Lecture 3: Statistical sampling uncertainty

Winter 2015

3.1 Central limit theorem (CLT)

Let $X_1, ..., X_N$ be a sequence of N independent identically-distributed (IID) random variables each with mean μ and standard deviation σ. Then

$$Z_N = \frac{X_1 + X_2 ... + X_N - N\mu}{\sigma\sqrt{N}} \to n(0,1) \qquad \text{as } N \to \infty$$

To be precise, the arrow means that the CDF of the left hand side converges pointwise to the CDF of the 'standard normal' distribution on the right hand side. An alternative statement of the CLT is that

$$\frac{X_1 + X_2 ... + X_N}{N} \sim n(\mu, \sigma/\sqrt{N}) \qquad \text{as } N \to \infty \tag{3.1.1}$$

where $\sim$ denotes asymptotic convergence; that is the ratio of the CDF of tM_N to the normal distribution on the right hand side tends to 1 for large N. That is, regardless of the distribution of the X_k, given enough samples, their sample mean is approximately normally distributed with mean μ and standard deviation $\sigma_m = \sigma/\sqrt{N}$. For instance, for large N, the mean of N Bernoulli random variables has an approximately normally-distributed CDF with mean p and standard deviation $\sqrt{p(1-p)/N}$. More generally, other quantities such as variances, trends, etc., also tend to have normal distributions even if the underlying data are not normally-distributed. This explains why Gaussian statistics work surprisingly well for many purposes.

Corollary: The product of N IID RVs will asymptote to a lognormal distribution as $N \to \infty$.

The Central Limit Theorem Matlab example on the class web page shows the results of 100 trials of averaging $N = 20$ Bernoulli RVs with $p = 0.3$ (note that a Bernoulli RV is highly non-Gaussian!). The CLT tells us that for large N, this average $\overline{x}$ for each trial is approximately normally distributed with mean $\overline{X} = p = 0.3$ and stdev $\sigma_m = \sqrt{p(1-p)/N} \approx 0.1$; the Matlab plot (Fig. 1) shows $N = 20$ is large enough to make this a good approximation, though the histogram of $\overline{x}$ suggests a slight residual skew toward the right.

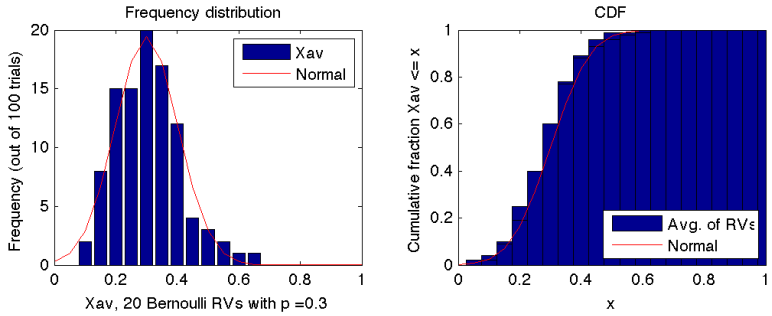

Figure 1: Histogram and empirical CDF of $\overline{x}$, compared to CLT prediction

3.2 Insights from the proof of the CLT

The proof of the CLT (see probability textbooks) hinges on use of the *moment generating function* $E[e^{tX}] = \Sigma_{n=0}^{\infty} t^n E[X^n]/n!$, which can be regarded as a Fourier-like transform of the PDF of X. One shows that the moment generating function for Z_N approaches the moment generating function $\exp(t^2/2)$ of a standard normal distribution. This argument can easily be loosely generalized to weighted averages of arbitrary (not identically-distributed) random variables, as long as no individual weight dominates the average; however, in this case, an effective sample size may need to be used in place of N in the $N^{-1/2}$ factor (e. g. Bretherton et al. 1999 it J. Climate). Thus, **it is generally true that weighted averages of large numbers of independent random variables (e. g. sample means and variances) have approximately normal distributions.**

