SAMPLING ERRORS

In the real world we can not obtain true statistics, but rather compute **sample statistics**. Our data ensembles are inevitably finite. Before physical conclusions can be drawn from the sample statistics the accuracy with which they approximate true statistics, computed from infinite ensembles, must be determined. The difference between constructs developed from finite ensembles and the analogous true statistic are called **sampling errors**.

1. Estimating statistics

The problem of finding a reasonable approximation of an average can be dealt with by **estimation theory**. Consider true averages, denoted by $\langle \cdot \rangle$, and sample averages, denoted by

$$\{X\} = \frac{1}{N} \sum_{n=1}^{N} X_n \tag{1}$$

which differ from the true averages only in that N is finite. The notation $langle \cdot \rangle$ and $\{\cdot\}$ will retain the same meaning throughout what follows.

In estimation theory one follows a simple general procedure: (a) define a measure of goodness of an estimate (the **beauty principle**), (b) select a general method of estimating (the method is the **estimator**), (c) use statistics to determine the typical quality of the estimator in terms of the beauty principle, and (d) optimize the estimator to maximize the quality. As simple as this sounds, it has four major limitations: (i) there is never an unambiguous beauty principle so that different analysts may well obtain truly different **optimal** estimators, (ii) it is almost never possible to make an exhaustive search of all plausible types of estimators so a true optimum can not be guaranteed, (iii) the measure of quality is statistical and can only be determined when the complete process of estimation is fully described and analyzed, and (iv) determination of quality requires knowledge of some true statistics of the process which are generally no better known than the thing being estimated. As was said before, real statistical analysis requires skill and creativity; it is not cookbook.

Consider now the simplest problem of estimating the mean of X using (1), that is using $\hat{m} = \{X\}$. To examine the performance of this estimator, imagine the same estimation process is applied an infinite number of times in each of which there is a completely new (independent) set of N observations X_n. It is then possible to find statistical descriptions of the resulting hypothetical estimates by imagining many estimates $\hat{m}_k$ and describe their statistics over an infinite ensemble of realizations denoted by different k. Two of the obvious measures of badness are the mean error and the mean square error

$$E_1 = \overline{\hat{m} - \langle X \rangle}, \qquad E_2 = \overline{[\hat{m} - \langle X \rangle]^2} \tag{2}$$

respectively. There are three kinds of average in this: the true average X, the sample average $\{X\}$ involved in $\hat{m}$, and the $\overline{\cdot}$ average over many estimates of the sample average. But since this last average involves an infinite number of realizations, it is really the same as the true average. For example,

$$E_1 = \overline{\{X\} - \langle X \rangle} = \langle \{X\} \rangle - \langle X \rangle, \qquad E_2 = \langle [\{X\} - X]^2 \rangle \tag{3}$$

Now it is a simple matter to substitute the definition of $\{X\}$ from (1) and find

$$E_1 = \left\langle \frac{1}{N} \sum_{n}^{N} X_n \right\rangle - \langle X \rangle = \frac{1}{N} \sum^{N} \langle X \rangle - \langle X \rangle = 0 \tag{4}$$

The estimator (1) is **unbiased**, that is the mean error is zero. Similarly,

$$E_2 = \left\langle \{X\}^2 \right\rangle - \langle X \rangle^2 = N^{-2} \sum^N \sum_{n,m}^N \langle X_n X_m \rangle - \langle X \rangle^2 = N^{-2} \sum_{n,m}^N \langle X'_n X'_m \rangle \tag{5}$$

E_2 depends on the covariance of X', and determining this is no simpler than estimating $\langle X \rangle$ (the process whose errors are being analyzed) but fortunately it is not critical to do this accurately. The simplest case is when the individual observations are independent, in which case E_2 becomes

$$E_2 = N^{-2} \sum_{n,m}^N \delta_{nm} \langle X'^2 \rangle = \langle X'^2 \rangle / N = \mu_2 / N \tag{6}$$

Thus the **root mean square** error decreases as $N^{-1/2}$ where N is the number of independent observations. Since the mean error is zero, this is also the **standard deviation** of the error.

