Chapter 3

Elementary statistical concepts

Demetris Koutsoyiannis
Department of Water Resources and Environmental Engineering
Faculty of Civil Engineering, National Technical University of Athens, Greece

Summary

This chapter aims to serve as a reminder and synopsis of elementary statistical concepts, rather than a systematic and complete presentation of the concepts. Statistics is the applied branch of probability theory which deals with real world problems, trying to draw conclusions based on observations. Two major tasks in statistics are estimation and hypothesis testing. Statistical estimation can be distinguished in parameter estimation and prediction and can be performed either on a point basis (resulting in a single value, the expectation), or on an interval basis (resulting in an interval in which the quantity sought lies, associated with a certain probability or confidence). Uses of statistical estimation in engineering applications include the estimation of parameters of probability distributions, for which several methods exist, and the estimation of quantiles of distributions. Statistical hypothesis testing is also an important tool in engineering studies, not only in typical decision making processes, but also in more analytical tasks, such as in detecting relationships among different geophysical, and particularly hydrological, processes. All this concepts are briefly discussed both in a theoretical level, to clarify the concepts and avoid misuses, and a more practical level to demonstrate the application of the concepts.

3.1 Introductory comments

Statistics is the applied branch of probability theory which deals with real world problems, trying to draw conclusions based on observations. The conclusions are only inferences based on induction, not deductive mathematical proofs; however, if the associated probabilities tend to 1, they become almost certainties. The conclusions are attributed to a *population*, while they are drawn based on a *sample*. Although the content of term population is not strictly defined in the statistical literature, loosely speaking we consider that the term describes any collection of objects whose measurable attributes are of interest. It can be an abstraction of a real world population or of the repetition of a real experiment. The population can be finite (e.g. the population of the annual flows of the hydrologic year 1990-91 for all hydrologic basins of earth with size greater than 100 km^2) or infinite and abstractively defined (e.g. the population of all possible annual flows of a hydrologic basin). The term sample describes a collection of observations from the particular population (see definition in section 3.2.1).

An important concept of statistics is the *estimation*. It is distinguished in *parameter estimation* and *prediction*. In order to clarify these concepts, we consider a population that is

represented by a random variable X with distribution function $F_X(x; \theta)$ where θ is a parameter. A parameter estimation problem is a problem in which the parameter is unknown and we seek an estimate of it. A prediction problem is a problem in which the parameter is known and we seek an estimate of the variable X or a function of X. As we will see below, these two problems are dealt with using similar methods of statistics, and thus they are both called estimation problems. The results of the estimation procedures are called *estimates*.

An estimate can be either a *point estimate*, i.e. a numerical value, or an *interval estimate*, i.e. one interval that contains the value sought with a given degree of certainty. Conversely, for a given interval, statistics can calculate the corresponding degree of certainty or, on the contrary, the degree of uncertainty, that the quantity sought lies within the interval.

Another important area of statistics is *hypothesis testing* that constitutes the basis of the decision theory. The process of hypothesis testing requires the formulation of two statements: the basic H_0, that is referred to as the null-hypothesis, and the alternative hypothesis H_1. We start the process of testing by considering that the null hypothesis is true and we use the observations to decide if this hypothesis should be rejected. This is done using of statistical methods. Although the hypothesis testing is based on the same theoretical background as the estimation, the difference lies in the examination of two alternative models, while in the estimation we use only one model.

The background for all these concepts is described in this chapter while in the next chapters several additional numerical examples are given. Of course, statistics include many other areas, such as the Bayesian analysis, but these are not covered in this text.

3.2 Concepts and definitions

3.2.1 Sample

We consider a random variable X with probability density function $f(x)$. The variable is defined based on a sample space Ω and is conceptualized with some population. A sample of X of size (or length) n of is a sequence of n *independent identically distributed* (IID random variables X_1, X_2, …, X_n (each having density $f(x)$) defined on the sample space $\Omega^n = \Omega \times \cdots \times \Omega$ (Papoulis, 1990, p. 238). Each one of the variables X_i corresponds to the possible results of a measurement or an observation of the variable X. After the observations are performed, to each variable there corresponds a numerical value. Consequently, we will have a numerical sequence $x_1, x_2, \ldots, x_n$, called the *observed sample*.

The concept of a sample is, therefore, related to two types sequences: an abstractive sequence of random variables and the corresponding sequence of their numerical values. It is common in engineering application to use the term *sample* indistinguishably for both sequences, omitting the term *observed* from the second sequence. However, the two notions are fundamentally different and we should be attentive to distinguish each time in which of the two cases the term sample refers to.

In statistics it is assumed that the construction of a sample of size n or the *sampling* is done by performing n repetitions of an experiment. The repetitions should be independent to each other and be performed under virtually the same conditions. However, in dealing with natural phenomena and in engineering it is not possible to repeat the same experiment, and thus sampling is a process of multiple measurements of the a natural process at different times. As a consequence, it is not possible to ensure that independence and same conditions will hold. Nonetheless, for certain situations we can assume that the previous conditions are approximately valid (an assumption equivalent to simultaneously assuming independence, stationarity and ergodicity, cf. chapters 2 and 4) and thus we can use classical statistical methods of statistics to analyse them. However, there are cases where these conditions (the independence in particular) are far from holding and the use of classical statistics may become dangerous as the estimations and inferences may be totally wrong (see chapter 4).

3.2.2 Statistic

A *statistic* is defined to be a function of a sample's random variables, i.e. $\Theta = g(X_1, \ldots, X_n)$ (in vector notation, $\Theta = g(\boldsymbol{X})$, where $\boldsymbol{X} := [X_1, \ldots, X_n]^T$ is known as the *sample vector*; note that the superscript T denotes the transpose of a vector or matrix). From the observations we can calculate the numerical value of the statistic, i.e. $\theta = g(x_1, \ldots, x_n)$. Clearly, the statistic Θ is not identical with its numerical value θ. In particular, the statistic, as a function of random variables, is a random variable itself, having a certain distribution function. Whereas the numerical value of the statistic is simply calculated from the mathematical expression $g(x_1, \ldots, x_n)$ using the sample observations, its distribution function is deducted based on theoretical considerations as we will see in later sections. Typical examples of commonly used statistics are given below.

3.2.3 Estimators and estimates

A statistics is used to estimate a population parameter. For any population parameter η, there exists one or more statistic of the form $\Theta = g(X_1, \ldots, X_n)$ suitable for the estimation of this parameter. In this case we say that $\Theta = g(X_1, \ldots, X_n)$ is an estimator of the parameter η and that the numerical value $\theta = g(x_1, \ldots, x_n)$ is an estimate of η.

There is not a unique criterion to decide if a statistic can be used for the estimation of a population parameter. Often the mathematical expression $g(X_1, \ldots, X_n)$ is formulated as if η was a population parameter of a finite sample space identical with the available sample. For example, if we wish to find an estimator of the mean value $\eta \equiv m_X$ of a variable X, based on the sample $(X_1, \ldots, X_n)$ with observations $(x_1, \ldots, x_n)$, we can think of the case where X is a discrete variable taking values $(x_1, \ldots, x_n)$, each with the same probability $P(X = x_i) = 1/n$. In this case, by definition of the mean (eq. (2.21) - (2.23)) we find that $\eta = (x_1 + \cdots + x_n)/n$. If in the latter equation we replace the numerical values with the corresponding variables, we obtain the statistic $\Theta = (X_1 + \cdots + X_n)/n$. As we will see, this is the estimator of the mean

value of any random variable, it is named *sample mean* and it is typically denoted as $\overline{X}$. However, this empirical approach does not give always a good estimator.

Whereas an estimator is not defined by in a strict mathematical procedure in the general case, several estimator categories have rigorous definitions. Thus:

1. A statistic Θ is an *unbiased* estimator of the parameter η if $E[\Theta] = \eta$. Otherwise, it is a biased estimator and the difference $E[\Theta] - \eta$ is called *bias*.
2. An estimator is Θ is a *consistent* estimator of the parameter η if the estimation error $\Theta - \eta$ tends to zero with probability 1 as $n \to \infty$. Otherwise, the estimator is inconsistent.
3. A statistic Θ is *the best* estimator of the parameter η if the mean square error $E[(\Theta - \eta)^2]$ is minimum.
4. A statistic Θ is the *most efficient* estimator of the parameter η if it is unbiased and with minimum variance (where due to unbiasednees the variance equals the estimation error).

It is easy to show that the estimator $\overline{X}$ of the previous example is an unbiased and consistent estimator of the population mean m_X (see section 3.3.1). Moreover, for certain distributions functions, it is best and most efficient.

