deviances of respectively the null model and the model with p explicative variables that we are considering. From what we have said about the relationship between residual deviance and number of explicative variables considered, we can understand that $\mathrm{SSR}_0 \geq \mathrm{SSR}_p$. The difference $\mathrm{SSR}_0 - \mathrm{SSR}_p$ measures the increase of explained deviance that we obtain by adding p explicative variables to the null model. It can be shown that the r.v. $\mathrm{SSR}_0 - \mathrm{SSR}_p$ has a χ^2 distribution with p degrees of freedom. Moreover, it is independent of SSR_p and that the test statistic

$$F = \frac{(\mathrm{SSR}_0 - \mathrm{SSR}_p)/p}{\mathrm{SSR}_p/(n-p-1)} \quad \sim \quad F_{p\,;\,n-p-1},$$

has a Snedecor F distribution with p and $n - p - 1$ degrees of freedom. Small values of the test statistic are in agreement with the null hypothesis. The rejection region of the test has the form $[c(\alpha), +\infty)$, where $c(\alpha)$ is the $(1 - \alpha)$-quantile of the $F_{p\,;\,n-p-1}$ distribution. If the null hypothesis is not rejected, then none of the variables has a linear influence on the response.

The null hypothesis involving all parameters is not rejected, it is possible to test for the significance of each slope parameter β_j by applying the test statistic

$$T(\hat{\beta}_j, \beta_j) = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{V}(\hat{\beta}_j)}} \quad \sim \quad t_{n-p-1},$$

where $\mathrm{V}(\hat{\beta}_j)$ is the variance of the estimator of β_j, which in the general case is not easy to write in a closed form. As before, small values of $|T(\hat{\beta}_j, \beta_j)|_{\beta_j=0}|$ are not significant against the null hypothesis $\beta_j = 0$.

A.5.2 Model Fitting

There are some situations where one wants to know if the probabilistic model that has been assumed to generate the data is the correct one or not. For instance, one of the assumptions in linear regression is the normality of errors. It is usual to check whether the errors can be considered as normally distributed or not because, in the latter case, the inferential results (e.g., tests on coefficients) may not be completely reliable.

One descriptive method to check the normality of error is the *QQ plot*, which is a graphical representation of points whose coordinates are theoretical and empirical (observed) quantiles. By theoretical quantiles we mean the quantiles of the distribution that is assumed to hold for errors. Here the null hypothesis is $H_0 : \varepsilon_i \sim N(0, \sigma^2)$ against the alternative that the errors follow a nonspecified distribution F_ε. Clearly, the evaluation of the null hypothesis will be based on the empirical distribution function of the *residuals* $\hat{\varepsilon}_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$, which are realizations (or can be assumed as estimates) of the errors.

To obtain a QQ plot, order the residuals in increasing order; the jth ordered residual $\hat{\varepsilon}_{(j)}$ has an *empirical distribution function*

$$\hat{F}_n(\hat{\varepsilon}_{(j)}) = \frac{1}{n} \sum_{i=1}^{n} I(\hat{\varepsilon}_i \leq \hat{\varepsilon}_{(j)}),$$

and $\hat{F}_n(\hat{\varepsilon}_{(j)}) = j/n$ if there are no ties in the residuals (we assume that this happens with zero probability). Note that $\hat{F}_n(x)$ is defined for all x in $\mathbb{R}$ and that it is a step function. Then the theoretical quantiles, if errors are assumed to be normally distributed, are

$$z_{(j/n)} = \Phi^{-1}(j/n) \qquad j = 1, \ldots, n.$$

The QQ plot represents the points whose coordinates are $[z_{(j/n)}, \hat{\varepsilon}_{(j)}]$. Since one of the properties of the normal distribution is that a linear combination of a normal r.v. is still normally distributed, if H_0 holds then the plotted points should lie along a line whose coefficients are approximately the coefficient of the linear combination linking the standard normal r.v. Z and the r.v. ε under testing. Thus, if one considers the standardized quantiles instead of the observed quantiles, what changes is just the equation of the theoretical line representing perfect agreement between the observed residuals and their theoretical quantiles.

