Technical Assistance
HYDROLOGY PROJECT

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

VOLUME 2
SAMPLING PRINCIPLES

REFERENCE MANUAL

Table of Contents

1 SAMPLING DISTRIBUTIONS

The moment and quantile statistics presented in Table 3.1 of Volume 2, Design Manual, Sampling Principles, 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 may, however, deviate significantly from normality. In addition to the normal distribution important sampling distributions are the Chi-square distribution, and the Student-t distribution. These 3 distributions will first be described, and subsequently be used to quantify the uncertainty in the sample mean and the sample variance.

1.1 NORMAL OR GAUSSIAN DISTRIBUTION FUNCTION

The mean μ_y and the standard deviation σ_Y fully describe the **normal** or **gaussian pdf** and **cdf**:

$$f_{N,Y}(y) = \frac{1}{\sigma_Y\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\frac{y-\mu_Y}{\sigma_Y})^2\right) \quad \text{and} \quad F_{N,Y}(y) = \frac{1}{\sigma_Y\sqrt{2\pi}}\int_{-\infty}^{y}\exp\left(-\frac{1}{2}(\frac{\xi-\mu_Y}{\sigma_Y})^2\right)d\xi \qquad (1.1)$$

The **standard normal pdf** and **cdf** follows from (1.1) when the **standardised random variable** Z is introduced, which is given by:

$$Z = \frac{Y-\mu_Y}{\sigma_Y} \qquad (1.2)$$

Then the standard normal pdf and cdf, where $\mu_Z = 0$ and $\sigma_Z = 1$, reads:

$$f_{N,Z}(z) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z^2) \quad \text{and} \quad F_{N,Z}(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp(-\frac{1}{2}\xi^2)d\xi \qquad (1.3)$$

The standard normal pdf is shown in Figure 1.1

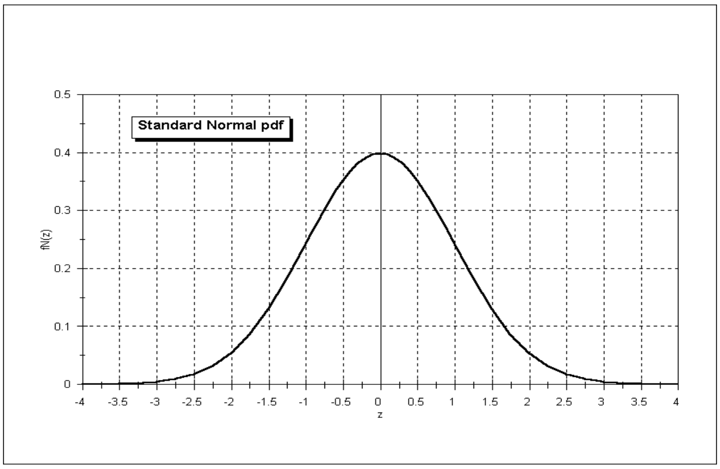


Figure 1.1
Standard normal pdf

The importance of the normal distribution in hydrology stems from the **Central Limit Theorem**, which states that this distribution results quite generally from the sum of a large number of random variables acting together. Let Y_S be the weighted sum of N **independent** random variables Y_i i = 1, N, each with mean μ_i and variance σ_i^2:

$$Y_S = \sum_{i=1}^{N} a_i Y_i \qquad (1.4)$$

Then, for sufficiently large N, **irrespective** of the distribution of the Y_i's, their weighted sum Y_S is normally distributed with mean μ_{Ys} and variance σ_{Ys}^2:

$$\mu_{Y_S} = \sum_{i=1}^{N} a_i \mu_i \quad \text{and} \quad \sigma_{Y_S}^2 = \sum_{i=1}^{N} a_i^2 \sigma_i^2 \qquad (1.5)$$

1.2 CHI-SQUARE DISTRIBUTION

Equation (1.2) defines the standard normal variate Z. Now let Z_1, Z_2, Z_3, …, Z_n be n independent standard normal random variables, then the Chi-square variable χ_n^2 with n degrees of freedom is defined as:

