Government of India & Government of The Netherlands

Technical Assistance
HYDROLOGY PROJECT

DHV CONSULTANTS &
DELFT HYDRAULICS
with HALCROW, TAHAL,
CES, ORG & JPS

VOLUME 2
SAMPLING PRINCIPLES

DESIGN MANUAL

Table of Contents

1 INTRODUCTION

The objective of this volume is to present a number of basic principles in relation with sampling of hydrological and hydro-meteorological variables in general. These principles deal with units to be applied to quantify the dimension of variable and with errors made in sampling a variable by using particular equipment at discrete moments in time at fixed locations in space, during a certain period.

These variables are being observed because one wants to be informed about their temporal and spatial characteristics for planning, design, operation and research purposes. The characteristics of the variables are generally expressed by statistical parameters describing the frequency distribution of the entire population of the variable or of features of the population like its minimum and maximum values. In view of the variation in time and space also the temporal and spatial correlation structure is of interest.

Hydrological and hydro-meteorological processes are continuous in time and space. This imposes a number of limitations on the quality with which statistical parameters can be determined by sampling, since:

1. The spatial continuous process is monitored at discrete locations in space
2. The temporal continuous process is monitored at discrete moments in time
3. The process is monitored during a limited period of time, and
4. The equipment with which the process is monitored at a fixed location at discrete moments in time during a certain period has a limited accuracy.

Due to these limitations errors are being made in the estimation of the statistical parameters. These errors differ from one parameter to another. The errors made due to point sampling in space are a function of the applied network density in relation to the spatial variation of the process. The errors made due to discrete sampling in time are a function of the sampling interval in relation to the temporal variation of the process. The errors made due to the limited duration of the monitoring period are a function of the representativeness of the sampled period relative to the population and of the correlation between successive observations. Another source of error is originating from the equipment being used for monitoring the variable at a fixed location at a particular moment in time.

Another source of error stems from the fact that the monitored processes are in a statistical sense generally inhomogeneous. Beside the obvious spatial inhomogeneity of climatic variables, climate change and variation of the basin's drainage characteristics make all hydrological and hydro-meteorological variables inhomogeneous with time. This implies that the statistical parameters are in principle not only a function of space co-ordinates but also of time.

Therefore, the monitoring system has to be designed in such a manner that with the data produced by the network an acceptable estimation can be given of the relevant statistical parameters or of their behaviour in time and/or space. To quantify 'an acceptable estimation' of a statistical parameter the uncertainty in the statistical estimate has to be known. This requires knowledge of some basic statistical principles and of the various sources of sampling errors involved. In this volume a number of common aspects of sampling of hydro-meteorological and hydrological quantity and quality variables are discussed, including:

- Units with which variables are quantified and unit conversions
- Sample statistics
- Measurement errors
- Sampling errors due to time discretisation, and
- Sampling errors due to spatial discretisation.

This information is used to arrive at general principles of monitoring network design, for which proper information is required with respect to the monitoring objectives, and of the physical characteristics of the monitored system.

2 UNITS

The use of standard methods is an important objective in the operation of the Hydrological Information System (HIS). Standard methods require the use of a coherent system of units with which variables and parameters are quantified. This chapter deals with the system of units used for the measurement of hydrological and hydro-meteorological quantities.

2.1 DEFINITIONS

A measurable property of an object, like the object's length or mass, is called a **quantity**. The object itself is not a quantity. The physical property described by the quantity is its **dimension**. In a measurement a quantity is expressed as a **number** times a reference quantity, the **unit**, i.e. the scale with which dimensions are measured. When quantities are being compared their **dimensions and units** should be the same. In any unit system some **base quantities** are (arbitrarily) defined with their associated **base units**. Any other quantity can be expressed as a product of base quantities and so can their units be derived from the base units without numerical factors. The latter property leads to a **coherent system** of units.

India officially adopted the International System of Units in 1972. Henceforth, the units to be used at all levels in the HIS should be in accordance with this unit system. It is abbreviated as SI, from the French **Le Système International d'Unités**. In this section the following unit-related topics are discussed:

- the base units of the International System of Units SI
- prefixes to units as allowed in SI, and
- summary of relevant derived units.

2.2 BASE UNITS OF SI

SI selected as base units the quantities displayed in Table 2.1.

Quantity	Symbol	SI Unit		Dimension
		Name	Symbol	
Time	t	second	s	T
Length	l	meter	m	L
Mass	m	kilogram	kg	M
Amount of substance	n	mole	mol	
Thermodynamic temperature	T	kelvin	K	Θ
Electric current	I	ampere	A	
Luminous intensity	I	candela	cd	

Table 2.1: SI Base Units

2.3 PREFIXES TO SI UNITS

Measures of variables or parameters may be several orders of magnitude larger or smaller than the base units. In order to avoid the use of powers of 10 prefixes for the units are in use. The prefixes adopted in SI are listed in Table 2.2.

Factor	Prefix	Symbol	Factor	Prefix	Symbol
10^{18}	exa-	E	10^{-1}	deci-	d
10^{15}	peta-	P	10^{-2}	centi-	c
10^{12}	tera-	T	10^{-3}	milli-	m
10^{9}	giga-	G	10^{-6}	micro-	μ
10^{6}	mega-	M	10^{-9}	nano-	n
10^{3}	kilo-	k	10^{-12}	pico-	p
10^{2}	hecto-	h	10^{-15}	femto-	f
10^{1}	deka-	da	10^{-18}	atto-	a

Table 2.2: SI Prefixes

In extension to the base units in Table 2.3 a summary is given of units for quantities derived from the base units which are relevant for hydrology. Though the use of SI units in HIS is in principle mandatory, a few non-SI units are accepted as well; these are shown in the last two columns of Table 2.3. Note that the units for the quantities typical to hydro-meteorology and hydrology are presented in the Chapter 1 of Volume 3, Design Manual, Hydro-meteorology along with a definition of those quantities.

Quantity	Symbol	SI Unit		Dimension	Also accepted Units	
		Name	Symbol		Name	Symbol
Geometric						
Area	A		m^2	L^2	hectare	ha
Volume	V		m^3	L^3	litre	L
Angle	$\alpha, \beta, \ldots$	radian	rad	1	degree	^{o}
Kinematic						
Time (base unit)	t		s	T	minute, hour, day, year	min, h, day, yr
Velocity	u, v, w, c		$m.s^{-1}$	LT^{-1}		
Acceleration	a		$m.s^{-2}$	LT^{-2}		
Angular velocity	ω		$rad.s^{-1}$	T^{-1}	revolutions/s	$rev.s^{-1}$
Angular acceleration	α		$rad.s^{-2}$	T^{-2}		
Frequency	f	Herz	Hz	T^{-1}		
Diffusivity	D, K		$m.s^{-2}$	LT^{-2}		
Kinematic viscosity	ν		$m.s^{-2}$	LT^{-2}		
Discharge rate	Q		$m.s^{-3}$	LT^{-3}		
Dynamic						
Mass density	ρ		$kg.m^{-3}$	ML^{-3}		
Force (weight)	F	Newton	N	MLT^{-2}		
Pressure	p	Pascal	Pa	$ML^{-1}T^{-2}$	Millibar	mb
Surface tension	σ		$N.m^{-1}$	MT^{-2}		
Momentum	M		N.s	MLT^{-1}		
Energy (work)	E, W, U	Joule	J	ML^2T^{-2}		
Power	P	Watt	W	ML^2T^{-3}		
Energy flux	q		$W.m^{-2}$	MT^{-3}		
Thermal						
Temperature	T		K	Θ	degree Celcius	^{o}C
Latent heat	L, λ		$J.kg^{-1}$	L^2T^{-2}		
Heat capacity	c		$J.kg^{-1}.K^{-1}$	$L^2T^{-2}\,\Theta^{-1}$		

Table 2.3: Derived SI and other accepted units

2.5 UNIT CONVERSIONS AND CONVERSION FACTORS

In this section the following topics are dealt with:

- conversion of units of one system into another, and
- conversion factors to be applied to transform data to SI.

Unit conversion

The procedure for converting a unit to another one with the same dimension is simply by replacing the original unit by a value expressed in the new unit of **exactly the same size**. This is illustrated in the following examples.

EXAMPLE 2.1

Wind run data are available in km/day; these values have to be transformed into m/s. Note that there are 1,000 m in a km and 86,400 s in a day, hence, to convert km/day into m/s, the following steps have to be taken:

$$1\frac{\text{km}}{\text{day}} = 1\frac{1000\text{m}}{86{,}400\text{s}} = \frac{1}{86.4}\frac{\text{m}}{\text{s}}$$

EXAMPLE 2.2

The solar constant is 2 ly/min (= langley/minute). Required is the solar constant expressed in W/m^2. Note that 1 langley = 1 cal/cm^2, 1 min = 60 s, 1 cal = 4.1868 J, 1 W = 1 J/s and 1 $cm^2 = 10^{-4} m^2$. The conversion is carried out as follows:

$$2\frac{\text{ly}}{\text{min}} = 2\frac{\text{cal}}{\text{cm}^2}\frac{1}{60\text{s}} = 2\frac{4.1868\text{J}}{10^{-4}\text{m}^2}\frac{1}{60\text{s}} = 2\frac{4.1868}{60.10^{-4}}\frac{\text{J}}{\text{s}}\frac{1}{\text{m}^2} = 1395.6\frac{\text{W}}{\text{m}^2}$$

Conversion factors to SI Units

In the past several unit systems have been applied. In India particularly the British system was used. Historical data may be available in those and other units. Therefore, in Table 2.4 a summary of a variety of units is given with the conversion factor to be applied to transform the unit into SI.

Unit	Symbol	Conversion to SI	Unit	Symbol	Conversion to SI
Geometric			**Dynamic**		
Inch	in	0.0254 m	Gram	g	1x10-3 kg
Foot	ft	0.3048 m	Slug	slug	14.5939 kg
Yard	yd	0.9144 m	Pound	lb	0.45359237 kg
Fathom	fath	1.8288 m	Dyne	dyn	1x10-5 N
Furlong	fur	201.168 m	Bar	b	105 Pa
Statute mile	mi	1609.344 m	Millibar	mb	102 Pa =1 hPa
Acre	ac	4046.86 m2	Poise	P	0.1 Pa.s
Hectare	ha	1x104 m2	Cm of water	cm H2O	101.99 Pa
Litre	L	1x10-3 m3	Mm of mercury	mm Hg	133.322 Pa
Gallon (UK)	(UK)gal	4.54609x10-3 m3	Erg	erg	1x10-7 J
Bushel (UK)	bu	36.3687x10-3 m3	Horsepower	hp	745.69987 W
Gallon (USA)	(US)gal	3.78541x10-3 m3	Voltampère	VA	W
Degree of angle	o	π/180 rad	Kilowatthour	kWh	3.6x106 J
			Thermodynamic		
Kinematic	min	60 s	Degree Celcius	^{o}C	+ 273.15 K
Minute	hr	3600 s	Degree Fahrenheit	^{o}F	+459.67/1.8 K
Hour	day	86400 s	British thermal unit	Btu	1055.06 J
Day	yr	31,557,600 s	Calorie (Int. Table)	cal	4.1868 J
Year	rev	2π rad			
Revolution					

Table 2.4: Conversion factors to SI

3 BASIC STATISTICS

This chapter deals with statistical descriptors of variables. Variables, whose values are entirely or in part determined by chance, are called **random variables**, and their behaviour can be described by probability distributions. Strictly speaking, to describe the behaviour of a random variable completely, full knowledge about its probability distribution is required. Practically, the dominant features of a distribution function can be described with a limited number of parameters quantifying the first few moments of the distribution function, like e.g. the mean, variance, covariance and skewness. This chapter deals with the following:

- a selection of relevant descriptors of distributions,
- procedures to estimate these parameters from samples with their uncertainty expressed by their sampling distribution, and
- the effect of serial correlation on the sampling distributions.

3.1 DISTRIBUTION FUNCTIONS AND DESCRIPTORS

The following descriptors of random variables are discussed:

- distribution functions
- mean, variance, standard deviation and coefficient of variation
- covariance and correlation coefficient
- skewness, and
- quantiles

Distribution functions

Let Y be a continuous random variable. Its **cumulative distribution function** or cdf $F_Y(y_p)$ expresses the probability that Y will be less or equal y_p:

$$F_Y(y_p) = \text{Prob}\,[Y \le y_p], \quad \text{with } 0 \le F_Y(y_p) \le 1 \text{ for all possible } y_p \tag{3.1}$$

Its derivative is the pdf or **probability density function** $f_Y(y)$; hence the relation between $F_Y(y)$ and $f_Y(y)$ becomes:

$$f_Y(y) = \frac{dF_Y(y)}{dy} \quad \text{and} \quad F_Y(y_p) = \int_{-\infty}^{y_p} f_Y(y)dy \qquad \text{with} \qquad \int_{-\infty}^{+\infty} f_Y(y)dy = 1 \tag{3.2}$$

Their relationship is shown in Figure 3.1. The following descriptors for $F_Y(y)$ and $f_Y(y)$ are now defined.

Mean, variance, standard deviation and coefficient of variation

The **mean** or central tendency μ_y, is the first moment of the pdf about the origin and reads:

$$\mu_Y = E[Y] = \int_{-\infty}^{+\infty} y f_Y(y)dy \tag{3.3}$$

The **variance** σ_Y^2 is the second central moment and gives the dispersion about the mean:

$$\sigma_Y^2 = E[(Y - \mu_Y)^2] = \int_{-\infty}^{+\infty} (y - \mu_Y)^2 f_Y(y)dy \tag{3.4}$$

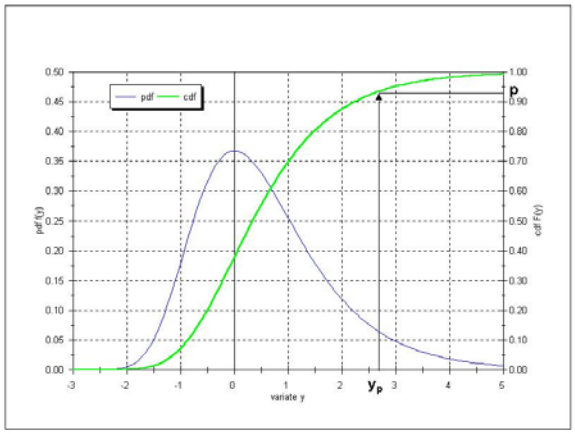

Figure 3.1:
Probability density function and cumulative distribution function.