The QQ plot is easy to interpret, but it lacks objectivity since the decision is made by visual inspection. There is a more scientific approach: the *Kolmogorov–Smirnov test.* hypothesis $H_0 : F_X(x) = F_0(x)$ against a nonspecified alternative $H_1 : F_X(x) \neq F_0(x)$, where $F_0(x)$ is a known distribution. The idea behind the test is that, if X is really distributed as $F_0(x)$, the theoretical and empirical distribution functions should be "close" to each other, and therefore the Kolmogorov–Smirnov test statistic

$$\mathrm{KS} = \max_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$

should be "small." Thus, this test has an acceptance region of the kind $[0, c(\alpha)]$, where $c(\alpha)$ is the $1 - \alpha$ quantile of the distribution of KS.

Recalling our starting example in Sect. A.1, we wish now to give a more precise indication of the runtime that the algorithm needs to solve an instance. To do this, we record the runtime of the algorithm on each specific instance (assuming it always finishes with a solution found). After the experiments we have a sample of run times, $y_1, y_2, \ldots y_n$, one for each of n instances. We saw that the sample mean and the sample variance are some indicators of the distribution of runtimes; nevertheless, if we could find a theoretical distribution that fit the data well, the description would be more complete and, depending on the context, would allow us to exploit properties of the theoretical distribution. Moreover, for some distributions, such as heavy-tailed distributions, not all the moments are finite, implying that the sample mean and sample variance are erratic and not reliable descriptors. Hence, a more complete insight is definitely needed in these cases.

The typical procedure is the following: select a theoretical model, estimate its parameters, and then test the goodness of fit. Two models that we encountered in Section A.1.2, the exponential and the Weibull distribution, are used to describe life data are therefore also appealing to describe runtime distributions of algorithms. In particular, the Weibull distribution exhibits large flexibility due to the presence of three parameters in its model. Parameters in this kind of application are conveniently estimated by the *maximum likelihood method*.

The *likelihood* function $L(\cdot)$ is basically a density/probability function which is considered as a function of the parameter(s) rather than a function of the data. This is why $L(\cdot)$ is not a probability/density function. This choice is due to the fact that usually we try to choose a (parametric) probabilistic model once the data have been observed. Thus it is reasonable to choose the probabilistic model that, a posteriori, maximizes the probability of observing the data.

Suppose that we have the runtime results $y_1, y_2, \ldots y_n$, and that we want to fit the distribution with an exponential model. The density of the exponential r.v. is

$$f_Y(y) = \lambda \exp\{-\lambda y\}, \qquad D_Y = (0, +\infty),\ \lambda > 0.$$

The density of a vector of i.i.d.[6] random variables $\mathbf{Y} = [Y_1, \ldots, Y_n]$ is the product of their densities. Therefore

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i) \qquad \mathbf{y} \in {D_Y}^n, f_{Y_i}(y_i) = f_Y(y_i)\ \forall\ i$$

where ${D_Y}^n$ is the n-dimensional Cartesian product of D_Y. The joint density of the vector in $y_1, y_2, \ldots y_n$ (the observed data), viewed as a function of the parameter λ, is equal to the likelihood function

$$L(\mu, \lambda|\mathbf{y}) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n \exp\left\{-\lambda \sum_{i=1}^{n} y_i\right\}.$$

According to the maximum likelihood method, the estimate of λ is the value that maximizes $L(\lambda)$. It is equivalent (and easier) to maximize the *log-likelihood*, i.e., the logarithmic transformation of $L(\lambda)$ (the logarithm function is monotone increasing)

---

[6] The requirement of identical distribution is not necessary in this definition. We have considered this case since the domain of $\mathbf{Y}$ is easier to describe and because it is part of the example.

$$\ell(\lambda|\mathbf{y}) = \log[L(\lambda|\mathbf{y})] = n\log(\lambda) - \lambda\sum_{i=1}^{n} y_i.$$

By differentiating $\ell(\lambda|\mathbf{y})$ with respect to λ:

$$\frac{\partial\ell(\lambda|\mathbf{y})}{\partial\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i = 0 \qquad \Rightarrow \qquad \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} y_i}.$$

The likelihood method sometimes gives biased estimators (e.g., the estimator of the variance of a normal random variable). Therefore it is advisable to check whether the estimators of the parameters are biased or not (and, if so, modify the estimators ad hoc).