$$\chi_n^2 = Z_1^2 + Z_2^2 + Z_3^2 + \ldots + Z_n^2 \qquad (1.6)$$

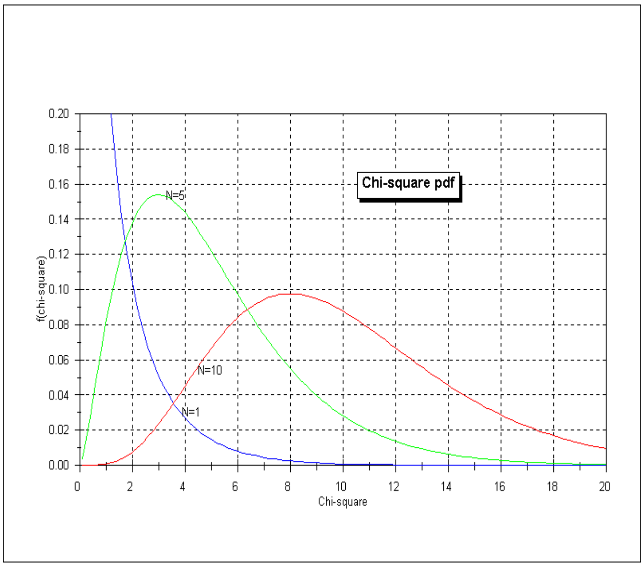
The number of degrees of freedom n represents the number of independent or 'free' squares entering into the expression. The pdf is given by:

$$f_C(\chi^2) = \frac{(\chi^2)^{n/2-1} \exp(-\chi^2/2)}{2^{n/2}\Gamma(n/2)} \quad \text{for :} \quad \chi^2 \geq 0$$

The function $f_C(\chi^2)$ for different degrees of freedom is depicted in Figure 1.2. The mean and the variance of the variable χ_n^2 are:

$$\mu_{\chi^2} = n \quad \text{and} \quad \sigma_{\chi^2}^2 = 2n \qquad (1.7)$$

The Chi-square distribution is a special case of the gamma distribution function. It approaches the normal distribution if N becomes large. This distribution function is particularly of importance to describe the sampling distribution of the variance.



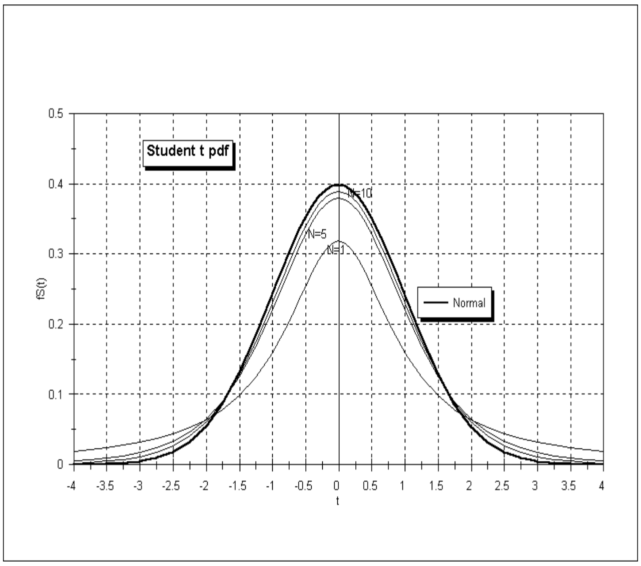


Figure 1.2 Chi-square distribution

Figure 1.3 Student t distribution

1.3 STUDENT T DISTRIBUTION

The Student t distribution results from a combination of a normal and a chi-square random variable. Let Y and Z be independent random variables, such that Y has a χ_n^2 distribution and Z a standard normal distribution then the variable T_n is the Student t variable with n degrees of freedom when defined by:

$$T_n = \frac{Z}{\sqrt{Y/n}} \tag{1.8}$$

The pdf of T_n is given by:

$$f_S(t) = \frac{\Gamma\{(n+1)/2\}}{\Gamma(n/2)} \frac{1}{\sqrt{\pi n}} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

The function $f_S(t)$ for different degrees of freedom is shown in Figure 1.3. The mean and the variance of the variable T_n are respectively:

$$\mu_t = 0 \quad \text{for } n > 1 \qquad \text{and} \quad \sigma_t^2 = \frac{n}{n-2} \quad \text{for } n > 2 \tag{1.9}$$

The Student t distribution approaches a standard normal distribution as the number of degrees of freedom becomes large. From (1.9) it is observed that the standard deviation is slightly larger than 1 particularly for small n. Hence, the dispersion about the mean is somewhat larger than in the standard normal case.

1.4 CONFIDENCE INTERVALS OF MEAN AND VARIANCE FOR UNCORRELATED DATA

The distribution functions dealt with in the previous section are used to arrive at confidence intervals for the mean and the variance or standard deviation. That is to say, that when an estimate of the mean or the variance is made based on a sample, a range around the estimate is determined in which the true mean or variance is likely to lie with a stated probability. To arrive at these confidence intervals in this sub-section it is assumed that the sample data are independent of each other, i.e. serially uncorrelated. Some adjustment is required when the sample values are serially correlated; in the next sub-section the required adjustment is dealt with. In the next the following cases are discussed:

- confidence limits for the mean when the population variance is known (use of normal distribution),
- confidence limits for the variance (use of χ^2 distribution), and
- confidence limits for the mean when the population variance is unknown (use of Student t distribution).