The parallel between accumulation of sampling errors and a random walk is strong. In each case we are concerned with adding up random numbers that may or may not be serially correlated. The rms size of the random walk sum grows as $N^{1/2}$ and, for fixed N and step variance, grows more rapidly if sequential steps are well correlated. The sampling error is a sum normalized by N^{-1} so the rms error decreases as $N^{-1/2}$. For fixed N and sample variance, the error increases as the correlation of successive samples increases.

Now consider estimating the variance $\mu_2 = \langle X'^2 \rangle$. If the mean is known, then X' is easily computed and the variance estimated as $\hat{\mu}_2 = \{X'^2\}$. The results above are then directly applicable and show this estimate to be unbiased with the variance $E_2 = (\mu_4 - \mu_2^2)/N$ where $\mu_4 = \langle X'^4 \rangle$. When the mean is estimated from the same data as the variance is to be found the problem is more complicated and gives us as simple example of **estimation theory**.

Given the sample mean $\hat{m}$ an appropriate estimate of the variance is

$$\hat{\mu}_2 = \frac{A}{N} \sum_n \left[X_n - \hat{m} \right]^2 \tag{7}$$

where the constant A is to adjusted to make a desirable variance estimator. There are two typical measures of what makes a good estimator: bias (the mean error) and mean square error. Substituting $X = X' + m$ and $\hat{m} = m + \epsilon$ where a little calculation shows that error of $\hat{m}$ is $\epsilon = \frac{1}{N} \sum_n^N X'_n$. The bias and mean square error of the variance estimate are then

$$F_1 \equiv \langle \hat{\mu}_2 \rangle - \mu_2 = A \left[\langle X'^2 \rangle - 2 \langle X' \epsilon \rangle + \epsilon^2 \right] - \mu_2 \tag{8}$$

$$F_2 \equiv \langle [\hat{\mu}_2 - \mu_2]^2 \rangle = \left[\frac{A}{N} \right]^2 \left\langle \left[\sum_m X'^2_m - 2\epsilon \sum_m X'_m + \epsilon^2 \right] \right\rangle - 2\mu_2 \hat{\mu}_2 + \mu_2^2 \tag{9}$$

The difficulty here is the error of the mean, ϵ, does not vanish and is correlated the data fluctuations X'_m.

Let us consider the case when the X's are serially independent (i.e. $X'_n X'_m = \delta_{nm} \mu_2$ so that

$$\langle \epsilon^2 \rangle \equiv N^{-2} \sum_{nm}^N \langle X'_n X'_m \rangle = \langle X' \epsilon \rangle \equiv \langle X'_n N^{-1} \sum_m^N X'_m \rangle = \mu_2 / N \tag{10}$$

so that

$$\langle \hat{\mu}_2 \rangle = \mu_2 A \frac{N-1}{N} \qquad F_1 = \mu_2 \left[A \frac{N-1}{N} - 1 \right] \tag{11}$$

In calculating the mean square error F_2 of $\hat{\mu}_2$, the first term on the right of (9) is $\langle\hat{\mu}_2^2\rangle$ and evaluating it is nontrivial.

$$\langle\hat{\mu}_2^2\rangle = A^2N^{-2}\left[\sum_{nm}\langle X_n'^2 X_m'^2\rangle - 2N^{-1}\sum_{n}\sum_{kl}\langle X_n'^2 X_k' X_l'\rangle + N^{-2}\sum_{nm}\sum_{kl}\langle X_n' X_m' X_k' X_l'\rangle\right] \tag{12}$$

Noting that the X' with different indices are independent, careful bookkeeping leads to

$$\hat{\mu}_2^2 = A^2N^{-2}[N(N-1)\mu_2^2+N\mu_4]-2\,A^2N^{-3}[N(N-1)\mu_2^2+N\mu_4]+A^2N^{-4}[3N(N-1)\mu_2^2+N\mu_4] \tag{13}$$