We know that the sum of n i.i.d. Gamma variables with parameters λ and $\nu = 1$ (recall that the exponential distribution is a Gamma with $\nu = 1$) has a Gamma distribution with parameters λ and n . Thus we may write the estimator as $\hat{\lambda} = n/W$, where $W \sim Ga(\lambda, n)$. Therefore

$$\mathrm{E}[\hat{\lambda}] = \mathrm{E}[nW^{-1}] = \frac{n}{\Gamma(n)}\int_{0}^{+\infty} w^{-1}\lambda^{n}w^{n-1}\mathrm{e}^{-\lambda w}\mathrm{d}w$$

$$= \frac{n\lambda}{\Gamma(n)}\int_{0}^{+\infty} \lambda^{n-1}w^{n-2}\mathrm{e}^{-\lambda w}\mathrm{d}w = n\lambda\frac{\Gamma(n-1)}{\Gamma(n)} = \lambda\frac{n}{n-1}.$$

In this case the maximum likelihood estimator of λ is biased, therefore an unbiased estimator of λ is $\hat{\lambda} = (n-1)/\sum_i y_i$. Unfortunately, if one wants to apply the Kolmogorov–Smirnov test, the null distribution $F_0(x)$ must be completely specified. That is, we cannot estimate the parameter(s) of the distribution $F_0(x)$ from data, otherwise the KS test becomes conservative.

The KS test can also be applied in a two-sample problem: let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two vectors of realizations of the r.v.s X_1 and X_2, respectively. The KS test can then be applied to assess the null hypothesis $X_1 \overset{d}{=} X_2$ against the alternative $X_1 \overset{d}{\neq} X_2$.

In this case the test statistic is equal to $\mathrm{KS} = \max_{x\in I\!R} |\hat{F}_{n_1}(x) - \hat{F}_{n_2}(x)|$, where $\hat{F}_{n_j}(x)$ is the empirical cdf of the jth sample at point x, $j = 1, 2$. Figure A.7 shows an example of the KS statistic when $\mathbf{x}_1$ are ten realizations from $X_1 \sim \mathrm{Exp}(1)$ and $\mathbf{x_2}$ are 20 realizations from $X_2 \sim \mathrm{Exp}(2)$. In this case the value of the test statistic is $\mathrm{KS} = 0.35$ and the related p-value is equal to 0.3686, so we do not reject the null hypothesis.

References

Davison AC (2008) Statistical Models. Cambridge University Press, New York

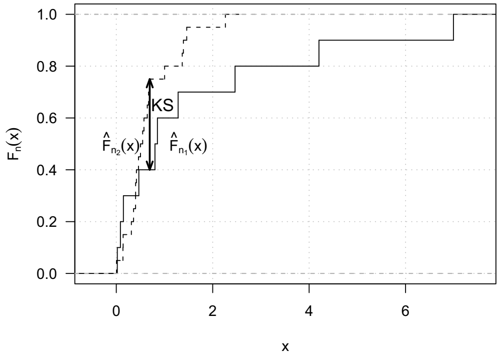

Fig. A.7: The Kolmogorov–Smirnov test statistic for two distributions of data $X_1 \sim \mathrm{Exp}(1)$ and $X_2 \sim \mathrm{Exp}(2)$

Johnson NL, Kotz S (1970) Distributions in statistics. Wiley Series in Probability and Mathematical Statistics, New York, NY

