

Lecture 3: Statistical sampling uncertainty

©Christopher S. Bretherton

Winter 2014

3.1 Central limit theorem (CLT)

Let X_1, \dots, X_N be a sequence of N independent identically-distributed (IID) random variables each with mean μ and variance σ^2 . Then

$$\frac{X_1 + X_2 \dots + X_N - N\mu}{\sigma\sqrt{N}} \rightarrow n(0, 1) \quad \text{as } N \rightarrow \infty$$

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

$$\frac{X_1 + X_2 \dots + X_N}{N} \sim n(\mu, \sigma/\sqrt{N}) \quad \text{as } N \rightarrow \infty \quad (3.1.1)$$

where \sim denotes asymptotic convergence; that is the ratio of the CDF of the left hand side 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 σ/\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 $N \rightarrow \infty$.

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

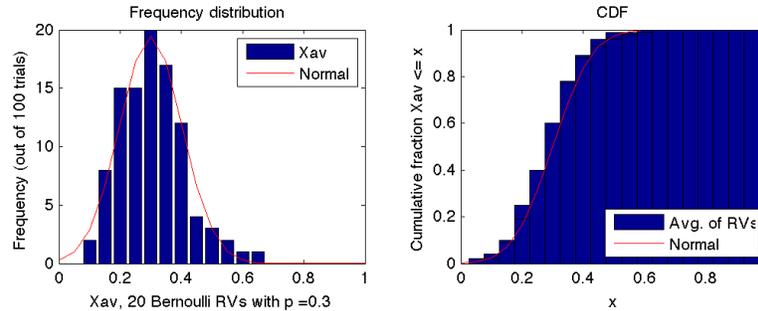


Figure 1: Histogram and empirical CDF of \bar{x} , compared to CLT prediction

3.2 Statistical uncertainty

In the above Matlab example, each trial of 20 samples of X gives an estimate \bar{x} of the true mean of the distribution (0.3). Fig. 1 shows that \bar{x} ranges from 0.1 to 0.65 (i. e. the \bar{x} estimated from each trial is quite uncertain). More precisely, the CLT suggests that given a single trial of $N \gg 1$ samples of a RV with true mean \bar{X} and true standard deviation σ_X , which yields a sample mean \bar{x} and a sample standard deviation $\sigma[x] \approx \sigma_X$:

$$\bar{X} - \bar{x} \approx n \left(0, \frac{\sigma_X}{N^{1/2}} \right) \approx n \left(0, \frac{\sigma[x]}{N^{1/2}} \right) \quad (3.2.1)$$

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

$$\bar{X} = \bar{x} \pm \frac{\sigma[x]}{\sqrt{N}}, \quad (3.2.2)$$

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 $\bar{X} = p$ to less than 0.01.

3.3 Normally distributed estimators

If we are trying to estimate some \bar{X} which we have reason to believe will be approximately $n(\bar{x}, \sigma_m[x])$, then the normal distribution tells us the probability \bar{X} lies within any given range (Fig. 2). About (2/3, 95%, 99.7%) of the time, \bar{X} will lie within (1, 2, 3) $\sigma_m[x]$ of \bar{x} . These estimates need to be used with caution since they are sensitive to (a) deviations of the estimator from normality, and (b) sampling uncertainties in the estimate of $\sigma_m[x]$.