It is very useful to be able to carry out the kind of calculation that leads to (13) and the student is encouraged to understand the steps. Combining (11) and (13), the mean square error in (9) can be evaluated:

$$F_2 = A^2\frac{N-1}{N^3}\Big[(N^2-2N+3)\mu_2^2+(N-1)\mu_4\Big] - 2\,A\,\frac{N-1}{N}\mu_2^2+\mu_2^2 \tag{14}$$

For the case of Gaussian X it is easy to show that $\mu_4 = 3\mu_2^2$ and

$$F_2 = A^2\mu_2^2\frac{(N-1)(N+1)}{N^2} - 2A\frac{(N-1)}{N}\mu_2^2+\mu_2^2 \tag{15}$$

We are now in a position to ask "What is the best value for A to use in estimating variance?" The most widely used choice is obtained by requiring the estimate be unbiased, that is $\langle\hat{\mu}_2\rangle = \mu_2$ or $F_1 = 0$; this corresponds to $A = \frac{N}{N-1}$. The "intuitive" variance estimator $\hat{\mu}_2 = \{(X - \{X\})^2\}$, corresponding to $A = 1$, is little used because people seem to insist on using the zero-bias estimator. This is usually more a matter of style than substance for the following reason. The bias is much smaller than the spread for any reasonable estimator. According to (11) the choice $A = 1$, for example, makes $\langle\hat{\mu}_2\rangle - \mu_2$ of order $1/N$ whereas from (15) F_2 is also $O(1/N)$ so that the rms error $\langle(\hat{\mu}_2 - \mu_2)^2\rangle^{1/2}$ is order $1/\sqrt{N}$. Driving the bias to zero increases the mean square error F_2 compared with its minimum value. Indeed the minimum square error can be obtained by adjusting A to minimize F_2 in (15) giving $A = \frac{N}{N+1}$. The minimum mean square error and the associated mean error (using $A = N/N+1$) are

$$F_2 = \frac{2\mu_2^2}{N}+1 \qquad\qquad F_1 = -\frac{2\mu_2}{N}+1 \tag{16}$$

Given the results for the two beauty principles zero bias and minimum mean square error, there is little reason not to simply use the intuitive estimator based on $A = 1$.

It is important to step back and take note of a few general facts. First, the process of estimating sampling errors involves us in an endless chain. To estimate the uncertainty in the mean we need to know the variance. To estimate the uncertainty in a sample variance we need to know higher moments. This chain can only be broken with a closure hypothesis that asserts specific connections between moments. Note that the Central Limit tendency associated with the number of terms in sample sums going to infinity does not help. One needs assumptions about the distribution of individual observations, not the sample average. Second, sample errors all depend on true statistics which, of course, will never be known. Thus describing sample errors always involves a "what if" story like "If the true statistics were ...... then the sample errors would be ..... ." There is fundamentally no way around this and a good analysis is one with a plausible "If the true statistics were ...." story that is backed up with observations where possible. Third, the particular construct (formula) used to approximate true statistics can be "tuned" to be optimal according to various different beauty principles. The particular goodness criterion should depend on the problem but samll bias and small mean square error are pretty common and useful criteria. The rule that one should divide by the number of samples minus the number of constraints (a) applies only to zero bias estimators and (b) makes assumptions about the distribution of variables. In general the error standard deviation of an estimate is a larger order in N than the mean error (bias). Some of us think it makes more sense to minimize the large error than the small one. Fourth, precision is neither possible nor necessary. For example, the error of an estimated mean, $\hat{m} - m$, is $\sqrt{\mu_2/N}$. All sensible variance estimators give rms values of the error $\hat{\mu}_2 - \mu_2$ of order $\mu_2/\sqrt{N}$ which would introduce a fractional error of $O(1/\sqrt{N})$ into the estimate of typical $\hat{m} - m$. Messing around with the variance estimator makes an $O(1/N)$ correction to the estimate of μ_2. Thus if N is large you can bootstrap your way to good answers and the precise formulas used won't matter. If N is not large then no statistical question can be answered well and fooling around with $O(1/N)$ corrections (like zero bias vs. minimum mean square error) are not going to fix up much of anything important.

