More on Confidence Intervals and Maximum Likelihood Estimation

Renato Lo Cigno

Simulation and Performance Evaluation 2014-15

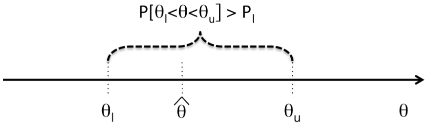
- Confidence Intervals (CI) are fundamental in measure-based analysis
- If possible they are even more important in simulations
  - When do I finish a simulation?
  - Once I have "numbers" from a simulation how much I can trust them?
- Even more than measures results of simulations can be correlated
- Care must be put to understand the correlation structure and to derive independent measures to estimate the reliability of results

Recall what is a Confidence Interval

- The **confidence interval** around the estimated value $\hat{\theta}$ is the interval (θ_l, θ_u) such that the true value θ falls within the interval (θ_l, θ_u) with a given probability $\mathbf{P}_l$ that we call the **confidence level**

$$\mathbf{P}[\theta_l \leq \theta \leq \theta_u \,|\, \hat{\theta}] \geq \mathbf{P}_l$$

- Often (θ_l, θ_u) is expressed as a fraction (percentage) of $\hat{\theta}$ around $\hat{\theta}$, assuming symmetry (which is not necessarily true)
- E.g., a confidence interval of $\pm 5\%$ with a confidence level $\mathbf{P}_l = 99\%$



- We have used Chebychev inequality to compute a CI for the average $\overline{X}$ of a dataset of size n given only its experimental variance s^2 and exploiting the fact that *displaystyle*$\text{Var}[\overline{X}] = \frac{\sigma^2}{n}$

$$\mathbf{P}[\mu - ks < X < \mu + ks] \geq 1 - \frac{1}{k^2}$$

- Letting $\epsilon = ks; \; k = \frac{\epsilon}{s} \simeq \frac{n\epsilon}{\sigma}$

$$\mathbf{P}[\mu - \epsilon < X < \mu + \epsilon] \geq 1 - \frac{s^2}{\epsilon^2} \simeq 1 - \frac{\sigma^2}{n\epsilon^2}$$

- The strength of Chebychev inequality is that it is completely independent from the distribution of X
- We can compute a CI without having any a-priori knowledge about the population we are measuring (or simulating)
- The limit is that it is a loose bound, so that a high level of confidence (normally $\mathbf{P}_I \leq 90\%$ is unacceptable for any practical purpose, while $\mathbf{P}_I \geq 95 - 99\%$ is highly desirable if not necessary for most applications) imply a very large CI
- Can we do better than this?
- Yes, if we know something about the distribution of the population we're measuring/simulating, or if we have large datasets of independent samples

Let's suppose we know that the population is normally distributed:

$$f_X(x) = N(\mu, \sigma^2)$$

In this case it is not difficult to show that the distribution of the sample mean $\overline{X}$ of a dataset with n independent points is also normally distributed

$$f_{\overline{X}}(x) = N(\mu, \sigma^2/n)$$

and finally

$$Z = \frac{\overline{X} - \mu}{(\sigma/\sqrt{n})}$$

is standard normal: $f_{\overline{Z}}(z) = N(0, 1)$

Assuming a symmetric interval of normalized half-width a and $\mathbf{P}_I = \gamma$ it is clear that for Z we have

$$\mathbf{P}[-a < Z < a] = \gamma$$

and that given γ a can be found on tables. Denormalizing to find the CI of our estimate $\overline{X}$ we have

$$\mathbf{P}[\overline{X} - \frac{a\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{a\sigma}{\sqrt{n}}] = \gamma$$

so the interval

$$\left(\overline{X} - \frac{a\sigma}{\sqrt{n}}, \overline{X} + \frac{a\sigma}{\sqrt{n}}\right)$$

is a $100\gamma\%$ CI for μ.

