# V. STATISTICAL ESTIMATOR UNCERTAINTIES

## A. General Principles

Consider a random variable u - with a true mean

$$\mu_u = E\{u\} = \int_{-\infty}^{\infty} uf(u)\,du \qquad (V.1)$$

and true variance

$$\sigma_u^2 = E\{u-\mu_u\}^2 = \int_{-\infty}^{\infty} (u-\mu_u^2)\,f(u)\,du \ , \qquad (V.2)$$

where f(u) is the probability density function.

Given that the PDF is not generally known, we must calculate instead the sample mean of the variable (with a size N) according to

$$\hat{\mu}_u = \frac{1}{N}\sum_{i=1}^{N} u_i \ ; \qquad (V.3)$$

and the corresponding sample variance according to

$$\hat{\sigma}_u^2 = \frac{1}{N}\sum_{i=1}^{N} (u_i - \hat{\mu}_u)^2 \ , \qquad (V.4)$$

where ..^.. signifies that we are computing a statistical estimator.

**An important question is ……How good is a particular estimator?**

There are three principal tests for establishing the quality of an estimator.

**TEST 1) Is the expected value of the estimator of a parameter $\hat{\phi}$ equal to the parameter $\phi$?**

That is $$E\{\hat{\phi}\} = \phi \ ? \qquad (V.5)$$

If so, then the parameter is said to be unbiased.

**TEST 2) Is the mean square error of the estimator smaller than that of other estimators**?

That is $$E\{(\hat{\phi}_1 - \phi)^2\} \overset{?}{\leq} E\{(\hat{\phi}_i - \phi)^2\} \qquad (V.6)$$

If so, then the estimator is said to be efficient.

**TEST 3)** **Does the estimator approach the parameter being estimated, with a probability approaching unity, as the sample size gets large?**

That is, does $$\lim_{N\to\infty} \text{Prob}\{|\hat{\phi} - \phi| \le \varepsilon\} = 0 \quad ? \tag{V.7}$$

If so, then the estimator is said to be <u>consistent</u>.

A sufficient (though not necessary) <u>condition for consistency</u> is that

$$\lim_{N\to\infty} E\{(\hat{\phi} - \phi)^2\} = 0 \tag{V.8}$$

(The latter two conditions are convergence requirements in the probability and mean-square sense.)

**********
Let's explore these issues concerning the mean and variance estimators.
**********

- Consider the ***quality*** **of the discrete <u>sample mean value</u> estimator**,

$$\hat{\mu}_u = \bar{u} = \frac{1}{N}\sum_{i=1}^{N} u_i \ .$$

**<u>Applying TEST 1</u>** leads to

$$E\{\hat{\mu}_u\} = E\left\{\frac{1}{N}\sum_{i=1}^{N} u_i\right\} = \frac{1}{N} E\left\{\sum_{i=1}^{N} u_i\right\} = \frac{1}{N}\sum_{i=1}^{N} E\{u_i\} = \frac{1}{N}(N\,\mu_u) = \mu_u \tag{V.9}$$

or $$E(\hat{\mu}_u) = \mu_u$$

**Hence $\hat{\mu}_u$ is an <u>unbiased</u> estimator!**

**Applying TEST 3**

The *mean square error* of the sample mean value estimator, $\hat{\mu}_u$ is

$$E\{(\hat{\mu}_u - \mu_u)^2\} = E\left\{\left(\frac{1}{N}\sum_{i=1}^{N} u_i - \mu_u\right)^2\right\}, \qquad \text{(V.10)}$$

or

$$= E\left\{\left(\frac{1}{N}\sum_{i=1}^{N} u_i - \frac{N\mu_u}{N}\right)^2\right\}, \qquad \text{(V.11a)}$$

which can be rewritten as

$$= \frac{1}{N^2} E\left\{\left[\sum_{i=1}^{N}(u_i - \mu_u)\right]^2\right\}, \qquad \text{(V.11b)}$$

since

$$E\left\{\left[\sum_{i=1}^{N}(u_i - \mu_u)\right]^2\right\} = E\{[(u_1 - \mu_u) + (u_2 - \mu_u) \ldots\ldots (u_N - \mu_n)]^2\} \qquad \text{(V.12b)}$$

or

$$E\{(u_1 - \mu_u)^2 + (u_2 - \mu_u)^2 \ldots\ldots (u_N - \mu_u)^2 + (u_1 - \mu_u)(u_2 - \mu_u) + \ldots\ldots.] \qquad \text{(V.12c)}$$

However all the cross-terms i.e.,

$$E\{(u_1 - \mu_u)(u_2 - \mu_u)\} \equiv 0 \quad ,$$

because the $u_i$ are statistically independent.

Taking the expectation process through $\sum$ of the only the squared terms

$$E\left\{\sum_{i=1}^{N}(u_i - \mu_u)^2\right\} \sum_{i=1}^{N} E\{(u_i - \mu_u)^2\} = \sum_{i=1}^{N} \sigma_u^2 = N\sigma_u^2. \qquad \text{(V.13a)}$$

Thus

$$E\left\{(\hat{\mu}_u - \mu)^2\right\} = \frac{1}{N^2} E\left\{\sum_{i=1}^{N}(u_1 - \mu_u)^2\right\} = \frac{1}{N^2} N\sigma_\mu^2 = \frac{\sigma_u^2}{N} \qquad \text{(V.13b)}$$

which $\rightarrow 0$ for $N \rightarrow \infty$ .

**Thus by Test 3 $\hat{\mu}_u$ is <u>consistent</u>.**

Also it can be shown that in the limit as $N \rightarrow \infty$, $\lim_{N\rightarrow\infty} E\{\hat{\mu}_u - \mu_u\} = 0$

**Thus by Test 2 $\hat{\mu}_u$ is <u>efficient</u>.**

- Now consider the ***quality*** **of the discrete <u>variance estimator</u>,** $\hat{\sigma}_u^2$ **.**

Applying TEST 1)

$$E\{\hat{\sigma}_u^2\} = E\left\{ \frac{1}{N}\sum_{i=1}^{N}(u_i - \hat{\mu}_u)^2 \right\}$$

$$= \frac{1}{N} E\left\{ \sum_{i=1}^{N}(u_i - \hat{\mu}_u)^2 \right\} \qquad \text{(V.14)}$$

Substituting into the summation in (V.14) as follows

$$\sum_{i=1}^{N}(u_i - \hat{\mu}_u)^2 = \sum_{i=1}^{N}(u_i[\overset{\text{true mean}}{-\mu_u + \mu_u}] - \overset{\text{sample mean}}{\hat{\mu}_u})^2$$

and rewriting as

$$= \sum_{i=1}^{N}[(u_i - \mu_u) - (\hat{\mu}_u - \mu_u)]^2 . \qquad \text{(V.15)}$$

Expanding (V.15)

$$= \sum_{i=1}^{N}(u_i - \mu_u)^2 - 2\sum_{i=1}^{N}[(\hat{\mu}_u - \mu_u)(u_i - \mu_u)] + \sum_{i=1}^{N}(\hat{\mu}_u - \mu_u)^2 . \qquad \text{(V.16)}$$

Rearranging (V.16)

$$= \sum_{i=1}^{N}(u_i - \mu_u)^2 - 2(\hat{\mu}_u - \mu_u)[\sum_{i=1}^{N}(u_i - \mu_u)] + N(\hat{\mu}_u - \mu_u)^2 . \qquad \text{(V.17)}$$

Since $\hat{\mu}_u = \frac{1}{N}\sum_{i=1}^{N} u_i$; and $\sum_{i=1}^{N} - \mu_u = - N\mu_u$

and $\sum_{i=1}^{N}(u_i - \mu_u) = N(\hat{\mu}_u - \mu_u)$, (V.17) becomes

$$\sum_{i=1}^{N}(u_i - \hat{\mu}_u)^2 = \sum_{i=1}^{N}(u_i - \mu_u)^2 - 2N(\hat{\mu}_u - \mu_u)^2 + N(\hat{\mu}_u - \mu_u)^2 \qquad \text{(V.18)}$$

Substituting back into (V.14) yields

$$E\{\hat{\sigma}_u^2\} = \frac{1}{N} E\left\{ \sum_{i=1}^{N} (u_i - \mu_u)^2 - N(\hat{\mu}_u - \mu_u)^2 \right\}. \tag{V.19}$$

Since we know from (V.13a) that

$$E\{(u_i - \mu_u)^2\} = \sigma_u^2 \Rightarrow E\left\{ \sum_{i=1}^{N} (u_i - \mu_u)^2 \right\} = N\,\sigma_u^2 \tag{V.20}$$

and from Eq. (V.13b) that

$$E\{(\hat{\mu}_u - \mu_u)^2\} = \frac{\sigma_u^2}{N} \ , \tag{V.21}$$

it follows that (V.19) can be written as

$$E\{\hat{\sigma}_u^2\} = \frac{1}{N}(N\,\sigma_u^2 - \sigma_u^2) = \frac{(N-1)}{N}\,\sigma_u^2 \quad . \tag{V.22}$$

**Thus $\hat{\sigma}_u^2$ is a BIASED estimator!**

The bias can be removed however by defining an **unbiased sample variance** as

$$s^2 = \hat{\sigma}_u^{2'} = \frac{1}{N-1} \sum_{i=1}^{N} (u_i - \hat{\mu}_u)^2 . \tag{V.23}$$

**[Though not shown here, $\hat{\sigma}_u^2$ is also <u>consistent</u> and <u>efficient.</u>]**

## B. Sampling Distributions

Consider a random variable u with a probability density function; $f_u$ (U). Collect <u>a discrete sample of observed values of u</u> (or $u_i$ , where $1 \le i \le N$). Any quantity computed from these values will also be a <u>random</u> variable. For example, the sample mean value estimator

$$\hat{\mu}_u = \bar{u} = \frac{1}{N} \sum_{i=1}^{N} u_i \, , \tag{V.24}$$

is a random variable which will have different values for different N. The distribution of the different sample mean values $\bar{u}$ is in accordance with its probability density function $f_u(\bar{u})$ .