1.4.1 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN WITH KNOWN VARIANCE

The mean m_Y is estimated by:

$$m_Y = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{1.10}$$

Then, from (1.4) and (1.5) it follows for the mean and standard error of m_Y with $a_i = 1/N$:

$$E[m_Y] = \mu_Y$$

and:

$$\sigma_{m_Y} = \frac{\sigma_Y}{\sqrt{N}}$$

Therefore, the following sampling distribution applies for the sample mean m_Y:

$$\frac{m_Y - \mu_Y}{\sigma_{m_Y}} = \frac{(m_Y - \mu_Y)\sqrt{N}}{\sigma_Y} = Z \qquad (1.11)$$

where Z has a standard normal distribution as defined by equation (1.3). According to the Central Limit Theorem as N becomes large (N>10), the sampling distribution of the sample mean approaches a normal distribution, **regardless** of the distribution of Y.

Confidence limits of the mean with known variance

Now we define the percentage points or quantiles $z_{\alpha/2}$ and $z_{1-\alpha/2}$, located symmetrically about the mean (Z=0), which enclose the range of possible realisations of Z with probability 1-α:

$$\Pr ob(z_{\alpha/2} < Z \le z_{1-\alpha/2}) = \int_{z_{\alpha/2}}^{z_{1-\alpha/2}} f_{NZ}(z)dz = 1 - \alpha$$

So:

$$\Pr ob(z_{\alpha/2} < \frac{(m_Y - \mu_Y)\sqrt{N}}{\sigma_Y} \le z_{1-\alpha/2}) = 1 - \alpha \qquad (1.12)$$

Given that a sample m_Y is available, then equation (1.12) can be used to indicate the range in which the **true** mean μ_Y is likely to lie with a probability of 100(1-α) percent (note that $z_{\alpha/2} = -z_{1-\alpha/2}$):

$$\Pr ob\left\{(m_Y - z_{1-\alpha/2}\frac{\sigma_Y}{\sqrt{N}}) < \mu_Y < (m_Y + z_{1-\alpha/2}\frac{\sigma_Y}{\sqrt{N}})\right\} = 1 - \alpha \qquad (1.13)$$

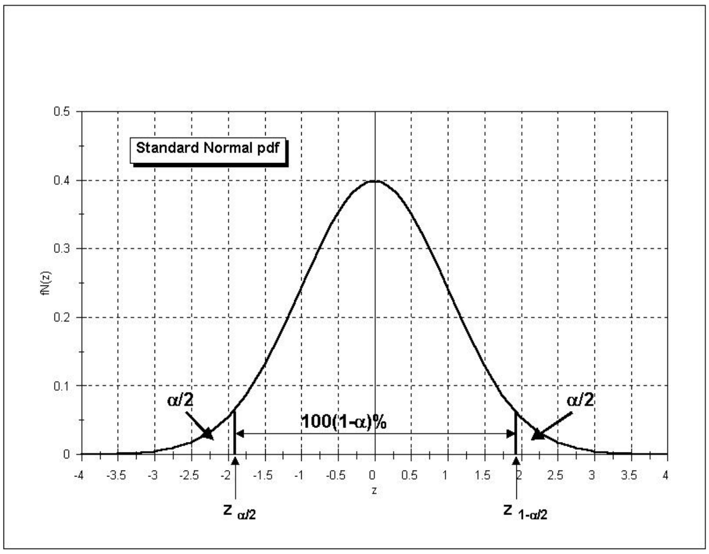


Figure 1.4:
Confidence limits of mean with known variance

The confidence statement expressed by equation (1.13) 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+ $z_{1-\alpha/2}$ $\sigma_Y/\sqrt{N}$) (see Figure 1.4). 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.

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 from the sample data:

$$s_Y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - m_Y)^2} \qquad (1.14)$$

Then, instead of the normal distribution the Student-t distribution has to be applied, which will be dealt with after discussing the distribution of the variance for explanatory reasons.

1.4.2 SAMPLING DISTRIBUTION OF THE SAMPLE VARIANCE

The square of (1.14) gives an unbiased estimate of the variance of Y. If Y has a normal distribution with mean μ_Y and standard deviation σ_Y then with (1.2) and (1.11) the sum term of (1.14) can be partitioned as follows:

$$\sum_{i=1}^{N}(y_i - m_y)^2 = \sum_{i=1}^{N}(y_i - \mu_Y)^2 - N(m_Y - \mu_Y)^2 = \sigma_Y^2\sum_{i=1}^{N}Z_i^2 - N\frac{\sigma_Y^2}{N}Z^2 = \sigma_Y^2\sum_{i=1}^{N-1}Z_i^2$$

