Dealing with stochastic processes

The stochastic process (SP)

- Definition (in the following material): A stochastic process is random process that happens over time, i.e. the process is dynamic and changes over time.
- An SP can be continuous- or discrete-time
  - If discrete-time, the events in the process are countable

Sampling refresher

- Let $X_1 \ldots . . X_n$ be independent random variables with same distribution function. Let θ denote the mean and σ^2 denote the variance. $\theta = E[X_i]$ and $\sigma^2 = \mathrm{Var}(X_i)$

$Let\ \overline{X}$ be the arithmetic average of X_i, then:

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

$$E[\overline{X}] = E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \sum_{i=1}^{n} \frac{E[X_i]}{n} = \frac{n\theta}{n} = \theta$$

this shows that $\overline{X}$ is an unbiased estimator of θ

Sampling refresher

How good an estimator is $\overline{X}$ of θ?

$$\begin{aligned} \text{Var}\ \overline{X} &= E[(\overline{X} - \theta)^2] \\ &= \text{Var}\left(\frac{1}{n}\sum_1^n X_i\right) \\ &= \frac{1}{n^2}\sum_1^n \text{Var}(X_i) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Therefore, $\overline{X}$ is a good estimator of θ when $\dfrac{\sigma}{\sqrt{n}}$ is small

Confidence Intervals

- We first need to revisit the Normal distribution. Consider $\mathcal{N}(0,1)$

$$P(Z \leq w) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{w} e^{-\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

- For example:

$$P(Z \leq 1) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{w} e^{-\left(\frac{x-\mu}{\sigma}\right)^2} dx = 0.8413$$

Confidence Intervals

- We use this to calculate w for different values of P such as:

$$P(Z \leq w) = 0.95 \Rightarrow w = 1.96$$
$$P(Z \leq w) = 0.99 \Rightarrow w = 2.58$$
$$P(Z \leq w) = 0.999 \Rightarrow w = 3.29$$

Confidence Intervals

If $\overline{X}$ is the mean of a random sample of size n from $\mathcal{N}(\mu, \sigma)$ then the sampling distribution of $\overline{X}$ is $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$

If $X_1, X_2, ...X_n$ constitute random samples from an infinite population with mean μ and standard deviation σ then the sampling distribution of

$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ as n $\rightarrow \infty$ is $\mathcal{N}(0, 1)$

This result is known as the Central Limit Theorem

Confidence Intervals

- What does this mean?
  - If we sample a random process with finite variance, our samples tend to follow a gaussian distribution and we can estimate the probability that our true value is within a range from the sample with a certain confidence
- We often want to find some mean value from a process e.g. the mean number of people in a queue with arrival rate λ and mean service time μ
- Working with means we need one more piece of information:

Confidence Intervals

- The standard deviation of the mean is calculated as follows:

$$Var(X) = \sigma_x^2$$
$$Var(cX) = c^2 Var(X)$$

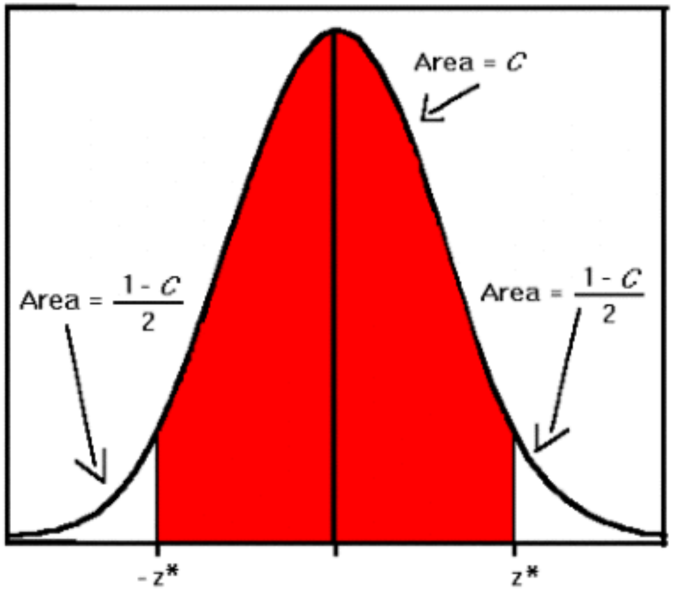
$$Var(mean) = Var\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2} Var\left(\sum_{i=1}^{N} X_i\right)$$
$$= \frac{1}{N^2}\sum_{i=1}^{N} Var(X_i) = \frac{N}{N^2} Var(X) = \frac{1}{N} Var(X)$$

$$\sigma_{mean} = \frac{\sigma}{\sqrt{N}}$$

Confidence Intervals

- We can now set the intervals that give us the desired confidence in our results for mean values. For example, if we want to find the mean average queue length with a confidence of 95% we set:

