Systems Simulation
ECE 597/697 S

Statistical Analysis of Simulated Data

Prof. Michael Zink
Overview

- Introduction
- Sample mean and variance
- Estimation of population variance
- When to stop generating new data?
- Interval estimates of population mean
- Bootstrapping technique
Introduction I

- Simulation study undertaken to determine value of some quantity $\theta$ connected with a particular stochastic model.
- Simulation results in data $X$ whose expected value is the quantity of interest $\theta$.
- Second simulation provides new and independent variable having mean $\theta$.
- After $k$ runs i.i.d random variables $X_1, \ldots, X_n$ with mean $\theta$.
- The average of these $k$ values is then used as an estimator, or approximator, of $\theta$. 
Introduction II

- Consider the problem of deciding when to stop the simulation study – that is, deciding on the appropriate value of $k$.
- Consider the quality of estimator $\theta$.
- Obtain an interval in which $\theta$ lies with a certain degree of confidence.
Sample Mean and Sample Variance

- Suppose that $X_1,\ldots, X_n$ are independent random variables having same distribution function.
- Let $\theta = E[X_i]$ and $\sigma^2 = \text{Var}(X_i)$ denote mean and variance.
- The sample mean is defined as $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ which is often used to estimate population mean $\theta$.
- Because

\[
E[\bar{X}] = E \left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{n\theta}{n} = \theta
\]
Sample Mean and Sample Variance

- $\bar{X}$ is an unbiased estimator for $\theta$.
- Consider mean square error to determine if $\bar{X}$ is a good estimator of mean $\theta$.

$$E\left[ (\bar{X} - \theta)^2 \right] = Var(\bar{X}) \quad (E[\bar{X}] = \theta)$$

$$= Var\left( \frac{1}{n} \sum_{i=1}^{n} X_i \right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) \quad (\text{by independence})$$

$$= \frac{\sigma^2}{n} \quad (Var(X_i) = \sigma^2)$$
Sample Mean and Sample Variance

- $\bar{X}$, the sample mean of $X_1,...,X_n$ is a random variable with mean $\theta$ and variance $\sigma^2/n$.
- Random variable is unlikely to be too many standard deviations from its mean it follows: $\bar{X}$ is a good estimator if $\frac{\sigma}{\sqrt{n}}$ is small.
Remark

- Justification of statement that “a random variable is unlikely to be too many standard deviations away from its mean”
- If \( n \) is large, which is usually the case in simulations, central limit theorem can be applied.
- Thus one can assert that \((\bar{X} - \theta)/(\sigma/\sqrt{n})\) is approximately distributed as a standard normal random variable; and thus

\[
P\left\{ \left| \bar{X} - \theta \right| > c\sigma/\sqrt{n} \right\} \approx P\{ |Z| > c \} = 2\left[ 1 - \Phi(c) \right]
\]

where \( \Phi \) is the standard normal distrib. function
Remark

- For example, since \( \phi(1.96) = 0.975 \) the equation above states that sample mean differs from \( \Theta \) by more than \( 1.96 \sigma / \sqrt{n} \) is approximately 0.05.

- Difficulty with directly using \( \sigma^2 / n \) of how well the sample mean of \( n \) data values estimates the population mean is:
  Population variance \( \sigma^2 \) is not usually known.

- Thus is also must be estimated.
Estimation of Population Variance $\sigma^2$

- Problem of using $\sigma^2$ as an indicator for quality of sample mean is that population variance is usually not known – need to estimate it.

- Since $\sigma^2 = E[(X - \theta)^2]$ is the average of the square of the differences of a datum value and its unknown mean.