In practice, efforts are taken to use unbiased and consistent estimators, while the calculation of the best and most effective estimator is more of theoretical interest. For a certain parameter it is possible to find more than one unbiased or consistent estimator. Often, the determination of unbiased estimators is difficult or impossible, and thus we may content with the use of biased estimators.

3.2.4 Interval estimation and confidence intervals

An *interval estimate* of a parameter η is an interval of the form (θ_1, θ_2), where $\theta_1 = g_1(x_1, \ldots, x_n)$ and $\theta_2 = g_2(x_1, \ldots, x_n)$ are functions of the sample observations. The interval (Θ_1, Θ_2) defined by the corresponding statistics $\Theta_1 = g_1(X_1, \ldots, X_n)$ and $\Theta_2 = g_2(X_1, \ldots, X_n)$ is called the interval estimator of the parameter η.

We say that the interval (Θ_1, Θ_2) is a *γ-confidence interval* of the parameter η if

$$P\{\Theta_1 < \eta < \Theta_2\} = \gamma \tag{3.1}$$

where γ is a given constant $(0 < \gamma < 1)$ called the *confidence coefficient*, and the limits Θ_1 and Θ_2 are called *confidence limits*. Usually we choose values of γ near 1 (e.g. 0.9, 0.95, 0.99, so as the inequality in (3.1) to become near certain). In practice the term confidence limits is often (loosely) used to describe the numerical values of the statistics θ_1 and θ_2, whereas the same happens for the term confidence interval.

In order to provide a general manner for the calculation of a confidence interval, we will assume that the statistic $\Theta = g(X_1, \ldots, X_n)$ is an unbiased point estimator of the parameter η and that its distribution function is $F_\Theta(\theta)$. Based on this distribution function it is possible to

calculate two positive numbers ξ_1 and ξ_2, so that the estimation error $\Theta - \eta$ lies in the interval $(-\xi_1, \xi_2)$ with probability γ, i.e.

$$P\{\eta - \xi_1 < \Theta < \eta + \xi_2\} = \gamma \tag{3.2}$$

and at the same time the interval $(-\xi_1, \xi_2)$ to be the as small as possible.* Equation (3.2) can be written as

$$P\{\Theta - \xi_2 < \eta < \Theta + \xi_1\} = \gamma \tag{3.3}$$

Consequently, the confidence limits we are looking for are $\Theta_1 = \Theta - \xi_2$ and $\Theta_2 = \Theta + \xi_1$.

Although equations (3.2) and (3.3) are equivalent, their statistical interpretation is different. The former is a *prediction*, i.e. it gives the confidence interval† of the random variable Θ. The latter is a *parameter estimation*, i.e. it gives the confidence limits of the unknown parameter η, which is not a random variable.

3.3 Typical point estimators

In this section we present the most typical point estimators referring to the population moments of a random variable X irrespectively of its distribution function $F(x)$. Particularly, we give the estimators of the mean, the variance and the third central moment of a variable. We will not extend to higher order moments, firstly because it is difficult to form unbiased estimators and secondly because for typical sample sizes the variance of estimators is very high, thus making the estimates extremely uncertain. This is also the reason why in engineering applications moments higher than third order are not used. Even the estimation of the third moment is inaccurate for a small size sample. However, the third moment is an important characteristic of the variable as it describes the skewness of its distribution. Moreover, hydrological variables are as a rule positively skewed and thus an estimate of the skewness is necessary.

Apart form the aforementioned moment estimators we will present the L-moment estimators as well as the covariance and correlation coefficient estimators of two variables that are useful for the simultaneous statistical analysis of two (or more) variables.

3.3.1 Moment estimators

The estimators of raw moments (moments about the origin) of one or two variables, i.e. the estimators of $m_X^{(r)}$ and $m_{XY}^{(rs)}$ (where r and s are chosen integers), formed according to the empirical method described in section 3.2.3, are given by the following relationships:

---

* If the distribution of Q is symmetric then the interval $(-\xi_1, \xi_2)$ has minimum length for $\xi_1 = \xi_2$. For non-symmetric distributions, it is difficult to calculate the minimum interval, thus we simplify the problem by splitting the (3.2) into the equations $P\{\Theta < \eta - \xi_1\} = P\{\Theta > \eta + \xi_2\} = (1 - \gamma) / 2$.

† The terms confidence limits, confidence interval, confidence coefficient etc. are also used for this prediction form of the equation.

$$\widetilde{M}_X^{(r)} = \frac{\sum_{i=1}^{n} X_i^{\,r}}{n}, \quad \widetilde{M}_{XY}^{(rs)} = \frac{\sum_{i=1}^{n} X_i^{\,r} Y_i^{\,s}}{n} \tag{3.4}$$

It can be proved (Kendall and Stewart, 1968, p. 229) that

$$E\left[\widetilde{M}_X^{(r)}\right] = m_X^{(r)}, \quad E\left[\widetilde{M}_{XY}^{(rs)}\right] = m_{XY}^{(rs)} \tag{3.5}$$

Consequently, the moment estimators are unbiased. The variances of these estimators are

$$\mathrm{Var}\left[\widetilde{M}_X^{(r)}\right] = \frac{1}{n}\left[m_X^{(2r)} - \left(m_X^{(r)}\right)^2\right], \quad \mathrm{Var}\left[\widetilde{M}_{XY}^{(rs)}\right] = \frac{1}{n}\left[m_{XY}^{(2r,2s)} - \left(m_{XY}^{(rs)}\right)^2\right] \tag{3.6}$$

It can be observed that if the population moments are finite, then the variances tend to zero as $n \to \infty$; therefore the estimators are consistent.

Typical central moment estimators, i.e. estimators of $\mu_X^{(r)}$ and $\mu_{XY}^{(rs)}$ of one and two variables, respectively, are those defined by the equations

$$\hat{M}_X^{(r)} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^r}{n}, \quad \hat{M}_{XY}^{(rs)} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^r \left(Y_i - \overline{Y}\right)^s}{n} \tag{3.7}$$

These have been formed based on the empirical method described in section 3.2.3.These estimators are biased (for $r + s > 1$).

3.3.2 Sample mean

The most common statistic is the sample mean. As we have seen in section 3.2.3, the sample mean is an estimator of the true (or population) mean $m_X = E[X]$ and is defined by

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.8}$$

which is a special case of (3.4) for $r = 1$. Its numerical value

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.9}$$

is called the *observed sample mean* or simply the *average*. The symbols $\overline{X}$ and $\overline{x}$ should not be conceptually confused with each other nor with the true mean of the random variable X, i.e. $m_X = E[X]$, which is defined based on the equations (2.20) or (2.21) and (2.23). Nevertheless, these three quantities are closely related. Implementation of equations (3.5) and (3.6) gives

$$E\left[\overline{X}\right] = E\left[X\right], \quad \mathrm{Var}\left[\overline{X}\right] = \frac{\mathrm{Var}\left[X\right]}{n} \tag{3.10}$$

regardless of the distribution function of X*. Thus, the estimator is unbiased and consistent.

* However, Var[$\overline{X}$] depends on the dependence structure of the variables X_i; the formula given in (3.10) holds only if X_i are independent. On the other hand, the formula for $E[\overline{X}]$ holds always.

3.3.3 Variance and standard deviation

A biased estimator of the true (population) variance σ_X^2 is:

$$S_X^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n} \tag{3.11}$$

It can be proved (Kendall and Stewart, 1968, p. 277) that

$$\begin{aligned} E\left[S_X^2\right] &= \frac{n-1}{n}\sigma_X^2 \\ \mathrm{Var}\left[S_X^2\right] &= \frac{\mu_X^{(4)} - \sigma_X^4}{n} - \frac{2\left(\mu_X^{(4)} - 2\sigma_X^4\right)}{n^2} + \frac{\mu_X^{(4)} - 3\sigma_X^4}{n^3} \end{aligned} \tag{3.12}$$

where $\mu_X^{(4)}$ is the fourth central population moment. The two last terms in the expression of $\mathrm{Var}\left[S_X^2\right]$can be omitted for large values of *n*. From the expression of $E\left[S_X^2\right]$ in (3.12) we observe that multiplication of S_X^2 by $n/(n-1)$ results in an unbiased estimator of σ_X^2, i.e.

$$S_X^{*2} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} \tag{3.13}$$

S_X^{*2} is known as *sample variance*. For large sample sizes, the two estimators S_X^2 and S_X^{*2} are practically the same. If the population is normally distributed, it can be shown that

$$\mathrm{Var}\left[S_X^2\right] = \frac{2(n-1)\sigma_X^4}{n^2}, \quad \mathrm{Var}\left[S_X^{*2}\right] = \frac{2\sigma_X^4}{(n-1)} \tag{3.14}$$

The standard deviation estimators in common use are the square roots of the variance estimators, namely the S_X and S_X^* and are not unbiased. Thus (Yevjevich, 1972, p. 193· Kendall and Stewart, 1968, p. 233),

$$E\left[S_X\right] = \sigma_X + O\left(\frac{1}{n}\right), \quad \mathrm{Var}\left[S_X\right] = \frac{\mu_X^{(4)} - \sigma_X^4}{4\sigma_X^2 n} + O\left(\frac{1}{n^2}\right) \tag{3.15}$$

where the terms $O(1/n)$ are $O(1/n^2)$ are quantities proportional to $1/n$ and $1/n^2$, respectively, and can be omitted if the sample size is large enough ($n \geq 20$).