$$ConfidenceInterval = \overline{z} \pm 1.96\sigma_{mean}$$

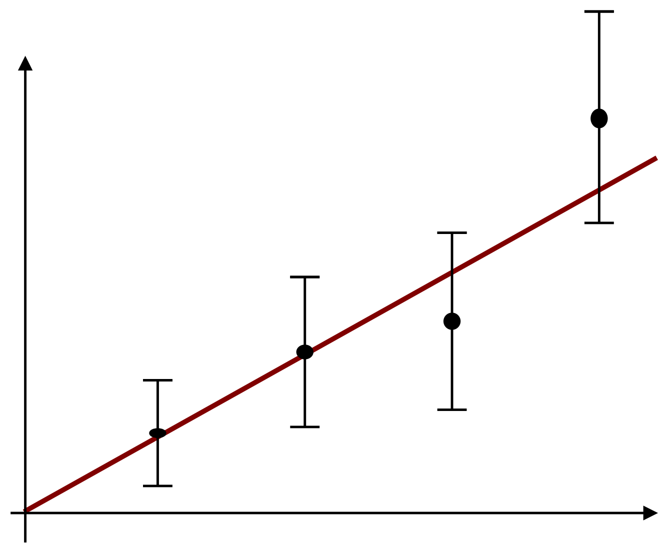


Confidence Intervals

- Process for the example with the queue
  1. Run the simulation to get mean q length
  2. Calculate mean of the means sampled
  3. Calculate stddev
  4. Calculate confidence (in this case based on σ_{mean})
  5. Change random seed
  6. Run again
  7. until confidence interval small enough
- Can also run as a single longer experiment

Confidence Intervals

- Plotting confidence intervals in graphs make interpretation easier



- Think about overlapping confidence intervals

Stopping condition

- Question: How long should the simulation run for?
  - "I simulated for 10 minutes", not a good answer.
- Better idea is to use variance measurements
  - Stopping condition: Variance < x

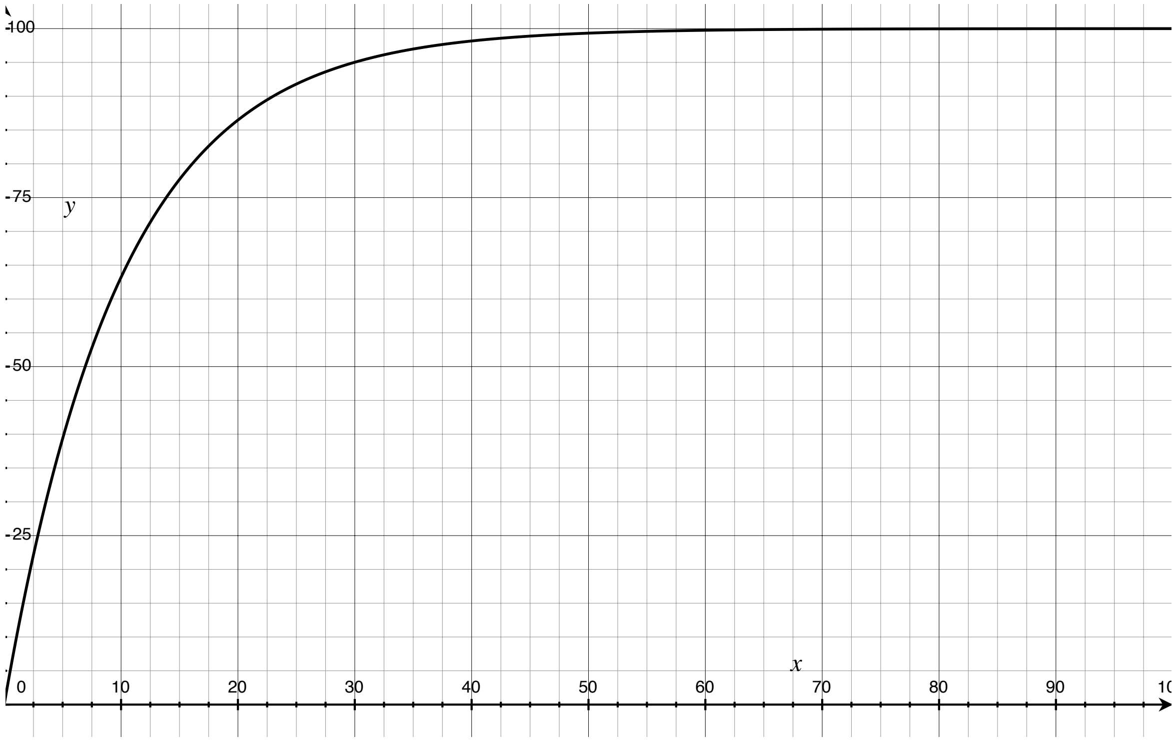
Warmup time

- Take as example a simulation of a M/M/1 queue
  - Say $\lambda = 9$, $\mu = 10$
- This system has a queue that will grow to large sizes over time
- There is a lag before the system reaches steady state
- Consider carefully when to start collecting statistics