Mood A, Graybill F, Boes D (1974) Introduction to the Theory of Statistics. McGraw-Hill, Auckland

Index

p-value, 22, 431
GAMBLETA, 168
2-edge-connectivity augmentation problem, 236

acceptance/rejection region, 428
accuracy of a linear model, 440
ACO/F-race, 327
ACOTSP, 321
active experimentation, 17
added-variable plots, 348
adjusted coefficient of determination, *see* multiple linear regression
Akaike information criterion, 348, 393
algorithm, 1
  cache-aware, 138
  cache-oblivious, 138
algorithm configuration problem, 308–310
algorithm design, 18, 337
algorithm engineering, 5, 16, 127
  cycle, 127
analysis of variance, 221, 348
  diagnostic plots, 240, 246
  fixed factors, 222, 229
  mixed effects design, 243
  random effects design, 238
  random factors, 223, 225
  residual plots, 246
  stratification factor, 233
  variance components, 226
analytic results, 191
ant colony system, 262
approximation algorithm, 2
assessment
  of an optimizer, 187
asymptotic test, 431
attained set, 106
attainment function, 99, 208
  k-th-order, 108
  empirical, 208
  empirical k-th-order, 116
    visualization, 116
  first-order, 109
attainment surface, 208
  k%-, 208
  median, 208

basic local alignment search tool, 141
Bernoulli distribution, 424
Beta family, 185
Binomial distribution, 417
biogas, 335
black-box optimization, 2, 336
blocks, 222
Box–Behnken design, 391
brute-force approach, 312
budget, 18

cache, 138
capacity functional, 107
cdf, *see* cumulative distribution function
ceiling effects, 37
central composite designs, 51
central limit theorem, 427
coding of variables, 344
coefficient of determination, *see* linear regression
computational budget, 309, 313
computer-generated imagery, 139
conditional sampling, 435
confidence intervals, 228, 436
confidence level, 436
configuration process, 321
confounded effects, 38

consistency of a statistical test, 431
consistency of an estimator, 426
contrasts, 353
convergence in distribution, 427
correlation coefficient, *see* linear regression
covariance function, 114
covariance matrix adaptation evolution strategy, 362
covering function, 109
critical value, 429
cumulative distribution function, 415
cumulative incidence function, 167

data
  library, 72
  random generation procedures, *see* data generation schemes
  real world, 71
  synthetic, 72
  synthetic generation procedures, *see* data generation schemes
  validation, 79
data censoring, 161
  type I, 161
  type II, 161
data decomposition, 139
data generation schemes, 71, 74
  applications, 81
    data mining, 87
    generalized assignment, 82
    graphs and networks, 85
    knapsack, 82
    routing, 86
    scheduling, 84
    stochastic programming, 88
    supply chains, 83
density function, 416
depth first search, 130
descriptive statistics, 8
design
  Kriging, 59
design and analysis of computer experiments, 50, 360
design and analysis of simulation experiments, 50
design of experiments (DOE), *see* experimental design
desirability function, 278
deviance of a linear model, 440
differential entropy, 293
distribution
  Beta family, 185
  endpoint, 183
  Fréchet, 183
  generalized Pareto, 185
  Gumbel, 183
  Weibull, 183
distribution function, *see* cumulative distribution function
DOE, *see* design of experiments
dummy regressors, 353

ECJ software library, 299
edge relaxation, 144
effect size, 268
effectivity, 18, 44
efficiency, 18, 44, 267
efficient global optimization, 360
entropy, 284, 290
estimate, 423
estimation, 101, 115
estimator, 423, 425
evaluation step, 312
evolution strategy, 29, 195
evolutionary algorithm, 283
EVT, *see* extreme value theory
exact test, 431, 433
exchangeability, 433
expectation, *see* mean of a distribution
expected improvement
  criterion, 360, 362, 370, 386–388
expected value, *see* mean of a distribution
experimental design, 8, 16, 50, 261
  2^k factorial, 263
  aliasing, 263
  blocking, 222, 315
  central composite (CCD), 266
  circumscribed central composite (CCC), 266
  face-centred composite (FCC), 267
  fractional factorial, 263
  full factorial, 263
  inscribed central composite (ICC), 266
  nested design, 223
  resolution, 264
  sparsity of effects principle, 263
experimental design notation, 225
exponential distribution, 421
external memory model, 137
extreme value theory (EVT), 182