- Upon using $\overline{X}$ as the estimator of the mean a natural estimator of the population variance would be $\sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n}$
Estimation of Population Variance $\sigma^2$

- **Definition:** The sample variance $S^2$ is defined by

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

- **Proposition:**

$$E[S^2] = \sigma^2$$
Estimation of Population Variance $\sigma^2$

- **Proof:** Using the algebraic identity

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

we see that

$$(n - 1)E[S^2] = E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2]$$

$$= nE[X_1^2] - nE[\overline{X}^2]$$

since all $X_i$ have the same distribution.
Estimation of Population Variance $\sigma^2$

- **Proof cont’d:** with

\[
E[Y^2] = \text{Var}(Y) + (E[Y])^2
\]

we obtain

\[
E[X_1^2] = \text{Var}(X_1) + (E[X_1])^2
\]

\[
= \sigma^2 + \theta^2
\]

\[
E[X^2] = \text{Var}(X) + (E[X])^2
\]

\[
= \frac{\sigma^2}{n} + \theta^2
\]
Estimation of Population Variance $\sigma^2$

- Thus we obtain:
  $$(n - 1)E[S^2] = n(\sigma^2 + \theta^2) - n\left(\frac{\sigma^2}{n} + \theta^2\right) = (n - 1)\sigma^2$$

- Sample variance $S^2$ is used as estimator for variance $\sigma^2$.

- $S=\sqrt{S^2}$ as sample standard deviation.

- When should we stop generating new data?
  - Chose acceptable standard deviation $d$.
  - Generate new data until $\sigma/\sqrt{n}$ –namely $S/\sqrt{n}$- is smaller than $d$.
  - Since sample standard deviation might not be a good estimate for small sample size, following procedure is recommended.
When to Stop Generating new Data?

1. Choose and acceptable value \( d \) for the standard deviation of the estimator.
2. Generate at least 100 data values.
3. Continue to generate additional data values, stopping when you have generated \( k \) values and \( \frac{S}{\sqrt{k}} < d \), where \( S \) is the sample standard deviation based on those \( k \) values.
4. The estimate of \( \theta \) is given by \( \overline{X} = \sum_{i=1}^{k} \frac{X_i}{k} \)
Example I

- Service system in which now customer are allowed after 5PM.
- Estimate time at which last customer exits system
- Be 95% certain that estimated answer will not differ from true answer by more than 15 seconds.
- Generate $k$ values by the equivalent number of simulation runs such that $1.96S/\sqrt{k} < 15$ – where $S$ is the sample standard deviation (in seconds) of these $k$ data values.
- Estimate of the expected time of last customer departure will be average of $k$ data values.
Recursive Computation

- Consider sequence of data values $X_1, X_2, \ldots, X_n$ and let
  \[
  \bar{X}_j = \sum_{i=1}^{j} \frac{X_i}{j} \quad \text{and} \quad S_j^2 = \sum_{i=1}^{j} \frac{(X_i - \bar{X}_j)^2}{j - 1}, \quad j \geq 2
  \]
denote the sample mean and the sample variance of the first $j$ values.
- Use following the following recursions for sample mean and variance. With $S_1^2 = 0$, $\bar{X}_0 = 0$
  \[
  \bar{X}_{j+1} = \bar{X}_j + \frac{X_{j+1} - \bar{X}_j}{j + 1} \quad \text{and} \quad S_{j+1}^2 = \left(1 - \frac{1}{j}\right)S_j^2 + (j + 1)\left(\bar{X}_{j+1} - \bar{X}_j\right)^2
  \]
Example II

If first three data values are: \( X_1 = 5, \ X_2 = 14, \ X_3 = 9, \)
then the two equations from above yield:

\[
X_1 = 5 \\
X_2 = 5 + \frac{9}{2} \\
S_2^2 = 2 \left( \frac{19}{2} - 5 \right)^2 = \frac{81}{2} \\
X_3 = \frac{19}{2} + \frac{1}{3} \left( 9 - \frac{19}{2} \right) = \frac{28}{3} \\
S_3^2 = \frac{81}{4} + 3 \left( \frac{28}{3} - \frac{19}{2} \right)^2 = \frac{61}{3}
\]
Case of Bernoulli Random Variables

- Somewhat modified analysis in case of Bernoulli random variables.
- Suppose random variable $X$ is created as follows:
  
  $$X_i = \begin{cases} 
  1 & \text{with probability } p \\
  0 & \text{with probability } 1 - p
  \end{cases}$$

  and we are interested in $E[X_i] = p$.

