Issues for the next couple of weeks

- Input of a simulation
  (Choice of input distributions, values of parameters)
- Output analysis of a simulation
  (how many runs, length of a run, confidence intervals)
- How do we generate uniform random variables?
- How do we generate arbitrarily distributed random variables?

INPUT OF A SIMULATION

Specifying distributions of random variables (e.g., interarrival times, processing times) and assigning parameter values can be based on:

- Historical numerical data
- Expert opinion

In practice, there is sometimes real data available, but often the only information of random variables that is available is their mean and standard deviation.

Empirical data can be used to:

- construct empirical distribution functions and generate samples from them during the simulation;
- fit theoretical distributions and then generate samples from the fitted distributions.

Fitting a distribution

Methods to determine the parameters of a distribution:

- Maximum likelihood estimation
- Moment fitting

Maximum likelihood estimation

Let $f(x;\theta)$ denote the probability density function with unknown parameter (vector) θ.
Let $X = (X_1, \ldots, X_n)$ denote a vector of *i.i.d.* observations from f.
Then

$$L(\theta, X) = \prod_{i=1}^{n} f(X_i, \theta)$$

is the *likelihood function* and $\hat{\theta}$ satisfying

$$L(\hat{\theta}, X) = \sup_{\theta} L(\theta, X)$$

is the *maximum likelihood estimator* of θ.

Examples:

- **Exponential distribution**

$$f(x, \mu) = \mu e^{-\mu x}$$

Then

$$\frac{1}{\hat{\mu}} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- **Uniform** (a, b)

$$f(x, a, b) = \frac{1}{b - a}$$

Then

$$\hat{a} = \min X_i, \qquad \hat{b} = \max X_i.$$

But for many distributions $\hat{\theta}$ has to be calculated numerically.

Moment fitting

Obtain an approximating distribution by fitting a *phase-type distribution* on the mean, $E(X)$, and the coefficient of variation,

$$c_X = \frac{\sigma_X}{E(X)},$$

of a given positive random variable X, by using the following simple approach.

Coefficient of variation less than 1

If $0 < c_X < 1$, then fit an $E_{k-1,k}$ distribution as follows. If

$$\frac{1}{k} \leq c_X^2 \leq \frac{1}{k-1},$$

for certain $k = 2, 3, \ldots$, then the approximating distribution is with probability p (resp. $1-p$) the sum of $k-1$ (resp. k) independent exponentials with common mean $1/\mu$. By choosing

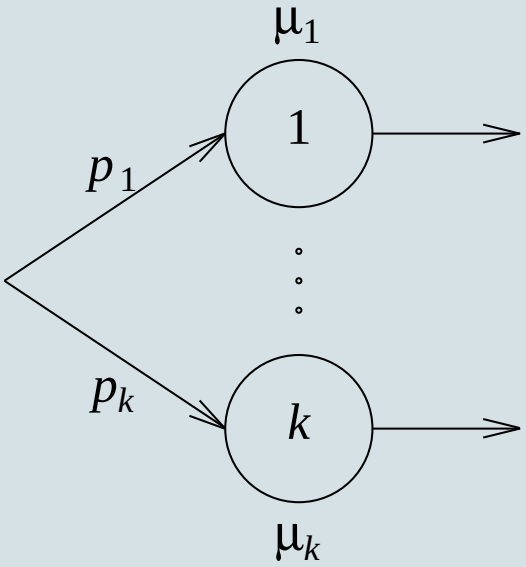
$$p = \frac{1}{1+c_X^2}[kc_X^2 - \{k(1+c_X^2) - k^2 c_X^2\}^{1/2}], \qquad \mu = \frac{k-p}{E(X)},$$

the $E_{k-1,k}$ distribution matches $E(X)$ and c_X.

Coefficient of variation greater than 1

In case $c_X \geq 1$, fit a $H_2(p_1, p_2; \mu_1, \mu_2)$ distribution.