If the population is normally distributed, the following approximate equations can be used for S_X

$$E\left[S_X\right] \approx \sigma_X\sqrt{\frac{n-1}{n}}, \quad \mathrm{Var}\left[S_X\right] \approx \frac{\sigma_X^2}{2n} \tag{3.16}$$

For $n \geq 10$ the errors of these equations are smaller than 2.5% and 2.7% respectively, while for $n \geq 100$ are practically zero. The corresponding equations for S_X^* are*

$$E\left[S_X^*\right] \approx \sigma_X, \quad \mathrm{Var}\left[S_X^*\right] \approx \frac{\sigma_X^2}{2(n-1)} \tag{3.17}$$

Finally, one of the two following estimators of the coefficient of variation can be used:

$$\hat{C}_{v_X} = \frac{S_X}{\overline{X}}, \quad \hat{C}_{v_X}^* = \frac{S_X^*}{\overline{X}} \tag{3.18}$$

If the variable X is positive, then it can be shown that these estimators are bounded from above ($\hat{C}_{v_X} \leq \sqrt{n-1}$) while the same does not hold for the corresponding population parameters. Obviously, this introduces bias.†

3.3.4 Third central moment and skewness coefficient

A biased estimator of the true (population) third central moment $\mu_X^{(3)}$ is given by

$$\hat{M}_X^{(3)} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{n} \tag{3.19}$$

for which it can be shown (Kendall and Stewart, p. 278-281) that

$$E\left[\hat{M}_X^{(3)}\right] = \frac{(n-1)(n-2)}{n^2}\mu_X^{(3)} \tag{3.20}$$

It immediately follows that an unbiased (and consistent) estimator of $\mu_X{}^{(3)}$ is

$$\hat{M}_X^{*(3)} = \frac{n\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{(n-1)(n-2)} \tag{3.21}$$

For large sample size n, the two estimators are practically the same.

The estimation of the skewness coefficient C_{s_X} is done using the following estimator

$$\hat{C}_{s_X} = \frac{\hat{M}_X^{(3)}}{S_X^3} \tag{22}$$

---

* More accurate approximations are given by

$$E\left[S_X^*\right] \approx \sigma_X \sqrt{\frac{n-\frac{5}{4}}{n-\frac{3}{4}}}, \quad Var\left[S_X^*\right] \approx \frac{\sigma_X^2}{2\left(n-\frac{3}{4}\right)}$$

the errors of which for $n \geq 10$ are less than 0.005% and 0.2%, respectively. The precise equations are

$$E\left[S_X^*\right] = \sigma_X \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)}, \quad Var\left[S_X^*\right] = \sigma_X^2\left[1 - \frac{\Gamma^2\left(\frac{n}{2}\right)}{\frac{n-1}{2}\Gamma^2\left(\frac{n-1}{2}\right)}\right]$$

† The expression of the estimator's variance is quite complex and is omitted (cf. Kendall and Stewart, 1968, p. 233). If X follows a normal distribution then

$$\mathrm{Var}\left[\hat{C}_{v_X}\right] \approx C_{v_X}^2 / 2n$$

which is not unbiased. The bias does not originate only from the fact that the two moment estimators (numerator and denominator) are not unbiased themselves, but also (mainly) from the fact that $\hat{C}_{s_X}$ is bounded both from above and from below, whilst the population C_{s_X} is not bounded. This is due to the finite sample size n, which determines the upper and lower limit. Thus, it has been shown (Kirby, 1974; Wallis *et al.*, 1974) that $|\hat{C}_{s_X}| \le (n-2)/\sqrt{n-1}$.

Several approximate bias correction coefficients have been proposed in the literature to be multiplied by $\hat{C}_{s_X}$ estimated from (3.22) to obtain a less biased estimate. None of them leads to a rigorous unbiased estimator of the coefficient of skewness. The four most common are:

$$\frac{\sqrt{n(n-1)}}{n-2},\ \frac{n^2}{(n-1)(n-2)},\ \frac{\sqrt{n(n-1)}}{n-2}\left(1+\frac{8.5}{n}\right),\ 1+\left(\frac{6.51}{n}+\frac{20.20}{n^2}\right)+\left(\frac{1.48}{n}+\frac{6.77}{n^2}\right)\hat{C}_{s_X}^2 \quad (3.23)$$

The first is obtained if in (3.22) the biased moment estimators are replaced by the unbiased ones. The second results if in (3.22) we replace the biased third moment estimator with the unbiased one (Yevjevich, 1978, p. 110). The third one has been proposed by Hazen and the last one has been proposed by Bobée and Robitaille (1975), based on results by Wallis *et al.* (1974).

3.3.5 L-moments estimates

Unbiased estimates $b_X^{(r)}$ of the probability weighted moments $\beta_X^{(r)}$ are given by the following relationship (Landwehr *et al.*, 1979):

$$b_X^{(r)} = \frac{1}{n}\frac{\sum_{i=1}^{n-r}\binom{n-j}{r}x_{(i)}}{\binom{n-1}{r}} = \frac{1}{n}\sum_{i=1}^{n-r}\frac{(n-i)(n-i-1)...(n-i+r+1)}{(n-1)(n-2)...(n-r)}x_{(i)} \quad (3.24)$$

where n is the sample size, and $x_{(i)}$ the ordered observations so that $x_{(n)} \le \ldots \le x_{(2)} \le x_{(1)}$*. The estimates† of the first four probability weighted moments are:

$$b_X^{(0)} = \frac{1}{n}\sum_{i=1}^{n}x_{(i)} = \bar{x}$$

$$b_X^{(1)} = \frac{1}{n}\sum_{i=1}^{n-2}\frac{n-i}{n-1}x_{(i)}$$

$$b_X^{(2)} = \frac{1}{n}\sum_{i=1}^{n-2}\frac{(n-i)(n-i-1)}{(n-1)(n-2)}x_{(i)} \quad (3.25)$$

* Notice that $x_{(1)}$ is the largest observation; the equations are somewhat simpler if the observations are ordered from smallest to largest but it has been the rule in engineering hydrology to put the observations in descending order.

† The estimators of the same quantities are obtained by replacing $x_{(i)}$ with the variable $X_{(i)}$, the so called *order statistic*.

$$b_X^{(3)} = \frac{1}{n}\sum_{i=1}^{n-3}\frac{(n-i)(n-i-1)(n-i-2)}{(n-1)(n-2)(n-3)}x_{(i)}$$

Accordingly, the estimates of the first four L moments are calculated by the equations relating L moments and probability weighted moments (see equation (2.32)), i.e.,

$$\begin{aligned} l_X^{(1)} &= b_X^{(0)} \ (= \bar{x}) \\ l_X^{(2)} &= 2\ b_X^{(1)} - b_X^{(0)} \\ l_X^{(3)} &= 6b_X^{(2)} - 6b_X^{(1)} + b_X^{(0)} \\ l_X^{(4)} &= 20b_X^{(3)} - 30b_X^{(2)} + 12\ b_X^{(1)} - b_X^{(0)} \end{aligned} \tag{3.26}$$

3.3.6 Covariance and correlation

A biased estimator of the covariance σ_{XY} of two variables X and Y is:

$$S_{XY} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n} \tag{3.27}$$

It can be shown (e.g. Papoulis, 1990, p. 295) that

$$E\left[S_{XY}\right] = \frac{n-1}{n}\sigma_{XY} \tag{3.28}$$

Therefore, an unbiased (and consistent) estimator of σ_{XY} is

$$S_{XY}^{*} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1} \tag{3.29}$$

known as *sample covariance*[*].