The **standard deviation** σ_Y is the root of the variance and is introduced to have a descriptor for the dispersion about the mean in the same units as the quantity itself. The ratio of the standard deviation and the mean is called the coefficient of variation. When expressed as a percentage it reads:

$$CV_Y = 100\frac{\sigma_Y}{\mu_Y} \quad (\%) \tag{3.5}$$

Covariance, cross- and auto-covariance functions

A measure for the linear association between two variables X and Y is the **covariance** C_{XY}, which is defined by (3.6). It is seen to be the expected or mean value of the product of deviations from the respective mean values:

$$C_{XY} = E[(X - \mu_X)(Y - \mu_Y)] \tag{3.6}$$

The linear association between the elements of two time series X(t) and Y(t) lagged time τ apart is the lag τ cross-covariance, see Figure 3.2. This covariance, expressed as a function of τ, is the **cross-covariance function** $C_{XY}(\tau)$ and is defined by:

$$C_{XY}(\tau) = E[(X(t) - \mu_X)(Y(t+\tau) - \mu_Y)] \tag{3.6a}$$

Similarly, the **auto-covariance function** $C_{YY}(\tau)$ describes the linear association between elements of a single time series Y(t) spaced time τ apart (see also Figure 3.2) :

$$C_{YY}(\tau) = E[(Y(t) - \mu_Y)(Y(t+\tau) - \mu_Y)] \tag{3.6b}$$

Note that for $\tau = 0$ equation (3.6b) is equivalent to (3.4), hence: $C_{YY}(0) = \sigma_Y^2$.

Correlation coefficient, cross- and auto-correlation function

When the covariance is scaled by the standard deviations of X and Y the **correlation coefficient** ρ_{XY} is obtained, which is a dimensionless measure for the degree of linear association between X and Y:

$$\rho_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y} \quad \text{with} \quad -1 \le \rho_{XY} \le 1 \tag{3.7}$$

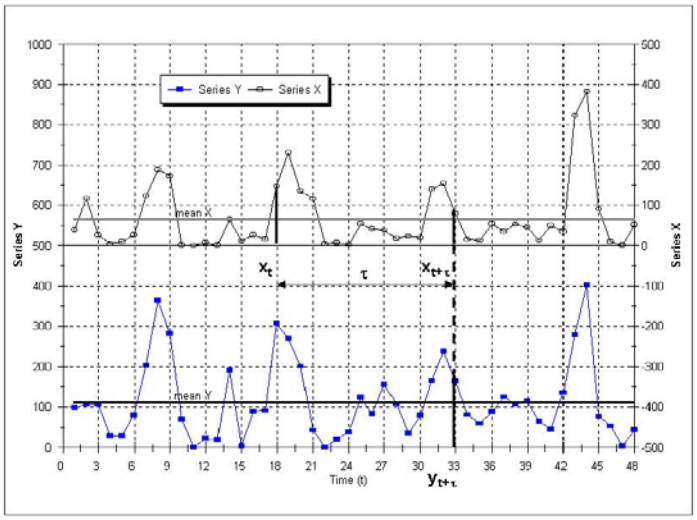

Figure 3.2:
Definition of cross- and autocovariance

Note that positive values of ρ_{XY} are obtained if $(X - \mu_X)$ and $(Y - \mu_Y)$ have the same sign, whereas negative values of ρ_{XY} follow when $(X - \mu_X)$ and $(Y - \mu_Y)$ have opposite sign. For time series, similar to the cross- and auto-covariance one defines the lag τ **cross-correlation** and **auto-correlation functions,** respectively:

$$\rho_{XY}(\tau) = \frac{C_{XY}(\tau)}{\sigma_X \sigma_Y} \qquad (3.7a)$$

$$\rho_{YY}(\tau) = \frac{C_{YY}(\tau)}{\sigma_Y^2} \qquad (3.7b)$$

The graphical displays of these functions are called cross-correlogram and auto-correlogram, respectively. Examples of these correlograms for monthly rainfall series are shown in Figures 3.3 and 3.4.

Note that since CYY(0) = σY2 , for the auto-correlogram at lag τ = 0 it follows: ρyy(0) = 1. Generally, the cross-correlogram at lag τ = 0 is ρyy(0) ≠ 1 unless series X and Y are identical.

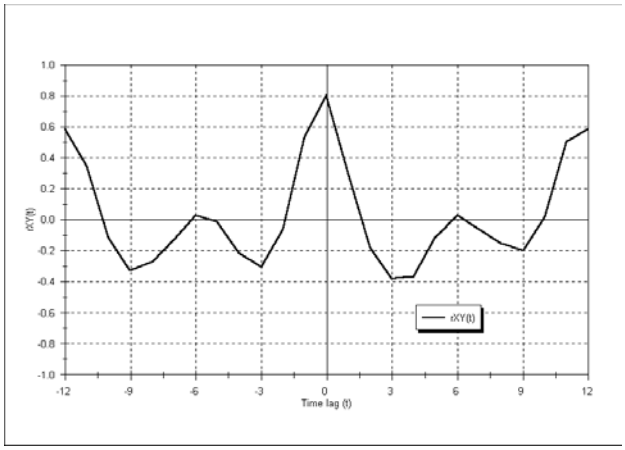

Figure 3.3: Crosscorrelogram

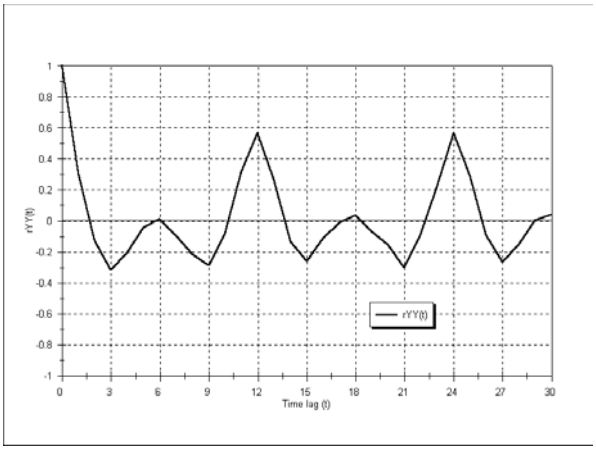

Figure 3.4: Autocorrelogram

Skewness

Distributions like the normal distribution are symmetrical about the mean. Many distribution functions, however, are skewed. A measure for this is the **skewness** γ_Y, which is defined as the third moment about the mean, scaled by the standard deviation:

$$\gamma_Y = \frac{E[(Y-\mu_Y)^3]}{\sigma_Y^3} = \frac{1}{\sigma_Y^3}\int_{-\infty}^{+\infty}(y-\mu_Y)^3 f_Y(y)dy \tag{3.8}$$

Hence, distributions with longer tails towards the right are positively skewed and vice versa.

Quantiles

The p^{th} **quantile** of the variable Y is the value y_p such that:

$$F_Y(y_p) = Prob[Y \le y_p] = \int_{-\infty}^{y_p} f_Y(y)dy = p \tag{3.9}$$

The quantile y_p is shown in Figure 3.1. Note that the quantile subscript 'p' indicates the probability of **non**-exceedance attached to it. Some commonly used quantiles are the **median** $y_{0.50}$ and the **lower** and **upper quartiles** $y_{0.25}$ and $y_{0.75}$, respectively.

3.2 PARAMETER ESTIMATION AND ESTIMATION ERROR

The distribution parameters as discussed in the previous sub-section are generally unknown as full information about the entire population/process of which the parameters are descriptors is not available. Therefore, these parameters can only be estimated from samples (measurements) of the process. Since the samples represent only a small portion of the total population, estimates for a particular parameter vary from one sample to another. The estimates are therefore random variables or **statistics** themselves with a frequency distribution, called sampling distribution. Parameters may be estimated in different ways, like by the method of moments, maximum likelihood method or mixed moment-maximum likelihood methods. These procedures are discussed in the Manual on Data Processing. To compare the quality of different estimators of a parameter, some measure of accuracy is required. The following measures are in use:

- mean square error and root mean square error
- error variance and standard error, and
- bias

The estimates used in this manual for the various parameters are based on unbiased estimators. This characteristic and other features of estimates of parameters are discussed in this subsection.

Mean square error

A measure for the quality of an estimator is the **mean square error,** mse. It is defined by:

$$mse = E[(\phi - \Phi)^2] \tag{3.10}$$

where ϕ is an estimator for Φ.

Hence, the mse is the average of the squared differences between the sample value and the **true** value. Equation (3.10) can be expanded to the following expression:

$$mse = E[(\phi - E[\phi])^2] + E[(E[\phi] - \Phi)^2]$$

$$\text{Since : } \quad E[(\phi - E[\phi])^2] = \sigma_\phi^2 \quad \text{and} \quad E[(E[\phi] - \Phi)^2] = b_\phi^2 \qquad \text{it follows that :}$$

$$mse = \sigma_\phi^2 + b_\phi^2 \qquad (3.11)$$

The mean square error is seen to be the sum of two parts:

- the first term is the **variance** of ϕ, i.e. the average of the squared differences between the sample value and the **expected** mean value of ϕ based on the sample values, which represents the **random** portion of the error, and
- the second term of (3.11) is the square of the **bias** of ϕ, describing the systematic deviation of expected mean value of ϕ from its true value Φ, i.e. the **systematic** portion of the error.

Note that if the bias in ϕ is zero, then mse = σ_ϕ^2. Hence, for **unbiased** estimators, i.e. if systematic errors are absent, mean square error and variance are equivalent.

Root mean square error

Instead of using the mse it is customary to work with its square root to arrive at an error measure, which is expressed in the same units as Φ, leading to the **root mean square (rms) error**:

$$rmse = \sqrt{E(\phi - \Phi)^2} = \sqrt{\sigma_\phi^2 + b_\phi^2} \qquad (3.11a)$$

Standard error

When discussing the frequency distribution of statistics like of the mean or the standard deviation, for the standard deviation σ_ϕ the term **standard error** is used, e.g. standard error of the mean and standard error of the standard deviation, etc.

$$\sigma_\phi = \sqrt{E[(\phi - E[\phi])^2}} \qquad (3.11b)$$

In Table 3.1, a summary of unbiased estimators for the distribution parameters is given, together with their standard error. With respect to the latter it is assumed that the sample elements y_i, i = 1, N are **serially uncorrelated**. If the sample elements are serially correlated a so-called **effective number of data** N_{eff} has to be applied in the expressions for the standard error in Table 3.1 (see also Section 3.4).

Note

From equations for the standard error, as presented in Table 3.1, it is observed that the standard error is inversely proportional with $\sqrt{(N)}$. This implies that the standard error reduces with increasing sample size. This is an important feature to reduce random errors in measurements as will be shown in the next chapter.

Parameter	Estimator	Standard error	Remarks
Mean (3.12)	$m_Y = \frac{1}{N}\sum_{i=1}^{N} y_i$	$\sigma_{m_Y} = \frac{\sigma_Y}{\sqrt{N}}$	The sampling distribution of m_Y is very nearly normal for N>30, even when the population is non-normal. In practice σ_Y is not known and is estimated by s_Y. Then the sampling distribution of m_Y has a Student distribution, with N-1 degrees of freedom
Variance (3.13)	$s_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - m_Y)^2$	$\sigma_{s_Y^2} = \sqrt{\frac{2}{N}}\sigma_Y^2$	Expression applies if the distribution of Y is approximately normal. The sampling distribution of s_Y^2 is nearly normal for N>100. For small N the distribution of s_Y^2 is chi-square (χ^2), with N-1 degrees of freedom
Standard deviation (3.14)	$s_Y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - m_Y)^2}$	$\sigma_{s_Y} = \frac{\sigma_Y}{\sqrt{2N}}$	The remarks made for the standard error of the variance apply here as well
Coefficient of variation (3.15)	$\hat{CV}_Y = \frac{s_Y}{m_Y}$ Sample value of CV_Y limited to: $CV_Y < \sqrt{(N-1)}$	$\sigma_{\hat{CV}} = \frac{\frac{\sigma_Y}{\mu_Y}}{\sqrt{2N}}\sqrt{1+2\left(\frac{\sigma_Y}{\mu_Y}\right)^2}$	This result holds if Y being normally or nearly normally distributed and N>100.
Covariance (3.16)	$\hat{C}_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)(y_i - m_Y)$		
Correlation coefficient (3.17)	$r_{XY} = \frac{C_{XY}}{s_X s_Y}$	$\sigma_W = \frac{1}{\sqrt{N-3}}$ where $W = \frac{1}{2}\ln\left(\frac{1+r_{XY}}{1-r_{XY}}\right)$	Rather then the standard error of r_{XY} the standard error of the transformed variable W is considered. The quantity W is approximately normally distributed for N>25.
Lag one auto-correlation coefficient (3.18)	$r_{YY}(1) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}(y_i - m_Y)(y_{i+1} - m_Y)}{s_Y^2}$	as for r_{XY} above	
Skewness (3.19)	$g_Y = \frac{\frac{N}{(N-1)(N-2)}\sum_{i=1}^{N}(y_i - m_Y)^3}{s_Y^3}$ Skewness limited to: $g_Y < \frac{N-2}{\sqrt{N-1}}$	$\sigma_{g_Y} = \sqrt{\frac{6}{N}}$	A reasonably reliable estimate of the skewness requires a large sample size. Standard error applies if Y is normally distributed.
Quantiles (3.20)	1. first rank the sample values in ascending order: $y_{(i)} < y_{(i+1)}$ 2. next assign to each ranked value a non-exceedance probability i/(N+1) 3. then interpolate between the probabilities to arrive at the quantile value $\hat{y}_p$ of the required non-exceedance level	$\sigma_{\hat{y}_p} = \frac{1}{f_Y(y_p)}\sqrt{\frac{p(1-p)}{N}}$ $\sigma_{\hat{y}_p} = \frac{\beta}{\sqrt{N}}\sigma_Y$	The denominator is derived from the pdf of Y. If Y is normally distributed then the standard error of the quantile is determined by the second expression. The coefficient β depends on the non-exceedance probability p. For various values of p the value of β can be obtained from Table 3.2.

