Introduction to Stochastic Simulation: 1. Ideas and Examples

Pierre L'Ecuyer
Université de Montréal, Canada

Operations Research Seminars, Zinal, June 2016

We focus on general ideas and methods that apply widely, not on specific applications.

Systems, models, and simulation

Model: simplified representation of a system.

Modeling: building a model, based on data and other information.

Simulation program: computerized (runnable) representation of the model. Some environments do not distinguish modeling and programming (no explicit programming).

Simulation: running the simulation program and observing the results. Much more convenient and powerful than experimenting with the real system.

Deterministic vs stochastic model.

Static vs dynamic model.

Analytic models vs simulation models

Analytic model: provides a simple formula. Usually hard to derive and requires unrealistic simplifying assumptions, but easy to apply and insightful.

Examples: Erlang formulas in queueing theory, Black-Scholes formula for option pricing in finance, etc.

Analytic models vs simulation models

Analytic model: provides a simple formula. Usually hard to derive and requires unrealistic simplifying assumptions, but easy to apply and insightful.

Examples: Erlang formulas in queueing theory, Black-Scholes formula for option pricing in finance, etc.

$M/M/s$ queue, arrival rate λ, service rate μ, $s > \lambda/\mu$ servers, steady-state. Let W the waiting time of a random customer.
There are exact expressions for $\mathbb{E}[W]$, $\mathbb{P}[W = 0]$, $\mathbb{P}[W > x]$, etc.
Conditionally on $W > 0$, W is exponential with mean $1/(s\mu - \lambda)$.

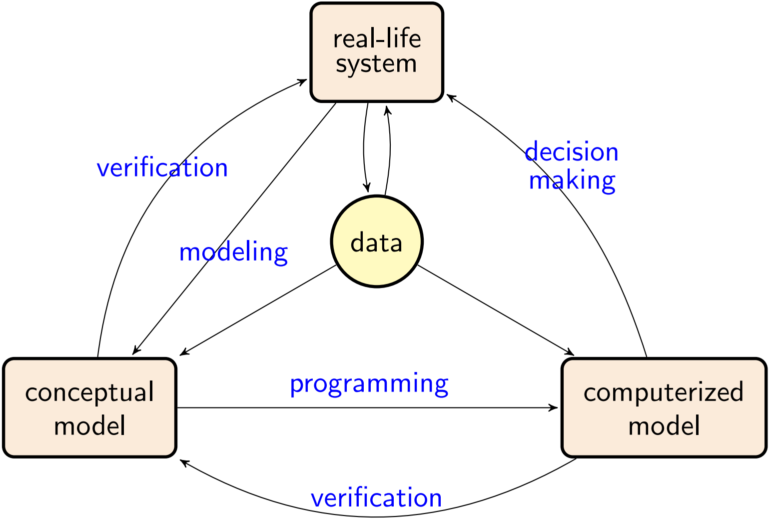
Analytic models vs simulation models

Analytic model: provides a simple formula. Usually hard to derive and requires unrealistic simplifying assumptions, but easy to apply and insightful.

Examples: Erlang formulas in queueing theory, Black-Scholes formula for option pricing in finance, etc.

$M/M/s$ queue, arrival rate λ, service rate μ, $s > \lambda/\mu$ servers, steady-state. Let W the waiting time of a random customer. There are exact expressions for $\mathbb{E}[W]$, $\mathbb{P}[W = 0]$, $\mathbb{P}[W > x]$, etc. Conditionally on $W > 0$, W is exponential with mean $1/(s\mu - \lambda)$.

Simulation model: can be much more detailed and realistic. Requires reliable data and information.



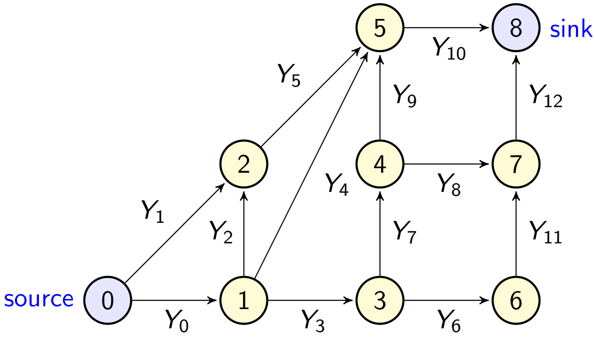
real-life system
data
conceptual model
computerized model
verification
modeling
decision making
programming
verification

Examples: A stochastic activity network

Gives precedence relations between activities. Activity k has random duration Y_k (also length of arc k) with known cumulative distribution function (cdf)

$$F_k(y) \stackrel{\text{def}}{=} \mathbb{P}[Y_k \leq y].$$

Project duration $T =$ (random) length of longest path from source to sink.



Simulation

Repeat n times:

generate Y_k with cdf F_k for each k, then compute T.

From those n realizations of T, we can estimate $\mathbb{E}[T]$, $\mathbb{P}[T > x]$ for some x, and compute confidence intervals on those values.

Better: construct a histogram to estimate the distribution of T.

Numerical illustration:

$Y_k \sim N(\mu_k, \sigma_k^2)$ for $k = 1, 2, 4, 11, 12$, and $V_k \sim \text{Expon}(1/\mu_k)$ otherwise.

$\mu_k, \ldots, \mu_{13}$: 13.0, 5.5, 7.0, 5.2, 16.5, 14.7, 10.3, 6.0, 4.0, 20.0, 3.2, 3.2, 16.5.

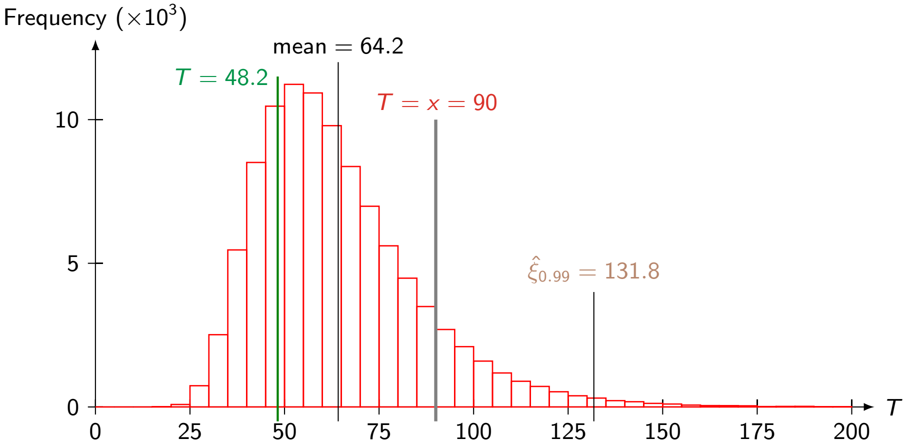
We may pay a penalty if $T > 90$, for example.