3.3 Statistical uncertainty

The CLT gives us a basis for assigning an uncertainty when using an N-independent-sample mean $\overline{x}$ of a random variable X as an estimator for its true mean $\overline{X}$. **It is important that all the samples can be reasonably assumed to be independent and have the same probability distribution!** In the above Matlab example. each trial of 20 samples of X gives an estimate $\overline{x}$ of the true mean of the distribution (0.3). Fig. 1 shows that $\overline{x}$ ranges from 0.1 to 0.65 over the 100 trials (i. e. the $\overline{X}$ estimated from each trial is rather uncertain). More precisely, the CLT suggests that given a single trial of $N \gg 1$ samples of a RV with true mean $\overline{X}$ and true standard deviation σ_X, which yields a sample mean $\overline{x}$ and a sample standard deviation $\sigma[x] \approx \sigma_X$:

$$\overline{x} - \overline{X} \approx n\left(0, \frac{\sigma_X}{N^{1/2}}\right) \tag{3.3.1}$$

We can turn this around into an equation for $\overline{X}$ given $\overline{x}$. One wrinkle is that we don't know σ_X so it is estimated using the sample standard deviation (this

is only a minor additional uncertainty for large N). Noting also that the unit normal distribution is an even function of x:

$$\overline{X} \approx \overline{x} + n(0, \sigma_m), \qquad \sigma_m = \sigma(x)/N^{1/2} \tag{3.3.2}$$

That is, we can estimate a ± 1 standard deviation uncertainty of the true mean of X from the finite sample as:

$$\overline{X} = \overline{x} \pm \sigma_m \tag{3.3.3}$$

For the Bernoulli example, the sample standard deviation will scatter around the true standard deviation of 0.1, so we'd have to average across more than $N = 100$ independent samples to reduce the $\pm 1\sigma$ uncertainty in the estimated $\overline{X} = p$ to less than 0.01.

3.4 Normally distributed estimators and confidence intervals

We have argued that the PDF of the $\overline{X}$ given the observed sample mean $\overline{x}$ is approximately $n(\overline{x}, \sigma_m)$, then the normal distribution tells us the probability $\overline{X}$ lies within any given range (Fig. 2).

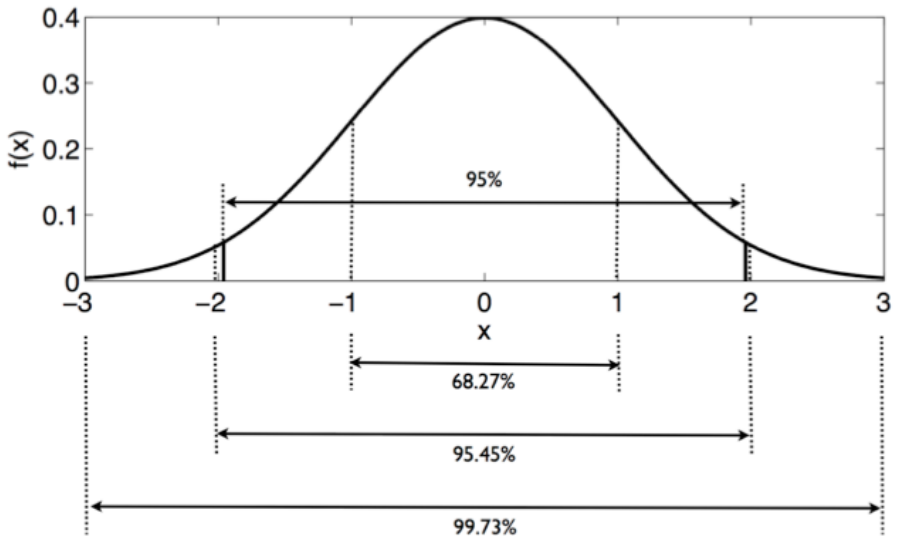

Figure 2: Probability within selected ranges of a unit normal distribution.

About (2/3, 95%, 99.7%) of the time, $\overline{X}$ will lie within (1, 2, 3)σ_m of $\overline{x}$. This allows us to estimate a **confidence interval** for the true mean of $\overline{X}$; for instance, we say that

$$\overline{x} + 2\sigma_m < \overline{X} < \overline{x} + 2\sigma_m \text{ with } 95\% \text{ confidence} \tag{3.4.1}$$

Such confidence intervals should be used with caution since they are sensitive to (a) deviations of the estimator from normality, especially in the tails of its PDF, and (b) sampling uncertainties if N is small.