Table 3.1: Estimators of sample parameters with their standard error

P	0.5	0.4/ 0.6	0.3/ 0.7	0.25/0.75	0.2/ 0.8	0.15/0.85	0.1/0.9	0.05/0.95
β	1.253	1.268	1.318	1.362	1.428	1.531	1.709	2.114

Table 3.2: $\beta(p)$ for computation of $\sigma_{\hat{y}_p}$ of quantiles if Y is normally distributed

3.3 CONFIDENCE LIMITS FOR MEAN AND VARIANCE

The moment and quantile statistics presented in Table 3.1 are asymptotically normally distributed. This implies that for large sample sizes N the estimate and the standard error fully describe the probability distribution of the statistic. For small sample sizes the sampling distributions, deviate from normality and sampling distributions like the Chi-square distribution, and the Student-t distribution become more appropriate. Reference is made to Volume 2, Reference Manual, Sampling Principles, for a description of these 3 distributions.

In this section use is made of the normal, Student-t and Chi-square distributions to quantify the uncertainty in the sample mean and the sample variance. The uncertainty is expressed by the confidence limits indicating the range in which the true value of the parameter is likely to lie with a stated probability.

Confidence limits of the mean

Given that a sample m_Y is available, the confidence limits for the mean of a process with known variance σ_Y^2 are given by:

$$\mathrm{Prob}\left\{(m_Y - z_{1-\alpha/2}\frac{\sigma_Y}{\sqrt{N}}) \le \mu_Y < (m_Y + z_{1-\alpha/2}\frac{\sigma_Y}{\sqrt{N}})\right\} = 1-\alpha \qquad (3.21)$$

The confidence statement expressed by equation (3.21) reads that: **'the true mean μ_Y falls within the indicated interval with a confidence of 100(1-α) percent'.** The quantity 100(1-α) is the **confidence level**, the interval for μ_Y is called the **confidence interval** enclosed by the **lower confidence limit** ($m_Y - z_{1-\alpha/2}\,\sigma_Y/\sqrt{N}$) and the **upper confidence limit** ($m_Y + z_{1-\alpha/2}\,\sigma_Y/\sqrt{N}$). The values for $z_{1-\alpha/2}$ are taken from tables of the standard normal distribution. E.g. if 100(1-α) = 95% then $z_{1-\alpha/2}$ = 1.96.

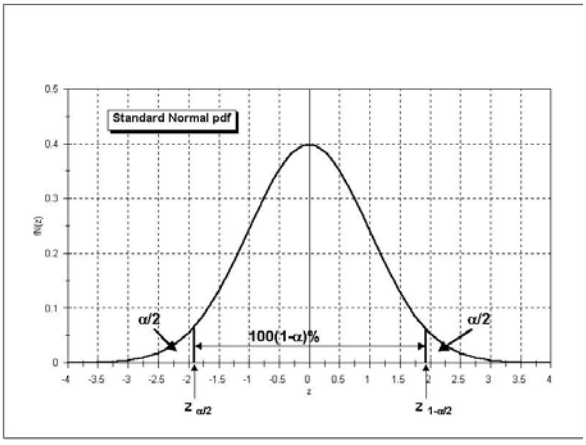

Figure 3.5:
Confidence limits of mean

Note that in the above procedure it has been assumed that σ_Y is known. Generally, this is not the case and it has to be estimated by s_Y according to equation (3.13). Then instead of the normal distribution the Student-t distribution has to be applied and the percentage points $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are replaced by $t_{n,\alpha/2}$ and $t_{n,1-\alpha/2}$., where n = N-1 is the number of degrees of freedom. The confidence limits then read:

$$\mathrm{Prob}\left\{(m_y - t_{n,1-\alpha/2}\frac{s_Y}{\sqrt{N}}) \le \mu_Y < (m_Y + t_{n,1-\alpha/2}\frac{s_Y}{\sqrt{N}})\right\} = 1-\alpha \qquad (3.22)$$

The values for the percentage point $t_{n,1-\alpha/2}$ can be obtained from statistical tables. For the confidence level 100(1-α) = 95 % percentage point $t_{n,1-\alpha/2}$ is presented in Table 3.3.

Confidence limits of the variance

Given an estimate of the sample variance computed by (3.13) the true variance σ_Y^2 will be contained within the following confidence interval with a probability of 100(1-α) %:

$$\text{Prob}\left\{\frac{ns_Y^2}{\chi^2_{n,1-\alpha/2}} \le \sigma_Y^2 < \frac{ns_Y^2}{\chi^2_{n,\alpha/2}}\right\} = 1-\alpha \quad \text{with} \quad n = N-1 \tag{3.23}$$

The values for $\chi^2_{n,\alpha/2}$ and $\chi^2_{n,1-\alpha/2}$ are read from the tables of the Chi-square distribution for given α and n. The Chi-square values defining the confidence intervals at a 100(1-α) = 95 % confidence level are presented in Table 3.3 as a function of the number of degrees of freedom n.

n	$t_{n,1-\alpha/2}$	$\chi^2_{n,\alpha/2}$	$\chi^2_{n,1-\alpha/2}$	n	$t_{n,1-\alpha/2}$	$\chi^2_{n,\alpha/2}$	$\chi^2_{n,1-\alpha/2}$
1	12.706	0.00098	5.02	21	2.080	10.28	35.48
2	4.303	0.0506	7.38	22	2.074	10.98	36.78
3	3.182	0.216	9.35	23	2.069	11.69	38.08
4	2.776	0.484	11.14	24	2.064	12.40	39.36
5	2.571	0.831	12.83	25	2.060	13.12	40.65
6	2.447	1.24	14.45	26	2.056	13.84	41.92
7	2.365	1.69	16.01	27	2.052	14.57	43.19
8	2.306	2.18	17.53	28	2.048	15.31	44.46
9	2.262	2.70	19.02	29	2.045	16.05	45.72
10	2.228	3.25	20.48	30	2.042	16.79	46.98
11	2.201	3.82	21.92	40	2.021	23.43	59.34
12	2.179	4.40	23.34	60	2.000	40.48	83.30
13	2.160	5.01	24.74	100	1.984	74.2	129.6
14	2.145	5.63	26.12	120	1.980	91.6	152.2
15	2.131	6.26	27.49				
16	2.120	6.91	28.85				
17	2.110	7.56	30.19				
18	2.101	8.23	31.53				
19	2.093	8.91	32.85				
20	2.086	9.59	34.17				

Table 3.3: Percentage points for the Student and Chi-square distributions at 95% confidence level for n = 1, 120 degrees of freedom

3.4 EFFECT OF SERIAL CORRELATION ON CONFIDENCE INTERVALS

Effect of correlation on confidence interval of the mean

In the derivation of the confidence interval for the mean, equation (3.22), it has been assumed that the sample series elements are independent, i.e. uncorrelated. In case persistency (i.e. non-zero correlation) is present in the data series the series size N has to be replaced with the effective number of data N_{eff}. Since persistence carries over information from one series element to another it reduces the information content of a sample series, hence N_{eff}<N. The value of N_{eff} is a function of the correlation structure of the sample:

$$N_{eff}(m) = \frac{N}{1+2\sum_{i=1}^{N-1}(1-\frac{i}{N})r_{YY}(i)} \approx N\frac{1-r_{YY}(1)}{1+r_{YY}(1)} \qquad \text{for}: \quad r_{YY}(1) > r^* \qquad \text{where}: \quad r^* = \frac{2}{\sqrt{N}} \tag{3.24}$$

The latter approximation in (3.24) holds if the correlation function can be described by its first serial correlation coefficient $r_{YY}(1)$ (which is true for a first order auto-regressive process).The condition mentioned on the right hand side of (3.24) is a significance test on zero correlation. If $r_{YY}(1)$ exceeds r* then persistence is apparent. The first serial correlation coefficient is estimated from (3.18), see Table 3.1.

The confidence interval to contain μ_Y with 100(1-α)% probability is now defined by equation (3.22) with N replaced by N_{eff} and the number of degrees of freedom given by $n = N_{eff} - 1$. An application of the above procedure is presented in Example 3.1 at the end of this section.

Effect of correlation on confidence interval of the variance or standard deviation

Persistence in the data also affects the sampling distribution of the sample variance or standard deviation. The effective number of data, however, is computed different from the way it is computed for the mean. Again, if the correlation function is described by its lag one auto-correlation coefficient the following approximation applies:

$$N_{eff}(s) \approx N \frac{1-[r_{YY}(1)]^2}{1+[r_{YY}(1)]^2} \qquad (3.25)$$

The 100(1-α)% confidence interval for σ_Y^2 follows from equation (3.23) with $n = N_{eff} - 1$.

EXAMPLE 3.1

Consider a series of 50 years of annual rainfall P with a mean value 800 mm and a coefficient of variation of 25%. The correlation coefficient $r_P(1) = 0.35$. From this it follows N = 50, m_P = 800 mm and s_P = 200 mm. To assess the uncertainties (95% confidence interval) in the estimate for the mean and the standard deviation the following steps are taken.

First the significance of $r_P(1)$ is tested. From (3.24) for r* one gets:

$r^* = 2/\sqrt{(50)} = 0.28$.

Since $r_P(1) = 0.35$ it follows $r_P(1) > r^*$ so $r_P(1)$ is significant at a 95% confidence level. It implies that N has to be replaced by with the effective number of data according to equations (3.24) and (3.25) respectively:

$$N_{eff}(m) = 50\frac{1-0.35}{1+0.35} = 24 \quad \text{and} \quad N_{eff}(s) = 50\frac{1-0.35^2}{1+0.35^2} = 39$$

The confidence limits for the mean are estimated from (3.22) for which $t_{n,1-\alpha/2}$ and $s_P/\sqrt{N_{eff}}$ have to be determined. With $N_{eff}(m) = 24$ and n = 23 one obtains from Table 3.3:

$T_{n,1-\alpha/2} = t_{23,0.975} = 2.07$ and $s_P/\sqrt{N_{eff}}(m) = 200/\sqrt{24} = 41$

Then the lower and upper 95% confidence limits for the mean become:

$m_Y - t_{Neff(m)-1,1-\alpha/2}\, s_{mP} = 800 - 2.07 \times 41 = 715$ mm

$m_Y + t_{Neff(m)-1,1-\alpha/2}\, s_{mP} = 800 + 2.07 \times 41 = 885$ mm

If the correction for effective number of data had not been made the limits would have been 743 and 857 respectively, which confidence interval is less than 70% of the above computed one. So, a too optimistic figure would have been produced.

The confidence limits for the **standard deviation** are estimated from (3.23) for which $\chi^2_{n,1-\alpha/2}$ and $\chi^2_{n,\alpha/2}$ have to be known. With $N_{eff}(s) = 39$ and n = 38 one obtains by interpolation from Table 3.3:

$\chi^2_{n,1-\alpha/2} = \chi^2_{38,0.975} = 56.9$ and $\chi^2_{n,\alpha/2} = \chi^2_{38,0.025} = 22.9$

Then the lower and upper 95% confidence limits for the **variance** are given by:

$$\frac{nsP^2}{\chi^2_{n,1-\alpha/2}} = \frac{38 \times 200^2}{56.9} = 163^2 \text{ mm}^2 \quad \text{and} \quad \frac{nsP^2}{\chi^2_{n,\alpha/2}} = \frac{38 \times 200^2}{22.9} = 258^2 \text{ mm}^2$$

So the 95% confidence limits for the **standard deviation** become 163 and 258 mm respectively. Clearly, the limits are not symmetrical around the estimated value of 200; the distribution is skewed towards the right. If the correction for effective number of data had not been applied then the limits would have been 167 and 249 mm respectively.

4 MEASUREMENT ERROR

4.1 DEFINITIONS

All measurements are subject to errors. Errors may be due to reading errors, scale resolution, instrument limitations, etc. A **measurement error** is defined as the difference between the **measured** and the **true** value of the observed quantity. The nature of an error may be different. Three types of errors are discerned, see Figure 4.1:

- **spurious errors**, due to instrument malfunctioning or human errors, which invalidate a measurement.
- **random errors**, also called precision or experimental errors, which deviate from the true value in accordance with the laws of chance, and can be reduced by increasing the number of measurements.
- **systematic errors**, due to e.g. incorrect calibration which essentially cannot be reduced by increasing the number of measurements. A distinction is made between (time) constant and variable systematic errors.