The estimator of the correlation coefficient ρ_{XY} is given by the next relationship, known as the *sample correlation coefficient*:

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{S_{XY}^{*}}{S_X^{*} S_Y^{*}} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \tag{3.30}$$

The precise distribution function of this estimator and its moments are difficult to determine analytically; however, this estimator is regarded approximately unbiased.

[*] In many books, the denominator of (3.29) has the term $n - 2$, which is not correct.

3.4 Typical confidence intervals

3.4.1 Mean – known population variance

Let X be a random variable with mean μ_X and standard deviation σ_X. According to the central limit theorem and equation (3.10), the sample mean $\overline{X}$ (the average of n random variables) will have normal distribution $N\left(\mu_X, \sigma_X / \sqrt{n}\right)$, if n is large enough. Moreover, it will have precisely this normal distribution irrespectively of the size n, if the random variable X is normal.

The problem we wish to study here is the determination of the confidence intervals of the mean μ_X for confidence coefficient γ. We denote $z_{(1+\gamma)/2}$ the $((1+\gamma)/2)$-quantile of the standard normal distribution $N(0, 1)$ (that is the value z that corresponds to non-exceedence probability $(1+\gamma)/2$). Apparently, due to symmetry, $z_{(1-\gamma)/2} = -z_{(1+\gamma)/2}$ (see Fig. 3.1). Thus,

$$P\left\{\mu_X - \frac{z_{(1+\gamma)/2}\sigma_X}{\sqrt{n}} < \overline{X} < \mu_X + \frac{z_{(1+\gamma)/2}\sigma_X}{\sqrt{n}}\right\} = \gamma \tag{3.31}$$

or equivalently

$$P\left\{\overline{X} - \frac{z_{(1+\gamma)/2}\sigma_X}{\sqrt{n}} < \mu_X < \overline{X} + \frac{z_{(1+\gamma)/2}\sigma_X}{\sqrt{n}}\right\} = \gamma \tag{3.32}$$

Equation (3.32) gives the confidence intervals sought. For the numerical evaluation we simply replace the estimator $\overline{X}$ in (3.32) with its numerical value $\bar{x}$.

For convenience, Table 3.1 displays the most commonly used confidence coefficients and the corresponding normal quantiles $z_{(1+\gamma)/2}$. We observe that as the confidence coefficient tends to 1, which means that the reliability of the estimate increases, the confidence interval becomes larger so that the estimate becomes more vague. On the contrary, if we choose a smaller confidence coefficient, a more "compact" estimate will result. In this case, the confidence interval will be narrower but the uncertainty will be higher.

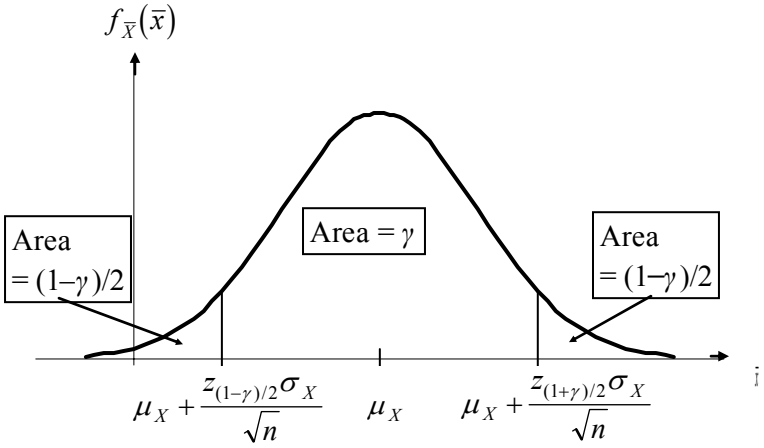

Fig. 3.1 Explanatory sketch for the confidence intervals of the mean.

Table 3.1 Typical values of the normal quantiles $z_{(1+\gamma)/2}$ useful for the calculation of confidence intervals.

γ	0.90	0.95	0.99	0.999
$(1+\gamma)/2$	0.95	0.975	0.995	0.9995
$z_{(1+\gamma)/2}$	1.645	1.960	2.576	3.291

We observe from (3.32) that the only way to increase the accuracy without increasing the length of the confidence interval is to increase the sample size n by taking additional measurements.

The previous analysis was based on the assumption of known population variance, which in practice it is not realistic, because typically all our information comes from a sample. However, the results are of practical interest, since (3.32) provides a good approximation if the sample size n is large enough (> 30) and if we replace the population variance with its sample estimate.

3.4.2 Mean – unknown population variance

The analysis that we present here can be used for unknown population variance and for any sample size. However, this analysis has a restrictive condition, that the random variable X is normal, $N(\mu_X, \sigma_X)$. In this case the following conclusions can be drawn:

1. The sample mean has a normal distribution $N\left(\mu_X, \sigma_X / \sqrt{n}\right)$. This conclusion is a consequence of a basic property of the normal distribution, specifically the normal distribution is closed under addition or, else, a *stable* distribution.
2. The function of the sample variance $(n-1)S_X^{*2}/\sigma_X^2$ follows the $\chi^2(n-1)$ distribution. This is concluded by the theorem of section 2.10.4, according to which the sum of the squares of a number of standard normal variables follows the χ^2 distribution.
3. The random variables $\overline{X}$ and S_X^{*2} are independent. This results form a statistical theorem (see e.g. Papoulis, 1990, p. 222).
4. The ratio $(\overline{X} - \mu_X)/(S_X^* / \sqrt{n})$ follows the Student $t(n-1)$ distribution. This results by a theorem of 2.10.5.

We denote $t_{(1+\gamma)/2}$ the $[(1+\gamma)/2]$-quantile of the Student $t(n-1)$ distribution (that is the point t that corresponds to exceedence probability $(1+\gamma)/2$, for $n-1$ degrees of freedom). Because of the symmetry, $t_{(1-\gamma)/2} = -t_{(1+\gamma)/2}$. Thus,

$$P\left\{-t_{(1+\gamma)/2} < \frac{\overline{X} - \mu_X}{S_X^* / \sqrt{n}} < t_{(1+\gamma)/2}\right\} = \gamma \tag{3.33}$$

or equivalently

$$P\left\{\overline{X}-\frac{t_{(1+\gamma)/2}S_X^*}{\sqrt{n}}<\mu_X<\overline{X}+\frac{t_{(1+\gamma)/2}S_X^*}{\sqrt{n}}\right\}=\gamma \tag{3.34}$$

Equation (3.34) provides the confidence interval sought. For its numerical evaluation we simply replace in the interval estimators of (3.34) the estimators $\overline{X}$ and S_X^* with the corresponding sample estimates $\bar{x}$ and s_X^*.

Even though (3.32) and (3.34) are considerably different regarding their theoretical grounds and the assumptions they rely upon, from a computational perspective they are quite similar. Furthermore, for large n (>30) they practically coincide taking into account that $t_{(1+\gamma)/2}\approx z_{(1+\gamma)/2}$ (more precisely $t_{(1+\gamma)/2}\approx z_{(1+\gamma)/2}\sqrt{(n-1)/(n-3)}$, for $n-1$ degrees of freedom).

The two previous analyses do not include the case of a small sample size, unknown variance and non-normal distribution. This case is not covered in statistics in a general and rigorous manner. However, as an approximation, often the same methodology is also used in these cases, provided that the population distribution is bell shaped and not too skewed. In general, the cases where precise confidence intervals can be determined based on a consistent theoretical procedure, are the exception rather than the rule. In most of the following problems we will use just approximations of the confidence intervals.

3.4.3 A numerical example of interval estimation of the mean

From a sample of annual inflows to a reservoir with length 15 (years), the sample mean is 10.05 hm^3 and the sample standard deviation 2.80 hm^3. We wish to determine (1) the 95% confidence interval of the annual inflow and (2) the sample size for 95% confidence coefficient that enables 10% precision in the estimation of the annual inflow.