Naive idea: replace each Y_k by its expectation. Gives $T = 48.2$.

Naive idea: replace each Y_k by its expectation. Gives $T = 48.2$.

Results of an experiment with $n = 100\,000$.
Histogram of values of T gives much more information than confidence interval on $\mathbb{E}[T]$ or $\mathbb{P}[T > x]$ or $\hat{\xi}_{0.99}$.

Values from 14.4 to 268.6; 11.57% exceed $x = 90$.

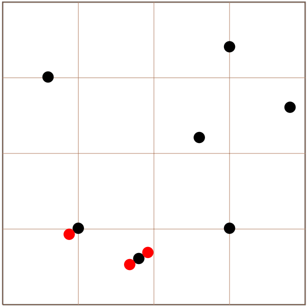


Collisions in a hashing system

In applied probability and statistics, Monte Carlo simulation is often used to estimate a distribution that is too hard to compute exactly.

Example: We throw M points randomly and independently into k squares.
C = number of collisions (when a point falls in square already occupied).
What is the distribution of C?

Exemple: $k = 100$, $M = 10$, $C = 3$.



One application: hashing. Large set of identifiers Φ, e.g., 64-bit numbers, only a few a used. Hashing function $h : \Phi \to \{1, \ldots, k\}$ (set of physical addresses), with $k \ll |\Phi|$.

Another application: testing random number generators.

One application: hashing. Large set of identifiers Φ, e.g., 64-bit numbers, only a few a used. Hashing function $h : \Phi \to \{1, \ldots, k\}$ (set of physical addresses), with $k \ll |\Phi|$.

Another application: testing random number generators.

Simplified analytic model:
We consider two cases: $M = m$ (constant) and $M \sim \mathrm{Poisson}(m)$.

Theorem:
If $k \to \infty$ and $m^2/(2k) \to \lambda$ (a constant) then $C \Rightarrow \mathrm{Poisson}(\lambda)$.
That is, in the limit, $\mathbb{P}[C = x] = e^{-\lambda}\lambda^x/x!$ for $x = 0, 1, 2, \ldots$.

But how good is this approximation for finite k and m?

Simulation:
Generate M indep. integers uniformly over $\{1, \ldots, k\}$ and compute C. Repeat n times independently and compute the empirical distribution of the n realizations of C. Can also estimate $\mathbb{E}[C]$, $\mathbb{P}[C > x]$, etc.

Simulation:
Generate M indep. integers uniformly over $\{1, \ldots, k\}$ and compute C. Repeat n times independently and compute the empirical distribution of the n realizations of C. Can also estimate $\mathbb{E}[C]$, $\mathbb{P}[C > x]$, etc.

Illustration:
$k = 10\,000$ and $m = 400$.
Analytic approx.: $C \sim \mathrm{Poisson}$ with mean $\lambda = m^2/(2k) = 8$.

We simulated $n = 10^7$ replications (or runs).
Empirical mean and variance:
7.87 and 7.47 for $M = m = 400$ (fixed);
7.89 and 8.10 for $M \sim$ Poisson(400).

c	M fixed	$M \sim$ Poisson	Poisson distrib.
0	3181	4168	3354.6
1	25637	32257	26837.0
2	105622	122674	107348.0
3	288155	316532	286261.4
4	587346	614404	572522.8
5	948381	957951	916036.6
6	1269427	1247447	1221382.1
7	1445871	1397980	1395865.3
8	1434562	1377268	1395865.3
9	1251462	1207289	1240769.1
10	978074	958058	992615.3
11	688806	692416	721902.0
12	442950	459198	481268.0
13	260128	282562	296164.9
14	141467	162531	169237.1
15	71443	86823	90259.7
16	33224	43602	45129.9
17	14739	20827	21237.6
18	5931	9412	9438.9
19	2378	3985	3974.3
20	791	1629	1589.7
21	310	592	605.6
22	79	264	220.2
23	26	83	76.6
24	5	33	25.5
≥ 25	5	15	11.7

What if we take smaller values?
$k = 100$ and $m = 40$.
We still have $\lambda = m^2/(2k) = 8$.

We simulated $n = 10^7$ replications (or runs).
Empirical mean, 95% confidence interval, variance:
6.90, (6.896, 6.899), 4.10, for $M = m = 40$ (fixed);
7.03, (7.032, 7.035), 8.48, for $M \sim \text{Poisson}(40)$.

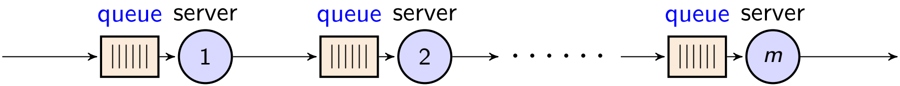
c	M fixed	$M \sim$ Poisson	Poisson distrib.
0	1148	17631	3354.6
1	14355	100100	26837.0
2	84046	294210	107348.0
3	302620	600700	286261.4
4	744407	951184	572522.8
5	1340719	1238485	916036.6
6	1828860	1379400	1221382.1
7	1941207	1353831	1395865.3
8	1634465	1194915	1395865.3
9	1103416	956651	1240769.1
10	602186	705478	992615.3
11	267542	485353	721902.0
12	97047	309633	481268.0
13	29161	187849	296164.9
14	7195	107189	169237.1
15	1363	58727	90259.8
16	224	30450	45129.9
17	35	15115	21237.6
18	4	7271	9438.9
19	0	3311	3974.3
20	0	1468	1589.7
21	0	607	605.6
22	0	258	220.2
23	0	110	76.6
24	0	44	25.5
≥ 25	0	30	11.7

$k = 25$ and $m = 20$.
Again, $\lambda = m^2/(2k) = 8$.