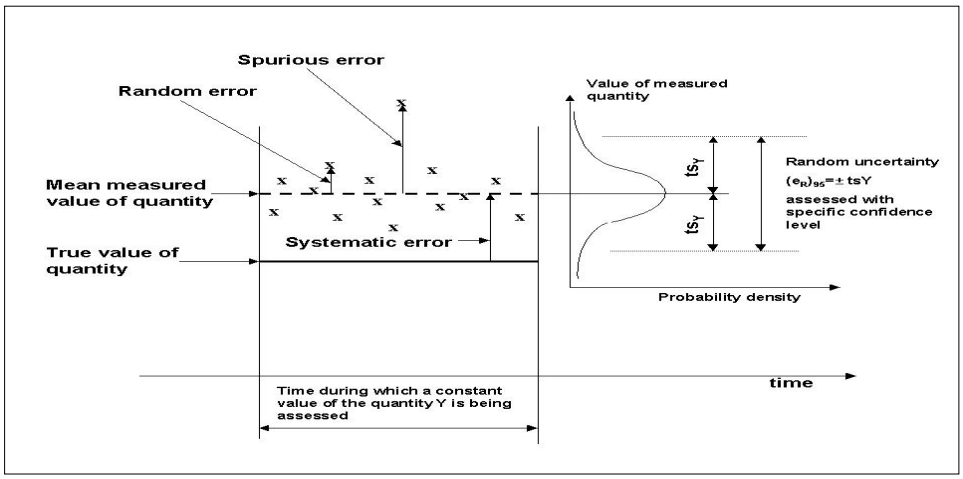

Figure 4.1: Nature of errors

By its very nature the size of an error is not exactly known. Instead, an interval is defined in which the true value of the measured quantity is expected to lie with a suitably high probability. The interval, which is likely to contain the true value, is called the **uncertainty** of the measurement. Associated with uncertainty is its **confidence level**, expressing the probability that the interval includes the true value. The interval is bounded by the confidence limits. It is noted that these confidence limits can only be calculated if the distribution of the measured values about the true value is known. For random errors this can be done, but for systematic errors generally not unless randomisation is possible. For systematic errors usually the **mean estimated error** is used to indicate the uncertainty range, which is defined as the mean of the maximum and minimum values a systematic error may have.

The uncertainty and confidence level are closely related: the wider the uncertainty, the greater is the confidence that the interval encloses the true value and vice versa. The confidence level is an essential part of the uncertainty statement and must always be included. In this manual uncertainty

statements are made at the 95% confidence level conformable to the International Organization for Standardization (ISO) standard ISO 5168-1978 (E). Hence, the **error limits of a measuring device** are defined as the maximum possible positive or negative deviations of a measured value from the true value; the interval between them characterises the range within which the true value will be found with 95% probability.

In addition to the above the following terms are in use when dealing with accuracy of measurements (WMO, 1994):

Measurement: an action intended to assign a number as the value of a physical quantity in stated units. No statement of the result of a measurement is complete unless it includes an estimate of the probable magnitude of the uncertainty.

Reference measurement: a measurement utilising the most advanced state of the science and the latest technology. The result is used to make a best approximation to the true state.

True value: the value which is assumed to characterise a quantity in the conditions which exist at the moment when that quantity is observed (or is the subject of a determination). It is an ideal value, which could be known only if all causes of error were eliminated.

Correction: the value to be added to the result of a measurement so as to allow for any known errors and thus to obtain a closer approximation to the true value.

Accuracy: the extent to which the true value of a quantity agrees with the true value. This assumes that all known corrections have to be applied.

Precision: the closeness of agreement between independent measurements of a single quantity obtained by applying a stated measurement procedure several times under prescribed conditions. (Note that accuracy has to do with closeness to the truth; precision has to do only with closeness together.)

Reproducibility: the closeness of agreement between measurements of the same value of a quantity obtained under different conditions, e.g. different observers, different instruments, different locations, and after intervals of time long enough for erroneous differences to be able to develop.

Repeatability: the closeness of agreement, when random errors are present, between measurements of the same value of a quantity obtained under the same conditions, i.e. the same observer, the same instrument, the same location, and after intervals of time short enough for real differences to be unable to develop (compare with reproducibility).

Resolution: the smallest change in a physical variable, which will cause a variation in the response of a measuring system. (In some fields of measurement resolution is synonymous with discrimination).

Response time: the time, which elapses, after a step-change in the quantity being measured, for the reading to show a stated proportion of the step-change applied. The time for 90 or 95% of the step-change is often given.

Lag error: the error, which a set of measurements may possess due to the finite response time of the observing instrument to variations in the applied quantity.

In this chapter the following topics will be presented:

- Spurious errors (Section 4.2).
- Random errors (Section 4.3).
- How to deal with various types of systematic errors (Section 4.4).
- Combination of random and systematic uncertainty (Section 4.5).
- Propagation of errors will be dealt with in Section 4.6. It covers errors in quantities being a function of several variables. In that case the uncertainty in the measurement of each individual variable determines the size of the composite error. Rules for assessing the uncertainty in such a quantity will be addressed.
- Identification of sources of errors (Section 4.7).
- Significant figures (Section 4.8).

Reference is also made to the Guidelines for evaluating and expressing the uncertainty of NIST measurement results presented in Chapter 2 of the Volume 2, Sampling Principles – Reference Manual.

4.2 SPURIOUS ERRORS

Spurious errors are errors due to instrument malfunctioning or human errors. Such errors invalidate the measurement and must either be eliminated if the source is known and the error is rectifiable or the measurement should be discarded. Errors of this type cannot be taken into consideration in a statistical analysis to assess the overall accuracy of a measurement.

Spurious errors can be detected by application of Dixon's "outlier" test, provided that the measurements are normally distributed. In the test the measurements $Y_1,\ldots Y_N$ are ranked as follows:

- when suspiciously high values are tested: $Y_1<Y_2<\ldots..<Y_N$
- when suspiciously low values are tested: $Y_N<Y_{N-1}<\ldots<Y_1$

Then the following test ratio is computed:

$$D = \frac{Y_N - Y_{N-K}}{Y_N - Y_L} \qquad (4.1)$$

where K and L vary with the sample size N, see Table 4.1. The ratio D is compared with its critical value D_c, presented in the same table as a function of N. If $D > D_c$, then Y_N is considered to be an outlier. The test may be repeated for subsequent outliers provided that the detected outlier is first removed from the sample.

N	K	L	D_c	N	K	L	D_c	N	K	L	D_c
3	1	1	0.941	11	2	2	0.576	19	2	3	0.462
4	1	1	0.765	12	2	2	0.546	20	2	3	0.450
5	1	1	0.620	13	2	2	0.521	21	2	3	0.440
6	1	1	0.560	14	2	3	0.546	22	2	3	0.430
7	1	1	0.507	15	2	3	0.525	23	2	3	0.421
8	1	2	0.554	16	2	3	0.507	24	2	3	0.413
9	1	2	0.512	17	2	3	0.490	25	2	3	0.406
10	1	2	0.477	18	2	3	0.475				

Table 4.1: Critical test value and ranks as function of sample size

4.3 RANDOM ERRORS

Random (or **stochastic**) **errors** originate from experimental and reading errors. They are caused by numerous, small, independent influences, which prevent a measurement system of producing the same reading under similar circumstances. Random errors determine the reproducibility of a measurement. Repeating the measurements under the same conditions produces a set of readings, which is distributed about the arithmetic mean in accordance with the laws of chance. The frequency distribution of these deviations from the mean approaches a **normal** distribution if the data set becomes large. For small sample sizes a **Student t** distribution applies. The sampling distributions are discussed at length in the Reference Manual.

To quantify the uncertainty in a measurement, assume that N measurements, comprising only random errors, are taken on a quantity Y during a period in which Y did not change. The sample mean m_Y and

sample standard deviation s_Y are computed by equation (3.12) and (3.14), respectively. From the Student t distribution it follows that 95% of the measurements will be contained in the interval:

$m_Y \pm t_{n,0.975} s_Y$ with: n = N – 1

where: n = is the number of degrees of freedom.
$t_{n,0.975}$ = 97.5% percentage point of the Student t distribution and is read from Table 3.3.

It implies, that if a single measurement Y_i on Y is made, then there is only 5% chance that the range:

$$Y_i \pm t_{n,0.975} s_Y \quad (4.2)$$

does not contain the true value of Y. The interval $\pm t_{n,0.975} s_Y$ or shortly $\pm t s_Y$ is defined as the random uncertainty $(e_R)_{95}$ in a measurement at a 95% confidence level:

$$(e_{R,Y})_{95} = \pm t_{n,0.975} s_Y \quad (4.3)$$

Or expressed as a relative random error in percent:

$$(X'_Y)_{95} = \pm 100 \frac{t_{n,0.975} s_Y}{Y} \quad (4.4)$$

Since the standard deviation or standard error of the sample mean of N independent measurements is according to (3.12) $\sqrt{N}$ times smaller than s_Y the random uncertainty in the mean value at a 95% confidence level becomes:

$$(e_{R,m_Y})_{95} = \pm t_{n,0.975} s_{m_Y} = \pm t_{n,0.975} \frac{s_Y}{\sqrt{N}} \quad (4.5)$$

It is observed that the random error in the mean value reduces with increasing number of measurements.

Note that in the rest of this section the parenthesis $(..)_{95}$ will be omitted; all errors will be considered at the 95% confidence level.

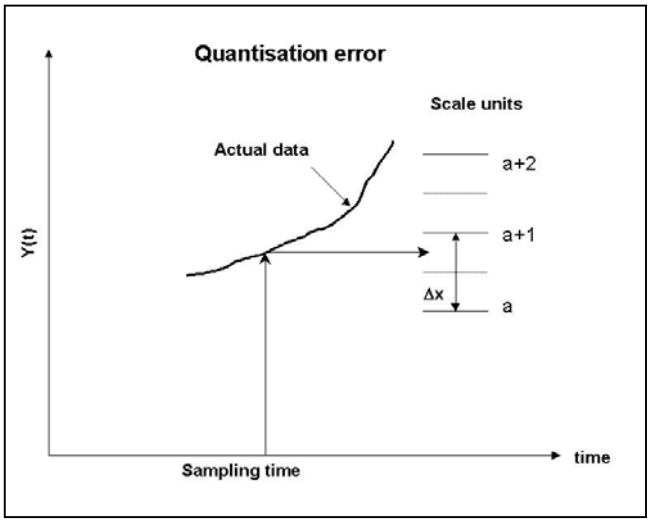

Figure 4.2: Quantisation error

A special type of random error is the **quantisation error**. Quantisation is the actual conversion of the observed value in numerical form, see Figure 4.2. An error is being made because of the scale unit. If the scale unit is Δx, then the true value is within $-½\Delta x$ and $+½\Delta x$ of the scale unit. Assuming a uniform distribution it can be shown that these errors have zero mean and a standard deviation of $\sqrt{(1/12)}$ = 0.29 scale unit.

4.4 SYSTEMATIC ERRORS

Systematic errors are errors, which cannot be reduced by increasing the number of measurements so long as equipment and conditions remain unchanged. In Figure 4.1 the systematic error is shown as the difference between the arithmetic mean value deduced from measurements and the true value of a quantity. Incorrect calibration and shift of scales are typical sources of systematic errors. Systematic errors may be divided into two groups:

- **constant** systematic errors, and
- **variable** systematic errors.

Constant systematic errors are typically calibration errors or result from incorrect setting of a scale zero. Constant errors do not vary with time. Dependent on the nature of the error they may or may not vary with the value of the measurement. Calibration errors vary usually with the instrument reading, whereas an incorrect zero-setting of an instrument leads to a reading independent systematic error.

Variable systematic errors result from inadequate control during an experiment. They may also arise when discrete measurements are taken on a continuously varying quantity. An example of this is the error made in a tipping bucket record. Only when the bucket is full a tipping is recorded. Hence, the error in the recording of a storm can be any value between zero and the rain depth equivalent with one tipping P_b. So, the uncertainty in the measurement is $\pm$ 0.5 P_b and if P mm has been recorded for the storm, the reading should be taken as P + 0.5 P_b.

Dependent on the information available, in ISO 5168 the following procedures are distinguished to arrive at the systematic uncertainty:

1. if the error has a unique known value then this should be added to or subtracted from the result of the measurement. Then the uncertainty is taken as zero
2. If the sign of the error is known but its magnitude is estimated subjectively, the mean estimated error should be added to the measurement and the uncertainty is taken as one-half of the interval within which the error is estimated to lie. If the measured value is denoted by M and the systematic error is estimated to lie between δ_1 and δ_2 then the estimated mean error is $(\delta_1 + \delta_2)/2$ and the result Y should read:

$$Y = M + \frac{\delta_1 + \delta_2}{2} \qquad (4.6)$$

with a systematic uncertainty of:

$$e_{S,Y} = \pm \frac{\delta_2 - \delta_1}{2} \qquad (4.7)$$

or expressed as a relative error in percent:

$$X''_Y = \pm 100 \frac{e_{S,Y}}{Y} \qquad (4.8)$$

3. If the magnitude of the systematic uncertainty can be assessed experimentally, the uncertainty is calculated as for random errors. The measured value is adjusted in accordance with (4.6).

4. If the sign of the error is unknown and its magnitude is assessed subjectively, the mean estimated error is zero and the uncertainty is taken as one-half of the estimated range of the error, see equation (4.7).

4.5 COMBINING RANDOM AND SYSTEMATIC UNCERTAINTIES

Generally, the measuring error has a random part and a systematic part. They can be combined to arrive at the uncertainty of a measurement. Spurious errors do not allow any statistical treatment. If the size of such errors are known measurements can be corrected; else those measurements have to be eliminated from the data set.

Notation convention for relative errors

The convention for notation of relative errors is to use one apostrophe for random uncertainty, two apostrophes for systematic uncertainty and no apostrophe for the total uncertainty:

- X'_Y = random uncertainty in measurement expressed in percent
- X''_Y = systematic uncertainty in measurement expressed in percent, and
- X_Y = total uncertainty of the measurement expressed in percent.