(1) We assume that the annual inflows are IID with normal distribution (section 2.10.2) and we use the equation (3.34). Using the table of the Student distribution (Appendix A3) or any computational method (see section 2.10.5) we find that for $n-1=14$ degrees of freedom $t_{(1+\gamma)/2}=t_{0.975}=2.14$. Consequently, the 95% confidence interval is*

$$10.05-2.14\times 2.80/\sqrt{15}<\mu_X<10.05+2.14\times 2.80/\sqrt{15}\ \text{(in hm}^3\text{)}$$

or

$$8.50<\mu_X<11.60\ \text{(in hm}^3\text{)}$$

For comparison, we will calculate the confidence interval using equation (3.32), even though this is not correct. From Table 3.1 we find $z_{(1+\gamma)/2}=z_{0.975}=1.96$. Thus, the 95% confidence interval is

---

* It would not be mathematically correct to write (3.34) replacing the estimators with their estimates, i.e.

$$P\{10.05-2.62\times 2.80/\sqrt{15}<\mu_X<10.05-2.62\times 2.80/\sqrt{15}\}=0.95$$

We note that μ_X is a (unknown) parameter (i.e. a number) and not a random variable, so it does not have a distribution function. Moreover, it is not correct to say e.g. that "with 95% probability the mean value lies in the interval (8.16, 11.94)". The correct expression would be "with 95% confidence".

$$10.05 - 1.96 \times 2.80/\sqrt{15} < \mu_X < 10.05 + 1.96 \times 2.80/\sqrt{15} \text{ (in hm}^3\text{)}$$

or

$$8.63 < \mu_X < 11.47 \text{ (in hm}^3\text{)}$$

The confidence interval is this case is a little smaller.

(2) Assuming that $n \geq 30$ we can use (3.32). The following equation must hold

$$1.96 \times 2.8 / \sqrt{n} = 10\% \times 10.05$$

so $n = 30$. We observe that the condition we have assumed ($n \geq 30$) is valid. (If it were not valid we should proceed with a trial-and-error procedure, using equation (3.34)).

3.4.4 Variance and standard deviation

As in the section 0, we will assume that the random variable X has a normal distribution $N(\mu_X, \sigma_X)$. As mentioned before, in this case the function of the sample variance $(n-1)S_X^{*2}/\sigma_X^2$ follows the $\chi^2(n-1)$ distribution.

We denote $\chi^2_{(1+\gamma)/2}$ and $\chi^2_{(1-\gamma)/2}$ the [(1+γ)/2]- and [(1–γ)/2]-quantiles, respectively, of the $\chi^2(n-1)$ distribution (the two are not equal because the χ^2 distribution is not symmetric). Thus, we have

$$P\left\{\chi^2_{(1-\gamma)/2} < \frac{(n-1)S_X^{*2}}{\sigma_X^2} < \chi^2_{(1+\gamma)/2}\right\} = \gamma \tag{3.35}$$

or equivalently

$$P\left\{\frac{(n-1)S_X^{*2}}{\chi^2_{(1+\gamma)/2}} < \sigma_X^2 < \frac{(n-1)S_X^{*2}}{\chi^2_{(1-\gamma)/2}}\right\} = \gamma \tag{3.36}$$

Equation (3.36) gives the confidence interval sought. It is easily obtained that confidence interval of the standard deviation is given by

$$P\left\{\frac{\sqrt{n-1}S_X^*}{\sqrt{\chi^2_{(1+\gamma)/2}}} < \sigma_X < \frac{\sqrt{n-1}S_X^*}{\sqrt{\chi^2_{(1-\gamma)/2}}}\right\} = \gamma \tag{3.37}$$

3.4.5 A numerical example of interval estimation of standard deviation

We wish to determine the 95% confidence interval of the standard deviation of annual inflow in the problem of section 3.4.3.

The sample standard deviation is 2.8 hm^3. With the assumption of normal distribution for the inflow, we utilize equation (3.37). Using the χ^2 distribution table (Appendix 2) or any computational method (see section 2.10.4) we find that for $n - 1 = 14$ degrees of freedom $\chi^2_{(1+\gamma)/2} = \chi^2_{0.975} = 26.12$ and $\chi^2_{(1-\gamma)/2} = \chi^2_{0.025} = 5.63$. Thus, the 95% confidence interval is

$$\frac{\sqrt{14} * 2.80}{\sqrt{26.12}} < \sigma_X < \frac{\sqrt{14} * 2.80}{\sqrt{5.63}} \text{ (in hm}^3\text{)}$$

or

$$2.05 < \sigma_X < 4.41 \text{ (in hm}^3\text{)}$$

3.4.6 Normal distribution quantile – Standard error

In engineering design and management (in engineering hydrology in particular), the most frequent confidence interval problem that we face, concerns the estimation of design values for quantities that are modelled as random variables. For instance, in hydrological design we may wish to estimate the reservoir inflow that corresponds to a non-exceedence probability 1%, that is the 1% quantile of the inflow. Let X be a random variable with distribution $F_X(x)$ representing a natural quantity, e.g. a hydrological variable. Here we assume that $F_X(x)$ is a normal distribution $N(\mu_X, \sigma_X)$, which can be easily handled, whereas in Chapter 6 we will present similar methods for a repertoire of distributions being commonly used in engineering applications. For a given non-exceedence probability $u = F_X(x)$, the corresponding value of the variable X (symbolically x_u, the u-quantile) will be

$$x_u = \mu_X + z_u \sigma_X \tag{3.38}$$

where z_u the u-quantile of the standard normal distribution $N(0, 1)$. However, in this equation the population parameters μ_X and σ_X are unknown in practice. Using their point estimates, we obtain an estimate $\hat{x}_u = \bar{x} + z_u s_X$, that can be considered as a value of the random variable

$$\hat{X}_u = \overline{X} + z_u S_X \tag{3.39}$$

This latter equation can be used to determine the confidence interval of x_u. The precise determination is practically impossible, due to the complexity of the distribution function of $\hat{X}_u$. Here we will confine our analysis in seeking an approximate confidence interval, based on the assumption that $\hat{X}_u$ has normal distribution.

The mean of $\hat{X}_u$ is given from equation (2.59), which can be combined with (3.10) and (3.15) to give

$$E\left[\hat{X}_u\right] = E\left[\overline{X}\right] + z_u E\left[S_X\right] \approx \mu_X + z_u \sigma_X = x_u \tag{3.40}$$

assuming that n is large enough and omitting the term $O(1/n)$ in $E[S_X]$.* Likewise, the variance of $\hat{X}_u$ is given by equation (2.61), which can be written as

$$\mathrm{Var}\left[\hat{X}_u\right] = \mathrm{Var}\left[\overline{X}\right] + z_u^2 \mathrm{Var}\left[S_X\right] + 2 z_u \mathrm{Cov}\left[\overline{X}, S_X\right] \tag{3.41}$$

Given that X has normal distribution, the third term of (3.41) is zero (as mentioned before, the variables $\overline{X}$ and S_X are independent). Combining (3.10) and (3.16), we write (3.41) as

* The analysis here has an approximate character, thus we do not discriminate between the estimators S_X and S_X^*, because for n large enough the two estimators are virtually identical.

$$\varepsilon_u^2 := \mathrm{Var}\left[\hat{X}_u\right] \approx \frac{\sigma_X^2}{n} + z_u^2 \frac{\sigma_X^2}{2n} = \frac{\sigma_X^2}{n}\left(1 + \frac{z_u^2}{2}\right) \tag{3.42}$$

The quantity ε_u is known in literature as *standard quantile error* or simply as *standard error*.

Assuming that $\hat{X}_u$ has a normal distribution $N(x_u, \varepsilon_u)$ we can write

$$P\left\{-z_{(1+\gamma)/2} < \frac{\hat{X}_u - x_u}{\varepsilon_u} < z_{(1+\gamma)/2}\right\} = \gamma \tag{3.43}$$

where γ is the confidence coefficient. Equivalently,

$$P\left\{\hat{X}_u - z_{(1+\gamma)/2}\varepsilon_u < x_u < \hat{X}_u + z_{(1+\gamma)/2}\varepsilon_u\right\} = \gamma \tag{3.44}$$

Replacing in the previous equation the term ε_u from (3.42), and then the standard deviation σ_X with its estimator, we obtain the following final relationship

$$P\left\{\hat{X}_u - z_{(1+\gamma)/2}\sqrt{1 + \frac{z_u^2}{2}}\frac{S_X}{\sqrt{n}} < x_u < \hat{X}_u + z_{(1+\gamma)/2}\sqrt{1 + \frac{z_u^2}{2}}\frac{S_X}{\sqrt{n}}\right\} = \gamma \tag{3.45}$$

The latter equation is an approximation, whose accuracy is increased a n increases. Moreover, it is valid only in the case of normal distribution. However, (3.44) is also used for other distributions of the variable X, but with a different expression of the standard error ε_u and a different calculation method. The interested reader for a general expression of the standard error may consult Kite (1988, p. 33-38).