Phase diagram for the $H_k(p_1, \ldots, p_k; \mu_1, \ldots, \mu_k)$ distribution:



But the H_2 distribution is not uniquely determined by its first two moments. In applications, the H_2 distribution with *balanced means* is often used. This means that the normalization

$$\frac{p_1}{\mu_1} = \frac{p_2}{\mu_2}$$

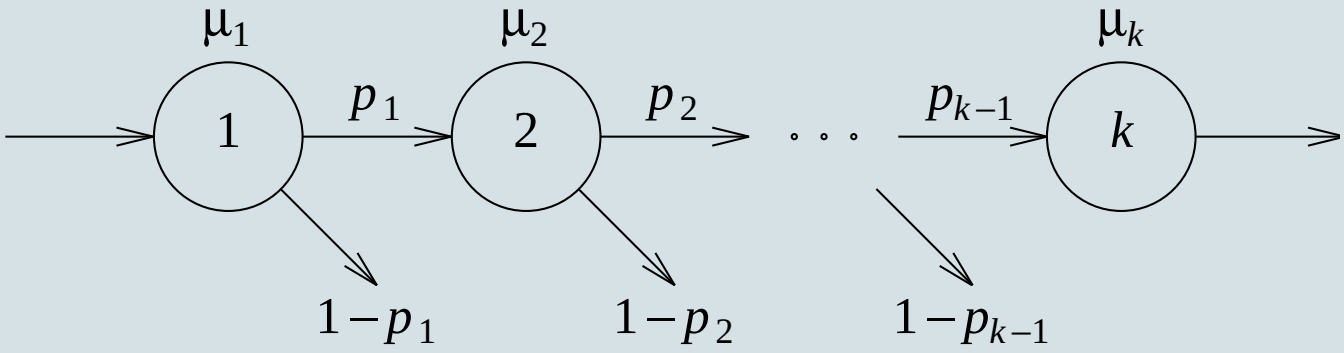
is used. The parameters of the H_2 distribution with balanced means and fitting $E(X)$ and c_X (≥ 1) are given by

$$p_1 = \frac{1}{2}\left(1 + \sqrt{\frac{c_X^2 - 1}{c_X^2 + 1}}\right), \qquad p_2 = 1 - p_1,$$

$$\mu_1 = \frac{2p_1}{E(X)}, \qquad \mu_1 = \frac{2p_2}{E(X)}.$$

In case $c_X^2 \geq 0.5$ one can also use a Coxian-2 distribution for a two-moment fit.

Phase diagram for the Coxian-k distribution:



The following parameter set for the Coxian-2 is suggested:

$$\mu_1 = 2E(X), \quad p_1 = 0.5/c_X^2, \quad \mu_2 = \mu_1 p_1.$$

Phase-type distributions may also naturally arise in practical applications.

Example:
The processing of a job involves performing several tasks, where each task takes an exponential amount of time; then the processing time can be described by an Erlang distribution.

Fitting nonnegative discrete distributions

Let X be a random variable on the non-negative integers with mean EX and coefficient of variation c_X. Then it is possible to fit a discrete distribution on $E(X)$ and c_X using the following families of distributions:

- Mixtures of Binomial distributions;
- Poisson distribution;
- Mixtures of Negative-Binomial distributions;
- Mixtures of geometric distributions.

This fitting procedure is described in Adan, van Eenige and Resing (see Probability in the Engineering and Informational Sciences, 9, 1995, pp 623-632).

Adequacy of fit

- Grapical comparison of fitted and empirical curves;
- Statistical tests (*goodness-of-fit tests*).

OUTPUT ANALYSIS OF A SIMULATION

Confidence intervals

Let $X_1, X_2, \ldots, X_n$ be independent *realizations* of a random variable X with unknown mean μ and unknown variance σ^2.

Sample mean

$$\bar{X}(n) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Sample variance

$$S^2(n) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}(n))^2$$

Clearly $\bar{X}(n)$ is an estimator for the unknown mean μ.
How can we construct a confidence interval for μ?

Central limit theorem states that for large n

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

is approximately a standard normal random variable, and this remains valid if σ is replaced by $S(n)$.
Hence, let $z_\beta = \Phi^{-1}(\beta)$ (e.g., $z_{1-0.025} = 1.96$), then

$$P(-z_{1-\delta/2} \leq \frac{\sum_{i=1}^{n} X_i - n\mu}{S(n)\sqrt{n}} \leq z_{1-\delta/2}) \approx 1 - \delta$$

or equivalently

$$P\left(\bar{X}(n) - z_{1-\delta/2}\frac{S(n)}{\sqrt{n}} \leq \mu \leq \bar{X}(n) + z_{1-\delta/2}\frac{S(n)}{\sqrt{n}}\right) \approx 1 - \delta$$