- Since in this case $Var(X_i) = p(1-p)$, there is no need to utilize sample variance to estimate $Var(X_i)$.
Case of Bernoulli Random Variables

- After generating \( n \) variables \( X_1, \ldots, X_n \), then an estimate of \( p \) will be
  \[
  \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i
  \]
  and a natural estimate of \( \text{Var}(X_i) \) is
  \[
  \bar{X}_n \left(1 - \bar{X}_n\right)
  \]

- Following method to decide when to stop:
  1. Chose an acceptable value \( d \) for the standard deviation of the estimator.
  2. Generate at least 100 data values.
  3. Continue to generate additional data values, stopping when \( k \) values generated and
     \[
     \left[\bar{X}_k \left(1 - \bar{X}_k\right)/k\right]^{1/2} < d
     \]
  4. The estimate of \( p \) is \( \bar{X}_k \)
Example III

- In Example I we are interested in estimating the probability that there was still a customer in the store at 5:30.

\[ X_i = \begin{cases} 
1 & \text{if there is a customer present at 5:30 on day } i \\
0 & \text{otherwise} 
\end{cases} \]

- Simulate 100 days and continue until \( k \)th day, where \( k \) is such that \([\overline{X}_k (1 - \overline{X}_k) / k]^{1/2} < d\)

- \( p_k = \overline{X}_k \) is the proportion of \( k \) days on which there is a customer present at 5:30 and \( d \) is an acceptable value for the standard deviation of the estimator \( p_k \).
Interval Estimates of Population Mean

- $X_1, X_2, \ldots, X_n$ are independent random variables from a common distribution having mean $\theta$ and variance $\sigma^2$.
- Although sample mean is an effective estimator of $\theta$ we do not expect that it will be equal to $\theta$ but rather be “close”.
- It is sometimes more valuable to specify an interval for which we have a certain degree of confidence that $\theta$ lies within.
Interval Estimates of Population Mean

- To obtain such interval we need (approximate) distribution of estimator $\bar{X}$.
- Since $E[\bar{X}] = \theta$, $Var(\bar{X}) = \frac{\sigma^2}{n}$

it follows from the central limit theorem that for large $n$

$$\sqrt{n} \left( \frac{\bar{X} - \theta}{\sigma} \right) \approx N(0,1)$$

which means it is approximately distributed as a standard normal.
Interval Estimates of Population Mean

- If $\sigma$ is replaced by its estimator $S$, the sample standard deviation, then it still remains the case that resulting quantity is approximately a standard normal.

- That is when $n$ is large

$$\sqrt{n} \frac{\bar{X} - \theta}{S} \approx N(0,1)$$

- For any $a$, $0 < a < 1$, let $z_a$ be such that $P\{Z > z_\alpha\} = \alpha$ where $Z$ is a standard normal random variable and if follows that $z_{1-a} = -z_a$. 
Interval Estimates of Population Mean

- **Definition:** *If the observed values of the sample mean and the sample standard deviation are \( \bar{X} = \bar{x} \) and \( S = s \), call the interval \( \bar{x} \pm z_{a/2} s / \sqrt{n} \) an (approximate) 100(1-\(a\)) percent confidence interval estimate of \( \Theta \).*
Example IV

- Assume simulation study in which additional values can be generated and question is when to stop.
- Initially choose values $\alpha$ and $l$ and continue generating data until approximate $100(1-\alpha)\%$ confidence interval estimate of $\theta$ is less than $l$. 
Example IV

- Since length of interval will be \(2z_{a/2}S/\sqrt{n}\) this can be accomplished by the following technique:

1. Generate at least 100 data values.
2. Continue to generate additional data values, stopping when the number of values you have generated \(k\) is such that \(2z_{a/2}S/\sqrt{n} < l\). where \(S\) is the sample standard deviation based on those \(k\) values. (Constantly updating \(S\) by using recursion.)
3. If \(\bar{x}\) and \(s\) are the observed values of \(\bar{X}\) and \(S\), then the \(100(1-a)\)% confidence interval estimate of \(\theta\), whose length is less than \(x \pm 2z_{a/2}S/\sqrt{k}\).
Bernoulli Random Variable Case

\[ X_i = \begin{cases} 
1 & \text{with probability } p \\
0 & \text{with probability } 1 - p 
\end{cases} \]