The estimates of the confidence limits are

$$\hat{x}_{u1} = \hat{x}_u - z_{(1+\gamma)/2}\sqrt{1 + \frac{z_u^2}{2}}\frac{s_X}{\sqrt{n}}, \quad \hat{x}_{u2} = \hat{x}_u + z_{(1+\gamma)/2}\sqrt{1 + \frac{z_u^2}{2}}\frac{s_X}{\sqrt{n}} \tag{3.46}$$

Clearly these estimates are functions of u or, equivalently, of the exceedence probability, $1 - u$. The depictions of those functions in a probability plot placed on either side of the x_u curve are known as confidence curves of x_u.

3.4.7 A numerical example of interval estimation of distribution quantiles

Further to the numerical example of section 3.4.3, we wish to determine the 95% confidence interval of the annual inflow that has exceedence probability (a) 1% and (b) 99%. We note that, because of the small sample size, we will not expect a high degree of accuracy in our estimates (recall that the theoretical analysis assumed large sample size).

We will calculate first the point estimates (all units are hm^3). For the annual inflow with exceedence probability $F^* = 0.01$ we have $u = 1 - F^* = 0.99$ and $z_u = 2.326$. Thus, the point estimate of $\hat{x}_u = 10.05 + 2.326 \times 2.80 = 16.56$. Likewise, for the annual inflow with exceedence probability $F^* = 0.99$ we have $u = 1 - F^* = 0.01$ and $z_u = -2.326$, thus $\hat{x}_u = 10.05 - 2.326 \times 2.80 = 3.54$.

We can now proceed in the calculation of the confidence limits. For γ = 95% and $z_{(1+\gamma)/2}$ = 1.96, the limits for the inflow with exceedence probability 1% are:

$$\hat{x}_{u1} = 16.56 - 1.96\sqrt{1+\frac{2.326^2}{2}}\frac{2.80}{\sqrt{15}} = 13.83$$

$$\hat{x}_{u2} = 16.56 + 1.96\sqrt{1+\frac{2.326^2}{2}}\frac{2.80}{\sqrt{15}} = 19.29$$

Likewise, the limits for exceedence probability 99% are:

$$\hat{x}_{u1} = 3.54 - 1.96\sqrt{1+\frac{2.326^2}{2}}\frac{2.80}{\sqrt{15}} = 0.81$$

$$\hat{x}_{u2} = 3.54 + 1.96\sqrt{1+\frac{2.326^2}{2}}\frac{2.80}{\sqrt{15}} = 6.27$$

3.4.8 Correlation coefficient

To calculate the confidence limits of the correlation coefficient ρ of a population described by two variables X and Y, we use the auxiliary variable Z, defined by the so-called Fisher transformation:

$$Z = \frac{1}{2}\ln\frac{1+R}{1-R} \leftrightarrow R = \frac{e^{2Z}-1}{e^{2Z}+1} = \tanh Z \tag{3.47}$$

where R the sample correlation coefficient. We observe that for $-1 < R < 1$ the range of Z is $-\infty < Z < \infty$, while for $R = 0$, $Z = 0$. It can be shown that if X and Y are normally distributed, then Z has approximately normal distribution $N(\mu_Z, \sigma_Z)$ where

$$\mu_Z = E[Z] \approx \frac{1}{2}\ln\frac{1+\rho}{1-\rho}, \quad \sigma_Z^2 = \mathrm{Var}[Z] \approx \frac{1}{n-3} \tag{3.48}$$

As a consequence, if $\zeta_{(1+\gamma)/2}$ is the $(1+\gamma)/2$-quantile of the standard normal distribution, we obtain

$$P\left\{\mu_Z - \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}} < Z < \mu_Z + \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}}\right\} = \gamma \tag{3.49}$$

or equivalently

$$P\left\{Z - \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}} < \mu_Z < Z + \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}}\right\} = \gamma \tag{3.50}$$

Replacing μ_Z from (3.48) into (3.50) and solving for ρ, and also taking into account the monotonicity of the transformation (3.47), we obtain

$$P\{R_1 < \rho < R_2\} \tag{3.51}$$

where

$$R_1 = \frac{e^{2Z_1}-1}{e^{2Z_1}+1} \qquad R_2 = \frac{e^{2Z_2}-1}{e^{2Z_2}+1}$$
$$\left.\begin{matrix} Z_2 \\ Z_1 \end{matrix}\right\} = Z \pm \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}} \qquad Z = \frac{1}{2}\ln\frac{1+R}{1-R} \tag{3.52}$$

To numerically evaluate the confidence limits we implement equations (3.52) replacing the estimators with the corresponding estimates (e.g. $R = r$, etc.).

3.5 Parameter estimation of distribution functions

Assuming a random variable X with known distribution function and with unknown parameters θ_1, θ_2, …, θ_r we can denote the probability density function of X as a function $f_X(x; \theta_1, \theta_2, \ldots, \theta_r)$. For known x, this, viewed as a function of the unknown parameters, is called the *likelihood function*. Here, we will examine the problem of the estimation of these parameters based on a sample X_1, X_2, …, X_n. Specifically, we will present the two most classical methods in statistics, namely the moments method and the maximum likelihood method. In addition, we will present a newer method that has become popular in hydrology, the method of L moments.

Several other general methods have been developed in statistics for parameter estimation, e.g. the maximum entropy method that has been also used in hydrology (the interested reader is referenced to Singh and Rajagopal, 1986). Moreover, in engineering hydrology in many cases, other types of methods like graphical, numerical, empirical and semi-empirical have been used. Examples of such methods will be given for certain distributions in chapter 6.

3.5.1 The method of moments

The method of moments is based on equating the theoretical moments of variable X with the corresponding sample moment estimates. Thus, if r is the number of the unknown parameters of the distribution, we can write r equations of the form

$$m_X^{(k)} = \hat{m}_X^{(k)}, \quad k = 1,2,\ldots,r \tag{3.53}$$

where $m_X^{(k)}$ are the theoretical raw moments, which are functions of the unknown parameters and are given by

$$m_X^{(k)} = \int_{-\infty}^{\infty} x^k f_X(x,\theta_1,\ldots,\theta_r)dx \tag{3.54}$$

whereas $\hat{m}_X^{(k)}$ are the estimates, calculated from the observed sample according to

$$\hat{m}_X^{(k)} = \frac{1}{n}\sum_{i=1}^{n} x_i^k \tag{3.55}$$

Thus, the solution of the resulting system of the r equations gives the unknown parameters $\theta_1, \theta_2, \ldots, \theta_r$. In general, the system of equations may not be linear and may not have an analytical solution. In this case the system can be solved only numerically.

Equivalently, we can use the central moments (for $k > 1$) instead of the raw moments. In this case, the system of equations is

$$\mu_X = \bar{x}, \quad \mu_X^{(k)} = \hat{\mu}_X^{(k)}, \quad k = 2,\ldots,r \tag{3.56}$$

where $\mu_X = m_X^{(1)}$ is the population mean, $\bar{x} = \hat{m}_X^{(1)}$ the sample mean, $\mu_X^{(k)}$ the theoretical central moments given by the

$$\mu_X^{(k)} = \int_{-\infty}^{\infty} (x - \mu_X)^k f_X(x;\theta_1,\ldots,\theta_r)dx \tag{3.57}$$

and $\hat{\mu}_X^{(k)}$ the corresponding sample estimates calculated by the relationship

$$\hat{\mu}_X^{(k)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k \tag{3.58}$$

We recall that the raw moments in (3.55) are unbiased estimates, while the central moments in (3.58) are biased estimates. Nevertheless, unbiased central moment estimates are often used instead of the biased. Regardless of using biased or unbiased estimates for moments, in general the method of moments does not result in unbiased estimates of the parameters $\theta_1, \theta_2, \ldots, \theta_r$ (except in special cases).

3.5.2 Demonstration of the method of moments for the normal distribution

As an example of the implementation of the method of moments, we will calculate the parameters of the normal distribution. The probability density function is:

$$f_X(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.59}$$

and has two parameter, μ and σ. Thus, we need two equations. Based on (3.56), these equations are

$$\mu_X = \bar{x}, \quad \sigma_X^2 = s_X^2 \tag{3.60}$$

where in the latter equation we have denoted the theoretical and sample variance (that is, the second central moment) of X, by the more common symbols σ_X^2 and s_X^2, respectively. We know (see section 2.10.2) that the theoretical moments are

$$\mu_X = \mu, \quad \sigma_X^2 = \sigma^2 \tag{3.61}$$

Consequently, the final estimates are

$$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \sigma = s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.62}$$

This estimation of σ biased, while that of μ is unbiased. If in the last equation we used the unbiased estimate of the variance, then we would have in the denominator (n-1) instead of n. Even in this case, the estimate of the σ would not be unbiased, for reasons explained in section 3.3.3.