F-Race, 312–315, 360
  applications, 325–327
  full factorial design, 316
  random sampling design, 316
  sampling strategy, 316
factor, 19, 222
  blocking, 222, 315
  design, 271

  fixed, 222
  held-constant, 270
  nuisance, 271
  random, 223
failures in experimentation, 39
fairness in parameter settings, 38
falsification, 23
first moment, *see* mean of a distribution
fixed parameter tractable problem, 130
floor effects, 37
Fréchet distribution, 183
Friedman test, 313
  post hoc test, 315
Friedman's two-way analysis of variance by ranks, 313
function evaluation, 312
fuzzy operator tree, 336
fuzzy operator trees, 336

gamma distribution, 422
Gaussian distribution, *see* normal distribution
Gaussian process model, 361
generalized extreme value distribution (GEV), 183
generic evolutionary algorithm scheme, 287
GEV, *see* generalized extreme value distribution (GEV)
goodness-of-fit test, 188
Graph Minors I–XX, 130
Gumbel distribution, 183

hazard function, 159
  cause-specific hazard, 166
  proportional hazards model, 164
heuristic, 2
hidden Markov model, 334
highway hierarchies, 145
hypothesis test problem, 118, 122
hypothesis testing, 117, 427
  multistage testing, 120
  permutation test, 119

i.i.d., *see* independent and identically distributed random variables
independent and identically distributed random variables, 414
  minimum, 182
inferential statistics, 8
instance class (or population), 224
instances, *see* data
intensification, 371, 379
interactive approach, 390
internal steps, 187
iterated distributed hypercube sampling, 376
iterated F-Race, 317–321

Kolmogorov–Smirnov test, 444
Kriging, 56, 337
  SPOT, 343

landmark concept, 145
least squares error estimation, 439
likelihood ratio test, 235, 242
linear mixed models, 223, 234
linear regression, 438
linearjet propulsion system, 335
local search algorithms, 222

max-stability, 183
maximum, 423, 425
maximum likelihood estimation, 58, 235, 445
mean of a distribution, 416
measures in metaheuristics, 40
median, 417
metaheuristic, 2, 206
metaphor, 284
min-stability, 183
minimization, 182
minimum, 182, 423, 425
mixed effects model, 229
modality, 417
model, 26
  data, 29
  experimental, 29
  fitting, 273, 443
  instantial, 26
  primary, 29
  representational, 26, 27
  selection, 393
  tree-based, 353
mold temperature control, 335
multi-objective evolutionary algorithm, 335
multiobjective optimizer, 101
multiple linear regression, 442
multivariate cumulative distribution function, 115

nested design, 223
new experimentalism, 25
nondominance, 105
nondominated set, 207
nonparametric test, 433
normal distribution, 419
normal quantile plot, 240, 246
null and alternative hypothesis, 427
null distribution of the test statistic, 430

objective function, 181

  examples, 188
  Rosenbrock, 188, 194, 197
  Sinc, 188, 193
obstruction set, 130
operator, 288
optimal computational budget allocation, 338, 380
optimization, 60, 181
optimizer, 181
  stochastic, 182
optimizer outcome distribution, 102
optimizer performance, 102
ordinary least squares, 52