- Here \( \text{Var}(X_i) \) can be estimated by \( \bar{X}(1 - \bar{X}) \) and it follows:

\[
\sqrt{n} \frac{(\bar{X} - p)}{\sqrt{\bar{X}(1 - \bar{X})}} \approx N(0,1)
\]

- And for any \( \alpha \)

\[
P \left\{ -z_{\alpha/2} < \sqrt{n} \frac{(\bar{X} - p)}{\sqrt{\bar{X}(1 - \bar{X})}} < z_{\alpha/2} \right\} = 1 - \alpha
\]

- Hence, \( 100(1-\alpha)\% \) confidence interval estimate of \( p \) is

\[
p_n \pm z_{\alpha/2} \sqrt{p_n(1 - p_n)/n}
\]
Bootstrapping Technique

- $X_1, X_2, ..., X_n$ are independent random variables having common distribution function $F$.
- Suppose we are interested in using them to estimate some parameter $\theta(F)$. (E.g. mean, median, variance, or any other parameter of $F$.)
- Also, an estimator of $\theta(F)$ – call it $g(X_1, ..., X_n)$ – has been proposed.
- To judge its quality as an estimator, estimate its mean square error:

$$MSE(F) \equiv \mathbb{E}_F \left[ \left( g(X_1, ..., X_n) - \theta(F) \right)^2 \right]$$
Bootstrapping Technique

- If distribution function $F$ were known one could compute expected square of the difference between $\theta$ and its estimator.
- However after $n$ data points, pretty good idea what distribution looks like.
- Assume observed values of the data are $X_i = x_i$, $i=1,...,n$.
- Now distribution function $F$ can be estimated by empirical distribution function $F_e$

$$F_e(x) = \frac{\text{number of } i : X_i \leq x}{n}$$
Bootstrapping Technique

- If $F_e$ is close to $F$, as it should be when $n$ is large, then $\theta(F_e)$ will probably be close to $\theta(F)$ and $\text{MSE}(F)$ should approximately be equal to

\[
\text{MSE}(F_e) \equiv E_{F_e} \left[ (g(X_1, \ldots, X_n) - \theta(F_e))^2 \right]
\]

- $X_i$ are independent random variables having distribution function $F_e$.
- The quantity $\text{MSE}(F_e)$ is called the bootstrap approximation to the mean square error $\text{MSE}(F)$. 
Example V

- Interested in estimating $\theta(F) = E[X]$ by using the sample mean $\bar{X}_n = \sum_{i=1}^{n} x_i / n$.
- If observed data are $x_i$, $i=1,...,n$ then empirical distribution puts weights $1/n$ on each $x_i$.
- $\theta(F_e) = \bar{x} = \sum_{i=1}^{n} x_i / n$
- The bootstrap estimator of the mean square error-call it $MSE(F_e)$-is given by

$$MSE(F_e) = E_{F_e} \left[ \left( \sum_{i=1}^{n} \frac{X_i}{n} - \bar{x} \right)^2 \right]$$
Example V

- Since

\[ E_{F_e} \left[ \sum_{i=1}^{n} \frac{X_i}{n} \right] = E_{F_e} [X] = \bar{x} \]

- It follows that

\[
MSE(F_e) = Var_{F_e} \left( \sum_{i=1}^{n} \frac{X_i}{n} \right) \\
= \frac{Var_{F_e} (X)}{n}
\]
Example V

- Now

\[ \text{Var}_{F_e}(X) = E_{F_e} \left[ (X - E_{F_e}[X])^2 \right] \]

\[ = E_{F_e} \left[ (X - \bar{x})^2 \right] \]

\[ = \frac{1}{n} \left[ \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] \]

- And so

\[ \text{MSE}(F_e) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n^2} \]
Example V

\[ \text{MSE}(F_e) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n^2} \]

- compares well to \( S^2/n \) because the observed values is

\[
\frac{S^2}{n} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n(n-1)}
\]
Example VI