Total error

The combined random and systematic uncertainty in a measurement on a quantity Y expressed as an absolute error e_Y or a relative error X_Y is obtained from:

$$e_Y = \pm\sqrt{e_{R,Y}^2 + e_{S,Y}^2} \tag{4.9}$$

$$X_Y = \pm\sqrt{(X'_Y)^2 + (X''_Y)^2} \tag{4.10}$$

The above equation follows from the rule that the variance of the sum is equal to the sum of the variances (3.25) assuming that the random errors and systematic uncertainties are independent.

Accuracy and precision

The **precision** of a measurement refers to its reproducibility and hence is determined by the random error; the smaller the random errors the higher the precision. The **accuracy** of a measurement indicates how close the measurement is likely to be to the true value. Hence, systematic and random errors **both** determine the **accuracy** of a measurement. So, accuracy differs from precision. Therefore, a measurement can be very precise but highly inaccurate if no corrections for systematic uncertainties are made.

4.6 PROPAGATION OF ERRORS

In many cases one cannot do observations on a variable itself, but rather on its constituting components. In such situations the propagation of errors made in observations on the components towards the compound variable is to be assessed. To derive the random uncertainty of a dependent variable $Z = F(Y_1, Y_2,\ldots, Y_n)$ a Taylor series expansion is used to arrive at the following expression for the absolute error $e_{R,Z}$, provided that the errors in the Y_i's are **independent**:

$$e_{R,Z} = \pm\left(\left(\frac{\partial F}{\partial Y_1} e_{R,Y_1}\right)^2 + \left(\frac{\partial F}{\partial Y_2} e_{R,Y_2}\right)^2 + \ldots + \left(\frac{\partial F}{\partial Y_n} e_{R,Y_n}\right)^2 \right)^{1/2} = \pm\left(\sum_{i=1}^{n} \left(\frac{\partial F}{\partial Y_i} e_{R,Y_i}\right)^2 \right)^{1/2} \tag{4.11}$$

The partial derivatives in (4.11) are called "sensitivity coefficients θ_i " which are a measure for the relative importance of each of the components. They represent the rate of change in Z due to a unit change in each of the Y_i's:

$$\theta_i = \frac{\partial F}{\partial Y_i} \tag{4.12}$$

The sensitivity coefficient may be rendered dimensionless by writing:

$$\theta_i^* = \frac{\partial F}{\partial Y_i}\frac{Y_i}{Z} = \theta_i \frac{Y_i}{Z} \tag{4.13}$$

which expresses the percentage change in Z due to 1% change in Y_i. From (4.11) and (4.13) for the relative random uncertainty X'_Z:

$$X'_Z = \pm\left((\theta_1^* X'_{Y_1})^2 + (\theta_2^* X'_{Y_2})^2 + ... + (\theta_n^* X'_{Y_n})^2\right)^{1/2} = \pm\left(\sum_{i=1}^{n}(\theta_i^* X'_{Y_i})^2\right)^{1/2} \tag{4.14}$$

This equation, as well as (4.11), is generally applicable provided that the errors in the Y_i's are **independent**. To ensure this the function Z = F(Y) should be partitioned into independent factors Y_1, Y_2, …Y_n contributing to the error in Z.

From (4.14) the following special cases are elaborated, which cover most applications:

- Z is a weighted sum of independent factors Y_i:

  $$Z = a_1Y_1 + a_2Y_2 + ...+a_nY_n \tag{4.15}$$

  Then with (4.13) one obtains for θ^*_i :

  $$\theta_i^* = a_i \frac{Y_i}{Z}$$

  and

  $$X'_Z = \pm\left((a_1 \frac{Y_1}{Z} X'_{Y_1})^2 + (a_2 \frac{Y_2}{Z} X'_{Y_2})^2 + ... + (a_n \frac{Y_n}{Z} X'_{Y_n})^2\right)^{1/2} \tag{4.16}$$

- Z is a product of Yi's of the following general form:

  $$Z = aY_1^{p_1}Y_2^{p_2}.....Y_n^{p_n} \tag{4.17}$$

  then:

  $$\theta_i^* = p_i$$

  and:

  $$X'_Z = \pm\left((p_1X'_{Y_1})^2 + (p_2X'_{Y_2})^2 + ... + (p_nX'_{Y_n})^2\right)^{1/2} \tag{4.18}$$

Above equations have been derived for random uncertainties. The same procedures also apply for systematic uncertainties. The total uncertainty is subsequently determined by equation (4.10). The use of above equations (4.16) and (4.18) is shown in the following examples.

EXAMPLE 4.1

Consider the variable R being the difference of a variable P and variable Q: R=P-Q. Random errors are present in both P and Q and a systematic uncertainty is expected in Q. To estimate the total uncertainty in R the following procedure can be used.

Note that R=P-Q is a form of (4.15) with $a_1 = 1$ and $a_2 = -1$and n = 2.

- The random uncertainty in R due to random uncertainty in P and Q is assessed from (4.16):

$$X'_R = \pm\left((\frac{P}{R} X'_P)^2 + (\frac{Q}{R} X'_Q)^2 \right)^{1/2}$$

- Similarly, for the systematic uncertainty in R due the systematic uncertainty in Q one finds with (4.16):

$$X''_R = \frac{Q}{R} X''_Q$$

Assume that the relative random errors in P and Q are both 5% and that the systematic uncertainty in Q is

$$X'_R = \pm\left((\frac{100}{20}5)^2 + (\frac{80}{20}5)^2 \right)^{1/2} = \pm 32\% \quad \text{and} \quad X''_R = \pm\frac{80}{20}1 = \pm 4\%$$

1%. Measurement on P gave a value of 100 and on Q a value of 80, so R = P-Q = 20.

Then from the above it follows for the total or combined uncertainty in R in percent according to (4.10):

$$X_R = \pm\sqrt{(X'_R)^2 + (X''_R)^2} = \pm\sqrt{32^2 + 4^2} = \pm 32\%$$

Hence, the value of R = 20 ± 6. It is observed that by subtracting two large figures of the same order of magnitude with moderate uncertainties an uncertainty in the result is obtained, which is much larger than the uncertainties in the two constituents.

EXAMPLE 4.2

Given is the following non-linear relation between variable R and the variables P and Q:

$R = CP^bQ^{-c}$

Measurements are available on P and Q. The powers b and c are deterministic values, i.e. without error, but the coefficient C has an error. The error in R due to random uncertainties in the coefficient C and in the measurements on P and Q can be deduced from equation (4.18), with a = 1, Y_1 = C, p_1 = 1, Y_2 = P, p_2 = b and Y_3 = Q, p_3 = -c with n = 3. It then immediately follows:

$$X'_R = \pm(X'^2_C + b^2 X'^2_P + c^2 X'^2_Q)^{1/2}$$

Also equation (4.14) could have been applied. This requires first determination of the sensitivity coefficients according to (4.13):

$$\theta_1^* = 1 \quad ; \quad \theta_2^* = b \quad ; \quad \theta_3^* = -c$$

Substitution in (4.14) then leads to the above equation for X'_R

If the measurements on P and Q contain systematic uncertainties as well then in the above equation the X'-s are replaced by X''-s. The combined error in R is subsequently derived from (4.10).

$$X'_R = \pm(1^2 + 2^2 3^2 + 0.5^2 3^2)^{1/2} = \pm 6.3\% \quad ; \quad X''_R = \pm(2^2 1^2 + 0.5^2 1^2)^{1/2} = \pm 2.1\% \quad ; \quad X_R = \pm(6.3^2 + 2.1^2)^{1/2} = \pm 6.6\%$$

Now let b=2 and c=0.5 and the random uncertainty in C be 1% and in P and Q be 3% and the systematic uncertainty be 1% in both P and Q, then the random, systematic and total error in R becomes:

Note that the total error is mainly determined by the uncertainty in the measurement of P, though the relative errors in P and Q are the same. So, it is the absolute value of the **power** to which a variable is raised that matters for its weight in the total error.

4.7 SOURCES OF ERRORS AND THEIR IDENTIFICATION

Each instrument and measuring method has its own sources of errors. Some typical sources of errors are listed below (WMO, 1994):

- **Datum or zero error**, which originates from the incorrect determination of the reference point of an instrument
- **Reading or observation error**, which results from the incorrect reading of the indication by the measuring instrument
- **Interpolation error**, which is due to inexact evaluation of the position of the index with reference to the two adjoining scale marks between which the index is located
- **Parallax error**, which is caused when the index of an instrument is at a distance from its scale and the observer's line of vision is not perpendicular to that scale.
- **Hysteresis error**, i.e. a different value given by the instrument for the same actual value depending on whether the value was reached by a continuously increasing change or by a continuously decreasing change of the variable
- **Non-linearity error** is that part of an error whereby a change of indication or response departs from proportionality to the corresponding change of the value of the measured quantity over a defined range
- **Insensitivity error** arises when the instrument cannot sense the given change in the measured element
- **Drift error** is due to the property of the instrument in which its measurement properties change with time under defined conditions of use, e.g. mechanical clockworks drift with time or temperature
- **Instability error** results from the inability of an instrument to maintain certain specified metrological properties constant
- **Out-of range error** is due to the use of an instrument beyond its effective measuring range, lower than the minimum or higher than the maximum value of the quantity, for which the instrument has been constructed, adjusted, or set
- **Out-of-accuracy class error** is due to the improper use of an instrument when the minimum error is more than the tolerance for that that measurement.

ISO 5168-1978 (E) lists the following procedure for error identification to be used before all the uncertainties are combined:

1. identify and list all independent sources of error
2. for each source determine the nature of the error
3. estimate the possible range of values which each systematic error might reasonably be expected to take, using experimental data whenever possible
4. estimate the uncertainty to be associated with each systematic error
5. compute, preferably from experimental data, the standard deviation of the distribution of each random error
6. if there is a reason to believe that spurious errors may exist, apply the Dixon outlier test
7. if the application of the outlier tests results in data points being discarded, the standard deviation should be recalculated where appropriate
8. compute the uncertainty associated with each random error at the 95% confidence level
9. calculate the sensitivity coefficient for each uncertainty
10. list, in descending order of value, the product of sensitivity coefficient and uncertainty for each source of error. As a general guide, any uncertainty, which is smaller then one-fifth of the largest uncertainty in the group being combined, may be ignored.

4.8 SIGNIFICANT FIGURES

In the previous sub-sections uncertainties in variables and parameters were discussed and specified. In many cases, though, uncertainties can not be stated explicitly, but only be indicated by the number of meaningful digits or **significant** figures.

> **Definition**: All of the digits that are known with certainty and the first uncertain or estimated digit are referred to as significant figures.

With respect to significant figures the following rules are to be applied:

- when **multiplying or dividing**, the number of significant figures in the product or the quotient should not exceed the number of significant digits in the least precise factor,
- when **adding or subtracting**, the least significant digit of the result (sum or difference) occupies the same relative position as the least significant digit of the quantities being added or subtracted. Hence, here the position rather than the number of significant figures is of importance.

These rules can easily be verified with the theory presented in Section 4.6. Applications are illustrated in Example 4.3.

EXAMPLE 4.3

Presented are the following calculations:

1. Multiplication: 3.1416 x 2.34 x 0.58 = 4.3
2. Division: 54.116/20.1 =2.69
3. Addition: 59.**7**

   1.2**0**

   0.33**7**

   ———

   61.**2**37 = 61.**2**
4. Subtraction: 10,**2**00

   8**5**0

   ———

   9,**3**50 = 9,**4**00

In the first calculation 0.58 has only two significant figures, hence the product should have two significant figures as well. In some cases this rule may relaxed, like in 9.8 x 1.06 = 10.4, which can easily be verified with the procedures presented in Sub-section 4.6. In the second calculation the denominator has three significant figures and so has the quotient. In calculation (3) the first doubtful digits are shown in boldface. Of the three values the position of the least significant in 59.7 is most to the left and determines the number of significant digits in the result, similarly for the fourth calculation.

In the calculations in Example 4.3 values were **rounded off** (and not truncated). For rounding off values the following rules apply. Let the objective be to round off to n significant digits, then:

- if the $(n+1)^{th}$ digit > 5 then add 1 to the last significant digit
- if the $(n+1)^{th}$ digit < 5 leave the last significant digit unchanged
- if the $(n+1)^{th}$ digit = 5 then:
  - add 1 to the last significant digit if the least significant digit is odd,
  - leave the last significant unchanged if the least significant digit is even.

Hence in calculation (3) in Example 4.3 the value 61.237 has three significant digits; the fourth digit is 3, which is less than 5, hence the least significant digit remains unchanged, so 61.237 becomes 61.2. With respect to the value 9,350 in calculation (4) it is seen that the 3 is the least significant figure and the digit behind it is 5, so 1 is added to the least significant digit.

Sometimes with trailing zeros there may be ambiguity with respect to the number of significant figures. For example the value 5200 may have two to four significant figures. To avoid this the **scientific notation** is to be applied, by expressing the value as a number between 1 and 10, multiplied by 10 raised to some power:

- 5.2×10^3 two significant figures
- 5.20×10^3 three significant figures
- 5.200×10^3 four significant figures

5 SAMPLING FREQUENCY

5.1 GENERAL

Hydrological and hydro-meteorological processes are generally continuous with time. For further processing the continuous records are discretised along the horizontal time axis (sampling) and the vertical axis (quantisation). The quantisation process and its inherent error is discussed in Section 4.3. Sampling determines the points at which the data are observed. Basically two sampling procedures are applied:

- Discrete point sampling, when the continuous process is observed as instantaneous values at discrete points in time, see Figure 5.1. This type of sampling is applied e.g. for water levels, temperature, etc.
- Average sampling, when the sampling is performed over a certain time interval and the result is the integral of the process over the interval. The result is generally presented as an average intensity in the time interval. For sampling of e.g. rainfall, pan evaporation and windrun this procedure is generally applied.