**What is the Sampling Density Function** $f_u(\bar{u})$ of $\bar{u}$ **?**

**First**, consider a normally-distributed random variable u, which has a *true mean* $\mu_u$ and *true variance* $\sigma_u^2$. Eq(V.9) shows that the expected mean of a sum of sample mean values ($\bar{u}$) is the true mean value of u, according to

$$\mu_{\bar{u}} = E\{\hat{\mu}_u\} = \mu_u \quad , \tag{V.25a}$$

and that the variance of the different sample mean values ($\bar{u}$) is

$$\sigma_{\bar{u}}^2 = E\{\hat{\mu}_u - \mu_u)^2\} = \frac{\sigma_u^2}{N} \quad \text{or} \quad \left[\ \sigma_{\bar{u}} = \frac{\sigma_u}{\sqrt{N}}\ \right] \ . \tag{V.25b}$$

**Thus $\bar{u}$ is normally distributed variable.**

**Second,** the sampling distribution of $\bar{u}$ is derived by considering the form of the normal probability density function of u, which is written in terms of the standardized variable z according to

$$f(z) = \left(\sqrt{2}\pi\right)^{-1} e^{-z_i^2/2} \quad \begin{pmatrix} \text{probability} \\ \text{density} \\ \text{function} \end{pmatrix} \ ,$$

in which $z_i = (u_i - \mu_u) / \sigma_u$ is the standardized variable in terms of u. This form (shown in Figure V.1) is compact and flexible.

Thus we redefine $z_i$ in terms of the sample mean, $\bar{u}$ is accomplished by replacing $u_i$ with $\bar{u}$, $\mu_u$ with $\mu_{\bar{u}}$, and $\sigma_u$ with $\sigma_{\bar{u}}$ according to

$$z = \frac{(\bar{u} - \mu_{\bar{u}})}{\sigma_{\bar{u}}} \quad , \tag{V.26}$$

where the i subscript has been dropped.

Upon the substitution of (V.25a & b), (V.26) becomes

$$z = \frac{(\bar{u} - \mu_u)\sqrt{N}}{\sigma_u} \quad \left[\text{or } \bar{u} = \frac{z\,\sigma_u}{\sqrt{N}} + \mu_u\right] \ ; \tag{V.27}$$

looking very much like the standardized variable for u.

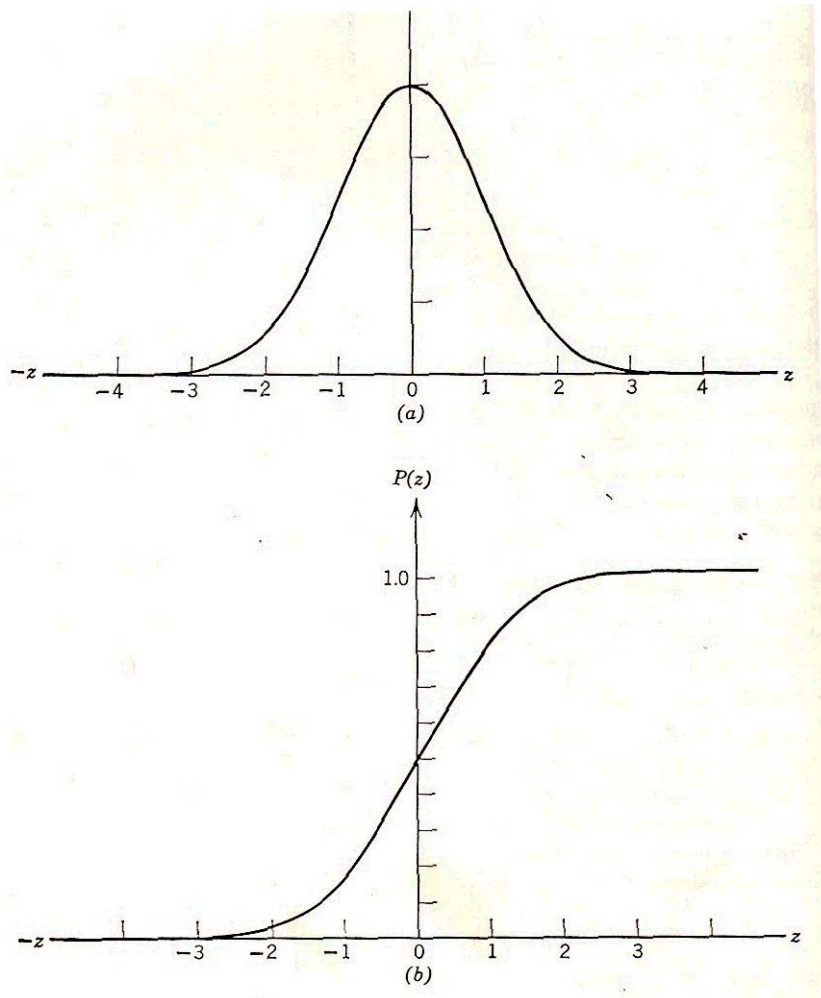


**Figure V.1** (a) Normal probability density function f(z) and (b) probability distribution function F(z).

**<u>What are the probabilities of past and future values of $\bar{u}$ ?</u>**

$$F(z) = (\sqrt{2\pi})^{-1} \int_{-\infty}^{z} e^{-\zeta^2/2}\, d\zeta \quad \begin{pmatrix} \text{probability} \\ \text{distribution} \\ \text{function} \end{pmatrix} .$$

As discussed earlier, such probabilities can be expressed in terms of the cumulative probability distribution function of z (see Figure V.1) by defining the value of z - say $z_\alpha$ - such that the <u>probability</u> concerning:

<u>Past Values</u> is a fixed amount <u>1-α</u> according to

$$F(z_\alpha) = \int_{-\infty}^{z_\alpha} f(z)\, dz = \text{Prob}\,[z \le z_\alpha] = 1 - \alpha \qquad ; \qquad \text{(V.28)}$$

and

Future Values is a fixed amount $\alpha$ (its complement) or:

$$1 - F(z_\alpha) = \int_{z_\alpha}^{+\infty} f(z)\, dz = \text{Prob}[z > z_\alpha] = \alpha \quad . \qquad \text{(V.29)}$$

This partitioning is depicted graphically in the Figure V.2 plot of a general probability density function.

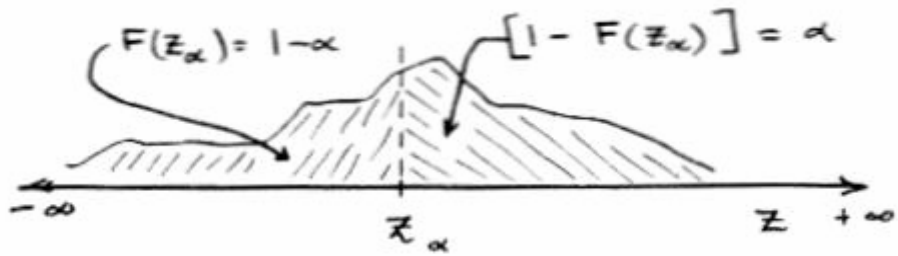


**Figure V.2** The total integral of the probability density function is partitioned as shown (see text).

Thus the probability statement concerning "future" values of $\bar{u}$ can be written as

$$\text{Prob}[\bar{u} > \bar{u}_\alpha] = \alpha \quad , \qquad \text{(V.30)}$$

where $$\bar{u}_\alpha = \frac{z_\alpha \sigma_u}{\sqrt{N}} + \mu_u \ .$$

[ Table A 2; B & P gives values of $z_\alpha$ for user specified values of $\alpha$ ]

The value of $z_\alpha$ which satisfies this relation is known as the 100α percentage point.

## C. Confidence Intervals

**In general, how close does your estimator come to the true value?**

Specifically for this case, what is the probability of finding $\bar{u}$ in a particular interval before collecting sample.

### 1. Confidence Interval: Sample Mean

**First**, consider this issue in terms of

$$z = (\bar{u} - \mu_u)\sqrt{N} / \sigma_u \quad ,$$

the z variable for $\bar{u}$ and the following probability statement

$$\text{Prob}\,[z_{(1-\frac{\alpha}{2})} < \frac{(\bar{u} - \mu_u)\sqrt{N}}{\sigma_u} \le z_{(\alpha/2)}] = 1 - \alpha\,, \qquad \text{(V.31)}$$

this is depicted graphically in Figure V.3.