The 95% confidence interval is a common choice; other common choices are 90% or 99%, called *very likely* and *virtually certain* in Intergovernmental Panel on Climate Change reports. These correspond to $\pm 1.6\sigma$ or $\pm 2.6\sigma$ for a normally-distributed estimator.

3.5 Serial correlation and effective sample size

Often, successive data samples are not independent. For instance, the daily-maximum temperature measured at Red Square in UW will be positively correlated between successive days, but has little correlation between successive weeks. Thus, each new sample has less new information about the true distribution of the underlying random variable (daily max temperature in this example) than if successive samples were statistically independent. After removing obvious trends or periodicities, many forms of data can be approximated as 'red noise' or first-order Markov random processes (to be discussed in later lectures) which can be characterized by the **lag-1 autocorrelation** r, defined as the correlation coefficient between successive data samples. Given r, an **effective sample size** N^* can be defined for use in uncertainty estimates (e. g. Bretherton et al. 1999 *J. Climate*).

Effective sample size for estimating uncertainty of a mean

$$\text{Sample mean: } N^* = N\frac{1-r}{1+r}; \qquad \sigma_m = \sigma(x)/N^{*1/2} \tag{3.5.1}$$

If $r = 0.5$ (fairly strong serial correlation), $N^* = N/3$. That is, it takes three times as many samples to achieve the same level of uncertainty about the mean of the underlying random process as if the samples were statistically independent. On the other hand, if $|r| < 0.2$ the effect of serial correlation is modest ($N^* \approx N$).

Fig. 3 shows examples of $N = 30$ serially correlated samples of a 'standard' normal distribution with mean zero and standard deviation 1, with different lag-1 autocorrelations r. In each case, the sample mean is shown as the red dashed line and the magenta lines $\overline{x} \pm \sigma(x)/N^{*1/2}$ give a ± 1 standard deviation uncertainty range for the true mean of the distribution, which is really 0 (the horizontal black line).

In the case with strong positive autocorrelation $r = 0.7$, successive samples are clearly similar, reducing $N^* \approx N/6$ and widening the uncertainty range by a factor of nearly 2.5 compared to the case $r = 0$. In the case $r = -0.5$, successive samples are anticorrelated and their fluctuations about the true mean tend to cancel out. Thus $N^* \approx 3N$ is *larger* than N, and the uncertainty of the mean is only 60% as large as if the samples were uncorrelated. In each case shown, the true mean is bracketed by the $\pm 1\sigma_m$ uncertainty range; given the statistics of a Gaussian distribution this would be expected to happen about 2/3 of the time.

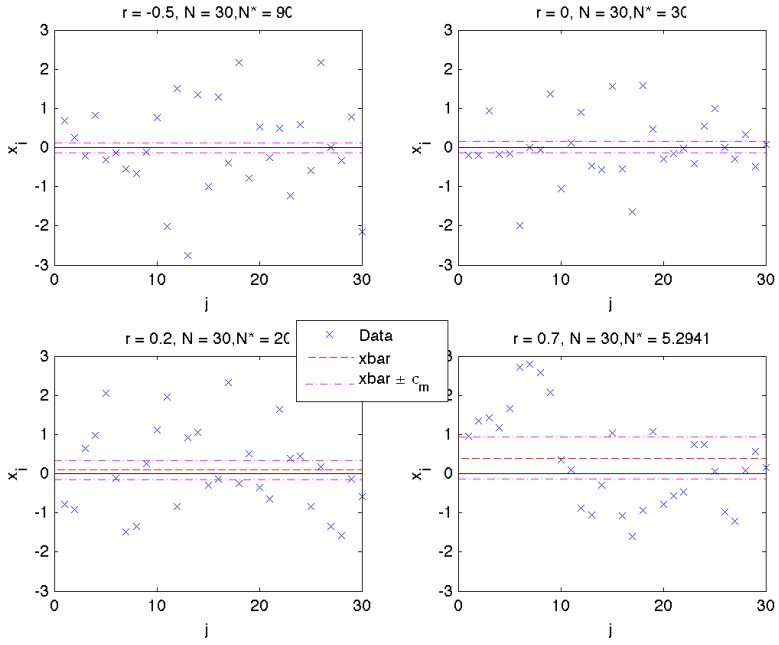

Figure 3: Random sets of $N = 30$ samples from a standard normal distribution with different lag-1 autocorrelations r, and the $\pm 1\sigma$ uncertainty range (magenta dashed lines) in estimating the true mean of 0 (black line) from the sample mean (red dashed line), based on the effective sample size N^*.