We make $n = 10^7$ runs.
Empirical mean, variance:
6.05, 2.16 for $M = m = 40$ (fixed);
6.23, 8.21 for $M \sim \text{Poisson}(40)$.

c	M fixed	$M \sim$ Poisson	Poisson distrib.
0	135	49506	3354.6
1	4565	220731	26837.0
2	52910	528292	107348.0
3	314067	905283	286261.4
4	1050787	1224064	572522.9
5	2130621	1400387	916036.6
6	2692369	1392858	1221382.1
7	2170849	1237244	1395865.3
8	1124606	997974	1395865.3
9	371366	743393	1240769.2
10	77190	513360	992615.3
11	9768	332413	721902.0
12	738	203692	481268.0
13	29	118015	296164.9
14	0	65363	169237.1
15	0	34266	90259.8
16	0	17384	45129.9
17	0	8600	21237.6
18	0	3984	9438.9
19	0	1861	3974.2
20	0	768	1589.7
21	0	345	605.6
22	0	128	220.2
23	0	56	76.6
24	0	15	25.5
25	0	18	11.7

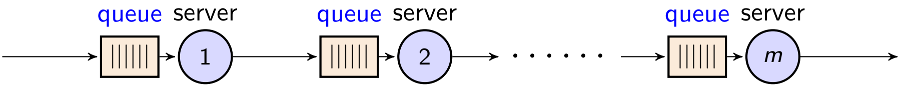
A tandem queue.



Station j: single-server FIFO queue with total capacity c_j $(c_1 = \infty)$.
A customer cannot leave server j when queue $j+1$ is full; it is blocked.

T_i = arrival time of customer i at first queue, for $i \geq 1$, and $T_0 = 0$;
$A_i = T_i - T_{i-1}$ = time between arrivals $i-1$ and i;
$W_{j,i}$ = waiting time in queue j for customer i;
$S_{j,i}$ = service time at station j for customer i;
$B_{j,i}$ = blocked time at station j for customer i;
$D_{j,i}$ = departure time from station j for customer i.

A tandem queue.



Station j: single-server FIFO queue with total capacity c_j ($c_1 = \infty$).
A customer cannot leave server j when queue $j+1$ is full; it is blocked.

T_i = arrival time of customer i at first queue, for $i \geq 1$, and $T_0 = 0$;
$A_i = T_i - T_{i-1}$ = time between arrivals $i-1$ and i;
$W_{j,i}$ = waiting time in queue j for customer i;
$S_{j,i}$ = service time at station j for customer i;
$B_{j,i}$ = blocked time at station j for customer i;
$D_{j,i}$ = departure time from station j for customer i.

First customer arrives time $T_1 = A_1$, leaves station 1 at $D_{1,1} = T_1 + S_{1,1}$ and starts its service at station 2, leaves station 2 at time $D_{2,1} = D_{1,1} + S_{2,1}$ to join queue 3, etc.
Customer 2 arrives at time $T_2 = T_1 + A_2$, starts service at queue 1 at $\max(T_2, D_{1,1})$, leaves station 1 at time $\max(T_2, D_{1,1}) + S_{1,2}$ if space is available at station 2, etc.

Recurence equations.

Suppose we can generate successive interarrival times A_i and service times $S_{j,i}$. Then we can compute the T_i's easily and the departure times $D_{j,i}$ via:

$$D_{j,i} = \max[D_{j-1,i} + S_{j,i},\ D_{j,i-1} + S_{j,i}\ ,\ D_{j+1,i-c_{j+1}}\ ]$$

for $1 \leq j \leq m$ and $i \geq 1$, where $D_{0,i} = T_i$, $D_{j,i} = 0$ for $i \leq 0$, and $D_{m+1,i} = 0$ for all i. Then:

$$\begin{aligned} W_{j,i} &= \max[0,\ D_{j,i-1} - D_{j-1,i}], \\ B_{j,i} &= D_{j,i} - D_{j-1,i} - W_{j,i} - S_{j,i}. \end{aligned}$$

Recurence equations.

Suppose we can generate successive interarrival times A_i and service times $S_{j,i}$. Then we can compute the T_i's easily and the departure times $D_{j,i}$ via:

$$D_{j,i} = \max[D_{j-1,i} + S_{j,i},\ D_{j,i-1} + S_{j,i}\ ,\ D_{j+1,i-c_{j+1}}\ ]$$

for $1 \leq j \leq m$ and $i \geq 1$, where $D_{0,i} = T_i$, $D_{j,i} = 0$ for $i \leq 0$, and $D_{m+1,i} = 0$ for all i. Then:

$$\begin{aligned} W_{j,i} &= \max[0,\ D_{j,i-1} - D_{j-1,i}], \\ B_{j,i} &= D_{j,i} - D_{j-1,i} - W_{j,i} - S_{j,i}. \end{aligned}$$

For infinite buffer sizes (no blocking), we always have $B_{j,i} = 0$ and

$$D_{j,i} = \max[D_{j-1,i},\ D_{j,i-1}] + S_{j,i}.$$

Recurence equations.

Suppose we can generate successive interarrival times A_i and service times $S_{j,i}$. Then we can compute the T_i's easily and the departure times $D_{j,i}$ via:

$$D_{j,i} = \max[D_{j-1,i} + S_{j,i},\ D_{j,i-1} + S_{j,i}\ ,\ D_{j+1,i-c_{j+1}}\ ]$$

for $1 \le j \le m$ and $i \ge 1$, where $D_{0,i} = T_i$, $D_{j,i} = 0$ for $i \le 0$, and $D_{m+1,i} = 0$ for all i. Then:

$$\begin{aligned} W_{j,i} &= \max[0,\ D_{j,i-1} - D_{j-1,i}], \\ B_{j,i} &= D_{j,i} - D_{j-1,i} - W_{j,i} - S_{j,i}. \end{aligned}$$

For infinite buffer sizes (no blocking), we always have $B_{j,i} = 0$ and

$$D_{j,i} = \max[D_{j-1,i},\ D_{j,i-1}] + S_{j,i}.$$

A single queue with no blocking: $GI/GI/1$ model, obeys Lindley recurrence:

$$W_{1,i} = \max[0,\ D_{1,i-1} - T_i] = \max[0,\ W_{1,i-1} + S_{1,i-1} - A_i].$$

Algorithm: Simulating N_c customers with production blocking.