2. Confidence limits

Confidence limits describe the range of true values consistent with the observation. Once the distribution function of sample values, here $\hat{m}$ or $\hat{\mu}_2$, are known for each true value, here m or μ_2, it is simple to place confidence limits on the estimate. For the case of mean values, for example, confidence limits describe the range of true means, m, which are likely to have led to the observation $\hat{m} = m_{OBS}$. Specifically, the upper $100 \cdot (1 - 2\epsilon)$ confidence limit is the largest true value for which there is an ϵ chance of getting an estimate as small, or smaller than, that observed. The lower confidence limit is the smallest true value that would give an observation as large as obtained with probability ϵ.

To make this precise, consider the general case of estimating a parameter m with an estimator $\hat{m}$ (the process by which estimates are made) which produced the value m_{OBS}. Let $F_{\hat{m}}(x|m)$ be the distribution function of estimates $\hat{m}$ when the true parameter is m (*i.e.* the probability that an observation will be less that x is $F_{\hat{m}}(x|m)$). The lower confidence limit m_L and the upper limit m_U are then defined by

$$F_{\hat{m}}(m_{OBS}|m_L) = 1 - \epsilon, \qquad F_{\hat{m}}(m_{OBS}|m_U) = \epsilon \tag{17}$$

Simply put, these say that for small ϵ very few observations are as large as m_{OBS} if $m < m_L$ and that almost all observations are greater than m_{OBS} if $m > m_U$.

First consider confidence limits for $\hat{m} = \{X\}$ By combining (4) and (6) with the results of the Central Limit Theorem the probability of an estimate being in error by any specified amount is easily found. The Central Limit Theorem says that if N is large then the pdf of (1) will be Gaussian. In practice the restriction on large N is not severe for two reasons: (i) except at the tails of low probability, the pdf of a finite sum quickly approaches the Gaussian form for fairly small N, and (ii) if the variable X is approximately normally distributed then this approach is accelerated (if X is normally distributed then so is $\{X\}$ for all N). The normal distribution is specified by two parameters and (4) says the mean error vanishes and (6) says that the error variance is $\sigma^2 = \langle X'^2 \rangle / N$. The distribution function $F_{\hat{m}}(x|m)$ is

$$F_{\hat{m}}(x + m|m) = \int_{-\infty}^{x} dy\, [2\pi\sigma^2]^{-1/2} \, \exp[-y^2/2\sigma^2] \; = \; Q(x/\sigma) \tag{18}$$

$$Q(x) \; = \; \int_{-\infty}^{x} dy\, [2\pi]^{-1/2} \, \exp[-y^2/2] \tag{19}$$

Some values of $Q(x)$ are tabulated here. Note that $Q(-x) = 1 - Q(x)$.

x	0.0	1.0	1.28	1.64	1.96	2.33	2.58	3.09
Q(x)	0.5	0.84	0.90	0.95	0.975	0.99	0.995	0.999

Note that a 2σ error occurs about 4% of the time and a 3σ error occurs only about 0.2%. Remember that $Q(x)$ is the probability of a deviation from the mean not exceeding x so that the probability of the absolute value exceeding $x > 0$ is $2(1 - Q)$.

Because the shape of $F_{\hat{m}}(x|m)$ does not depend on m it is easy to find the confidence limits. For example, from the table we see there is less than a 1% chance of obtaining an observation m_{OBS} if $m < m_{OBS} - 2.33\sigma$ and less than a 1% chance of obtaining m_{OBS} if $m > m_{OBS} + 2.33\sigma$; thus $\hat{m} \pm 2.33\sigma$ are the 98% confidence limits. These are sometimes called the 2% confidence limits (they are not the 99% or 1% limits). The discussion of the confidence limits of the sample mean may seem cumbersome, but this complexity is necessary to describe the general concept of confidence limits and to find them in more complicated examples.