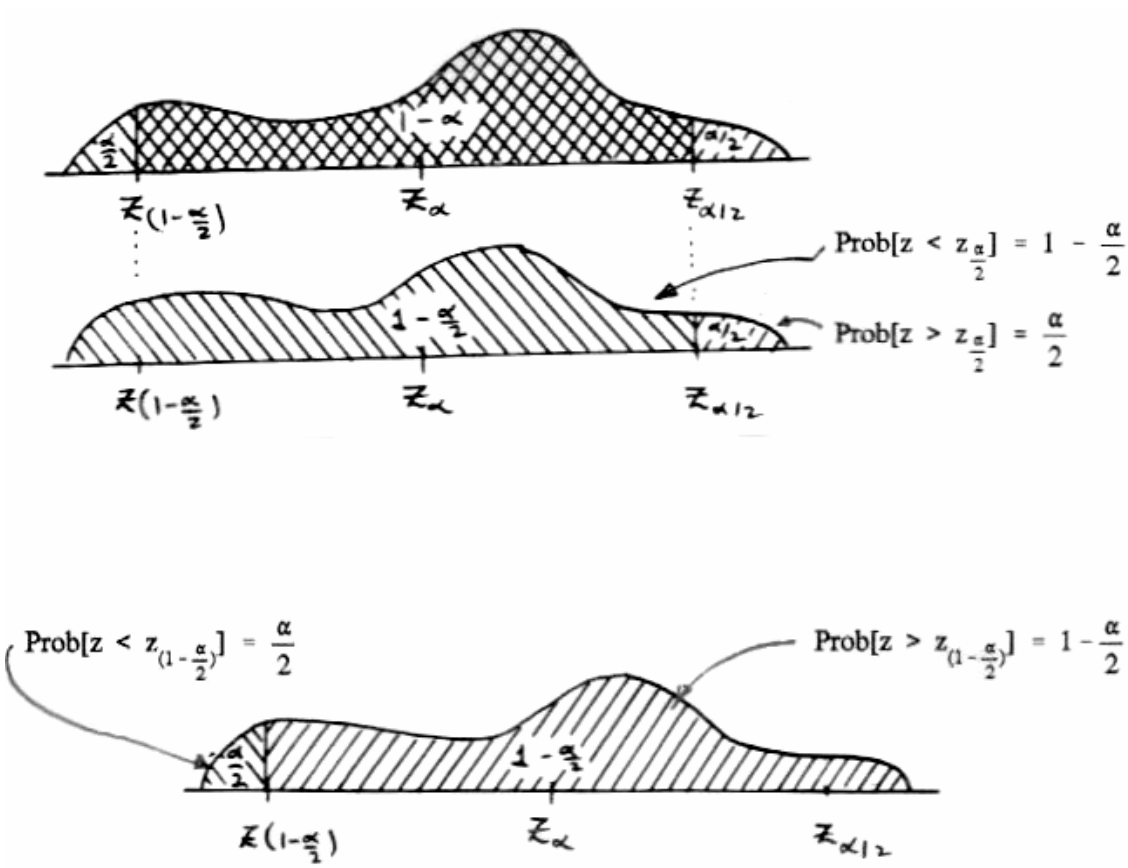


**Figure V.3**. (top panel) The schematic of the 1-α probability of finding the specified true z value between the defined limits in (V.31) can be constructed by summing the middle and lower panel specifications.

After $\bar{u}$ is computed, it is no longer a random number and is fixed and corresponding z value is either in the prescribed interval or not! That is

$$\text{Prob}\,[\,z_{(1-\frac{\alpha}{2})} < \frac{(\bar{u} - \mu_u)\sqrt{N}}{\sigma_u} < z_{(\frac{\alpha}{2})}\,] = \begin{cases} 0 \\ 1 \end{cases} \qquad \text{(V.32)}$$

Note that as $\alpha \to 0$, the interval gets wider and Prob [ ] $\to 1$.

For repeated $\bar{u}$'s from new sets of samples, the value of the probability (V.32) approaches $1 - \alpha$ .

**********

We want to make a confidence statement about $\bar{u}$ falling within a confidence interval with a degree of trust we call the confidence coefficient. Since

$$Z_{(1-\frac{\alpha}{2})} = -Z_{(\frac{\alpha}{2})} ,$$

the terms in the [ ] of (V.32) can be rearranged for the true mean according to

- *Confidence Interval* $$\bar{u} - \frac{\sigma_u Z_{\alpha/2}}{\sqrt{N}} \le \mu_u < \bar{u} + \frac{\sigma_u Z_{\alpha/2}}{\sqrt{N}} , \qquad (V.33)$$

with a

- *Confidence Coefficient* $= (1-\alpha)$

lead to the

- *Confidence Statement* The true value of $\mu_u$ falls within the indicated interval with a confidence coefficient of 1-α; or 100(1-α) percent.

*Note that this says that* the true mean $\mu_u$ lies in a range of values based on the "observed" or sample mean value $\bar{u}$ and the statistical properties of the process.

## 2. Confidence Interval: Unbiased Sample Variance

Consider the Eq(V.23) unbiased sample variance ($s^2$) of N statistically independent observations of a random, normally-distributed variable $u_i$:

$$s^2 = (\hat{\sigma}_u^{2'}) = \frac{1}{N-1}\sum_{i=1}^{N}(u_i - \bar{u})^2 . \qquad (V.34)$$

From (V.17)

$$\sum_{i=1}^{N}(u_i - \bar{u})^2 = \sum_{i=1}^{N}(u_i - \mu_u)^2 - N(\bar{u} - \mu_u)^2 . \qquad (V.35)$$

(V.35) can be written in terms of a normal distribution standardized variable for $u_i$ and $\bar{u}$ as

$$= \sigma_u^2 \sum_{i=1}^{N} z_i^2 - N\frac{\sigma_u^2 z^2}{N} , \qquad (V.36)$$

where $$z_i = \frac{u_i - \mu_u}{\sigma_u} \text{ and } z = \frac{(\bar{u} - \mu_u)}{\sigma_u}\sqrt{N} .$$

Thus (V.35) becomes

$$\sum_{i=1}^{N}(u_i - \bar{u})^2 = \sigma_u^2 \sum_{i=1}^{N-1} z_i^2 . \qquad (V.37)$$

Note the reduction of the number of summed terms in (V.37) relative to (V.36).

Substituting (V.37) into (V.34) leads to an **unbiased sample variance**, which is defined in terms of a new variable $z_i^2$ according to

$$s^2 = \frac{\sigma_u^2}{N-1} \sum_{i=1}^{N-1} z_i^2 \tag{V.38}$$

At this point we define a **new random variable called "chi square"** $\chi_n^2$, which is the sum of a set of standardized random variables according to

$$\chi_n^2 = z_1^2 + z_2^2 + \ldots z_n^2$$

or

$$\chi_n^2 = \sum_{i=1}^{n} z_i^2 \quad , \tag{V.39}$$

where $z_i^2$ are normally distributed. So $\chi_n^2$ is the **chi-square variable** with **"n" degrees of freedom (dof)**, which is the number of independent (or free) squares in the expression.

The <u>probability density function</u> for $\chi_n^2 \geq 0$ is a family of curves depending on the parameter n (see Figure V.4):

$$f(\chi^2) = [2^{n/2}\,\Gamma(n/2)]^{-1} (\chi^2)^{((n/2)-1)}\, e^{-(\frac{\chi^2}{2})} \quad , \tag{V.40}$$

where $\Gamma(n/2)$ is the gamma function.

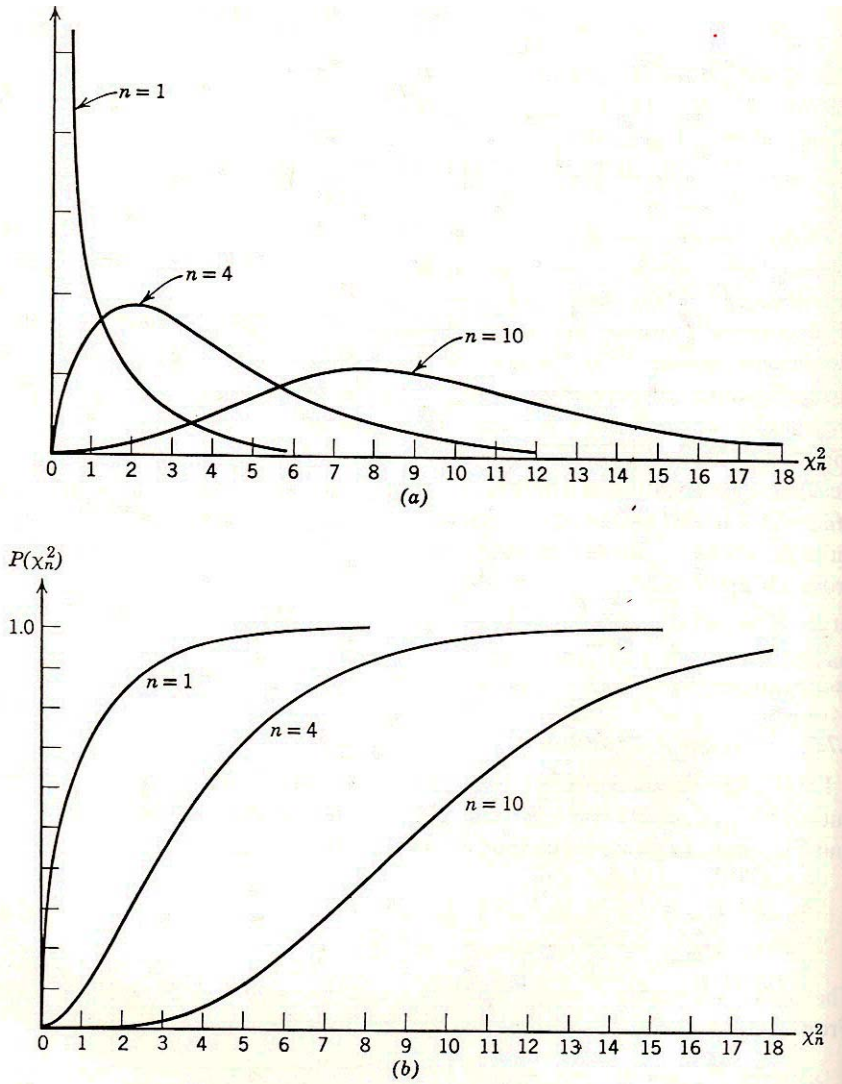


**Figure V.4** The family of (a) $\chi_n^2$ probability density functions according to the number of degrees of freedom = n. and (b) $\chi_n^2$ probability distribution functions according to the number of degrees of freedom = n.