Let $T_0 = 0$ and $D_{0,0} = D_{1,0} = 0$;
For $j = 2, \ldots, m$,
  For $i = -c_j + 1, \ldots, 0$, let $D_{j,i} = 0$;
For $i = 1, \ldots, N_c$,
  Generate A_i from its distribution and let $D_{0,i} = T_i = T_{i-1} + A_i$;
  Let $W_i = B_i = 0$;
  For $j = 1, \ldots, m$,
    Generate $S_{j,i}$ from its distribution;
    Let $D_{j,i} = \max[D_{j-1,i} + S_{j,i},\ D_{j,i-1} + S_{j,i},\ D_{j+1,i-c_{j+1}}]$;
    Let $W_{j,i} = \max[0,\ D_{j,i-1} - D_{j-1,i}]$ and $W_i = W_i + W_{j,i}$;
    Let $B_{j,i} = D_{j,i} - D_{j-1,i} - W_{j,i} - S_{j,i}$ and $B_i = B_i + B_{j,i}$;
Compute and return the averages

$$\bar{W}_{N_c} = (W_1 + \cdots + W_{N_c})/N_c \text{ and } \bar{B}_{N_c} = (B_1 + \cdots + B_{N_c})/N_c.$$

Algorithm: Simulating N_c customers with production blocking.

Let $T_0 = 0$ and $D_{0,0} = D_{1,0} = 0$;
For $j = 2, \ldots, m$,
  For $i = -c_j + 1, \ldots, 0$, let $D_{j,i} = 0$;
For $i = 1, \ldots, N_c$,
  Generate A_i from its distribution and let $D_{0,i} = T_i = T_{i-1} + A_i$;
  Let $W_i = B_i = 0$;
  For $j = 1, \ldots, m$,
    Generate $S_{j,i}$ from its distribution;
    Let $D_{j,i} = \max[D_{j-1,i} + S_{j,i},\ D_{j,i-1} + S_{j,i}\ ,\ D_{j+1,i-c_{j+1}}\ ]$;
    Let $W_{j,i} = \max[0,\ D_{j,i-1} - D_{j-1,i}]$ and $W_i = W_i + W_{j,i}$;
    Let $B_{j,i} = D_{j,i} - D_{j-1,i} - W_{j,i} - S_{j,i}$ and $B_i = B_i + B_{j,i}$;
Compute and return the averages

$$\bar{W}_{N_c} = (W_1 + \cdots + W_{N_c})/N_c \text{ and } \bar{B}_{N_c} = (B_1 + \cdots + B_{N_c})/N_c .$$

What if we want to simulate all customers who arrive before time T?
Replace "For $i = 1, \ldots, N_c$" by "For ($i = 1,\ T_i < T,\ i{+}{+}$)".

Algorithm: Simulating N_c customers with production blocking.

Let $T_0 = 0$ and $D_{0,0} = D_{1,0} = 0$;
For $j = 2, \ldots, m$,
    For $i = -c_j + 1, \ldots, 0$, let $D_{j,i} = 0$;
For $i = 1, \ldots, N_c$,
    Generate A_i from its distribution and let $D_{0,i} = T_i = T_{i-1} + A_i$;
    Let $W_i = B_i = 0$;
    For $j = 1, \ldots, m$,
        Generate $S_{j,i}$ from its distribution;
        Let $D_{j,i} = \max[D_{j-1,i} + S_{j,i},\ D_{j,i-1} + S_{j,i},\ D_{j+1,i-c_{j+1}}]$;
        Let $W_{j,i} = \max[0,\ D_{j,i-1} - D_{j-1,i}]$ and $W_i = W_i + W_{j,i}$;
        Let $B_{j,i} = D_{j,i} - D_{j-1,i} - W_{j,i} - S_{j,i}$ and $B_i = B_i + B_{j,i}$;
Compute and return the averages

$$\bar{W}_{N_c} = (W_1 + \cdots + W_{N_c})/N_c \text{ and } \bar{B}_{N_c} = (B_1 + \cdots + B_{N_c})/N_c.$$

What if we want to simulate all customers who arrive before time T?
Replace "For $i = 1, \ldots, N_c$" by "For ($i = 1$, $T_i < T$, i++)".

What we just saw is production blocking. In communication blocking, service at station j starts only when queue $j+1$ is not full. Then:

$$D_{j,i} = \max[D_{j-1,i},\ D_{j,i-1},\ D_{j+1,i-c_{j+1}}] + S_{j,i}.$$

When things get more complicated: **Discrete-event simulation**.

Example: Pricing a financial derivative.

Consider a financial contract based on a single asset (e.g., one share of a stock), whose market price at time t is $S(t)$. We assume $\{S(t),\ t \geq 0\}$ is a stochastic process with known probability law (in practice, estimated from data).

Example: Pricing a financial derivative.

Consider a financial contract based on a single asset (e.g., one share of a stock), whose market price at time t is $S(t)$. We assume $\{S(t),\ t \geq 0\}$ is a stochastic process with known probability law (in practice, estimated from data).

Suppose the contract owner receives a net payoff $g(S(t_1), \ldots, S(t_d))$ at time T, for some function $g : \mathbb{R}^d \to \mathbb{R}$, where $t_1 \ldots, t_d$ are fixed observation times, with $0 \leq t_1 < \cdots < t_d = T$.

Suppose also that money left in the bank account (or borrowed, when negative) yields interest at compounded rate r (the short rate). This means that one dollar placed in the account at time 0 is worth e^{rt} dollars at time t. Equivalently, an amount Y to be received in t units of time is worth $e^{-rt}Y$ today. That is, it must be multiplied by the discount factor e^{-rt}.

Example: Pricing a financial derivative.

Consider a financial contract based on a single asset (e.g., one share of a stock), whose market price at time t is $S(t)$. We assume $\{S(t),\ t \geq 0\}$ is a stochastic process with known probability law (in practice, estimated from data).

Suppose the contract owner receives a net payoff $g(S(t_1), \ldots, S(t_d))$ at time T, for some function $g : \mathbb{R}^d \to \mathbb{R}$, where $t_1 \ldots, t_d$ are fixed observation times, with $0 \leq t_1 < \cdots < t_d = T$.

Suppose also that money left in the bank account (or borrowed, when negative) yields interest at compounded rate r (the short rate). This means that one dollar placed in the account at time 0 is worth e^{rt} dollars at time t. Equivalently, an amount Y to be received in t units of time is worth $e^{-rt}Y$ today. That is, it must be multiplied by the discount factor e^{-rt}.