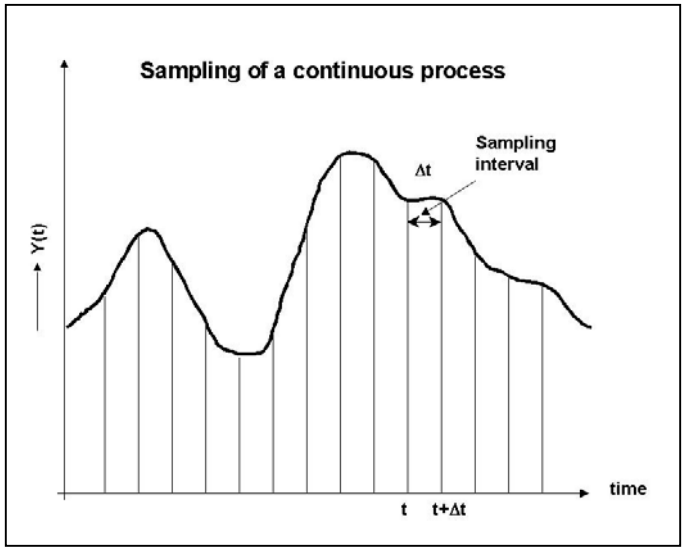

Figure 5.1:
Digitisation of a continuous process

By time-discretisation information might be lost with respect to characteristics of the process. Intuitively, if the variability of the process is large then a small sampling interval has to be applied to reproduce all features of the continuous process. Errors will be made if the sampling interval is chosen too large. On the other hand a too small sampling interval will yield correlated and highly redundant data.

5.2 NYQUIST FREQUENCY

Theoretically, it requires an infinite number of harmonics to reproduce a continuous process without loss of information. Experience has shown that hydrological and hydro-meteorological processes can be considered approximately frequency limited; the higher harmonics do not contribute to the reproduction of the continuous process. This limiting frequency is called the Nyquist or cut-off frequency f_c. If the continuous process is sampled at an interval Δt apart, according to the sampling theorem there will be no loss of information if the sampling interval fulfils the following criterion:

$$\Delta t < \frac{1}{2f_c} \quad \text{or :} \quad \Delta t < \frac{1}{2} T_c \tag{5.1}$$

where: T_c = the period of the highest significant harmonic, $T_c = 1/f_c$. This condition stems from the fact that **more than two** samples are required to reproduce a harmonic. In practice it is assumed that by discrete point sampling at a frequency of $2f_c$ the continuous process can be fully recovered.

Applying a sampling frequency smaller than fc leads to:

- A larger sampling interval, which will reduce the correlation between the observations, making then more independent. This feature will increase the information per sample point.
- Reduction of the number of data during the total sampling period, which reduces the information content in the sample.
- The high frequency components may be missed if the sampling interval is chosen too wide.

It has been shown that for a proper reproduction of extreme values and of run properties (run length, run sum and number of runs) a sampling frequency close to the Nyquist frequency have to chosen (Dyhr-Nielsen, 1972). Similarly, when rates of rise and of fall have to be reproduced properly, an interval very close to (5.1) should be used. Generally, by enlarging the sampling interval the standard errors of estimate and bias of statistical parameters will increase. However, the standard error of the mean is generally least affected, particularly when the data are highly correlated; the standard error and bias of the variance is more sensitive to the sampling interval. In general, for preservation of the lower order moments of the frequency distribution one can go much beyond the Nyquist interval without significant loss of information.

It is obvious that by applying average sampling the mean value of the continuous process can be properly reproduced, without loss of information, no matter what interval is being applied. For other statistical parameters, like the variance, average sampling introduces extra information loss compared to discrete point sampling.

In this chapter the following topics will be dealt with:

- Estimation of the Nyquist frequency,and
- Errors due to discrete point sampling below the Nyquist frequency

5.3 ESTIMATION OF NYQUIST FREQUENCY

The Nyquist frequency f_c as defined by (5.1) can easily be determined from the power spectrum, which displays the variance of the harmonic components as a function of frequency. The spectrum can be estimated via the covariance or auto-correlation function.

Covariance and auto-correlation function of a single harmonic

Consider a time series Y(t) of a single harmonic component, see also Figure 5.2:

$$Y(t) = A\sin(\omega t + \varphi) \qquad \text{or with : } \omega = 2\pi f \quad \text{and} \quad f = \frac{1}{T}:$$

$$Y(t) = A\sin(2\pi f t + \varphi) = A\sin(2\pi \frac{t}{T} + \varphi) \tag{5.2}$$

where: A = amplitude

ω = angular frequency in radians per unit of time
f = ordinary frequency in cycles or harmonic periods per unit of time
φ =phase angle with respect to time origin or phase shift
T = period of harmonic

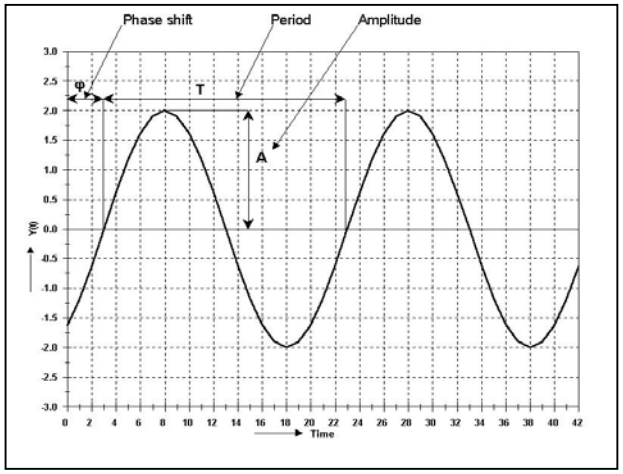

Figure 5.2:
Series of a single harmonic

The covariance of a single harmonic process is derived from:

$$C_{YY}(\tau) = E\left[(Y(t) - \mu_Y)(Y(t+\tau) - \mu_Y)\right] = E\left[A\sin(2\pi f t + \varphi).A\sin(2\pi f(t+\tau) + \varphi)\right]$$

$$\text{or with : } \quad \sin(a).\sin(b) = \frac{1}{2}\left\{\cos(a-b) - \cos(a+b)\right\} \quad \text{it follows :}$$

$$C_{YY}(\tau) = \frac{1}{2}A^2\left\{E\left[\cos(2\pi f\tau)\right] - E\left[\cos(2\pi f t(2+\tau) + 2\varphi)\right]\right\}$$

$$C_{YY}(\tau) = \frac{1}{2}A^2\cos(2\pi f\tau) \qquad \text{so :} \quad C_{YY}(0) = \sigma_Y^2 = \frac{1}{2}A^2 \tag{5.3}$$

Hence it is observed that:

- the covariance of a harmonic remains a harmonic of the same frequency f, so the frequency information in the original series is preserved in the covariance function (or its scaled alternative the auto-correlation function $\rho_{YY}(\tau)=C_{YY}(\tau)/C_{YY}(0)$).
- information about the phase shift has vanished in the covariance and correlation function.
- The amplitude of the periodic covariance function is seen to be equal to the variance of Y(t), see Figure 5.3.

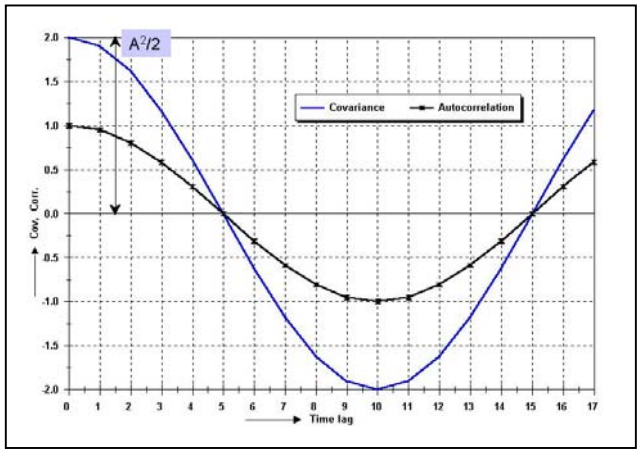

Figure 5.3: Covariance and auto-correlogram

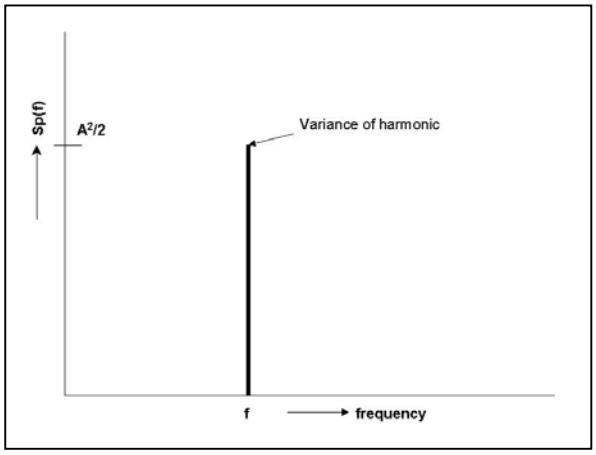

Figure 5.4: Variance or power spectrum

Power or variance spectrum

The plot of the variance of the harmonic against its frequency is called power or variance spectrum and is shown for this particular case in Figure 5.4. Generally, a large number of harmonics is required to describe a process. Let $S_p(f)$ be the ordinate of the continuous spectrum, then the variance contributed by all frequencies in the frequency interval df is given by $S_p(f).df$. Hence, when the process Y(t) is frequency limited with highest frequency f_c, then according to the sampling theorem it follows:

$$\int_0^{f_c} S_p(f).df = \sigma_Y^2 \qquad (5.4)$$

Generally, the spectrum is scaled by the variance (like the covariance function) to make spectra of processes with different scales comparable. It then follows:

$$S_d(f) = \frac{S_p(f)}{\sigma_Y^2} \qquad \text{hence :} \quad \int_0^{f_c} S_d(f).df = 1 \qquad (5.5)$$

where: $S_d(f)$ = spectral density function

The spectral density function is the Fourier transform of the auto-correlation function and can be computed from:

$$S_d(f) = 2\left[1 + 2\sum_{k=1}^{\infty} \rho_{YY}(k)\cos(2\pi fk)\right] \qquad (5.6)$$

To arrive at an estimate for the spectrum, first, $\rho_{YY}(k)$ is to be replaced by its sample estimate $r_{YY}(k)$, with k = 1,2,..,M, where M = maximum lag up to w hich the correlation function is estimated. It will be shown that M is to be carefully selected.

Estimating $S_d(f)$ by just replacing the auto-correlation function by its sample estimate creates a spectral estimate which has a large sampling variance. To reduce this variance a smoothing function is to be applied. With this smoothing function the spectral density at frequency f_k is estimated as a weighted average of the spectral density at surrounding frequencies e.g. f_{k-1}, f_k and f_{k+1}. This smoothing function is called a spectral window, which dimension has to be carefully designed.

An appropriate window for hydrological time series is the Tukey window, which has the following form in the time domain:

$$T_w(k) = \frac{1}{2}\left(1 + \cos(\frac{\pi k}{M})\right) \qquad \text{for :} \quad k \le M$$
$$T_w(k) = 0 \qquad \text{for :} \quad k > M \tag{5.7}$$

The Tukey window has a bandwidth B (indicative for the width over which smoothing takes place in the spectrum) and associated number of degrees of freedom n of :

$$B = \frac{4}{3M\Delta t} \qquad \text{and :} \quad n = \frac{8N}{3M} \tag{5.8}$$

The smoothed spectral estimate then reads:

$$s(f) = 2\left[1 + 2\sum_{k=1}^{M-1} T_w(k) r_{YY}(k) \cos(2\pi f k)\right] \qquad \text{for :} \quad f = 0,....,\frac{1}{2} \tag{5.9}$$

where: s(f) = smoothed estimator for $S_d(f)$

It is often mentioned that the spectrum is to be computed for the following frequencies (Haan (1977):

$$f_k = \frac{k f_c}{M} \qquad \text{for :} \quad k = 0,1,.....,M \tag{5.10}$$

Jenkins and Watts (1968) argue that the spacing following from (5.10) is too wide and suggest that a number of frequency points equal to 2 to 3 times (M+1) is more appropriate.

A (1-α)100% confidence interval for $S_d(f)$ is obtained from:

$$\frac{n.s(f)}{\chi^2_n(1-\frac{\alpha}{2})} \le S_d(f) \le \frac{n.s(f)}{\chi^2_n(\frac{\alpha}{2})} \tag{5.11}$$

In Figure (5.5) the 95% confidence limits are shown for s(f) = 2. From Figure 5.5 it is observed, that the variance of the spectral estimate reduces with increasing number of degrees of freedom, i.e. according to (5.8) with decreasing number of lags M in the computation of the auto-correlation function.

Hence, to reduce the sampling variance the maximum lag M should be taken small. From Figure 5.5 it is observed that for say n > 25 the sampling variance reduces only slowly, so little further improvement in the estimate is obtained beyond that point. Consequently, a value for M of about 10 to 15% of N will do in line with (5.8). But a small value of M leads, according to (5.8), to a large bandwidth B. The value of B should be smaller than the frequency difference between two successive significant harmonics in the spectrum. Say a water level is sampled at hourly intervals, hence Δt = 1 hour. One expects significant harmonics with periods of 16 and 24 hours. The frequency difference between the two is 1/16 - 1/24 = 1/48. Hence, it is requested that: B < 1/48, so the condition for M becomes: M > 4x48/3 = 64. If one chooses M to be 10% of N then at least 640 data points should be available for the analysis, i.e. some 27 days or about one month of hourly data. It is to be noted though that since it is not known in advance which harmonics will be significant, that this process is repeated for different values of M.