Let $\gamma = 1 - \alpha$ for convenience. Since the normal distribution is symmetric we have that

$$\mathbf{P}[Z < -a] = \mathbf{P}[Z > a] = \frac{\alpha}{2}$$

normally this specific value of a is called $z_{\frac{\alpha}{2}}$ and can be found in tables as the following one, derived from the normal standard distribution $N(0, 1)$

$1 - \alpha$	0.90	0.95	0.99
$z_{\frac{\alpha}{2}}$	1.645	1.96	2.576

As we have a $100(1-\alpha)\%$ CI given by

$$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)$$

it is immediate to compute the number of samples n that we need to measure or simulate to have an estimate $\overline{X}$ that deviates less than

$$\epsilon = \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}$$

from the true value μ

$$n = \left\lceil \left(\frac{z_{\frac{\alpha}{2}}\sigma}{\epsilon}\right)^2 \right\rceil$$

- What if the population is not Gaussian?
    - Easy if we have many samples and they are i.i.d.
- What if the measures/simulations are not i.i.d.?
    - More complex, but we can still "survive" with batch means (sometimes)

- Given any set of i.i.d. RV, the central limit theorem guarantees that under fairly mild assumptions the statistics of

$$Z = \frac{\overline{X} - \mu}{(\sigma/\sqrt{n})}$$

  is $N(0, \mu)$
- This means that we can still use the improved technique described above to compute the CI given that we have enough samples (say more than 30–50)
- In general (also for Gaussian populations) we do not know σ so we have to use its dataset estimation s

- If the sample set is small (say $n < 30$–50), then we should use the Student-t distribution with $n-1$ degree of freedom

- With modern simulation techniques having enough samples is normally not a problem, so the Student-t use is limited to "difficult" experiments, where getting many measures is difficult (e.g., medical studies)

- In simulations it is not easy to guarantee that the output is i.i.d.
- In general we are exploring a DTMC, where the evolution is controlled by the states, so that the "next" sample cannot be independent from the previous one
- Consider once more a queuing station, anyone, say a G/G/m/K/LIFO
  - Let $N(t)$ be the process describing the number of customers in the queue sampled whenever a customer leaves
  - $N(t+1)$ is obviously **very** dependent (not only correlated) on $N(t)$
- Batch means techniques can help in these cases

- Thanks to the linearity of the average operator we can compute $\overline{X}$ in *batches* splitting the sample of dimension n in k smaller subsets

$$\overline{X} = \frac{1}{k}\sum_{i=1}^{k}\left[\frac{k}{n}\sum_{j=1}^{n/k} x_{(ki+j)}\right] = \frac{1}{k}\sum_{i=1}^{k}\left[\frac{k}{n}\overline{X_i}\right]$$

- This was originally meant to reduce numerical problems with large datasets ...
- ... so how can we exploit this to our advantage in computing CI with correlated processes and simulations in particular?

Consider a generic queue (e.g., the G/G/m/K/LIFO)

- Let's define a new process $N'(k)$ defined as the average number of customers in the queue between two successive time instances k when a leaving customer leaves the queue empty

$$N'(k) = \frac{1}{n_s} \sum_{i=1}^{n_s} N(i)$$

  where n_s is the number of customers arrived (and served) between two instances that left the queue empty

- It is not difficult to realize that when the queue empties it loses all its memory so that $N'(k)$ is by construction an i.i.d. process
- Moreover $\overline{N} = \overline{N'}$, so we can compute not only the average value of N, but also its confidence interval based on N'

Use of Batch Means

- Whenever we can identify a renewal process (back to processes definition for it)
- Whenever we can estimate some parameters with a subset of the samples we have and we can use/define at least 30–50 subsets
- With this method we can estimate CIs also for parameters that are not the mean (including variance, general parameters of a distribution, . . .)
- If the process identified is not strictly renewal
  - Make all efforts to guarantee that it is identically distributed
  - Verify that the output samples are reasonably independent
- A powerful verification tool is checking that the process of the errors is actually Gaussian