Our aim is to estimate the fair (present) value of the financial contract.
To do this, it is common to assume a no-arbitrage market, which means that one cannot make more money than with the interest rate r without taking risks. That is, if there exists an investment strategy with which one dollar at time 0 can be worth more than e^{rt} dollars at time t with positive probability, then there must be a positive probability that it is worth less than e^{rt}.

Under this no-arbitrage assumption, it can be proved (this is beyond the scope of this course) that the present value (or fair price) of the financial contract (at time 0), when $S(0) = s_0$, can be written as

$$v(s_0, T) = \mathbb{E}^*\left[e^{-rT} g(S(t_1), \ldots, S(t_d))\right],$$

where $\mathbb{E}^*$ is the mathematical expectation under a certain probability measure $\mathbb{P}^*$ called the risk-neutral measure.

Changing the probability measure amounts to a change of variable to account for our utility function.

Under this no-arbitrage assumption, it can be proved (this is beyond the scope of this course) that the present value (or fair price) of the financial contract (at time 0), when $S(0) = s_0$, can be written as

$$v(s_0, T) = \mathbb{E}^* \left[e^{-rT} g(S(t_1), \ldots, S(t_d)) \right],$$

where $\mathbb{E}^*$ is the mathematical expectation under a certain probability measure $\mathbb{P}^*$ called the risk-neutral measure.

Changing the probability measure amounts to a change of variable to account for our utility function.

Under this measure $\mathbb{P}^*$, the process $\{e^{-rt}S(t),\ t \geq 0\}$ is a martingale, which means that for any $t \geq 0$ and $0 < \delta \leq T - t$, we have

$$\mathbb{E}^*[e^{-\delta t} S(t + \delta)] = S(t).$$

Under this no-arbitrage assumption, it can be proved (this is beyond the scope of this course) that the present value (or fair price) of the financial contract (at time 0), when $S(0) = s_0$, can be written as

$$v(s_0, T) = \mathbb{E}^* \left[e^{-rT} g(S(t_1), \ldots, S(t_d)) \right],$$

where $\mathbb{E}^*$ is the mathematical expectation under a certain probability measure $\mathbb{P}^*$ called the risk-neutral measure.

Changing the probability measure amounts to a change of variable to account for our utility function.

Under this measure $\mathbb{P}^*$, the process $\{e^{-rt}S(t),\ t \geq 0\}$ is a martingale, which means that for any $t \geq 0$ and $0 < \delta \leq T - t$, we have

$$\mathbb{E}^*[e^{-\delta t} S(t + \delta)] = S(t).$$

The measure $\mathbb{P}^*$ generally differs from the true measure under which the process $\{S(t),\ t \geq 0\}$ evolves in real life.
Note that a risk-neutral measure does not always exist, and is not always unique. Here, we assume that it does exist and is known (or has been estimated).

Except for very simple models, we do not know how to compute $v(s_0, T)$ exactly. But we can estimate it by Monte Carlo if we know how to generate a path $S(t_1), \ldots, S(t_d)$ from $\mathbb{P}^*$ and to compute g. We simulate n independent copies of $X = e^{-rT} g(S(t_1), \ldots, S(t_d))$, say $X_1, \ldots, X_n$, and our estimator is the average $\bar{X}_n = (X_1 + \cdots + X_n)/n$.

Except for very simple models, we do not know how to compute $v(s_0, T)$ exactly. But we can estimate it by Monte Carlo if we know how to generate a path $S(t_1), \ldots, S(t_d)$ from $\mathbb{P}^*$ and to compute g. We simulate n independent copies of $X = e^{-rT} g(S(t_1), \ldots, S(t_d))$, say $X_1, \ldots, X_n$, and our estimator is the average $\bar{X}_n = (X_1 + \cdots + X_n)/n$.

What are the appropriate models for S?
In the popular model of Black and Scholes (1973), $S(t)$ is assumed to evolve as a geometric Brownian motion (GBM), which means that its logarithm is a Brownian motion. This implies that under $\mathbb{P}^*$, one has

$$S(t) = S(0)e^{(r-\sigma^2/2)t+\sigma B(t)}$$

where r is the short rate, σ is a constant called the volatility, and $B(\cdot)$ is a standard Brownian motion: for any $t_2 > t_1 \geq 0$, $B(t_2) - B(t_1)$ is a normal random variable with mean 0 and variance $t_2 - t_1$, independent of the behavior of $B(\cdot)$ outside the interval $[t_1, t_2]$.

An European call option gives the right to buy one unit of the asset at price K (the strike price) at time T (the expiration date). The net payoff at time T is

$$g(S(T)) = \max[0, S(T) - K].$$

In this simple case, under the GBM model, the Black-Scholes formula gives:

$$v(s_0, T) \quad = \quad s_0 \Phi(-z_0 + \sigma\sqrt{T}) - Ke^{-rT}\Phi(-z_0),$$

where

$$z_0 = \frac{\ln(K/s_0) - (r - \sigma^2/2)T}{\sigma\sqrt{T}}$$

and Φ is the standard normal cdf.

When g or S is more complicated, there is usually no analytic formula for $v(s_0, T)$. For a discretely-monitored Asian call option, for example, we have

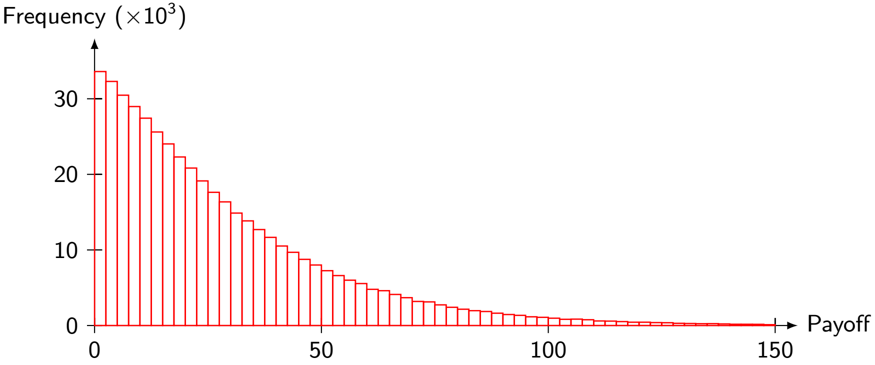
$$g(S(t_1), \ldots, S(t_d)) = \max\left(0, \frac{1}{d}\sum_{j=1}^{d} S(t_j) - K\right),$$

and no closed-form formula is available for $v(s_0, T)$. We can estimate it by simulation, as follows (for a general g):