- The <u>probability distribution function</u> (also a family of curves; Figure V.5) is

$$F(\chi^2) = \int_0^{\chi^2} f(\chi'^2)\, d(\chi'^2)\,, \tag{V.41}$$

such that

$$\int_{\chi^2_{n;\alpha}}^{\infty} f(\chi'^2)\, d\chi'^2 = \text{Prob}\,[\chi^2_n > \chi^2_{n;\alpha}] = \alpha \tag{V.42}$$

The chi-squared random variable $\chi_n^2$ has the convenient properties that its mean is

$$E\{\chi_n^2\} = n \ . \tag{V.43a}$$

and its variance is

$$E\{(\chi_n^2 - \mu_{\chi^2})^2\} = \sigma_{\chi^2}^2 = 2n \ . \tag{V.43b}$$

The $\chi_n^2$ distribution is the general case of the following well-known special case probability distribution functions:

1. *Rayleigh distribution function* $(\chi_2)$ - the square root of the chi-square with two degrees of freedom;
2. *Maxwell distribution function* $(\chi_3)$ - the square root of chi-square with three degrees of freedom.
3. *Normal or Gaussian distribution function* - for large n (e.g., for n > 30) the chi-square distribution $\sqrt{2\chi_n^2}$ is distributed approximately as a normal variable with $\mu = \sqrt{2n\text{-}1}$ and $\sigma^2 = 1$.

To obtain the unbiased sample variance in terms of chi-squared distribution, rewrite Eq(V.38) as

$$(N\text{-}1)\, s^2 = \sum_{i=1}^{N} (u_i - \bar{u})^2 = \sigma_u^2 \sum_{i=1}^{N\text{-}1} z_i^2 \quad , \tag{V.44}$$

and substitute (V.39) to obtain

$$n\, s^2 = \sum_{i=1}^{N} (u_i - \bar{u})^2 = \sigma_u^2\, \chi_n^2 \quad , \tag{V.45}$$

where n = N-1 degrees of freedom and (V.45) becomes

$$\left| \frac{ns^2}{\sigma_u^2} = \chi_n^2 \ . \right. \tag{V.46}$$

As a result, the <u>probability statement concerning future values of $s^2$</u> is

$$\text{Prob}\,[s^2 > s_{n,\alpha}^2] = \alpha \tag{V.47a}$$

or

$$\text{Prob}\,[s^2 > \frac{\sigma_u^2\, \chi_{n;\alpha}^2}{n}] = \alpha \tag{V.47b}$$

(See Table A.3 B&P gives $\chi^2_{n;\alpha}$ values such that $\text{Prob}[\chi^2_n > \chi_{n;\alpha}] = \alpha$ )

The corresponding 1-α confidence interval for the true variance $\sigma^2_u$, based on the unbiased sample variance $s^2$ from sample size N, is

$$\left[ \frac{ns^2}{\chi^2_{n;\alpha/2}} \le \sigma^2_u < \frac{ns^2}{\chi^2_{n;1-\alpha/2}} \right] \cdot , \qquad \text{(V.48)}$$

where n = N-1.

Note that in the above it has been assumed that $\sigma_u$ was known (so as to compute $z_i$ and z). If however $\sigma_u$ is unknown, then we must use a **new random variable $t_n$, called *Student's t***, which is defined as

$$t_n = \frac{z}{\sqrt{y/n}} = \frac{z\sqrt{N\text{-}1}}{\sqrt{y}}, \qquad \text{in which } n = N\text{-}1 \text{ are the degrees of freedom;}$$

z is a normally-distributed variable with $\bar{z} = 0$ and $\sigma^2_z = 1$ ; and

y is a $\chi^2_n$-distributed variable.

The probability density function of $t_n$ is

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

The probability-distribution function (with n degrees of freedom) of $t_n$ is

$$F(t_n) = \int_0^{t_n} f(t')\, d(t') \quad .$$

The expectancy of $t_n$ is

$$E\{t_n\} = \mu_t = 0 \qquad \text{for } n > 1 \ .$$

$$E\{t_n\} = \mu_t = 0 \qquad \text{for } n > 1 \ .$$

The variance of $t_n$ is

$$E\left\{(t_n - \mu_t)^2\right\} = \sigma^2_t = \frac{n}{n\text{-}2} \qquad \text{for } n > 2 \ .$$

Thus, the sampling distribution for $\bar{u}$, with $\sigma_u$ unknown, is

$$t_n = \frac{(\bar{u} - \mu_u)\sqrt{N}}{s} \qquad ; n = N\text{-}1 \quad .$$

The corresponding probability statement concerning future values of $\bar{u}$ , which is

$$\text{Prob}\left[\ \bar{u} > \left(\frac{st_{n,\alpha}}{\sqrt{N}} + \mu_u\right)\right] = \alpha$$

$$\text{Prob}\left[\ \bar{u} < \left(\frac{st_{n,\alpha}}{\sqrt{N}} + \mu_u\right)\right] = 1 - \alpha \quad ,$$

translates into the $1 - \alpha$ confidence intervals according to

$$\left[\bar{u} - \frac{st_{n,\alpha/2}}{\sqrt{N}} \le \mu_u \le \bar{u} + \frac{st_{n,\alpha/2}}{\sqrt{N}}\right] \quad .$$

## D. General Definition of Errors (i.e., Uncertainties)

Here <u>generalize</u> the previous considerations in a discussion of the errors in the calculation of a general estimator $\hat{\phi}$ for random data.

First expand the <u>mean square error</u> (MSE) as follows –

$$E\{(\hat{\phi} - \phi)^2\} = E\{[(\hat{\phi} - E\{\hat{\phi}\}) + (E\{\hat{\phi}\} - \phi)]^2\}$$

$$= E\{(\hat{\phi} - E\{\hat{\phi}\})^2\} + 2\,E\{(\hat{\phi} - E\{\hat{\phi}\})\,(E\{\hat{\phi}\} - \phi)\} + E\{(E\{\hat{\phi}\} - \phi)^2\}\,. \qquad \text{(V.49)}$$

But since $E\{\hat{\phi} - E\{\hat{\phi}\}\} = 0$ , (V.49) becomes

$$\text{MSE} = \underbrace{\frac{E\{(\hat{\phi} - E\{\hat{\phi}\})^2\}}{\text{random part of the error}}}_{} + \underbrace{\frac{E\{(E\{\hat{\phi}\} - \phi)^2\}}{\text{square of the bias part of the error}}}_{} \qquad \text{(V.50)}$$

$$= \sigma^2[\hat{\phi}] + b^2[\hat{\phi}]$$

The <u>root mean square error</u> (RMS ERROR)

$$\text{RMS ERROR} \quad \sqrt{E\{(\hat{\phi} - \phi)^2\}} = \sqrt{\sigma^2(\hat{\phi}) + b^2(\hat{\phi})} = \sqrt{\text{Var}[\hat{\phi}] + b^2[\hat{\phi}]} \qquad \text{(V.51a)}$$

has 2 components, namely:

1. RANDOM ERROR (or STANDARD ERROR)

$$\sigma[\hat{\phi}] = \sqrt{E\{(\hat{\phi} - E\{\hat{\phi}\})^2\}} = \sqrt{E\{\hat{\phi}^2\} - E\{\hat{\phi}\}^2}\,, \qquad \text{(V.51b)}$$

square
of
mean

2. BIAS ERROR

$$b[\hat{\phi}] = E\{\hat{\phi}\} - \phi \quad . \qquad \text{(V.51c)}$$

The normalized forms of the (V.51) errors are:

<u>normalized random error</u> = $\varepsilon_r = \dfrac{\sigma[\hat{\phi}]}{\phi} = \dfrac{\sqrt{E\{\hat{\phi}^2\} - E^2(\hat{\phi})}}{\phi}$ (V.52a)

normalized bias error = $\varepsilon_b = \dfrac{b[\hat{\phi}]}{\phi} = \dfrac{E\{\hat{\phi}\}}{\phi} - 1$ (V.52b)

normalized rms error = $\varepsilon = \dfrac{\sqrt{\sigma^2[\hat{\phi}] + b^2[\hat{\phi}]}}{\phi} = \dfrac{\sqrt{E\{(\hat{\phi} - \phi^2)\}}}{\phi}$ (V.52c)