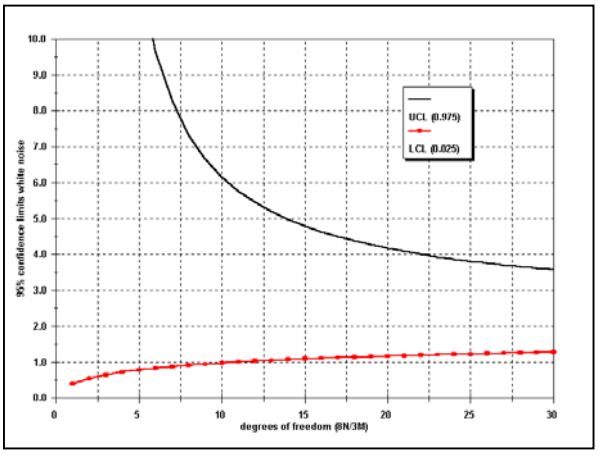

Figure 5.5:
Confidence limits for spectral density function (displayed for s(f)=2, i.e. white noise)

To investigate which harmonics are significant its variance should be outside the confidence limits for white noise. A white noise process means a random process without any correlation between successive data points. According to (5.6) with $\rho_{YY}(k) = 0$ for $k \geq 0$ it follows that $S_d(f) = 2$ for $0 \leq f \leq ½$ The confidence limits for a white noise spectrum are obtained from (5.11) by substituting for s(f) the average spectral density estimated for $0 \leq f \leq ½$. This average should be ≈ 2. The 95% confidence levels are shown in Figure 5.5.

The Nyquist frequency can now be estimated as the frequency of the highest harmonic showing a significant variance, and the sampling interval should then be calculated according to (5.1). If the period of this component is T_c, then $\Delta t < ½ T_c$. Subsequently, the consequences of this choice are to be evaluated.

In case no significant harmonics are apparent, i.e. when the series are white noise, then the standard errors in estimating the statistical parameters can directly be obtained from Table 3.1. An assessment of the effect of reduction of the sampling frequency is then straightforward.

Pre-filtering

Before carrying out the spectral analysis it is advisable to eliminate, if apparent, low frequency components from the series first, for in the determination of the Nyquist frequency interest is on the higher frequencies. Trends will be interpreted as a harmonic with a frequency near zero. A strong trend produces large power in the spectrum at the lowest frequency's, that scales down the visibility at the higher frequencies. Elimination of the low frequencies will expose the higher frequency components in the spectrum better.

The low frequency component can easily be modelled by a moving average scheme of an appropriate length. Say, a series of hourly values is considered and from the series it is obvious that fluctuations within the day do exist. Hence harmonics with periods of one day or more are not of concern. Then the moving average can be performed with a sliding width of one day.

It is important to display the original series together with the moving average series and the residual. For an assessment of the Nyquist frequency the residuals are subjected to spectral analysis. If the variation of the original series around the moving average series is found to be very small, then the latter series may be investigated in a spectral analysis. In any case the consequences for the statistics should always be confirmed.

5.4 DISCRETE POINT SAMPLING BELOW THE NYQUIST FREQUENCY

In 5.1 it was mentioned that when the measuring objective is to obtain estimates of the lower moments of the frequency distribution one can generally choose a sampling interval which is much larger than the Nyquist interval. In this section it will be shown how the trade off between sampling interval and uncertainty in the statistical parameter can be determined by considering an estimate for the mean as the measuring objective.

According to equations (3.39) and (3.40) the 100(1 – α)% confidence interval for the mean is approximately:

$$d = 2t_{n,1-\alpha/2}\frac{s_Y}{\sqrt{N_{eff}}} \quad \text{with :} \quad N_{eff} \approx N\frac{1-r_{YY}(1)}{1+r_{YY}(1)} \tag{5.12}$$

where: d = (1-α) confidence interval of the mean.

For a specified value of 'd' the number of data required becomes:

$$N = \frac{1+r_{YY}(1)}{1-r_{YY}(1)}\left(2t_{n,1-\alpha/2}\frac{s_Y}{d}\right)^2 \tag{5.13}$$

From (5.13) it is observed that evidently a small value of d requires a large number of data, the more if the serial correlation is high.

The correlation dependence, shown in equation (5.13), is valid for a first order auto-regressive process for which the correlation function reads:

$$r_{YY}(k) = r_{YY}^k(1) \tag{5.14}$$

This correlation function for some values of $r_{YY}(1)$ is shown in Figure 5.6.

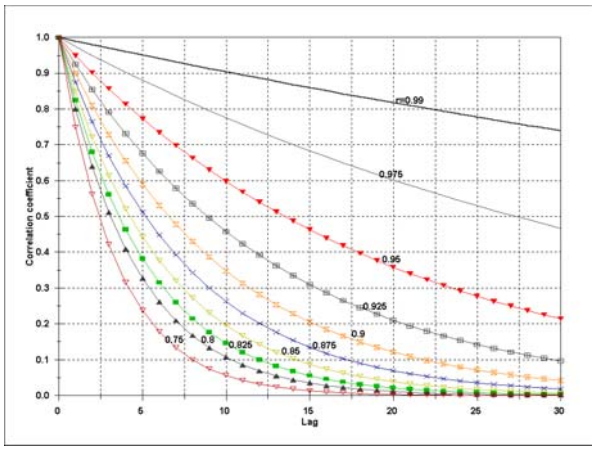

Figure 5.6: Correlation function for a fist order auto-regressive process

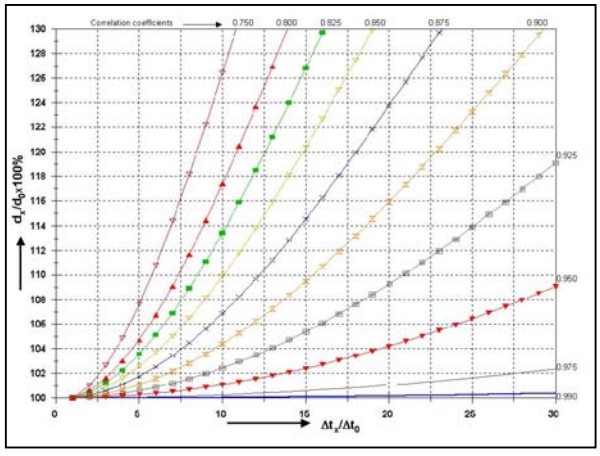

Figure 5.7: Relative increase in width of confidence interval for estimating the mean as a function of increase in sampling interval and of correlation coefficient

For processes with a correlation structure as shown in Figure 5.6 it can now be determined what the impact will be on the width of the confidence interval for estimating the mean. Assume that lag 1 is equivalent to the Nyquist interval, hence $r_{YY}(1)$ is the correlation between successive series elements. In Figure 5.7 the effect of enlarging the sampling interval on the width of the confidence interval relative to the width for sampling at the Nyquist interval is shown (disregarding the effect of N on the t-value). For example, let the Nyquist interval be 1 hour and $r_{YY}(1) = 0.95$. If we would enlarge our sampling interval with a factor 24 i.e. applying a sampling interval of 1 day the width of the confidence interval would only increase by about 6%. This shows that when the correlation between the series elements is high a sampling frequency much lower than the Nyquist frequency can be applied without significant loss of information. Generally, for higher order moments of the frequency distribution the effect of enlargement of the sampling interval is more severe, but can still be feasible. Typically, when the objective is trend detection a tendency similar to estimating the mean is observed. Hence it really pays off to determine beforehand the measuring objective proper.

5.5 SUMMING UP

To carry out an analysis on the sampling frequency, following steps are to be taken:

1. Specify which features of the series are of importance, and which statistical parameters are to be preserved.
2. Select a part of the series, which is representative for a particular season; the sampling frequency applied for this series should be smaller than the expected Nyquist interval
3. Select the series length N and the maximum lag of the auto-correlation function M for use in the spectral computations:
   - first M is determined based on the difference in frequencies between expected nearby peaks in the spectrum; if this difference is df, then $M > 4/(3.df)$
   - then, in order to keep the sampling variance low, select $N > 10M$.
4. Pre-filtering: if the selected part of the series possesses a trend or another low frequency component, model that component by a moving average scheme of appropriate length.
5. Apply spectral analysis on the residual and repeat the spectral analysis for different values of M. Determine the Nyquist frequency fc from the spectrum.
6. Apply data reduction on the original series according to $\Delta t < ½$ fc and analyse the effect of reduction on the selected statistical parameters and process features.
7. If the previous data reduction did not affect the selected parameters or features, analyse effects of further data reduction, i.e. beyond the interval determined at step 6.
8. Select the sampling interval, which fulfils above set criteria. Different sampling intervals may be used from one season to another, hence repeat the analysis for series of other seasons.
9. Evaluate the applicability of the interval in operational practice and in further use of the data.
10. Implement the sampling frequency.

6 SAMPLING IN SPACE

6.1 GENERAL

Processes like rainfall, evaporation, are sampled at fixed locations. Network design involves the determination at how many locations and with what frequency these processes have to be observed. Though sampling in space and in time are strongly interrelated when long yearly averages have to be determined, here the objective is get a proper estimate at one moment in time. Discrete point sampling in space is carried out, generally for one of the following objectives:

- To get a proper areal average, or
- To get a value of the process at any location.

With respect to the areal average, the network will be based on an admissible error in the areal estimate. For the second objective the values will be obtained by means of interpolation between the point sampling stations; the network will be based on an admissible interpolation error.

Nowadays, kriging techniques are widely used to solve the areal estimation and interpolation problem providing best linear estimates of the two quantities. For excellent treatises on kriging reference is made to Isaaks and Srivastava (1989) and Delhomme (1978). Kriging requires a description of the spatial correlation structure, generally described in a semi-variogram. Typical examples of semi-variograms are the gaussian model, the exponential model and the spherical model either or not including a nugget effect. The models are examples of stationary models. For such models there is a relationship between the semi-variogram and the spatial covariance function:

$$\gamma(d) = \sigma^2 - C(d) \qquad (6.1)$$

where: $\gamma(d)$ = semi-variogram
d = distance between points in space
σ^2 = variance of the point process also called the sill of the semi-variogram
C(d) = covariance of the process of points in space at distance d

Note that for large distance the semi-varigram equals the sill: $\gamma(d) = \sigma^2$.

In hydrology and hydro-meteorology the exponential model has obtained great popularity and hence this spatial correlation structure will be used in this chapter. To arrive at transparent equations our analysis will first focus on the design of a uniformly spaced network. In this way complex kriging equations can be avoided and the relevant parameters can be made visible. Thereafter the general kriging variant of the equations will be presented. The network design will be dealt with using one of the following two criteria:

- Minimisation of estimation error in areal estimate, and
- Minimisation of interpolation error.

6.2 SPATIAL CORRELATION STRUCTURE

The exponential model used to describe the spatial correlation structure of e.g. rainfall in spatially homogeneous areas is of the following type, see also figure 6.1:

$$r(d) = r_0 \exp(-d / d_0) \qquad (6.2)$$

where: r(d) = correlation coefficient as a function of distance
d = distance
r_0 = correlation coefficient at d = 0
d_0 = characteristic correlation distance: $r(d_0) = r_0 e^{-1} = 0.368 r_0$

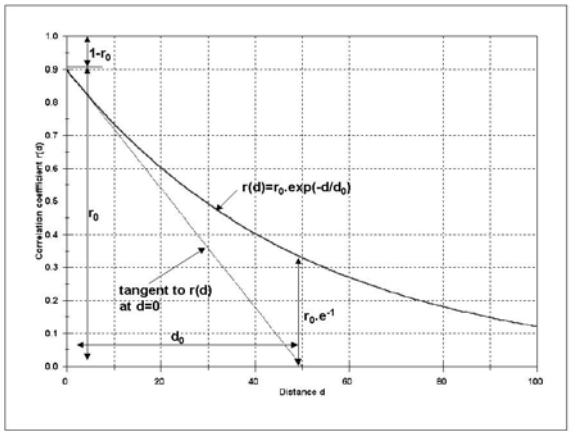

Figure 6.1:
Exponential spatial correlation function

This model includes two parameters, r_0 and d_0. The first one gives the correlation at zero distance. One would expect that the correlation at d = 0 is 1 rather than r_0. However, this deviation is typical for spatial correlation functions describing e.g. the rainfall field. The difference 1 - r_0 is generally ascribed to microclimatic effects and measurement errors of the spatial process at a fixed point. In semi-variogram terminology this difference is called 'nugget effect'. It will be shown that the deviation of r_0 from 1 has important consequences for the accuracy of estimates of areal averages and of interpolated values. The second parameter is a measure for the areal extent of the spatial correlation structure and indicates the distance at which the correlation has reduced to 37% (e^{-1}) of its value at d = 0. It can be obtained as the intersect of the tangent to the correlation function for small d. More appropriate is to get the parameters from linear regression between ln(r(d)) and d.

6.3 STANDARD ERROR OF AREAL ESTIMATE

Let the true areal value of the spatially homogeneous process Y in a basin be denoted by y^*_A and its estimate, based on N point values of Y, by y_A then the error R in estimating y^*_A reads:

$$R = y_A - y^*_A \tag{6.3}$$

If y_A is an unbiased estimate of y^*_A then the mean square error in y_A is the error variance σ^2_R.

The error variance is defined by:

$$\sigma^2_R = E[(y_A - y^*_A)^2] \tag{6.4}$$

Further, let the (time) average value of Y measured at a point be denoted by y_{av} then the root mean square error Z_{areal} in estimating the areal value of Y, expressed as a fraction of y_{av}, is defined by:

$$Z_{areal} = \frac{\sigma_R}{y_{av}} \tag{6.5}$$

This relative root mean square error is equivalent to the relative standard error. As will be elaborated below, the relative root mean square error is a function of:

- the coefficient of variation of **time** series of Y at a point,
- the **spatial** correlation structure of the process,
- the **size** of the basin for which an areal estimate has to be made, and
- the **number** of point data considered in estimating the areal value.