Algorithm: Option pricing under a GBM model by Monte Carlo
For $i = 1, \ldots, n$,
    Let $t_0 = 0$ and $B(t_0) = 0$;
    For $j = 1, \ldots, d$,
        Generate $Z_j \sim N(0, 1)$;
        Let $B(t_j) = B(t_{j-1}) + \sqrt{t_j - t_{j-1}}\, Z_j$;
        Let $S(t_j) = s_0 \exp\left[(r - \sigma^2/2)t_j + \sigma B(t_j)\right]$;
    Compute $X_i = e^{-rT} g(S(t_1), \ldots, S(t_d))$;
Compute and return the average $\bar{X}_n$ as an estimator of $v(s_0, T)$, with a c.i.;
Optional: Plot a histogram of $X_1, \ldots, X_n$.

Numerical illustration: Bermudean Asian option with $d = 12$, $T = 1$ (one year), $t_j = j/12$ for $j = 0, \ldots, 12$, $K = 100$, $s_0 = 100$, $r = 0.05$, $\sigma = 0.5$.

We perform $n = 10^6$ independent simulation runs. Mean: **13.1**. Max = 390.8
In 53.47% of cases, the payoff is 0. Histogram of the 46.53% positive values:



Other models

Better models for S: Variance-gamme, Heston, etc.

Several types of payoff functions g.

Improving the accuracy when estimating the mean

If we replace the arithmetic average by a geometric average, we obtain

$$C = e^{-rT} \max\left(0, \prod_{j=1}^{d} (S(t_j))^{1/d} - K\right),$$

whose expectation $\nu = \mathbb{E}[C]$ has a closed-form formula.

One can then use C as a control variate to estimate $\mathbb{E}[X]$: Replace the estimator X by the "corrected" version $X - \beta(C - \nu)$, for some well-chosen β.

This can provide a huge variance reduction, by factors of over a million in some examples.

American options

Here the option owner can decide when to exercise the option.

We now have a dynamic stochastic (stopping time) optimization problem.

Solution methods involve variants of approximate dynamic programming combined with simulation. Interesting and challenging.

In some cases (e.g., bonds with options), both the issuer and the owner can decide dynamically when to exercise the option.

Discrete choice with multinomial mixed logit probability, max likelihood estimation

Want to estimate a model of choices between alternatives, for when users buy a product, choose a route or a transportation mode, choose a restaurant, etc.

Utility of alternative j for individual q is

$$
\begin{aligned}
U_{q,j} &= \sum_{\ell=1}^{s} \beta_{q,\ell} x_{q,j,\ell} + \epsilon_{q,j} = \boldsymbol{\beta}_q^{\mathrm{t}} \mathbf{x}_{q,j} + \epsilon_{q,j}, \quad \text{where} \\
\boldsymbol{\beta}_q^{\mathrm{t}} &= (\beta_{q,1}, \ldots, \beta_{q,s}) \text{ gives the tastes of individual } q, \\
\mathbf{x}_{q,j}^{\mathrm{t}} &= (x_{q,j,1}, \ldots, x_{q,j,s}) \text{ attributes of alternative } j \text{ for individual } q, \\
\epsilon_{q,j} & \quad \text{noise; Gumbel of mean 0 and scale parameter } \lambda = 1.
\end{aligned}
$$

Individual q selects alternative y_q with largest utility $U_{q,j}$.

Can observe the $\mathbf{x}_{q,j}$ and choices y_q, but not the rest.

Logit model: for $\boldsymbol{\beta}_q$ fixed, j is chosen with probability

$$L_q(j \mid \boldsymbol{\beta}_q) = \frac{\exp[\boldsymbol{\beta}_q^{\mathrm{t}} \mathbf{x}_{q,j}]}{\sum_{a \in \mathcal{A}(q)} \exp[\boldsymbol{\beta}_q^{\mathrm{t}} \mathbf{x}_{q,a}]}$$

where $\mathcal{A}(q)$ are the available alternatives for q.

Logit model: for $\boldsymbol{\beta}_q$ fixed, j is chosen with probability

$$L_q(j \mid \boldsymbol{\beta}_q) = \frac{\exp[\boldsymbol{\beta}_q^{\mathsf{t}} \mathbf{x}_{q,j}]}{\sum_{a \in \mathcal{A}(q)} \exp[\boldsymbol{\beta}_q^{\mathsf{t}} \mathbf{x}_{q,a}]}$$

where $\mathcal{A}(q)$ are the available alternatives for q.

Mixed logit. For a random individual, suppose $\boldsymbol{\beta}_q$ is random with density $f_{\boldsymbol{\theta}}$, which depends on (unknown) parameter vector $\boldsymbol{\theta}$. We want to estimate $\boldsymbol{\theta}$ from the data (the $\mathbf{x}_{q,j}$ and y_q).

Given $\boldsymbol{\theta}$, the unconditional probability of choosing j is

$$p_q(j, \boldsymbol{\theta}) = \int L_q(j \mid \boldsymbol{\beta}) f_{\boldsymbol{\theta}}(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta}.$$

It depends on $\mathcal{A}(q)$, j, and $\boldsymbol{\theta}$.

Maximum likelihood: Maximize the log of the joint probability of the sample, w.r.t. $\boldsymbol{\theta}$:

$$\ln L(\boldsymbol{\theta}) = \ln \prod_{q=1}^{m} p_q(y_q, \boldsymbol{\theta}) = \sum_{q=1}^{m} \ln p_q(y_q, \boldsymbol{\theta}).$$

Maximum likelihood: Maximize the log of the joint probability of the sample, w.r.t. $\boldsymbol{\theta}$:

$$\ln L(\boldsymbol{\theta}) = \ln \prod_{q=1}^{m} p_q(y_q, \boldsymbol{\theta}) = \sum_{q=1}^{m} \ln p_q(y_q, \boldsymbol{\theta}).$$

No formula for $p_q(j, \boldsymbol{\theta})$, but can estimate it by simulation.