---

**<u>Statistical Estimator Errors: Relatively Small Amplitudes and Zero Bias</u>**

When the random errors $\varepsilon_r$ are relatively small and the bias error $\varepsilon_b$ is zero, then one can write

$$\underset{\substack{\uparrow \\ \text{general} \\ \text{estimator}}}{\hat{\phi}^2} = \underset{\substack{\uparrow \\ \text{true}}}{\phi^2}\,(1 \pm \varepsilon_r) \qquad \text{(V.53)}$$

or

$$\hat{\phi} = \phi\left(1 \pm \varepsilon_r\right)^{1/2} \;\_\; \phi\left(1 \pm \frac{\varepsilon_r}{2}\right) \quad . \qquad \text{(V.54)}$$

Thus generally (except for spectral density and frequency response functions) with a negligible bias error $b[\hat{\phi}]$ and small normalized rms error

$$\varepsilon = \frac{\sigma[\hat{\phi}]}{\phi}\,, \qquad \text{(V.55)}$$

(e.g., $\varepsilon \le 0.20$), the <u>probability density function</u> of $\hat{\phi}$ ($f_{\hat{\phi}}$) can be approximated by a normal distribution according to

$$f_{\hat{\phi}} = \frac{1}{\varepsilon\phi\sqrt{2\pi}} \exp\left[ \frac{-(\hat{\phi} - \phi)^2}{2(\varepsilon\phi)^2} \right] \quad , \tag{V.56}$$

where $E\{\hat{\phi}\} \cong \phi$ and $\sigma[\hat{\phi}] = \varepsilon\phi$ .

This gives the following two probability statements for future values of $\hat{\phi}$ ($\varepsilon \le 0.20$):

$$\text{Prob}\left[ \phi(1-\varepsilon) \le \hat{\phi} \le \phi(1+\varepsilon) \right] \approx 0.68 \tag{V.57a}$$

$$\text{Prob}\left[ \phi(1-2\varepsilon) \le \hat{\phi} \le \phi(1+2\varepsilon) \right] \approx 0.95 \tag{V.57b}$$

The corresponding 68% confidence intervals are

$$\left[ \frac{\hat{\phi}}{1+\varepsilon} \le \phi \le \frac{\hat{\phi}}{1-\varepsilon} \right] \quad , \tag{V.58a}$$

and 95% confidence intervals are

$$\left[ \frac{\hat{\phi}}{1+2\varepsilon} \le \phi \le \frac{\hat{\phi}}{1-2\varepsilon} \right] \quad . \tag{V.58b}$$

For $\varepsilon \le 0.10$ the 68%confidence intervals are

$$\left[ \hat{\phi}(1-\varepsilon) \le \phi \le \hat{\phi}(1+\varepsilon) \right] \quad , \tag{V.59a}$$

and the 95% confidence intervals are

$$\left[ \hat{\phi}(1-2\varepsilon) \le \phi \le \hat{\phi}(1+2\varepsilon) \right] \quad . \tag{V.59b}$$

# E. Time Series Statistical Estimator Uncertainties

## 1. Mean Value Estimator Uncertainty

For stationary, ergodic $\{u(t)\}$, the sample mean is

$$\bar{u} = \hat{\mu}_u = \frac{1}{T}\int_0^T u(t)\,dt \quad . \tag{V.60}$$

The true mean is

$$\mu_u = E\{u(t)\} \tag{V.61}$$

and is statistically independent of t (i.e., stationary). It follows that the

expectation of the series mean is

$$E\{\hat{\mu}_u\} = E\left\{\frac{1}{T}\int_0^T u(t)\,dt\right\} = \frac{1}{T}\int_0^T E\{u(t)\}\,dt \tag{V.62}$$

and equals the true mean according to

$$E\{\hat{u}_u\} = \frac{1}{T}\int_0^T \mu_u\,dt = \mu_u \quad ; \tag{V.63}$$

just as (V.9) showed for a statistically independent sample $u_i$.

Hence $\hat{\mu}_u$ is an unbiased estimate of a series mean value that is independent of T. (i.e. $b(\hat{\mu}_u) \equiv 0$).

Therefore, since

$$E\{-2\,\hat{\mu}_u\,\mu_u\} = -2\,\mu_u^2 \ ,$$

the mean square error is due to the random error only according to

$$\mathrm{Var}[\hat{\mu}_u] = E\{(\hat{\mu}_u - \mu_u)^2\} = \underset{(a)}{E\{\hat{\mu}_u^2\}} - \underset{(b)}{\mu_u^2} \ , \tag{V.64}$$

For a stationary, ergodic process part (a) is

$$E\{\hat{\mu}_u^2\} = \frac{1}{T^2}\int_0^T\int_0^T E\{u(\xi)\,u(\eta)\}\,d\eta\,d\xi \quad . \tag{V.65}$$

However, since by our definition, the auto-covariance function of a random variable is

$$R_{uu}(\tau) = E\{u(t)\,u(t+\tau)\} - \mu_u^2 \tag{V.66}$$

then the random error (V.64) can be written

$$\mathrm{Var}[\hat{\mu}_u] = \frac{1}{T^2}\int_0^T\Big[\int_0^T \underbrace{R_{uu}(\eta - \xi)}_{\substack{\downarrow \\ \text{time} \\ \text{difference only}}}\,d\eta\Big]\,d\xi \tag{V.67a}$$

Detailed algebra presented in Appendix 1 shows that

$$\mathrm{Var}[\hat{u}_u] \approx \frac{1}{T}\int_{-\infty}^{\infty} R_{uu}(\tau)\,d\tau \tag{V.67b}$$

Physically this means that the mean square error (i.e., uncertainty) in computing the mean value of a time series is proportional to the integral time scale of the observations. In particular, the larger the integral time scale, the fewer statistically independent samples go into estimating the mean making it more uncertain.

**********

# EXAMPLE

**********

**Variance of Mean Value Estimate of Bandwidth-Limited White Noise with $\mu_u$ □ 0.**

It is assumed that the one-sided auto-spectral density is (see Figure V.8)

(Note how the mean value appears in the spectrum.)

$$G_{uu}(f) = \frac{1}{B} + \mu_u^2\,\delta(f) \quad 0 \le f \le B$$
$$= 0 \qquad\qquad f > B \tag{V.68}$$

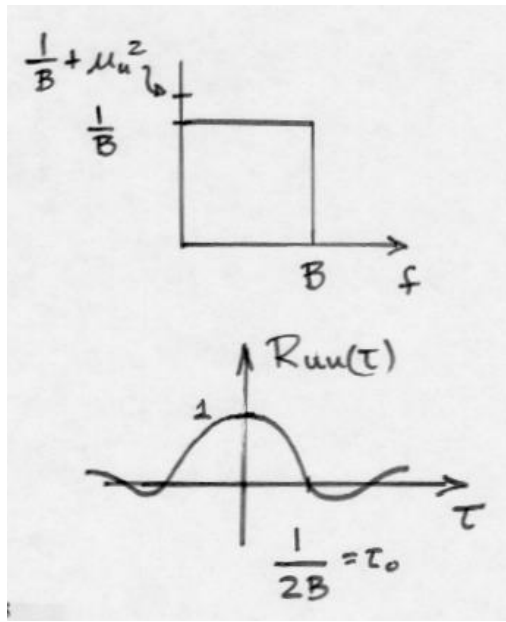


**Figure V.8** The auto-covariance of band-limited noise.

The auto-covariance function is

$$R_{uu}(\tau) = \int_0^\infty G_{uu}(f)\cos(2\pi f\,\tau)\,df - \mu_u^2 \quad . \tag{V.69}$$

$$= \int_0^B \left(\frac{1}{B}\right)\cos 2\pi f\,\tau\,df = \frac{\sin(2\pi B\,\tau)}{(2\pi B\,\tau)} \quad , \tag{V.70}$$

which is the now-familiar *sinc function* (Figure V.8). Note that (a) $R_{uu}(0) = 1$ and (b) the first zero of $R_{uu}(\tau)$ occurs at $\tau = ½\,B$ or $\tau = \tau_o$. Thus

$$2B\tau_o = 1. \tag{V.71}$$

Thus under the condition that T must be sufficiently large so $\tau << T$ for this particular case, the **mean square error associated with** $\hat{\mu}_u$ is

$$\mathrm{Var}[\hat{\mu}_u] \approx \frac{1}{T}\int_{-\infty}^{\infty} \frac{\sin(2\pi B\tau)}{2\pi B\tau}\, d\tau = \frac{1}{2BT} \ . \tag{V.72}$$

**What are the conditions for large T?**

They are that

$$T \geq 10\,|\,\tau_o\,| \ . \tag{V.73}$$

Since $\tau_o = \frac{1}{2B}$,

$$\frac{B}{\Delta f} = BT > 5 \tag{V.74}$$

More generally for the case in which $R_{uu}(0) \neq 1$, the <u>random or mean square error</u> for $\hat{\mu}_u$ is

$$\mathrm{Var}[\hat{\mu}_u] \approx \frac{R_{uu}(0)}{2BT} = \frac{\sigma_u^2}{2BT} \ . \tag{V.75}$$

When $\mu_u$ □ 0, the normalized <u>mean square error</u> for $\hat{\mu}_u$ is

$$\varepsilon^2 = \frac{\mathrm{Var}[\hat{\mu}_u]}{\mu_u^2} \approx \frac{1}{2BT}\frac{\sigma_u^2}{\mu_u^2} \ . \tag{V.76}$$

**********
**********

## <u>2. Mean Square Value Estimator Uncertainty</u>

[Note that we are considering $u^2$ **not** $(u - \mu_u)^2$ -the variance ]

Assume a stationary, ergodic, random $\{u(t)\}$ for which the mean square value estimator is

$$\widehat{\overline{u^2}} = \frac{1}{T}\int_0^T u^2(t)\, dt \ . \tag{V.77}$$

The true mean square value $\overline{u^2} = E\{u^2(t)\}$ is independent of t, because $\{u(t)\}$ is stationary. Thus

$$E\{\widehat{\overline{u^2}}\} = \frac{1}{T}\int_0^T E\{u^2(t)\}\, dt = \overline{u^2} \tag{V.78}$$

which is the true mean square.