As we have seen in this example, the application of the method of moments is very simple and this extends to other distribution functions.

3.5.3 The method of L moments

The logic of the method of L moments is the same as in the method of moments. If the distribution has r unknown parameters, we can write r equations of the form

$$\lambda_k = l_k,\ k = 1, 2, \ldots, r \tag{3.63}$$

where λ_k are the theoretical L-moments, which are functions of the unknown parameters, and l_k their sample estimates. Solving this system of equations we obtain the L-moment estimates of the unknown parameters of the distribution. Because L moments are linear combinations of the probability weighted moments, (3.63) can be written equivalently as

$$\beta_k = b_k,\ k = 0, 2, \ldots, r - 1 \tag{3.64}$$

where β_k is the probability weighted moment of order k and b_k is its estimate (see section 3.3.5).

Estimates based on L-moments are generally more reliable than those based on classical moments. Moreover, the L-moment estimators have some statistically desirable properties e.g. they are robust with respect to outliers, because contrary to standard moments, they do not involve squaring, cubing, etc., of the sample observations. In hydrology, the L moments have been widely used as descriptive statistics and in parameter estimation of several distributions. Examples of applications of the method can be found, among others, in Kjeldsen *et al.* (2002), Kroll and Vogel (2002), Lim and Lye (2003) and Zaidman *et al.* (2003).

3.5.4 The maximum likelihood method

Let X be a random variable with probability function $f_X(x, \theta_1, \theta_2, \ldots, \theta_r)$ where $\theta_1, \theta_2, \ldots, \theta_r$ are parameters, and $X_1, X_2, \ldots, X_n$ a sample of the variable. Let $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots\theta_r)$ be the joint distribution function of the sample vector $\boldsymbol{X} := [X_1, X_2, \ldots, X_n]^T$. Our entire observed sample can be thought of as single observation of the vector variable $\boldsymbol{X}$. The idea behind the maximum likelihood method is that the probability density $f_{X_1,\ldots,X_n}(\ )$ at this single point will be as high as possible (it is natural to expect an observation to lie in an area with high probability density). We can thus find $\theta_1, \theta_2, \ldots, \theta_r$, so that the function $f_{X_1,\ldots,X_n}(\ )$ have a value as high as possible at the point $(x_1, x_2, \ldots, x_n)$.

In a random sample, the variables $X_1, X_2, \ldots, X_n$ are independent and the joint probability density function is

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots\theta_r)=\prod_{i=1}^{n} f_X(x_i,\theta_1,\ldots\theta_r) \tag{3.65}$$

and, viewed as a function of the parameters θ_1, θ_2, …, θ_r (for values of random variables equal to the observations x_1, …, x_n) is the likelihood function of these parameters.

Assuming that $f_{X_1,\ldots,X_n}(\ )$ is differentiable with respect to its parameters, the condition that maximizes it is

$$\frac{\partial f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_r)}{\partial\theta_k}=0,\quad k=1,\ldots,r \tag{3.66}$$

Using these r equations, the r unknown parameters will result. However, the manipulation of these equations may be complicated and, instead of maximizing the likelihood, we may attempt to maximize its logarithm

$$L(x_1,\ldots,x_n;\theta_1,\ldots\theta_r):=\ln f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots\theta_r)=\sum_{i=1}^{n}\ln f_X(x_i;\theta_1,\ldots\theta_r) \tag{3.67}$$

The function $L(\ )$ is called the log-likelihood function of the parameters θ_1, θ_2, …, θ_r. In this case, the condition of maximum is

$$\frac{\partial L(x_1,\ldots,x_n;\theta_1,\ldots\theta_r)}{\partial\theta_k}=\sum_{i=1}^{n}\frac{1}{f_X(x_i;\theta_1,\ldots\theta_r)}\frac{\partial f_X(x_i;\theta_1,\ldots\theta_r)}{\partial\theta_k}=0,\quad k=1,\ldots,r \tag{3.68}$$

Solving these r equations we obtain the values of the r unknown parameters.

3.5.5 Demonstration of the maximum likelihood method for the normal distribution

We will calculate the parameters of the normal distribution using the maximum likelihood method. The probability density function of the normal distribution is

$$f_X(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.69}$$

Based on (3.65) we form the likelihood function

$$f_{\boldsymbol{X}}(x_1,\ldots,x_n;\mu,\sigma)=\frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \tag{3.70}$$

and taking its logarithm we form the log-likelihood function:

$$L(x_1,\ldots,x_n;\mu,\sigma)=-\frac{n}{2}\ln(2\pi)-n\ln\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{3.71}$$

Taking the derivatives with respect of the unknown parameters μ and σ and equating them to 0 we have

$$\frac{\partial L}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-m)=0,\quad \frac{\partial L}{\partial\sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2=0 \tag{3.72}$$

and solving the system we obtain the final parameter estimates:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}, \qquad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = s_X \tag{3.73}$$

The results are precisely identical with those of the section 3.5.2, despite the fact that the two methods are fundamentally different. The application of the maximum likelihood method is more complex than that of the method of moments. The identical results found here are not the rule for all distribution functions. On the contrary, in most cases the two methods yield different results.

3.6 Hypothesis testing

A statistical hypothesis is a hypothesis related to the value of one or more parameters of a statistical model, which is described by a distribution function. The hypothesis testing is a process of establishing the validity of a hypothesis. The process has two possible outcomes: either the hypothesis is rejected or accepted (more precisely: not rejected).

In this section we present very briefly the related terminology and procedure, while in next chapters we will present some applications. The reader interested for a more detailed presentation of the theory should consult statistics books (e.g. Papoulis, 1990, p. 321-387, Freund *et al.*, 1988, p. 310-542), while for a presentation for hydrological applications is referenced to Hirsch *et al.* (1993, p. 17.11-29).

3.6.1 Terminology

- *Null hypothesis* is the hypothesis to be tested (symbolically H_0). Usually, it is a hypothesis of the form $\theta = \theta_0$, where θ is parameter related to a distribution function of a given variable and θ_0 is a numerical value.
- *Alternative hypothesis* is a second hypothesis that should not be true at the same time with the null hypothesis (symbolically H_1). It can be *simple*, such as $\theta = \theta_1$, or (more commonly) *composite*, such as $\theta \neq \theta_0$, $\theta > \theta_0$ or $\theta < \theta_0$.
- *Test statistic* is an appropriately chosen sample statistic, that is used for the test (symbolically Q).
- *Critical region* is an interval of real values. When the test statistic value lies in the critical region then the null hypothesis is rejected (symbolically R_c; see Fig. 3.2).
- *One-sided test* is a test where the alternative hypothesis is of the form $\theta > \theta_0$ or $\theta < \theta_0$. In this case the critical region is a half line of the form ($q > q_C$) or ($q < q_C$), respectively.
- *Two-side test* is a test where the alternative hypothesis if of the form $\theta \neq \theta_0$. In this case the critical region consists of two half lines ($q < q_L$) and ($q > q_U$).

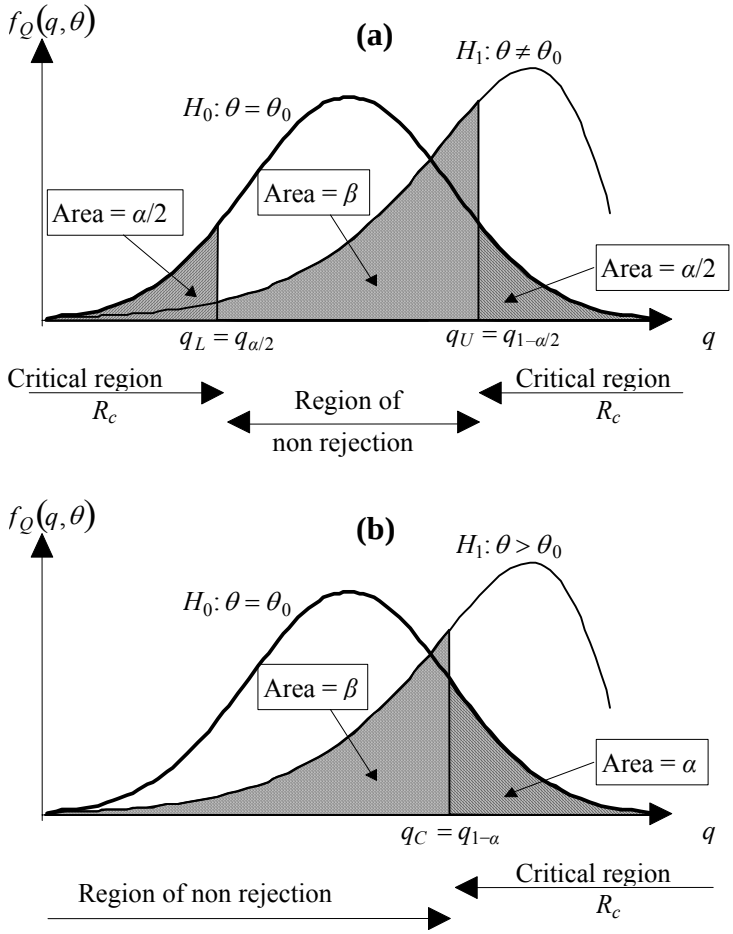

Fig. 3.2 Explanatory sketch of the concepts related to statistical testing: (a) a two-sided test, (b) an one-sided test.