Warmup time

$$\overline{X} = \frac{1}{100}\int_{0}^{100} f(x)dx = 90$$

$$\overline{X} = \frac{1}{40}\int_{60}^{100} f(x)dx = 99.95$$



Variance reduction techniques

- There are ways of reducing variance to speed up convergence of confidence intervals from several runs
- One common approach is the use of "Antithetic variates"
- Consider the following. We have generated two identically distributed variables X_1 and X_2 with mean θ. Then:

$$Var(\frac{X_1 + X_2}{2}) = \frac{1}{4}[Var(X_1) + Var(X_2) + 2Cov(X_1, X_2)]$$

Antithetic variables

- Variance will be reduced if X_1,X_2 negatively correlated (covariance negative)
- Consider $X_1 = f(U_1,U_2,U_3 \ldots U_m)$ where U are m independent random numbers. Then so are $1-U_m$
- Therefore, $X_2 = f(1-U_1,1-U_2,1-U_3 \ldots 1-U_m)$ has the same distribution as X_1
- Good chance that X_1 and X_2 are negatively correlated when using U_m and $1-U_m$ for generation

Discussion

- For each seed, the two antithetic runs combined produce results closer to real mean with high probability.
- Not to be confused with correlation between samples in a series

Sample Correlation

- Calculating confidence intervals assumes uncorrelated samples
  - If this is not the case, there is a problem
- Covariance again

Let the sampled mean from a simulation run be $\hat{a}$

Then, $cov(\hat{a}, \hat{a}) = cov\left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{j=1}^{n} x_j\right)$ should be close to 0

Sample Correlation

- If the samples are not uncorrelated, the variance will be underestimated since:

$$V(\hat{a}) = cov\left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{j=1}^{n} x_j\right) = \frac{1}{n^2}\sum_i \sum_j cov(x_i, x_j)$$

$cov(\hat{a}_i, \hat{a}_j)$ should be 0 when $i \neq j$

When $i = j$, this reduces to the variance

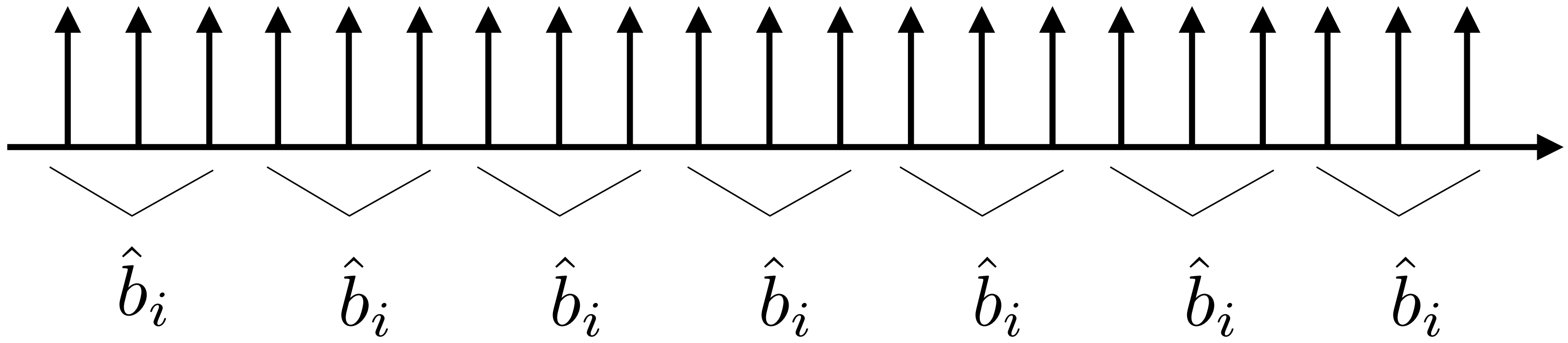
Reducing covariance (1)

- Reducing covariance simplest way:
  - sample less frequently
- Instead of sampling all events, generate a random variable for inter-sample times i.e.

$$\text{next sample} = \text{getsampleTime}(\ 1 - e^{-\lambda t})$$

- This will also lower the data volume for a simulation run. could be useful
- Reduces risk of correlation because of cyclic phenomena

Reducing covariance (2) Batch means



- Another fast way of lowering correlation between successive samples, use method of batch means.
- Take average from a sequence of samples and use as a sample
- Selecting current batch size a problem. Too large, confidence intervals become too large; too small, variance underestimated. From literature, 10 < size < 20 common