**Thus by Test 1 $\widehat{\overline{u^2}}$ an <u>unbiased</u> estimate of $\overline{u^2}$ and is independent of T.**

**What is the mean square error [i.e.** $\mathrm{Var}(\widehat{\overline{u^2}})$ of $\widehat{\overline{u^2}}$ **]?**

Since $E\{-2\,\overline{\hat{u}^2}\;\overline{u^2}\} = -2\,\overline{u^2}^{\,2}$ ,

$$\mathrm{Var}\left[\overline{\hat{u}^2}\right] = E\left\{(\overline{\hat{u}^2}-\overline{u^2})^2\right\} = E\left\{\overline{\hat{u}^2}^{\,2}\right\} - \overline{u^2}^{\,2} \tag{V.79a}$$

$$= \frac{1}{T^2}\int_0^T\int_0^T [E\{u^2(\xi)\,u^2(\eta)\} - \overline{u^2}^{\;2}]\,d\eta\,d\xi \tag{V.79b}$$

It can be shown that (Appendix 1) for large T where $|\tau| \gg T$, the **variance of the mean square value estimator** is

$$\mathrm{Var}[\overline{\hat{u}^2}] \approx \frac{2}{T}\int_{-\infty}^{\infty} [R_{uu}^2(\tau) + 2\,\mu_u^2\,R_{uu}(\tau)]\,d\tau \tag{V.80}$$

**********

**EXAMPLE- Variance of the Mean Square Estimator for White Noise**

**********

For the <u>special case</u> of bandwidth-limited, Gaussian white noise with T □ $10\tau_o$ and BT □ 5;

$$R_{uu}(\tau) = R_{uu}(0)\left(\frac{\sin 2\pi B\tau}{2\pi B\tau}\right) \tag{V.84}$$

and

$$\mathrm{Var}\left[\overline{\hat{u}^2}\right] \approx \frac{R_{uu}^2(0)}{BT} + \frac{2}{BT}\,\mu_u^2\,R_{uu}(0) \tag{V.85}$$

For $\mu_u = 0$

$$\mathrm{Var}\left[\overline{\hat{u}^2}\right] \approx \frac{R_{uu}^2(0)}{BT} \tag{V.86a}$$

as compared to

$$\mathrm{Var}[\hat{\mu}_u] \approx \frac{R_{uu}(0)}{2\,BT} \tag{V.86b}$$

The **normalized mean square error of the mean square estimator** (for $\mu_u = 0$) includes only the Var[ ] component, because the bias error is zero. Thus because $B = \frac{1}{2\tau_o}$ (with $\tau_o$ being a measure of the **integral time scale**),

$$\varepsilon^2 = \frac{\mathrm{Var}[\overline{\hat{u}^2}]}{R_{uu}^2(0)} \approx \frac{1}{BT} = \frac{2\tau_o}{T} \qquad \text{(V.87)}$$

**********
**********

## 3. Cross-Covariance Function Estimator Uncertainty

Given **two sample histories u(t), v(t)** with **zero means** from **stationary, ergodic** processes {u(t)}, {v(t)}. What are the errors associated with their cross-covariance $R_{uv}(\tau)$ and respective auto-covariance functions $R_{uu}(\tau)$ and $R_{vv}(\tau)$. Focusing on the more general $R_{uv}(\tau)$ [$R_{uu}(\tau)$ and $R_{vv}(\tau)$ are just special cases], we consider the sample cross-covariance defined as

$$\hat{R}_{uv}(\tau) = \frac{1}{T-\tau}\int_0^{T-\tau} u(t)\, v(t+\tau)\, dt \qquad \text{for } 0 \le \tau < T \qquad \text{(V.88a)}$$

$$= \frac{1}{T-|\tau|}\int_{|\tau|}^{T} u(t)\, v(t+\tau)\, dt \qquad \text{for } -T < \tau \le 0 \qquad \text{(V.88b)}$$

In order to estimate the errors we will

- consider positive τ only; and
- let data exist for $0 \le \tau \le T + \tau$ .

Thus so that (V.88a) can be written

$$\hat{R}_{uv}(\tau) = \frac{1}{T}\int_0^{T} u(t)\, v(t+\tau)\, dt \qquad 0 \le \tau < T \qquad \text{(V.89)}$$

(Note that the mean square estimates of u(t) or v(t) are only special cases of the above for τ = 0).

**Is $\hat{R}_{uv}(\tau)$, as defined in (V.89), biased?**

Applying Test 1 we find that the expected value of $\hat{R}_{uv}$ is

$$E\{\hat{R}_{uv}\} = \frac{1}{T}\int_0^{T} E\{u(t)\, v(t+\tau)\}\, dt = \frac{1}{T}\int_0^{T} R_{uv}(\tau)\, dt = R_{uv}(\tau) \qquad \text{(V.90)}$$

because the expectation process passes through an integral with constant limits and $R_{uv}(\tau)$ is independent of t.

**Hence $\hat{R}_{uv}(\tau)$ is an <u>unbiased</u> estimate of $R_{uv}(\tau)$ and independent of T.**

**What is the mean square error (i.e. $\mathrm{Var}[\hat{R}_{uv}(\tau)]$ of $\hat{R}_{uv}(\tau)$ )?**

Since $E\{-2\,\hat{R}_{uv}\,R_{uv}\} = -2\,R_{uv}^2$ ,

$$\mathrm{Var}[\hat{R}_{uv}(\tau)] = E\{(\hat{R}_{uv}(\tau) - R_{uv}(\tau))^2\} = E\{\hat{R}^2_{uv}(\tau)\} - R^2_{uv}(\tau) \quad ; \tag{V.91}$$

which in turn can be written

$$= \frac{1}{T^2}\int_0^T\int_0^T [E\{u(\xi)\, v(\xi+\tau)\, (u(\eta)\, u(\eta+\tau)]\} - R_{uv}{}^2(\tau)]\, d\eta\xi \tag{V.92}$$

It is shown in Appendix 1 that the ***mean square error*** **of the cross-covariance estimator** (i.e. $\mathrm{Var}[\hat{R}_{uv}]$) is

$$\mathrm{Var}[\hat{R}_{uv}(\tau)] \approx \frac{1}{T}\int_{-\infty}^{\infty} [R_{uu}(\alpha)\, R_{vv}(\alpha) + R_{uv}(\alpha+\tau)\, R_{vu}(\alpha-\tau)]\, d\alpha\, ; \tag{V.93}$$

while that for the special case of the **auto-covariance estimator** $R_{uu}(\tau)$ [or $R_{vv}(\tau)$] is

$$\mathrm{Var}[\hat{R}_{uu}(\tau)] \approx \frac{1}{T}\int_{-\infty}^{\infty} [R^2_{uu}(\alpha) + R_{uu}(\alpha+\tau)\, R_{uu}(\alpha-\tau)]\, d\alpha \ , \tag{V.94}$$

Exploring the behavior of (V.94) gives:

1. **for $\tau = 0$**

$$\mathrm{Var}[\hat{R}_{uu}(0)] \approx \frac{2}{T}\int_{-\infty}^{\infty} R^2_{uu}(\alpha)\, d\alpha\, ; \tag{V.95}$$

2. **as $\tau$ gets large ($\rightarrow \infty$)**

$$R_{uu}(\tau) \rightarrow 0\ ; \tag{V.96}$$

3. **for large + $\tau$**

$$R^2_{uu}(\alpha) \geq R_{uu}(\alpha+\tau)\, R_{uu}(\alpha-\tau)\, , \tag{V.97}$$

so that (V.94) can be written

$$\mathrm{Var}[\hat{R}_{uu}(\tau)] \approx \frac{1}{T}\int_{-\infty}^{\infty} R^2_{uu}(\alpha)\, d\alpha = \frac{1}{2}\mathrm{Var}[\hat{R}_{uu}(0)] \tag{V.98}$$

**********

## Example

**********

**<u>For Band Limited White Noise -</u>** (with zero mean)

$$R_{uu}(\tau) = R_{uu}(0)\frac{\sin 2\pi B\tau}{2\pi B\tau} \tag{V.99}$$

For $\tau = 0$ and from Eq(V.87)

$$\mathrm{Var}[\hat{\overline{u^2}}] \ \cong \frac{R^2_{uu}(0)}{BT} \tag{V.100}$$

Thus the $\mathrm{Var}[\hat{R}_{uu}]$ for all $\tau$ is <u>conservatively</u>

$$\mathrm{Var}[\hat{R}_{uu}(\tau)] \approx \frac{1}{2BT}\left(R_{uu}^2(0) + R_{uu}^2(\tau)\right) \tag{V.101}$$

and

$$\mathrm{Var}[\hat{R}_{uv}(\tau)] \approx \frac{1}{2BT}\left(R_{uu}(0)\,R_{vv}(0) + R_{uv}^2(\tau)\right), \tag{V.102}$$

with $T \geq 10|\tau_o|$ and $BT \geq 5$.