3.6 Confidence intervals for correlations

Another important application of confidence intervals is to the correlation coefficient between two variables. Given N independent samples X_i and Y_i of RVs with true correlation coefficient R_{XY}, what will be the PDF of their sample correlation coefficient R_N? For reasonably large values $N > 10$ this can be estimated using the **Fisher Transformation**:

$$z = F(r) = \tanh^{-1}(r); \qquad r = \tanh(z) \tag{3.6.1}$$

(http://en.wikipedia.org/wiki/Fisher_transformation). For small $|r|, z \approx r$, but as $r \to \pm 1, z \to \infty$. Letting $Z_N = F(R_N)$ and $Z_{XY} = F(R_{XY})$, one can show

$$Z_N \sim n(Z_{XY}, \sigma_N), \qquad \sigma_N = (N-3)^{-1/2}, \qquad N \gg 1. \tag{3.6.2}$$

This formula is typically derived assuming Gaussian RVs, but application of the CLT to the sample covariance and variance between X and Y implies that it will work for arbitrary PDFs if N is sufficiently large.

Thus, suppose the observed correlation coefficient between N samples of X and Y is r, with Fisher transform $z = F(r)$. Then

$$Z_{xy} \sim n(z, \sigma_N), \qquad \sigma_N = (N-3)^{-1/2} \tag{3.6.3}$$

with a 95% confidence interval

$$z - 2\sigma_N < Z_{xy} < z + 2\sigma_N \tag{3.6.4}$$

Taking the inverse Fisher transform gives the corresponding confidence interval for R_{XY}.

For instance, if $N = 30$ independent samples of two variables give a sample correlation coefficient $r = 0.7$, then $z = F(r) = 0.87, \sigma_N = 27^{-1/2} = 0.19$, and the 95% confidence interval is $0.48 < Z_{XY} < 1.25$, or $0.45 < R_{XY} < 0.85$; note this is slightly asymmetric about the sample r.

Note that if $R_{XY} = 0$, Fisher's transformation implies that $R_N \sim n(0, \sigma_N)$ for large $N \geq 10$, which gives confidence intervals on how large we expect the sample correlation coefficient to be if the variables are actually uncorrelated. For arbitrary N but uncorrelated and Gaussian-distributed variables, $(N-2)^{1/2} R_N/(1-R_N^2)^{1/2}$ can be shown to have at-distribution with $N-2$ DOF (http://en.wikipedia.org/wiki/Student's_t-distribution); this provides more exact confidence intervals for this case for small $N < 10$, but gives essentially the same results as the easier-to-use Fisher transformation for larger N.

Effective sample size for the correlation coefficient of serially-correlated data A dataset is statistically **stationary** if its statistical characteristics (mean, variance) do not systematically change across the samples, The ESS can be calculated for two stationary AR1 random variables X_1 and X_2 with respective estimated lag-1 autocorrelations r_1 and r_2 (Bretherton et al. 1999 *J. Climate*):

$$\text{Correlation coefficient: } N^* = N \frac{1 - r_1 r_2}{1 + r_1 r_2} \tag{3.6.5}$$

If *either* $|r_1|$ or $|r_2|$ is less than 0.2, the effect of serial correlation is small ($N^* \approx N$), e. g. if $r_1 = 0.9$ but $r_2 = 0$, $N^* = N$. At the heart of the correlation coefficient is the covariance $X_1' X_2'$, whose serial correlation requires serial correlation of *both* X_1 and X_2.

If either or both variables have clear trends vs,. sample index, these must be removed before this analysis, e. g. via linear regression (Lecture 5). If one of the variables, say X_2, is a space or time index with a fixed increment between successive samples, then we detrend X_1 and set $r_2 = 1$ in (3.6.5).