Now consider placing confidence limits on an estimate of a variance when X is normally distributed and the mean is known and has been subtracted. The estimate $\hat{\mu}_2$ is then approximately the sum of N squares of identically and normally distributed variables with zero mean and variance μ_2/N. Formally

$$\frac{N\,\hat{\mu}_2}{\mu_2} = \sum_{m=1}^{N} x^2 \tag{20}$$

where the x are normally distributed, $\langle x \rangle = 0$, and $\langle x^2 \rangle = 1$. The sum of the N such squares of Gaussian variables is called a **CHI squared variable with N degrees of freedom**. For large N this becomes a normally distributed variable with mean N and variance $2N$ (since $\langle x^4 \rangle = 3$). The distribution of χ^2_N differs strongly from a Gaussian for N up to $O(10)$.

Let the distribution function of a χ^2_N variable be defined so that $F_{\chi^2_N}(y)$ is the probability that $\chi^2_N < y$. This function is given in many math tables for various N and is tabulated below for $N = 10$. Also shown is the value of χ^2_{10} corresponding to F if χ^2_N were normally distributed with mean N and variance $2N$. You will note that the more common occurrences correspond well between the exact distribution and the normal distribution, but the tails of the distribution are significantly different (the normal distribution indicates a 1.5% probability of a negative sum of squares).

$F(y)$	0.01	0.05	0.10	0.90	0.95	0.99
y, χ squared	2.56	3.94	4.86	15.99	18.31	23.21
y, normal	-.42	2.67	4.28	15.72	17.33	20.42

Since $N\,\hat{\mu}/\mu$ is distributed like χ^2_N, it is simple to find the confidence limits for $\hat{\mu}$ from (17). Since $N\hat{\mu}/\mu > 18.3$ only 5% of the time, the lower 90% confidence limit is $\mu_L = (10/18.3 = 0.55)\mu_{OBS}$. Similarly $N\hat{\mu}/\mu$ will be smaller than 3.94 only 5% of the time so the upper limit is $\mu_U = (10/3.94 = 2.54)\mu_{OBS}$. The asymmetry of these limits is the result of the convoluted question they answer: “How far from μ_{OBS} could μ be and still be apt to give the observed value?” Since the variability of $\hat{\mu}$ increases with μ, the confidence limit in the positive error direction will be farther from the sample than is the negative error limit. The asymmetry of the chi-squared probability density actually counteracts the asymmetry caused by the increase in variance of with increasing true value. To see this the student should repeat the confidence limit calculation using the symmetric Gaussian form for Q.

3. Correlated observations

The discussion above provides the method for deducing sampling errors for statistical estimates when the observations are independent. In this case the "size" of the sample is simply N, the number of independent observations. Now consider the more likely case that the observations are correlated. This happens when the observations are more closely spaced, in time and/or space, than the appropriate correlation scale.

Consider first the estimator $\hat{m}$ of the mean m in (1). This will be unbiased regardless of the correlation of observations. The mean square error (5), however, depends on the covariance $\langle X'_n X'_m \rangle$. In arriving at (6) it was assumed that this vanishes unless $n = m$. Suppose instead that X has stationary statistics and the samples X_n are equally spaced so that $\langle X'_n X'_m \rangle = \langle X'^2 \rangle \rho(n - m)$ where $\rho(n - m)$ is the correlation of observations separated by the interval $n - m$. A little manipulation of (5) then shows that

$$E_2 = \mu_2 N^{-2} \sum_n^N \sum_m^N \langle X'_n X'_m \rangle = \mu_2 N^{-1} \sum_{n=-N}^{n=N} \left[1 - \frac{|n|}{N} \right] \rho(n) \tag{21}$$

If the observations are uncorrelated then the only nonzero term in the sum is $\rho(0) = 1$ and (5) is recovered. Otherwise the error may be larger or smaller, depending on the sum over $\rho(n)$. Correlation of observations means that the real size of the sample is generally less than the number of observations N.

The extreme of observation correlation comes from a continuous time series. For this case the mean would be estimated as

$$\frac{1}{T} \int_0^T dt\, X(t) \tag{22}$$

which will be unbiased. The mean square error would be

$$E_2 = \frac{1}{T^2} \int_0^T \int_0^T dt_1\, dt_2 \langle X'(t_1) X'(t_2) \rangle = \mu_2 \frac{1}{T} \int_{-T}^T dt \left[1 - \frac{|t|}{T} \right] \rho(t) \tag{23}$$

where $\rho(t)$ is the serial correlation of two observations separated by time t.