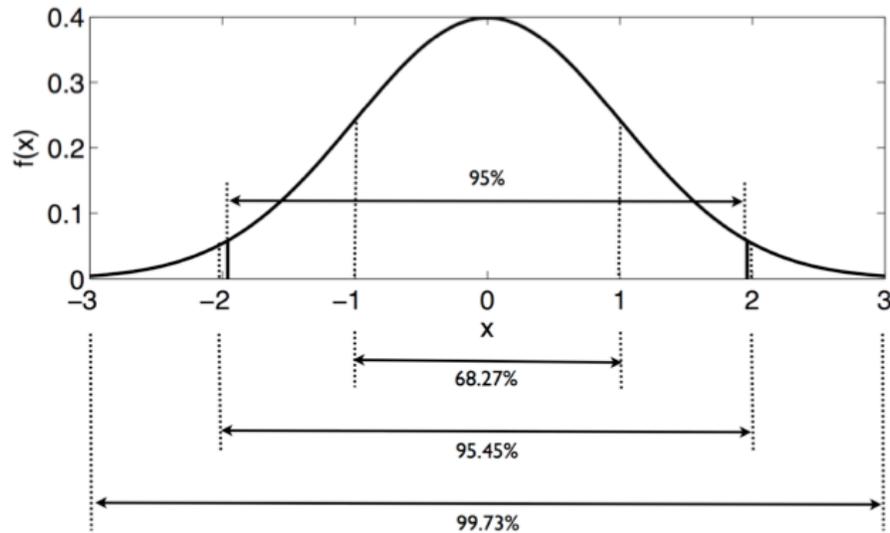


Figure 2: Probability within selected ranges of a unit normal distribution.

3.4 Serially-correlated data 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.

Effective sample size for estimating uncertainty of a mean

$$\text{Sample mean: } N^* = N \frac{1-r}{1+r} \quad (3.4.1)$$

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 $\bar{x} \pm \sigma_x/\sqrt{N^*}$ give a ± 1 standard deviation uncertainty range of \bar{x} as an estimate of the true mean of the distribution, which is 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.

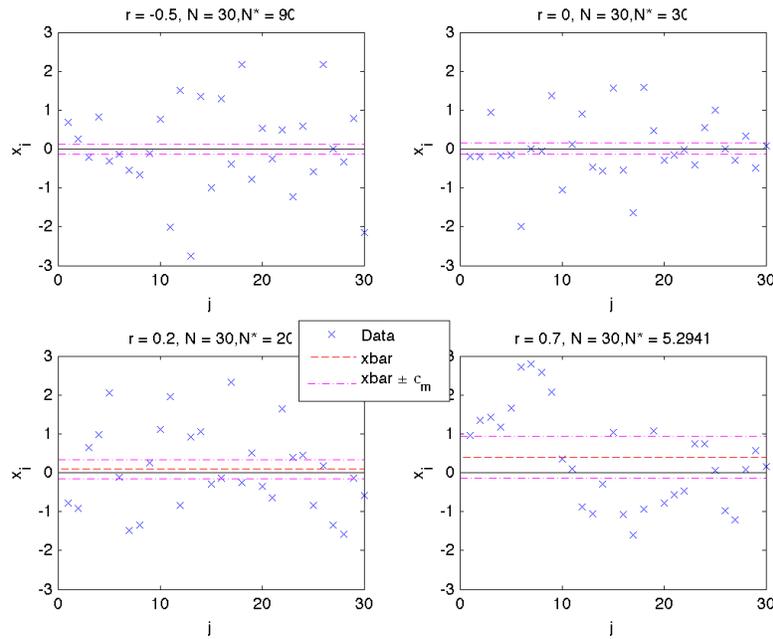


Figure 3: Random sets of $N = 30$ samples from a standard normal distribution with different lag-1 autocorrelations r , and the $\pm 1\sigma$ uncertainty range (magenta dashed lines) in estimating the true mean of 0 (black line) from the sample mean (red dashed line), based on the effective sample size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

Effective sample size to use for estimating uncertainty of the correlation coefficient between two random variables X_1 and X_2 with respective

estimated lag-1 autocorrelations r_1 and r_2 ,

$$\text{Correlation coefficient: } N^* = N \frac{1 - r_1 r_2}{1 + r_1 r_2} \quad (3.4.2)$$

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 product $X_1 X_2$, whose serial correlation requires serial correlation of *both* X_1 and X_2 .

3.5 Estimating uncertainty using multiple trials or resampling

A real dataset may consist of a single time series or trial. If the time series is **stationary**, i. e. its statistical characteristics do not systematically change across the samples, then a useful strategy to estimate statistical uncertainty is to break the time series up into chunks which are long enough not to be strongly correlated with each other, perform the data analysis on each chunk separately, and compare the results for different chunks.

In testing for the uncertainty of trends, it can also be useful to resample the data (known in statistics as **bootstrapping**, e. g. by making a new time series in which the data or blocks of data have been randomly reordered, a procedure which would remove the trend. We perform the same estimate of a trend to these resampled time series as was used on the original time series. The standard deviation of these trends around zero gives a guess at the standard deviation of the original trend about its estimated value.