paired comparison plot, 231, 246
parameter
  categorical, 311
  conditional, 311
  continuous, 311
  control, 289
  numeric, 288
  numerical, 311
  ordinal, 311
  pseudo-ordinal, 311
  qualitative, 288
  quantitative, 288
  quasi-continuous, 311
  relevance, 284
  symbolic, 288
  tuning, 41, 289, 360
  types, 311
parametric test, 433
ParamILS, 360
Pareto
  generalized, 185
  generalized, fitting, 197
Pareto distribution, 417
Pareto optimality, 206
Pareto-optimal front, 206
peaks over threshold (POT), 185
performance
  final, 364
permutation test, 433
pitfalls of experimentation, 36
pivotal quantity, 436
planarity testing, 129
point estimation, 422
Poisson distribution, 418
POT for minima, *see* peaks over threshold (POT)
practical differences, 223
pre-experimental planning phase, 391
predictive equation, 343
probability function, 415
probability model, 414
problem design, 18, 337
problem instance generator, 270

QQ plot, 444
quadratic assignment problem
  biobjective, 213
quantile, 417

r.v., *see* random variable
racing, 312–313
racing approach, 312
random nondominated point set (RNP set), 105
random search, 188
random variable, 414, 415
  categorical, 415
  continuous, 415
  discrete, 415
random-effects model, 225
region of interest, 340, 345, 363
regression variable, 343
reparameterization, 421
reporting experiments, 42
residuals, 344, *see* model fitting
resilent algorithm, 142
response, 343
response surface methodology, 60, 266, 346, 360, 362
response variable, 270
restricted maximum likelihood, 235
REVAC, 295
robust optimization, 143
robustness, 19, 44, 142
Rosenbrock function, 188
run-length distributions, 20
runtime distribution, 158
  cumulative distribution function, 159
  probability density function, 159

sample, 413
sample size, 268
satisfiability problems, 171
scalar quality indicators, 207
scale parameter, *see* gamma distribution
scaling and probabilistic smoothing, 363
scatter plot, 346
screening, 19, 262, 284
second central moment, *see* variance
selection, 285
sensitivity analysis, 19, 51, 143
sequence alignment, 141
sequential Kriging optimization, 361
sequential parameter optimization, 17, 32, 361
  interactive, 390

  open issues, 43
sequential parameter optimization toolbox, 17, 333, 361
  automated, 341
  interactive, 341
severity, 32
shannon entropy, 291
shape parameter, *see* gamma distribution
shipbuilding, 335
significance level, 429
similarity, 27
simulation
  away from optimum, 191
  near the optimum, 189
software package
  Xtremes, 187
software testing, 74
SPO, *see* sequential parameter optimization
SPO$^+$, 382
SPOT, *see* sequential parameter optimization toolbox
standard normal distribution, 420
standard template library for extra large data sets, 138
standardized variables, 344
statistical differences, 223
statistical power, 228, 267
statistical test, 428
steepest descent, 61, 350, 394
  adapted, 61
stochastic programming, 143
stopping criterion, 270
succinct data structures, 141
survival analysis, 158
survival function, 159
  joint survival function, 166
  marginal survival function, 166
  subsurvival function, 167

Taguchian robust optimization, 64
task decomposition, 139
test statistics, 428
testing the slope coefficient, 441
time-critical applications, 141
transformation, 378
traveling salesman problem, 321
  biobjective, 216
tree-based regression, 354, 355
  SPOT, 343
Tukey's multiple comparison procedure, 231
tuning, 18, 19, 38, 44, 360, 362
  algorithm, 18
  amount of, 44
  automated, 342
  automated versus interactive, 362
  example, 354
  interactive, 342, 398
  local, 391
  manual, 39, 40
  problem, 18
  sequential parameter optimization, 339
  understanding, 19
two-phase local search, 215
type I and type II errors, 428

unbiasedness of an estimator, 425
uncertainty analysis, 19
uniform distribution, 418
utility, 295

variance, 417
variance function, 112
variation, 286
von Neumann model, 128
Vorob'ev median, 111

waterfall model, 131
Weibull distribution, 183, 422
weighted robust tabu search, 215
Wilcoxon matched-pairs signed-ranks test, 315
Wilkinson-Rogers notation, 343

Xtremes software package, 187