Let there be N gauging stations in a basin with area S, equally distributed over the basin and the process Y is spatially statistically homogeneous. The areal average of Y over S, y_A , is estimated as the arithmetic average of the observations at the N point stations:

$$y_A = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{6.6}$$

where: y_i = point value of Y observed at gauge station i.

Since the gauges are equally distributed over S, each gauge covers an area s_i=S/N. Let the sub-areas be a square with one gauge located at the centre of the sub-area. Kagan (1972) derived the following expression for the mean square error in the areal average of Y over sub-area s_i when estimated by one point observation:

$$\sigma^2_{R,i} = \sigma^2_y\left[1 - r_0 + 0.23\frac{\sqrt{s_i}}{d_0}\right] = \sigma^2_y\left[1 - r_0 + \frac{0.23}{d_0}\sqrt{\frac{S}{N}}\right] \tag{6.7}$$

Now, if it is assumed that the errors made in estimating the areal average of Y in each sub-areas s_i in S are statistically independent, then the error variance σ^2_R in the areal average of Y for the entire area S, when y_A is estimated by equation (6.6), follows from:

$$\sigma^2_R = \sum_{i=1}^{N}\frac{1}{N^2}\sigma^2_{R,i} = \frac{\sigma^2_y}{N}\left[1 - r_0 + \frac{0.23}{d_0}\sqrt{\frac{S}{N}}\right] \tag{6.8}$$

If the coefficient of variation of series Y at any fixed point in S is denoted by Cv= σ_y/y_{av} then, by substituting equation (6.8) in (6.5), the standard error in the areal average of Y over S, expressed as a fraction of the (time) average point process, finally becomes:

$$Z_{areal} = \frac{\sigma_R}{y_{av}} = Cv\sqrt{\frac{1}{N}\left(1 - r_0 + \frac{0.23}{d_0}\sqrt{\frac{S}{N}}\right)} \tag{6.9}$$

By stating the permissible value of Z_{areal}, one obtains an estimate for the required minimum number of stations N in a basin with area S. It should be recalled:

- that Z is the root of the **mean** square error and, in specific cases, errors twice and even three times as high as Z are possible.
- In the above derivation a **uniformly** spaced network was assumed. If the distribution is less even, the error variance will be somewhat larger and so will Z.

The error variance in case of a **non-uniformly** spaced network can be determined with block kriging for a selected basin. Using ordinary kriging, available in HYMOS-4, the areal average is estimated by:

$$y_A = \sum_{i=1}^{N} w_i y_i \quad \text{with} \quad \sum_{i=1}^{N} w_i = 1 \tag{6.10}$$

The stations weights w_i are determined by minimising the estimation error variance. In this general case, where the observation points are not necessarily uniformly distributed over S, the error variance of the estimate of the areal average follows from an equation similar to (6.8) (see e.g. Isaaks, et.al.,1989):

$$\sigma^2_R = \overline{C}_{AA} + \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j C_{ij} - 2\sum_{i=1}^{N} w_i \overline{C}_{iA} \tag{6.11}$$

where: $\bar{C}_{AA}$ = average covariance within the entire area S
C_{ij} = covariance between the gauging stations at locations i and j
$\bar{C}_{iA}$ = average covariance between the gauging stations and the area S

The effect of non-uniformity of the distribution of gauges on the error variance is observed from the second term on the right hand side of equation (6.11). If the stations are clustered, the values for Cij will be higher then in the case of a uniform distribution and so will σ^2_R be. However, to get proper insight in the factors affecting the required network density use of equation (6.9) is preferred, as it explicits the various functional relationships. Once a choice has been made on the required network density according to equation (6.9), kriging may be applied for a final verification of the error variance.

6.4 INTERPOLATION ERROR

Similar to the error in estimating the areal average, the relative root mean square interpolation error made in a **uniformly distributed** gauging network is a function of:

- the **time variability** of the variable at a point, expressed in the coefficient of variation
- the **spatial correlation** structure
- network **density** S/N

Using the exponential correlation function (6.2) to describe the spatial correlation structure of spatial process Y the maximum root mean square error in the estimated point values of Y expressed as a fraction of the time average point value becomes (Kagan, 1972):

$$Z_{int} = Cv\sqrt{1/3(1-r_0) + 0.52\frac{r_0}{d_0}\sqrt{\frac{S}{N}}} \quad \text{with} \quad Cv = \frac{\sigma_y}{y_{av}} \tag{6.12}$$

with: σ_y = standard deviation of point the point process
y_{av}= average of point process
r_0,d_0,S,N are defined as before, see Section (6.3)

The error variance in case of a **non-uniformly** spaced network can be determined with point kriging. If the value of Y at a point, y_0 is estimated by:

$$y_0 = \sum_{i=1}^{N} w_i y_i \quad \text{with} \quad \sum_{i=1}^{N} w_i = 1 \tag{6.13}$$

Then the error variance of the estimate follows from an equation similar to (6.11) (see e.g. Isaaks, et.al.,1989):

$$\sigma_R^2 = \sigma_y^2 + \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j C_{ij} - 2\sum_{i=1}^{N} w_i C_{i0} \tag{6.14}$$

where: C_{ij} = covariance between the gauging stations at locations i and j
C_{i0} = covariance between the gauging stations and the location of the point estimate.

By expressing σ_R as a fraction of y_{av} the relative root mean square error Z_{int} is obtained.

7 NETWORK DESIGN AND OPTIMISATION

7.1 INTRODUCTION

A monitoring network is based upon two considerations, namely:

- the monitoring objectives, and
- the physical characteristics of the systems to be monitored.

The identification of the monitoring objectives is the first step in the design and optimisation of monitoring systems. Related to this is the identification of the potential data users and their future needs. If there is more than one objective, priorities need to be set. Identification of monitoring objectives is also important because they determine the scale of changes to be detected in the data, the kind of information to be extracted from the data and therefore the way the data are analysed.

The analysis of the data, obtained from the network, is also determined by the dynamics of the measured processes. The physical basis of the relevant processes must be known in order to be able to make preliminary guesses of the scale of the variability with respect to space and time.

To enable an optimal design of a monitoring network, a measure is required, which quantifies the effectiveness level. Which measure is adequate depends on the monitoring objectives. Often, this measure is related to statistical concepts like errors in areal estimates, interpolation error, trend detectability, etc, and can be formulated as a function of sampling variables (what), sampling locations (where), sampling frequencies (when) and sampling accuracy (with what (technique/equipment)). These quantities also determine the cost of establishing and running of the network, like the costs related to land acquisition, station construction, equipment procurement and installation, station operation, maintenance, data processing and storage and staffing of field stations and data centres. Once the relationship between the chosen effectiveness measure and costs have been established, the optimal network can be found, in principle, by weighting the two in a cost – effectiveness analysis. The optimisation process is depicted in Figure 7.1

It is stressed that once the network is operational, it has to be evaluated regularly to see whether (revised) objectives still match with the produced output in a cost-effective manner. A network, therefore, is to be seen as a **dynamic system** and should never be considered as a static entity. This requires some flexibility in establishing new stations and closing down others.

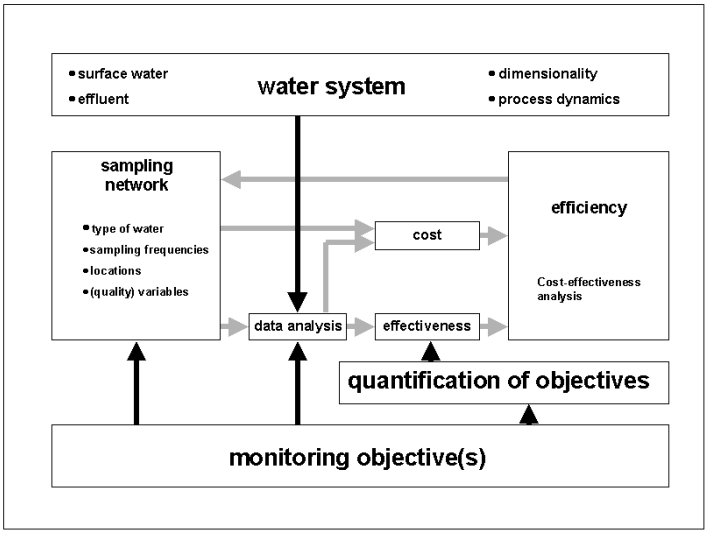

Figure 7.1: Optimisation of network

7.2 TYPES OF NETWORKS

It is necessary to distinguish between the following network levels:

- **basic** or **primary** network, with a low network density, where measurements are continued for a long period of time,
- **secondary** network, with a density supplementary to basic network to meet accuracy demands, and where stations are kept operational for a shorter period of time,
- **dedicated** networks, put in place for a certain project, where the project objectives determine the network density and period of operation, and
- networks for **representative basins**, to study certain phenomena in detail.

Despite the necessary flexibility in the network layout as stipulated above, part of the network should have a permanent character, to ensure that some basic information is continually be gathered. The network used/maintained by IMD or CWC can be considered as such the primary or basic network. This network has a large coverage, though the density is limited and is in operation for a long period of time.

In addition to that network, stations may be established to better cope with the spatial variability of the observed variable. Once sufficient data have been collected from the secondary network to be able to establish relations with the primary stations, the added value of keeping the secondary station operational should be re-examined. This is particularly so if one is interested in reliable long term mean monthly, seasonal or annual values rather than in each individual value. Spatial correlation reduces the information content in a set of data from the network taken at a particular moment in time. For variables like rainfall, where any temporal correlation is fairly non-existing, one more year of data adds on much more information to the data set to compute some long-term average than one extra station does in case of non-zero spatial correlation.

The concept of representative basins is particularly useful when phenomena have to be studied in detail. The representativeness in this case particularly refers to the hydro-meteorological boundary conditions. Small basins may be selected to study e.g. the spatial and temporal variability of short duration rainfall for design purposes.

7.3 INTEGRATION OF NETWORKS

In the Hydrological Information System the following networks are operational:

- hydro-meteorological network of rainfall and full climatic stations,
- hydrometric network,
- surface water quality network,
- geo-hydrological network, and
- groundwater quality network.

These networks are operated by various State and Central agencies. To avoid duplication of work and to reduce cost the networks operated by the various agencies have to be integrated, technically and organisationally.

The hydro-meteorological network has to be considered in conjunction with the surface water and groundwater networks. The former should have sufficient spatial coverage so that all discharge stations in the hydrometric network are fully covered, i.e. that dependent on the objectives, rainfall-runoff computations can be made or water balances can be established. Similar water balance and

resource assessment considerations apply also for the hydro-meteorological network in relation to the groundwater network.

Organisational integration of the networks implies that the networks are complimentary and that regular exchange of field data takes place to produce authenticated data of high quality. Review of the networks is also to be done in close collaboration.

7.4 STEPS IN NETWORK DESIGN

The sequence of steps to be carried out for network review and redesign include:

1. **Institutional set-up**: review of mandates, roles and aims of the organisations involved in the operation of the HIS. Where required communication links should be improved to ensure co-ordination/integration of data collection networks.
2. **Data need identification**: with the aid of the questionnaire 'Data needs assessment' presented in the Part III of Volume 1, Field Manual, HIS, the existing and potential future data users have to be approached to review their data needs.
3. **Objectives of the network**: based on the outcome of step 2 a Hydrological Information Need (HIN) document is to be prepared which lists out a set of objectives in terms of required network output. The consequences of not meeting the target are to be indicated.
4. **Prioritisation**: a priority ranking among the set of objectives is to be made in case of budget constraints.
5. **Network density**: based on the objectives the required network density is determined using an effectiveness measure, taking in view the spatial (and temporal) correlation structure of the variable(s).
6. **Review existing network**: the review covers existing network density versus the required one as worked out in step 5, spreading of the stations in conjunction with the hydrometric and groundwater network, available equipment and its adequacy for collecting the required information, and adequacy of operational procedures and possible improvements. Deficiencies have to be reported upon.
7. **Site and equipment selection**: if the existing network is inadequate to meet the information demands additional sites have to be selected as well as the appropriate equipment.
8. **Cost estimation**: costs involved in developing, operating and maintaining the existing and new sites as well as the data centres have to be estimated.
9. **Cost-effectiveness analysis**: cost and effectiveness are compared. The steps 5 to 8 have to be repeated in full or in part if the budget is insufficient to cover the anticipated costs.
10. **Implementation**: once the network design is approved the network is to be implemented in a planned manner where execution of civil works, equipment procurement and installation and staff recruitment and training is properly tuned to each other. The use of HIDAP is a necessity.
11. The network has to be **reviewed** after 3 years or at a shorter interval if new data needs do develop. The above listed procedure should then be executed again.

8 REFERENCES

1. Delhomme, J.P., (1978)
Kriging in the Hydrosciences.
Adv. Water Resources. Vol. 1, No. 5, pp. 251-266.

2. Dyhr-Nielsen, M. (1972)
Loss of information by discretizing hydrologic series.
Hydrology Papers No 54, Colorado State University, October.

3. Haan, C.T. (1979)
Statistical methods in hydrology.
The Iowa State University Press.
Edition. E&F.N. Spon, London.

4. Isaaks, E.H,. and R.M. Srivastava (1989)
Applied Geostatistics.
Oxford University Press, New York

5. ISO 5168 (1978)
Calculation of uncertainty of a measurement of flow-rate.
International Organization for Standardization, Geneva.

6. Jenkins, G.M., and D.G. Watts (1969)
Spectral Analysis and its Applications.
Holden-Day, San Francisco.

7. Kagan, R.L. (1972)
Planning the spatial distribution of hydrological stations to meet an error criterion.
Casebook on hydrological network design practice III.1.2, WMO No 324, Geneva

8. WMO (1994)
Guide to hydrological practices.
Data acquisition and processing, analysis, forecasting and other applications
WMO-No. 168. World Meteorological Organization, Geneva.