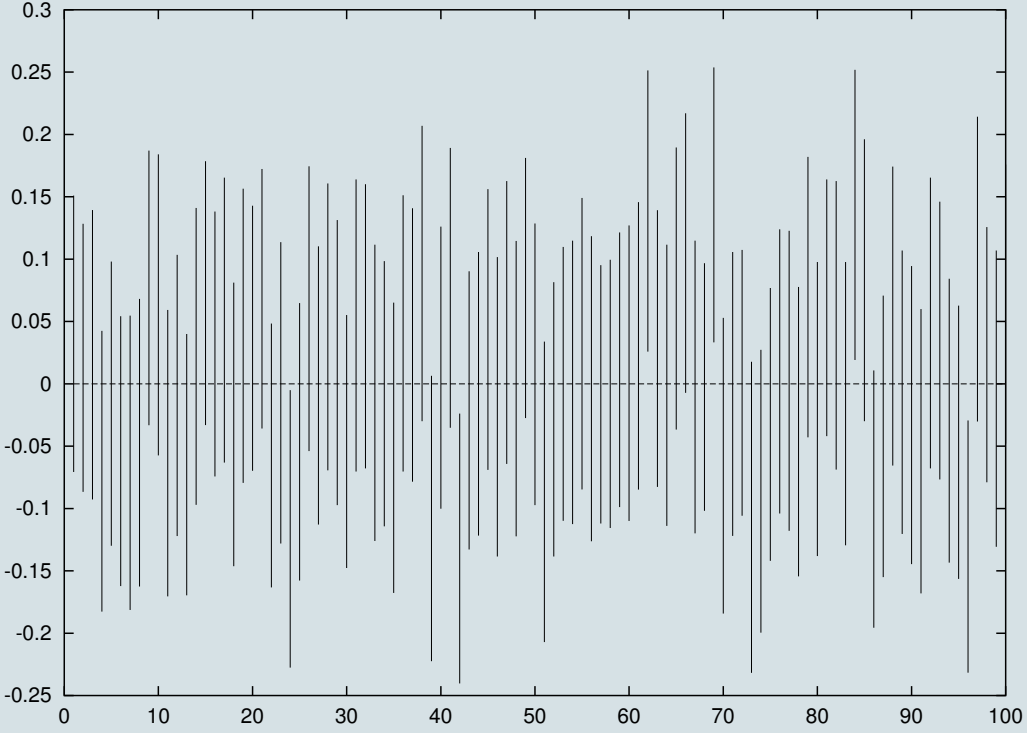
Conclusion:

An approximate $100(1-\delta)\%$ confidence interval for the unknown mean μ is given by

$$\bar{X}(n) \pm z_{1-\delta/2}\frac{S(n)}{\sqrt{n}}$$

As a consequence, to obtain one extra digit of the parameter μ, the required simulation time increases with approximately a factor 100.

100 confidence intervals for the mean of uniform random variable on $(-1, 1)$; each interval is based on 100 observations.



Remark:
If the observations X_i are normally distributed, then

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{S(n)\sqrt{n}}$$

has *for all* n a Student's t distribution with $n-1$ degrees of freedom; so an *exact* confidence interval can be obtained by replacing $z_{1-\delta/2}$ by the corresponding quantile of the t distribution with $n-1$ degrees of freedom.

Remark:
The width of a confidence interval can be reduced by

- increasing the number of observations n;
- decreasing the value of $S(n)$.

The reduction obtained by halving $S(n)$ is the same as the one obtained by producing four times as much observations. Hence, *variance reduction techniques* are important.

Remark:
Recursive computation of the sample mean and variance of the realizations $X_1, \ldots, X_n$ of a random variable X:

$$\bar{X}(n) = \frac{n-1}{n}\bar{X}(n-1) + \frac{1}{n}X_n$$

and

$$S^2(n) = \frac{n-2}{n-1}S^2(n-1) + \frac{1}{n}\left(X_n - \bar{X}(n-1)\right)^2$$

for $n = 2, 3, \ldots$, where

$$\bar{X}(1) = X_1, \qquad S^2(1) = 0.$$

OUTPUT ANALYSIS OF A SIMULATION

Method of independent replications

Example: Long-term ("steady-state") mean waiting time $E(W)$ in the single-stage production line