The normalized mean square error for $\hat{R}_{uv}$ with $\mu_u = \mu_v = 0$ is

$$\varepsilon^2 = \frac{\mathrm{Var}(R_{uv}(\tau)]}{R_{uv}^2(\tau)} \approx \frac{1}{2BT}\left(1 + \frac{R_{uu}(0)\,R_{vv}(0)}{R_{uv}^2(\tau)}\right). \tag{V.103}$$

**********

**********

## 4. Linear Regression Estimator Uncertainty

**Linear Regression**

The linear relationship

$$\tilde{y} = A + Bx \quad , \tag{V.104}$$

relates the predicted value of the dependent variable ( $\tilde{y}$ ) to the value of the independent variable x, where A and B are the true coefficients. **How are A and B estimated?**

Given N pairs of random variable values $x_i$ and $y_i$, where

$$\bar{x} = N^{-1}\sum_{i=1}^{N} x_i \qquad \& \qquad \bar{y} = N^{-1}\sum_{i=1}^{N} y_i \tag{V.105}$$

we can compute a linear fit or regression line for y (the dependent variable) on x (the independent variable)

$$\hat{y} = a + bx\,, \tag{V.106}$$

where the coefficients are

$$a = \bar{y} - b\,\bar{x} \qquad (\overline{.....}\text{ mean values})$$

$$b = \frac{\sum_{i=1}^{N}(x_i - \bar{x})\,y_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2} - \frac{\sum_{i=1}^{N} x_i\,y_i - N\,\bar{x}\,\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\,\bar{x}^2}.$$

Thus the predicted dependent variable is based on the independent variable x according to

$$\hat{y} = \bar{y} + b(x - \bar{x}) \ .$$

By switching dependent and independent variables above, we can calculate the following regression line for x on y

$$\hat{x} = \bar{x} + b'(y - \bar{y}) \ , \qquad \text{(V.107)}$$

where

$$b' = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\,\bar{y}}{\sum_{i=1}^{N} y_i^2 - N\bar{y}^2} \ .$$

The $\hat{\ldots}$ in (V.106) and (V.107) indicate that the **computed dependent variable is an estimate**. The slopes of the two regression lines are related to the same correlation coefficient of x and y which is

$$\hat{\rho}_{xy} = [b'b]^{1/2} \ .$$

or

$$\hat{\rho}_{xy} = \frac{\sum_{i}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i}^{N}(x_i - \bar{x})^2 \sum_{i}^{N}(y_i - \bar{y})^2\right]^{1/2}} = \frac{R_{xy}}{\sigma_x \sigma_y} \qquad \text{(V.108)}$$

**What are the uncertainties associated with linear regression coefficient estimators a and b?** Considering the accuracy of estimates a and b in (V.106). We note that the Student-t probability distribution is pertinent for a ratio between **z** -a normally distributed random variable, with zero mean and unit variance - and $\mathbf{y}^*$ - a $\chi_n^2$-distributed random variable ($y^*$).

Recall that the Student $t_n$ random variable with n degrees of freedom is

$$t_n = \frac{z}{\sqrt{y^*/n}} \ ,$$

whose respective and mean and variance are

$$E\{t_n\} = \mu_t = 0 \qquad \text{for } n > 1$$

and

$$E\{t_n - \mu_t)^2\} = \sigma_t^2 = n/(n-2) \text{ for } n > 2 \quad .$$

The respective probability density and distribution functions of $t_n$ are (with n degrees of freedom) are

$$f_t(t_n) = \frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\,\Gamma(n/2)}\left[1 + \frac{t^2}{n}\right]^{-(n+1)/2}$$

and

$$F(t_n) = \int_0^{t_n} f_t(t_n)\,dt \quad .$$

Assuming that $y_i$ is normally distributed, it can be shown that a and b are <u>unbiased estimates of A and B</u> with sampling distributions related to the <u>Student-t distribution</u> according to

$$\frac{\overset{\text{true values}}{\overset{\downarrow}{a - A}}}{\left(\frac{1}{N} + \frac{\overline{x^2}}{\sum_{i=1}^{N}(x_i - \overline{x})^2}\right)^{1/2}} = S_{y/x}\, t_{N-2}\,; \quad \frac{\overset{\text{true values}}{\overset{\downarrow}{b - B}}}{\left(\frac{1}{\sum_{i=1}^{N}(x_i - \overline{x})^2}\right)^{1/2}} = S_{y/x}\, t_{N-2} \qquad \text{(V.109)}$$

The **sampling distribution of** $\hat{y}$**,** associated with specific values of $x = x_o$, is

$$\frac{\overset{\text{true predicted value}}{\overset{\downarrow}{\hat{y} - \tilde{y}}}}{\left(\frac{1}{N} + \frac{(x_o - \overline{x^2})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}\right)^{1/2}} = S_{y/x}\, t_{N-2}\,, \qquad \text{(V.110)}$$

where $S_{y/x}$ is the <u>sample standard deviation</u> of the observed values of $y_i$ about the prediction

$$\hat{y}_i = a + bx_i$$

and is given by either

$$S_{y/x} = \left[ \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N-2} \right]^{1/2} \tag{V.111a}$$

or the more computationally convenient equation

$$S_{y/x} = \left[ \frac{1}{N-2} \left( \sum_{i=1}^{N} (y_i - \overline{y})^2 - \frac{\left[ \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}) \right]^2}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \right)^{1/2} \right], \tag{V.111b}$$

where $\sum_{i=1}^{N} (y_i - \overline{y})^2 = \sum_{i=1}^{N} y_i^2 - Ny^{-2}$ and $\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{N} x_i y_i - N\overline{xy}$ .

## 5. Correlation Coefficient Uncertainty

Recall that the **cross-correlation function** is defined as

$$\rho_{uv} = \frac{R_{uv}}{\sigma_u \sigma_v} , \tag{V.112}$$

where $R_{uv}$ is the cross-covariance function.

**What is the uncertainty of the *sample cross-correlation coefficient* estimator?**

The relevant sample correlation coefficient for N pairs of statistically-independent samples of $u_t$, $v_t$ is

$$\hat{\rho}_{uv} \equiv \frac{\hat{R}_{uv}}{\hat{\sigma}_u \hat{\sigma}_v} = \frac{\sum_{i=1}^{N} (u_i - \hat{\mu}_u)(v_i - \hat{\mu}_v)}{\left[ \sum_{i=1}^{N} (u_i - \hat{\mu}_u)^2) \sum_{i=1}^{N} (v_i - \hat{\mu}_v)^2 \right]^{1/2}} \tag{V.113a}$$

$$= \frac{\sum_{i=1}^{N} u_i v_i - N \mu_u \mu_v}{\left[ \left( \sum_{i=1}^{N} u_i^2 - N \mu_u^2 \right) \left( \sum_{i=1}^{N} v_i^2 - N \mu_v^2 \right) \right]^{1/2}} \tag{V.113b}$$

The uncertainty of the sample cross-correlation coefficient $\hat{\rho}_{uv}$ can be evaluated in terms of the following function

$$w = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}_{uv}}{1-\hat{\rho}_{uv}}\right). \quad , \qquad \text{(V.114)}$$

which is known to be approximately normally-distributed, with a

**mean**

$$\mu_w = \frac{1}{2}\ln\left[\frac{1+\rho_{uv}}{1-\rho_{uv}}\right]; \qquad \text{(V.115a)}$$

and **variance**

$$\sigma_w^2 = \frac{1}{N-3} \quad . \qquad \text{(V.115b)}$$

Confidence limits can be established for these values according to ways we have discussed for other normally-distributed variables. Because correlations are highly variable, we will focus on verifying that $\boldsymbol{\rho_{uv}}$ is statistically significantly non-zero.

**Explicitly, we test the hypothesis that $\boldsymbol{\rho_{uv} = 0}$.**

Given $\rho_{uv} = 0$, the sampling distribution of **w** is normal with zero mean (or $\mu_w = 0$) and $\sigma_w^2 = 1/(N-3)$. Thus the acceptable region for the hypothesis of zero correlation is given by

$$-z_{\alpha/2} \le \frac{\sqrt{N-3}}{2}\ln\left[\frac{1+\hat{\rho}_{uv}}{1-\hat{\rho}_{uv}}\right] < z_{\alpha/2} \quad , \qquad \text{(V.116)}$$

where z is again the standardized normal variable. From this perspective, a value outside the above interval would constitute a statistically significant correlation at the $\alpha$-level of significance.