Testing correlation, deciding on batch size

- Mean of a batch = $\bar{b_i}$ mean of all batches = $\bar{\bar{b}}$ batch size = N
- Test for correlation magnitude as follows:

$$C = 1 - \frac{\sum_{i=1}^{N-1} (\bar{b}_i - \bar{b}_{i+1})^2}{\sum_{i=1}^{N-1} (\bar{b}_i - \bar{\bar{b}})^2}$$

- Then, if: $\dfrac{C}{(N-2)(N^2-1)} \approx \mathcal{N}(0,1)$

- then batch size well chosen

Compensate for covariance contribution

- The below is typically more difficult but a possibility!
- There are statistical methods in literature e.g.

$$\sigma_{\overline{x}}^2 = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \cdot var(x_i) + 2 \cdot \sum_{j=i+1}^{n} \left(\frac{1}{n}\right)^2 \cdot cov(x_i, x_j)$$

If the process is stationary in regards to covariance and k-dependent, loosely, if a sample is dependent on k previous samples this becomes:

$$\sigma_{\overline{x}}^2 = \frac{1}{n} \cdot \left\{ var(x_i) + 2 \cdot \sum_{i=1}^{k} cov(x_i, x_{i+1}) \right\}$$

We use the following estimate of the covariance from the literature:

$$cov(w_i, w_{i+s}) = \tfrac{1}{n-s} \textstyle\sum_{i=1}^{n-1} (w_i - \overline{w})(w_{i+s} - \overline{w}) \qquad \bar{w} = \frac{1}{M} \times \sum_{i=1}^{M} w_i$$

In plain text, estimate mean from long run

- Sample mean values w_i and calculate sampled average taking into account dependency
- Using $\bar{w} = \frac{1}{M} \times \sum_{i=1}^{M} w_i$, estimate covariance and calculate adjusted σ^2.
- Use this to estimate confidence interval with significance 1-α as: $\bar{w} \pm \lambda_{\alpha/2} \cdot \sigma_{\bar{w}}$

Compensate for covariance contribution

- Problem, estimate k. Can be done using the *autoregressive approach.* Translate {*w*} to a set {*u*} of independent variables

$$u_i = \sum_{i=0}^{k} b_s \cdot (w_{i-s} - \overline{w}), \quad i = k+1, \ldots, n$$

We estimate $\hat{b}_s$ from a set of linear equations:

$$\sum_{s=1}^{k} \hat{b}_s \hat{R} r - s = -\hat{R}_r \quad r = 1, \ldots, k$$

Where $\hat{R}_s$ is the estimated covariance $cov(w_t, w_{t+s})$

$and\ \hat{b}_s\ are\ estimated\ coefficients\ of\ b_s$

Compensate for covariance contribution

The variance at the k_{th} order is estimated to be:

$$\hat{\sigma}^2_{u,k} = \sum_{s=0}^{k} \hat{b}_s \hat{R}_s \quad \hat{b}_0 \equiv 1$$

pick $k_{max} = K$ and calculate $\hat{\sigma}^2_{u,k}$ for $k = 1, \ldots, K$

formulate a χ^2 test where the density Q is $Q = n \cdot \left(1 - \frac{\hat{\sigma}_{u,K}}{\hat{\sigma}_{u,k}}\right)$

The hypothesis "sequence $w_1, \ldots, w_n$ is of order k" is discarded with significance α if

$Q > \chi^2_\alpha(f)$ where $f = K - k$ signifies degree of freedom

After k is found: $\sigma^2_{\overline{w}} = \frac{\sigma^2_{u,k}}{nb^2}$ where $b = \sum_{s=0}^{k} \hat{b}_s$

We have to use a t-distribution instead of the normal distribution for confidence limits where: $\overline{w} \pm t_{a/2}(f) \cdot \sigma_{\overline{w}}$ and

$$f = \frac{nb}{2 \cdot \sum_{s=0}^{k} (k - 2s)\hat{b}_s}$$

Sample Correlation

- Important!
  - your aim is to capture the correlation between different tests on the process
  - you do not want to capture correlation from dependencies between successive samples
- often, it is therefore useful to run many simulation runs with varying seeds rather than few long simulation runs
- not always though, depends on the nature of the simulation

Tradeoff

- Each simulation run brings warmup time (wasted time)
  - Many shorter simulations are more costly in time
- Estimating and compensating for covariance possible but takes time
- can be tedious work to do the stats….
- However, if statistical analysis not carried out, results don't have meaning.

More information

- Very good resource is
  - Sheldon Ross, “Simulation” 4th ed. academic press 2006
- Books on time series analysis, especially sections on auto regression for methods on estimating and compensating for covariance