Produce n *independent* sample paths of waiting times $W_1^{(i)}, W_2^{(i)}, \ldots, W_N^{(i)}$ and compute

$$\bar{W}_N^{(i)} = \frac{1}{N} \sum_{j=1}^{N} W_j^{(i)}, \quad i = 1, \ldots, n.$$

Then, for large N, an approximate $100(1-\delta)\%$ confidence interval for the mean waiting time $E(W)$ is

$$\bar{W}_{n,N} \pm z_{1-\delta/2}\frac{S_{n,N}}{\sqrt{n}}$$

where $\bar{W}_{n,N}$ and $S^2_{n,N}$ are the sample mean and variance of the realizations $\bar{W}_N^{(1)}, \ldots, \bar{W}_N^{(n)}$;

$$\begin{aligned} \bar{W}_{n,N} &= \frac{1}{n}\sum_{i=1}^{n} \bar{W}_N^{(i)} \\ S^2_{n,N} &= \frac{1}{n-1}\sum_{i=1}^{n}(\bar{W}_N^{(i)} - \bar{W}_{n,N})^2 \end{aligned}$$

Results for $\lambda = 0.5, \mu = 1$ and 10 runs, each of $N = 10^4$ waiting times

i	$\bar{W}_N^{(i)}$
1	0.995
2	1.002
3	0.959
4	1.037
5	0.902
6	1.011
7	1.125
8	1.007
9	1.075
10	1.044

$E(W) = 1.016 \pm 0.036$ (95% confidence interval)

Results for $\lambda = 0.9$, $\mu = 1$ and 10 runs, each of $N = 10^4$ waiting times

i	$\bar{W}_N^{(i)}$
1	7.373
2	8.496
3	8.574
4	7.752
5	8.637
6	7.404
7	9.556
8	8.863
9	8.537
10	11.000

$E(W) = 8.619 \pm 0.632$ (95% confidence interval)

Clearly, a more congested system is harder to simulate! To obtain a more accurate estimate should we increase the number of runs and/or the length of each run? And, how much?

Problem of the initialization effect

We are interested in the long-term behaviour of the system and maybe the choice of the initial state of the simulation will influence the quality of our estimate.
One way of dealing with this problem is to choose N very large and to neglect this initialization effect. However, a better way is to throw away in each run the first k observations, i.e. we set

$$\bar{W}_N^{(i)} = \frac{1}{N-k} \sum_{j=k+1}^{N} W_j^{(i)}.$$

We call k the length of the *warm-up period* and it can be determined by a graphical procedure.
Disadvantage of the independent replication method is that we have the initialization effect in each simulation run.

OUTPUT ANALYSIS OF A SIMULATION

Batch means

Instead of doing n independent runs, we try to obtain n independent observations by making a *single long run* and, after deleting the first k observations, dividing this run into n subruns.

The advantage is that we have to go through the warm-up period only once.

Let $W_1, W_2, \ldots, W_{nN}$ be the output of a single run, where we have already deleted the first k observations and renumbered the remaining ones. Hence $W_1, W_2, \ldots, W_{nN}$ will be representative for the steady-state. We divide the observations into n batches of length N. Thus, batch 1 consists of

$$W_1, W_2, \ldots, W_N;$$

batch 2 of

$$W_{N+1}, W_{N+2}, \ldots, W_{2N},$$

and so on. Let $\bar{W}_N^{(i)}$ be the sample (or batch) mean of the N observations in batch i, so

$$\bar{W}_N^{(i)} = \frac{1}{N} \sum_{j=(i-1)N+1}^{iN} W_j$$

The $\bar{W}_N^{(i)}$'s play the same role as the ones in the independent replication method. Unfortunately, the $\bar{W}_N^{(i)}$'s will now be *dependent*.

But, under mild conditions, for *large* N the $\bar{W}_N^{(i)}$'s will be approximately independent, each with the same mean $E(W)$.
Hence, for N large enough, it is reasonable to treat the $\bar{W}_N^{(i)}$'s as i.i.d. random variables with mean $E(W)$; thus

$$\bar{W}_{n,N} \pm z_{1-\delta/2}\frac{S_{n,N}}{\sqrt{n}}$$

provides again a $100(1-\delta)\%$ confidence interval for $E(W)$, with $\bar{W}_{n,N}$ and $S^2_{n,N}$ again the sample mean and variance of the realizations $\bar{W}_N^{(1)}, \ldots, \bar{W}_N^{(n)}$;.