## Cross-Correlation Coefficient Uncertainty – An Alternative

In time/space series analysis there is rarely statistical independence between successive terms (or samples). The time between statistically independent terms in the series is called the **integral time (or space) scale** (conceptually the same as our $\tau_o$). (See the **l_correl** appendix for additional discussion). Sciremammano (1979) argues that, when there is a significant integral time scale, $\tau^*$, defined by Davis (1976; 1977) as

$$\tau^* = \sum_{j=-\infty}^{\infty} \rho_{uu}(j\Delta t)\, \rho_{vv}(j\Delta t)\Delta t \ , \tag{V.117}$$

it is **desirable to normalize cross-correlation functions by the large lag standard error $\sigma^*$.** Bartlett (1978, p. 352) defines $\sigma^*$ as

$$\sigma^{*2} = N^{-1} \sum_{j=-\infty}^{\infty} \rho_{uu}(j\Delta t)\, \rho_{vv}(j\Delta t)\,, \tag{V.118}$$

which is approximated by

$$\sigma^{*2} \cong (N - j)^{-1} \sum_{j=-Mn}^{Mn} \rho_{uu}(j\Delta t)\, \rho_{vv}(j\Delta t) \ . \tag{V.119}$$

(Note that Mn must be large enough so that $\tau^*$ is approximated well.) It is shown by Sciremammano (1979) that, for computations with **equivalent degrees of freedom** $N^* = N\Delta t/\tau^* \geq 10$, the **zero cross-correlation levels** are

$$\begin{aligned} \rho_{99} &= 2.6 \\ \rho_{95} &= 2.0 \\ \rho_{90} &= 1.7 \end{aligned} \tag{V.120}$$

for probability levels of $\rho$ = 99%, 95%, and 90%, respectively. Note that

$$\sigma^{*2} = \frac{\tau^*}{N\Delta t} = \frac{1}{N^*} \ . \tag{V.121}$$

**References:**


- Bartlett, M.S., 1978. An Introduction to Stochastic Processes, Cambridge Univ. Press, 362 pp.
- Davis, R.E., 1976. Predictability of sea surface temperature and sea level pressure anomalies in the north Pacific, *J. Phys. Oceanogr.*, 6, 249-266.
- Davis, R.E., 1977. Techniques for statistical analysis of prediction of geophysical fluid systems, *Geophys. Astrophys. Fluid Dyn.*, 8, 245-277.
- Sciremammano, F., 1979. A suggestion for the presentation of correlations and their significance levels, *J. Phys. Oceanogr.*, 9, 1273-1276.


**********

# Example

**********

*Auto- and Cross-Correlation Analysis Uncertainty*

Calculation of auto-correlogram implies zero crossing

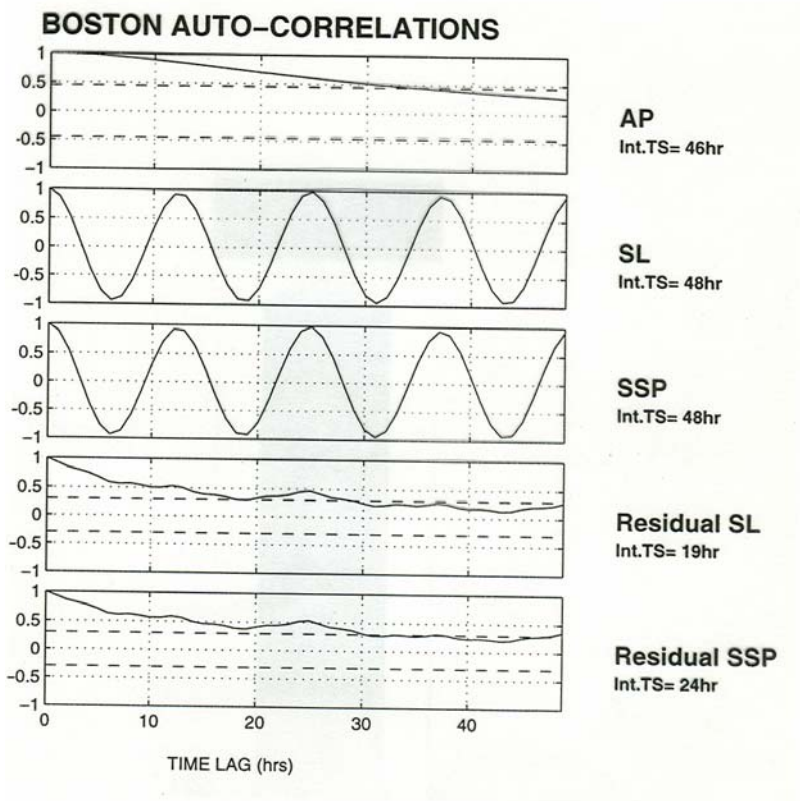


**Figure V.9** Auto-correlation (AC) functions of (top panel down) Portland atmospheric pressure; Boston sea level (SL), Boston synthetic subsurface pressure (SSP), Boston non-tidal residual SL, and Boston non-tidal residual SSP. The 95% confidence intervals for zero AC (dashed) are consistent with integral time scale (Int. TS) estimates, calculated according to (V.117) for the "random series" – i.e., Portland AP, Boston residual SL and Boston residual SSP.

**********

**MAR 670 - Data Analysis Methods EXERCISES**

(due XXX March 2006)

**********

## Exercise 3: Measured Ocean Pressures

You will find the following set of real data records (see Appendix below for more detail) at: /hosts/iselin/data01/users/wsbrown/classes.dir/mar670_2008/CLASS_DIR_2008/EXERCIZES_2008/EX3_REAL_DATA:

(1) An hourly sea level (SL) time series measured at the Boston National Ocean Service (NOS) station between December 1983 and February 1984
(2) An hourly atmospheric pressure (AP) measured by the National Data Buoy Center (NDBC) buoy near Portland, ME between December 1983 and February 1984.
(3) An hourly synthetic subsurface pressure (SSP=AP+SL) record. (For reasons, which will become clearer from your analysis, we have added SL to AP records to produce the SSP record).
(4) Three bottom pressure records from near Hampton, NH.

The filenames of time series (units of decibars = pressure equivalent meters) are:

| **variable** | **file name** |
|---|---|
| Portland atmospheric pressure (AP) | aprsdb.mhr |
| Boston sea level (SL) | bosslp.mhr |
| Boston synthetic subsurface pressure (SSP) | bosssp.mhr |
| Hampton bottom pressure #1 | pktpr1.mhr |
| Hampton bottom pressure #2 | pktpr2.mhr |
| Hampton bottom pressure #3 | pktpr3.mhr |

(a) Use **l_header** and **l_dates** can be used to determine the Gregorian date/time of the start and end times of these data.

(b) Sea level variations reflect, among other things, the effects of both astronomical tidal forcing and atmospheric pressure forcing. To examine the atmospherically-forced ocean response remove the tides from these basic pressure/sea level series to produce "residual series".

- Extract the first 50 hours (**l_piece**) of each of the 3 Boston/Portland pressure series and plot (**Matlab**). Suggested scales: vertical = 1 inch per decibar; horizontal = 10 hours per inch.
    1. *What do you notice visually about the relation of variability in observed sea level to those in the predicted sea level and residual sea level?*
    2. *What do you notice visually about the relation of residual sea level to the atmospheric pressure and subsurface pressure?*

- Plot (**Matlab**) the residual Boston SSP series and Hampton BP series. Suggested scales: vertical = 1 inch per decibar; horizontal = 10 days (240hrs) per inch
  *What do you notice visually about the relation of variability in observed residual pressure records?*

- Compute the statistics of the Boston/Portland pressure series using **l_lstats**.
  *What can you conclude about these series by looking at statistics alone?*
  [Hint: Compare theoretical variance with actual variance and try to explain differences.]

- Compute the statistics of the Boston SSP series and Hampton BP series using **l_lstats.**
  *What can you conclude about these series by looking at statistics alone?*

## Appendix: Hampton, NH Data Description

```
Hampton bottom pressure #1    - pktpr1.mhr
Hampton bottom pressure #2    - pktpr2.mhr
Hampton bottom pressure #3    - pktpr3.mhr
```

These time series are from a 57-day deployment of bottom instrument Picket in 20.7m (68 feet) of water near buoy "N2" off Hampton, NH (42 deg 54.0 min N, 70 deg 47.1 min W;42.900 deg N,70.785 deg W) between 16 December 1983 and 7 February 1984.

Three Paroscientific Inc. pressure sensors, a Sea Data Inc. temperature sensor embedded in the data logger cap, and a Savonius rotor at an elevation of about 1 meter above the bottom were mounted on the instrument frame along with a camera and lights for taking pictures of sediment scouring around the base of the instrument.

The raw data was processed as described in the trip book. The original data series (*.eng) consist of 20340 terms at time intervals of 0.062500481 hours (approx. 3.75 minutes), starting at 16:48:45.604 UTC 16 December 1983 (735952.81266786 Julian hours - referenced to 0000 UTC 1 January 1900). The original pressure series were lowpass-filtered and subsampled to the hourly Ocean Format files (*.hr) that are also provided.

**********
YOUR EXERCISE 0 was to:
**********

(1) Compute the basic statistics of the hourly pressure time series (**l_lstats**)

(2) Plot the 3 different hourly pressure series versus the same time axis on one
page..make a "stackplot".
(Use a modified version of the Matlab plotting routine in
/hosts/davis/data/home/wsbrown/classes.dir/time_series/std_plot.dir/prs_stkplt.dir)

(3) Do a harmonic analysis of the 3 time series (**l_tiharc98**) and save the results as presented on the screen.

NOTE: Use a modified version of the "parameter file", *.PAR, provided with **l_tiharc98** to produce the requested results.