According to definition (1.6) the last sum is a Chi-square variable with n = N-1 degrees of freedom:

$$\sigma_Y^2\sum_{i=1}^{N-1}Z_i^2 = \sigma_Y^2\chi_n^2$$

Hence, by inserting the above expression in the formula for the sample variance (1.14), it follows for

$$\frac{ns_Y^2}{\sigma_Y^2} = \chi_n^2 \quad \text{with} \quad n = N - 1 \qquad (1.15)$$

Confidence limits of the variance

Similar to the procedure followed for the sampling distribution of the mean above, here the percentage points points $\chi^2_{n,\alpha/2}$ and $\chi^2_{n,1-\alpha/2}$ are defined:

$$\Pr ob\left\{\chi^2_{n,\alpha/2} < \chi_n^2 \le \chi^2_{n,1-\alpha/2}\right\} = 1 - \alpha$$

Hence, given an estimate of the sample variance computed by (1.14), from (1.15) and above expression, it follows that the true variance σ_Y^2 will be contained within the following confidence interval with a probability of 100(1-α) %:

$$\Pr ob\left\{\frac{ns_Y^2}{\chi^2_{n,1-\alpha/2}} \le \sigma_Y^2 < \frac{ns_Y^2}{\chi^2_{n,\alpha/2}}\right\} = 1 - \alpha \quad \text{with} \quad n = N - 1 \qquad (1.16)$$

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 1 as a function of the number of degrees of freedom n.

1.4.3 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN WITH UNKNOWN VARIANCE

The sampling distribution of the sample mean was given by equation (1.11) under the assumption that the variance of Y was known. This, however, is usually not the case. Then σ_Y in (1.11) is estimated by the sample standard deviation s_Y using equation (1.14). Then by combining (1.11) and (1.15) according to (1.8) the following Student T variate is obtained:

$$\frac{(m_Y-\mu_Y)\sqrt{N}}{s_Y}=\frac{(m_Y-\mu_Y)\sqrt{N}}{\sigma_Y}.\frac{\sigma_Y}{\sqrt{\frac{\sigma_Y^2\chi_n^2}{n}}}=\frac{(m_Y-\mu_Y)\sqrt{N}}{\sigma_Y}.\frac{1}{\sqrt{\chi_n^2/n}}=\frac{Z}{\sqrt{\chi_n^2/n}}=T_n$$

Hence, the sampling distribution of the sample mean m_Y when σ_Y is unknown is given by:

$$\frac{(m_Y-\mu_Y)\sqrt{N}}{s_Y}=T_n \quad \text{with} \quad n=N-1 \tag{1.17}$$

where T_n has a Student t distribution with n = N - 1 degrees of freedom.

Confidence limits of the mean with variance unknown

The percentage points $t_{n,\alpha/2}$ and $t_{n,1-\alpha/2}$ define the 100(1-α) % confidence range of T_n:

$$\Pr ob\{t_{n,\alpha/2}<T_n\le t_{n,1-\alpha/2}\}=1-\alpha \tag{1.18}$$

Note that $t_{n,\alpha/2}$ = - $t_{n,1-\alpha/2}$. Given that an estimate for the sample mean is available, then from (1.17) and (1.18) it follows that the true mean will fall in the following interval at a 100(1-α) % confidence level:

$$\Pr ob\left\{(m_y-t_{n,1-\alpha/2}\frac{s_Y}{\sqrt{N}})\le\mu_Y<(m_Y+t_{n,1-\alpha/2}\frac{s_Y}{\sqrt{N}})\right\}=1-\alpha \tag{1.19}$$

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 1.

1.5 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 (1.19), it has been assumed that the sample series elements are independent. In case persistency 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:

$$N_{eff}(m) = \frac{N}{1+2\sum_{i=1}^{N-1}(1-\frac{i}{N})r_{YY}(i)} \approx N\frac{1-r_{YY}(1)}{1+r_{YY}(1)} \quad \text{for :} \quad r_{YY}(1) > r^* \quad \text{where :} \quad r^* = \frac{2}{\sqrt{N}} \qquad (1.20)$$

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

The latter approximation in (1.20) 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 (1.20) 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 (1.21):

$$r_{YY}(1) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}(y_i - m_y)(y_{i+1} - m_Y)}{s_Y^2} \qquad (1.21)$$

The confidence interval to contain μ_Y with 100(1-α)% probability is now defined by equation (1.19) with N replaced by N_{eff} and the number of degrees of freedom given by $n = N_{eff} - 1$.

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:

$$N_{eff}(s) \approx N\frac{1-[r_{YY}(1)]^2}{1+[r_{YY}(1)]^2} \qquad (1.22)$$

The 100(1-α)% confidence interval for σ_Y^2 follows from equation (1.16) with n = N_{eff}–1.

2 GUIDELINES FOR EVALUATING AND EXPRESSING THE UNCERTAINTY OF NIST MEASUREMENT RESULTS