- Illustration of the bootstrap technique in analyzing the output of a queuing simulation.
- In Example I we are interested in estimating the long-run average amount of time a customer spends in the system.
- That is, letting $W_i$ be the amount of time the $i$th customer spends in the system $i \geq 1$, we are interested in

$$\theta \equiv \lim_{n \to \infty} \frac{W_1 + W_2 + \cdots + W_n}{n}$$
Example VI

- To show that above limit does exist (random variables $W_i$ are neither independent nor identically distributed) let $N_i$ denote customers that arrive at day $i$, and let

$$D_1 = W_1 + \cdots + W_{N_1}$$

$$D_2 = W_{N_1+1} + \cdots + W_{N_1+N_2}$$

and in general, for $i \geq 2$,

$$D_i = W_{N_1 + \cdots + N_{i-1} + 1} + \cdots + W_{N_1 + \cdots + N_i}$$

- $D_i$ is the sum of the times in the system of all arrivals on day $i$. 
Example VI

- $\theta$ can now be written as

$$\theta = \lim_{m \to \infty} \frac{D_1 + D_2 + \cdots + D_m}{N_1 + N_2 + \cdots + N_m}$$

$$\theta = \lim_{m \to \infty} \frac{D_1 + D_2 + \cdots + D_m}{m/N_1 + N_2 + \cdots + N_m / m}$$

This ratio is just the average time in the system of all customers.
Example VI

- Now each day follows the same probability law and random variables $D_1, \ldots, D_m$ and $N_1, \ldots, N_m$ are i.i.d.

- By strong law of large numbers average of the first $m$ of the $D_i$ will, with probability 1, converge to their common expectation. (Same is true for $N$.)

$$\theta = \frac{E[D]}{E[N]}$$

where $E[N]$ is the expected number of customers to arrive in a day, and $E[D]$ is the expected sum of the times those customers spend in the system.
Example VI

- To estimate $\Theta$ simulate system over $k$ days.
- Because $E[D]$ and $N[D]$ can be estimated by

$$\overline{D} = \frac{D_1 + D_2 + \cdots + D_k}{k}$$

$$\overline{N} = \frac{N_1 + N_2 + \cdots + N_k}{k}$$

- $\Theta = E[D]/N[D]$ can be estimated by

$$\overline{\frac{D}{N}} = \frac{D_1 + D_2 + \cdots + D_k}{N_1 + N_2 + \cdots + N_k}$$

which is just the average time in the system of all arrivals during the first $k$ days.
Example VI

- To estimate
  \[ MSE = E \left[ \left( \frac{\sum_{i=1}^{k} D_i}{\sum_{i=1}^{k} N_i} - \theta \right)^2 \right] \]
  we employ the bootstrap approach.

- Suppose that the simulation resulted in \( n_i \) arrivals on day \( i \) spending a total time \( d_i \) in the system.

- Under the empirical distribution function we have
  \[ P_{F_e} \{ D = d_i, N = n_i \} = \frac{1}{k}, \quad i = 1, \ldots, k \]
Example VI

Hence, \( E_{F_e}[D] = \bar{d} = \sum_{i=1}^{k} \frac{d_i}{k}, \quad E_{F_e}[N] = \bar{n} = \sum_{i=1}^{k} \frac{n_i}{k} \)

\[
\theta(F_e) = \frac{\bar{d}}{\bar{n}}
\]

\[
MSE(F_e) = E_{F_e} \left[ \left( \frac{\sum_{i=1}^{k} D_i}{\sum_{i=1}^{k} N_i} - \frac{\bar{d}}{\bar{n}} \right)^2 \right]
\]
Example VI

- Exact computation of the MSE would require computing the sum of $k^k$ terms
- Simulation experiment to approximate it.
- Compute
  \[
  Y_1 = \left( \frac{\sum_{i=1}^{k} D_i^1}{\sum_{i=1}^{k} N_i^1} - \frac{\bar{d}}{\bar{n}} \right)^2
  \]
- Then compute $Y_2, \ldots, Y_r$ ($r=100$ should be sufficient)
- The average of these $r$ values is then used to estimate $MSE(F_e)$
Summary

- Sample mean and variance
- Estimation of population variance
- When to stop generating new data?
- Interval estimates of population mean
- Bootstrapping technique