Equations (21–23) tell how to find the equivalent number of independent samples, N_E, provided by a record of correlated data. These say

$$E_2 = \mu_2 / N_E, \quad N_E = \frac{N}{\sum_{-N}^{N} [1 - |n|/N] \rho(n)} \quad \text{or} \quad N_E = \frac{T}{\int_{-T}^{T} dt\, [1 - |t|/T]\, \rho(t)} \tag{24}$$

In the limits that N and T go to infinity (that is are much larger than the value required to make the correlation ρ vanish), the equivalent degrees of freedom are

$$N_E = \frac{N}{\sum_{-N}^{N} \rho(n)} \qquad \text{or} \qquad N_E = \frac{T}{\int_{-T}^{T} dt\, \rho(t)} \tag{25}$$

It usually takes several samples or a specified continuous sample time to gain as much information from a serially correlated series as is provided by a single independent sample. It is, however, possible that the serial correlation of observations will help averaging and N_E will be larger than N. If, for example, the process were narrowband noise centered around a period of T, the average 4 equally spaced samples at t_0, $t_0 + T/4$, $t_0 + T/2$ and $t_0 + 3T/4$ would be very much better than a much larger sample taken at irreularly spaced intervals. If the process had zero bandwidth, the sample mean would be exact.

In order to find the sampling errors of higher moments it is necessary to estimate even higher order moments of the process, as we did in going from (14) to (15). If it is assumed that the process is joint-normally distributed then results are obtained from the two-time correlation, or for stationary processes from the time-lagged correlation.

Much of oceanographic analysis concerns the relation between different variables. Consequently, it is worth considering the errors in a sample covariance. Let the sample covariance be

$$\hat{C}_{XY} = \{XY\} = \frac{1}{T}\int_0^T X(t)\,Y(t)\,dt \tag{26}$$

If C_{XY} is the true covariance then $Y = \alpha X + Z$ where Z is the unexplained part of Y and $\alpha = C_{XY}/\langle XX\rangle$. Hence

$$\hat{C}_{XY} = \frac{\{X^2\}}{\langle X^2\rangle}C_{XY} \; + \; \frac{1}{T}\int_0^T X(t)Z(t)dt\;. \tag{27}$$

and a measure of the mean square error is

$$E_2 = (\hat{C}_{XY-}\frac{\{X^2\}}{\langle X^2\rangle}C_{XY})^2 = T^{-2}\int ds\int dt\langle X(t)X(s)Z(t)Z(s)\rangle \tag{28}$$

While we know that X and Z are uncorrelated, they may not be independent and we can not proceed further. Thus sampling error estimates must be based on the model that X and Z are independent. It is typical that sampling error estimates are based on such implicit assumptions. If X and Z are independent

$$E_2 = T^{-2}\int ds\int dt C_{XX}(t-s)C_{ZZ}(t-s) \approx \frac{1}{T}\int_{-\infty}^{\infty} dt C_{XX}(t)C_{ZZ}(t) \tag{29}$$

The infinite integral of the product $C_{XX}C_{ZZ}$ has the units of time and represents the extra record length required to obtain another effectively independent sample. Thus in the spirit of the “effective degrees of freedom” in (24–25), for a sample covariance

$$E_2 = \frac{\langle XX\rangle\langle ZZ\rangle}{N_E}, \qquad N_E = \frac{N}{\sum_{-\infty}^{\infty}\rho_{XX}(k)\rho_{ZZ}(k)}, \quad \text{or} \quad N_E = \frac{T}{\int_{-\infty}^{\infty}\rho_{XX}(t)\rho_{ZZ}(t)dt}. \tag{30}$$

where ρ_{XX} and ρ_{ZZ} are time lagged correlations. Recall that Z is the part of Y which is uncorrelated with X (and assumed independent of it) so that ZZ is the unexplained variance of Y.