Generate n realizations of $\boldsymbol{\beta}$ from $f_{\boldsymbol{\theta}}$, say $\boldsymbol{\beta}_q^{(1)}(\boldsymbol{\theta}), \ldots, \boldsymbol{\beta}_q^{(n)}(\boldsymbol{\theta})$, and estimate $p_q(y_q, \boldsymbol{\theta})$ by

$$\hat{p}_q(y_q, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} L_q(j, \boldsymbol{\beta}_q^{(i)}(\boldsymbol{\theta})).$$

Then we can find the maximizer $\hat{\boldsymbol{\theta}}$ of $\sum_{q=1}^{m} \ln \hat{p}_q(y_q, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$.
The latter is a deterministic nonlinear nonconvex (difficult) optimization problem. Also **Biased!**

Maximum likelihood: Maximize the log of the joint probability of the sample, w.r.t. $\boldsymbol{\theta}$:

$$\ln L(\boldsymbol{\theta}) = \ln \prod_{q=1}^{m} p_q(y_q, \boldsymbol{\theta}) = \sum_{q=1}^{m} \ln p_q(y_q, \boldsymbol{\theta}).$$

No formula for $p_q(j, \boldsymbol{\theta})$, but can estimate it by simulation.

Generate n realizations of $\boldsymbol{\beta}$ from $f_{\boldsymbol{\theta}}$, say $\boldsymbol{\beta}_q^{(1)}(\boldsymbol{\theta}), \ldots, \boldsymbol{\beta}_q^{(n)}(\boldsymbol{\theta})$, and estimate $p_q(y_q, \boldsymbol{\theta})$ by

$$\hat{p}_q(y_q, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} L_q(j, \boldsymbol{\beta}_q^{(i)}(\boldsymbol{\theta})).$$

Then we can find the maximizer $\hat{\boldsymbol{\theta}}$ of $\sum_{q=1}^{m} \ln \hat{p}_q(y_q, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$.
The latter is a deterministic nonlinear nonconvex (difficult) optimization problem. Also **Biased!**

Simulation is usually combined with bias reduction and variance reduction methods to improve the accuracy.

Monte Carlo to estimate an expectation

Suppose we want to estimate the mathematical expectation (the mean) of some random variable X:

$$\mu = \mathbb{E}[X].$$

Monte Carlo estimator:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

where $X_1, \ldots, X_n$ are independent replicates of X.

We have $\mathbb{E}[\bar{X}_n] = \mu$ and $\mathrm{Var}[\bar{X}_n] = \sigma^2/n = \mathrm{Var}[X]/n$.

Convergence

Theorem. Suppose $\sigma^2 < \infty$. When $n \to \infty$:
(i) Strong law of large numbers: $\lim_{n\to\infty} \hat{\mu}_n = \mu$ with probability 1.

Convergence

Theorem. Suppose $\sigma^2 < \infty$. When $n \to \infty$:
(i) Strong law of large numbers: $\lim_{n\to\infty} \hat{\mu}_n = \mu$ with probability 1.
(ii) Central limit theorem (CLT):

$$\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma} \Rightarrow N(0,1), \text{ i.e.,}$$

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma} \leq x\right] = \Phi(x) = \mathbb{P}[Z \leq x]$$

for all $x \in \mathbb{R}$, where $Z \sim N(0,1)$ and $\Phi(\cdot)$ its cdf.

Convergence

Theorem. Suppose $\sigma^2 < \infty$. When $n \to \infty$:
(i) Strong law of large numbers: $\lim_{n\to\infty} \hat{\mu}_n = \mu$ with probability 1.
(ii) Central limit theorem (CLT):

$$\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma} \Rightarrow N(0,1), \text{ i.e.,}$$

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma} \leq x\right] = \Phi(x) = \mathbb{P}[Z \leq x]$$

for all $x \in \mathbb{R}$, where $Z \sim N(0,1)$ and $\Phi(\cdot)$ its cdf.

Property (ii) still holds if we replace σ^2 by its unbiased estimator

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n(\bar{X}_n)^2\right),$$

where $X_i = f(\mathbf{U}_i)$. We have $\sqrt{n}(\hat{\mu}_n - \mu)/S_n \Rightarrow N(0,1)$.

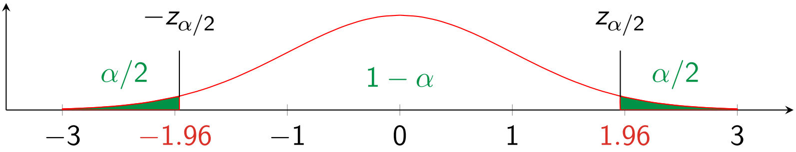
For large n and confidence level $1-\alpha$, we have

$$\mathbb{P}[\hat{\mu}_n - \mu \le x S_n/\sqrt{n}] = \mathbb{P}[\sqrt{n}(\hat{\mu}_n - \mu)/S_n \le x] \approx \Phi(x).$$

Confidence interval at level α (we want $\Phi(x) = 1-\alpha/2$):

$$(\hat{\mu}_n \pm z_{\alpha/2} S_n/\sqrt{n}), \text{ where } z_{\alpha/2} = \Phi^{-1}(1-\alpha/2).$$

Example: $z_{\alpha/2} \approx 1.96$ for $\alpha = 0.05$.



The width of the confidence interval is asymptotically proportional to $\sigma/\sqrt{n}$, so it converges as $O(n^{-1/2})$. Relative error: $\sigma/(\mu\sqrt{n})$.
We want a small σ.
To add one more decimal digit of accuracy, we must multiply n by 100.

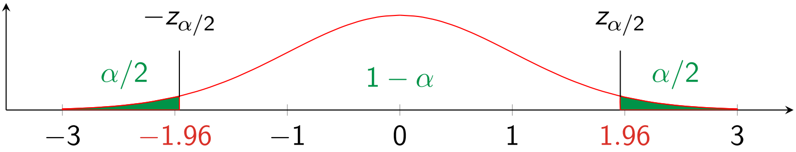
For large n and confidence level $1-\alpha$, we have

$$\mathbb{P}[\hat{\mu}_n - \mu \le x S_n/\sqrt{n}] = \mathbb{P}[\sqrt{n}(\hat{\mu}_n - \mu)/S_n \le x] \approx \Phi(x).$$

Confidence interval at level α (we want $\Phi(x) = 1-\alpha/2$):

$$(\hat{\mu}_n \pm z_{\alpha/2} S_n/\sqrt{n}), \text{ where } z_{\alpha/2} = \Phi^{-1}(1-\alpha/2).$$

Example: $z_{\alpha/2} \approx 1.96$ for $\alpha = 0.05$.