- *Parametric* is a test whose hypotheses include specification of the population distribution function.
- *Non parametric* is a test valid for every population distribution function.
- *Decision rule* is the rule to reject or not the null hypothesis, expressed as:

$$\text{reject } H_0 \text{ if } q \in R_c$$

- *Type I error* is the rejection (based on the decision rule) of a true null hypothesis.
- *Type II error* is the acceptance (based on the decision rule) of a false null hypothesis
- *Significance level* of a test is the probability of type I error, namely the probability to reject a true null hypothesis. Symbolically

$$\alpha = P\{Q \in R_c \mid H_0\} \tag{3.74}$$

- *Power* of a test is the probability of rejecting a false null hypothesis. Symbolically,

$$p = 1 - \beta = P\{Q \in R_c \mid H_1\} \tag{3.75}$$

where β is the probability of type II error, that is

$$\beta = P\{Q \notin R_c \mid H_1\} \tag{3.76}$$

3.6.2 Testing procedure

The testing procedure consists of the following steps:

1. Formulation of the null hypothesis H_0 and of the alternative H_1.
2. Choice of the test statistic $Q = g(X_1, \ldots, X_n)$ and determination of the probability density function of the $f_Q(q; \theta)$.
3. Choice of the significance level α of the test and determination of the critical region R_c.
4. Calculation of the value $q = g(x_1, \ldots, x_n)$ of Q from the sample.
5. Application of the decision rule and rejection or acceptance of H_0.
6. Calculation of the power p of the test.

The last step is usually omitted in practice, due to its complexity. All remaining steps are clarified in the following section.

3.6.3 Demonstration of significance testing for the correlation coefficient

As an example of the above procedure we will present the significance testing of the correlation coefficient of two random variables X and Y, according to which we can decide whether or not the variables are linearly correlated.

If the variables are not linearly correlated then their correlation coefficient will be zero. Based on this observation, we proceed in the following steps of the statistical testing.

1. The null hypothesis H_0 is $\rho = 0$ and the alternative hypothesis H_1 is $\rho \neq 0$. As a consequence we will proceed with a two-sides test. (If we wanted to decide on the type of correlation, positive or negative, the alternative hypothesis would be formulated as $\rho > 0$ or $\rho < 0$, and we would perform an one-sided test).
2. We choose the test statistic as

$$Q = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)\sqrt{n-3} = Z\sqrt{n-3} \tag{3.77}$$

   where R is the sample correlation coefficient and Z is the auxiliary Fisher variable (section 3.3.6), which, if H_0 is true, has approximately a normal distribution with mean 0 and standard deviation $1/\sqrt{n-3}$. Consequently, Q has standard normal distribution $N(0, 1)$.
3. We choose a significance level $\alpha = 0.05$. If $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the normal distribution, then the corresponding critical region R_c is the $|q| > z_{1-\alpha/2}$ or $|q| > z_{0.975}$, or finally $|q| > 1.96$, given that

$$P(|Q| > z_{1-\alpha/2}) = P(Q < -z_{1-\alpha/2}) + P(Q > z_{1-\alpha/2})$$
$$= 2P(Q < z_{\alpha/2}) = 2\,\alpha / 2 = \alpha$$

(We recall that, because of the symmetry of the normal probability density function, $z_{1-u} = z_u$. In the case of the one-side test with alternative hypothesis $\rho > 0$, the critical region would be $q > z_{1-\alpha}$).

4. The numerical value of q is determined from the observed sample by the following equations

$$q = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)\sqrt{n-3}, \qquad r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.78}$$

5. The decision rule will be:

$$\text{reject } H_0 \text{ if } |q| > z_{1-\alpha/2}$$

and for $\alpha = 0.05$

$$\text{reject } H_0 \text{ if } |q| = \frac{1}{2}\left|\ln\left(\frac{1+r}{1-r}\right)\right|\sqrt{n-3} > 1.96$$

At the limit of this inequality, solving for r, we find the critical value r_c of the sample correlation coefficient, that determines the critical region R_c of the statistic R, that is,

$$r_c = \frac{e^{3.92/\sqrt{n-3}} - 1}{e^{3.92/\sqrt{n-3}} + 1} \tag{3.79}$$

A simple formula easy to remember that provides a very good approximation of (3.79) is:

$$r_c \approx \frac{2}{\sqrt{n}} \tag{3.80}$$

As a consequence, we can conduct the hypothesis testing in a more direct manner, by comparing the absolute value of r with the critical value r_c. If $|r| > r_c$ then we conclude that there is statistically significant correlation between the two variables.

3.6.4 A numerical example of significance testing for the correlation coefficient

From a 18-year-long record of measurements of concurrent annual rainfall and runoff at a catchment, we have calculated the correlation coefficient equal to 0.58. Is there a linear correlation between the annual rainfall and runoff?

We calculate the critical value r_c using one of (3.79) or (3.80). Here for comparison we use both:

$$r_c = \frac{e^{3.92/\sqrt{15}} - 1}{e^{3.92/\sqrt{15}} + 1} = 0.470, \qquad r_c = \frac{2}{\sqrt{18}} = 0.471$$

Indeed, the two equations give practically the same result. Since 0.58 > 0.47 we conclude that there is statistically significant correlation between the annual rainfall and runoff.

Acknowledgement I thank Simon Papalexiou for translating into English an earlier Greek version of this text.

References

Bobée, B., and R. Robitaille, Correction of bias in the estimation of the coefficient of skewness, *Water Resour. Res.*, 11(6), 851-854, 1975.

Freund, J. E., F. J. Williams, and B. M. Perles, *Elementary Business Statistics, The Modern Approach*, Prentice-Hall, 1988.

Hirsch, R. M., D. R. Helsel, T. A. Cohn, and E. J. Gilroy, Statistical analysis of hydrologic data, in *Handbook of Hydrology*, edited by D. R. Maidment, McGraw-Hill, 1993.

Kendall, M. G., and A. Stuart, *The Advanced Theory of Statistics*, Vol. 2, Inference and relationship, Third edition, Charles Griffin & Co., London, 1973.

Kite, G. W., *Frequency and Risk Analyses in Hydrology*, Water Resources Publications, Littleton, Colorado, 1988.

Kirby, W., Algebraic boundness of sample statistics, *Water Resour. Res.*, 10(2), 220-222, 1974.

Kjeldsen, T.R., J.C. Smithers and R.E. Schulze, Regional flood frequency analysis province, South Africa, using the index-flood method, *J. Hydrol.*, 255, 194–211, 2002.

Kroll, C.N., and R.M. Vogel, Probability distribution of low streamflow series in the United States, *J. Hydrol. Eng.*, 7, 137–146, 2002.

Landwehr, J.M., N.C. Matalas and J.R. Wallis, Probability weighted moments with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resour. Res.*, 15, 1055–1064, 1979.

Lim, Y.H., and L.M. Lye, Regional flood estimation for ungauged basins in Sarawak, Malaysia, *Hydrol. Sci. J.*, 48, 79–94, 2003.

Papoulis, A., *Probability and Statistics*, Prentice-Hall, New Jersey, 1990.

Singh, V. P., and A. K. Rajagopal, A new method of parameter estimation for hydrologic frequency analysis, *Hydrological Science and Technology*, 2(3) 33-44, 1986.

Wallis, J. R., N. C. Matalas, and J. R. Slack, Just a moment!, *Water Resour. Res.*, 10(2), 211-219, 1974.

Yevjevich, V., *Probability and Statistics in Hydrology*, Water Resources Publications, Fort Collins, Colorado, 1972.

Zaidman, M.D., V. Keller, A.R. Young and D. Cadman, Flow-duration-frequency behaviour of British rivers based on annual minima data, *J. Hydrol.*, 277, 195–213, 2003.