The width of the confidence interval is asymptotically proportional to $\sigma/\sqrt{n}$, so it converges as $O(n^{-1/2})$. Relative error: $\sigma/(\mu\sqrt{n})$.
We want a small σ.
To add one more decimal digit of accuracy, we must multiply n by 100.

Warning: If the X_i have an asymmetric law, these confidence intervals can have very bad coverage (convergence can be very slow).

Example: Stochastic activity network. For $\mu = \mathbb{P}[T > x]$, we have

$$\begin{aligned} X_i &= \mathbb{I}[T_i > x], \\ \hat{\mu}_n &= \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{Y}{n}, \\ S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - Y/n)^2 = \frac{Y(1-Y/n)}{n-1}. \end{aligned}$$

Confidence interval at 95%: $(Y/n \pm 1.96 S_n/\sqrt{n})$.
Reasonable if $\mathbb{E}[Y] = n\mu$ is not too close to 0 or 1.

In fact, $Y \sim \text{Binomial}(n, \mu)$.

Suppose $n = 1000$ and we observe $Y = 882$. Then:

$\bar{X}_n = 882/1000 = 0.882$;

$S_n^2 = \bar{X}_n(1 - \bar{X}_n)n/(n-1) \approx 0.1042$.

Suppose $n = 1000$ and we observe $Y = 882$. Then:

$\bar{X}_n = 882/1000 = 0.882$;

$S_n^2 = \bar{X}_n(1 - \bar{X}_n)n/(n-1) \approx 0.1042$.

This gives the 95% confidence interval:

$(\bar{X}_n \pm 1.96 S_n/\sqrt{n}) \approx (0.882 \pm 0.020) = (0.862, 0.902)$.

Suppose $n = 1000$ and we observe $Y = 882$. Then:

$\bar{X}_n = 882/1000 = 0.882$;

$S_n^2 = \bar{X}_n(1 - \bar{X}_n)n/(n-1) \approx 0.1042$.

This gives the 95% confidence interval:

$(\bar{X}_n \pm 1.96 S_n/\sqrt{n}) \approx (0.882 \pm 0.020) = (0.862, 0.902)$.

Thus, our estimator of μ has two significant digits: $\mu \approx 0.88$.
The "2" in 0.882 is not significant.

Simulation efficiency

We define the efficiency of the estimator by

$$\mathrm{Eff}(X) = \frac{1}{c(X) \cdot \mathrm{Var}(X)}$$

where $c(X)$ is the (expected) computing cost of X.

This measure does not depend on the computing budget:

$$\mathrm{Eff}(\bar{X}_n) = \frac{1}{n \cdot c(X) \cdot \mathrm{Var}(X)/n} = \mathrm{Eff}(X).$$

Simulation efficiency

We define the efficiency of the estimator by

$$\mathrm{Eff}(X) = \frac{1}{c(X) \cdot \mathrm{Var}(X)}$$

where $c(X)$ is the (expected) computing cost of X.

This measure does not depend on the computing budget:

$$\mathrm{Eff}(\bar{X}_n) = \frac{1}{n \cdot c(X) \cdot \mathrm{Var}(X)/n} = \mathrm{Eff}(X).$$

With a 10-fold efficiency improvement, we we need a computing budget 10 times smaller for the same accuracy.

Simulation efficiency

We define the efficiency of the estimator by

$$\mathrm{Eff}(X) = \frac{1}{c(X) \cdot \mathrm{Var}(X)}$$

where $c(X)$ is the (expected) computing cost of X.

This measure does not depend on the computing budget:

$$\mathrm{Eff}(\bar{X}_n) = \frac{1}{n \cdot c(X) \cdot \mathrm{Var}(X)/n} = \mathrm{Eff}(X).$$

With a 10-fold efficiency improvement, we we need a computing budget 10 times smaller for the same accuracy.

In the presence of bias $\beta = \mathbb{E}[X] - \mu$, we define

$$\mathrm{Eff}(X) = \frac{1}{c(X) \cdot \mathrm{MSE}(X)} = \frac{1}{c(X) \cdot (\mathrm{Var}[X] + \beta^2)}.$$

In this case, $\mathrm{Eff}(\bar{X}_n)$ depends on n.

Other sources of error

The confidence intervals based on the previous CLTs only take the simulation error (or noise) into account.

They ignore the approximations made in building the model and the errors made in estimating the model parameters.

Ideally, confidence intervals should take all sources of error into account.

CI for a function of several means

Want a CI on $g(\boldsymbol{\mu})$ where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ and g is smooth but generally nonlinear. We estimate μ_j by Y_{jn}.

CI for a function of several means

Want a CI on $g(\boldsymbol{\mu})$ where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)$ and g is smooth but generally nonlinear. We estimate μ_j by Y_{jn}.

If $\mathbf{Y}_n = (Y_{1n}, \dots, Y_{dn}) \to \boldsymbol{\mu} = (\mu_1, \dots, \mu_d)$ and $g : \mathbb{R}^d \to \mathbb{R}$ is continuous, then $g(\mathbf{Y}_n) \to g(\boldsymbol{\mu})$.

CI for a function of several means

Want a CI on $g(\boldsymbol{\mu})$ where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ and g is smooth but generally nonlinear. We estimate μ_j by Y_{jn}.

If $\mathbf{Y}_n = (Y_{1n}, \ldots, Y_{dn}) \to \boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ and $g : \mathbb{R}^d \to \mathbb{R}$ is continuous, then $g(\mathbf{Y}_n) \to g(\boldsymbol{\mu})$.

Delta theorem. Suppose $g : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable in a neighborhood of $\boldsymbol{\mu}$, and ∇g its gradient. If $r(n)(\mathbf{Y}_n - \boldsymbol{\mu}) \Rightarrow \mathbf{Y}$ when $n \to \infty$, then

$$r(n)(g(\mathbf{Y}_n) - g(\boldsymbol{\mu})) \Rightarrow (\nabla g(\boldsymbol{\mu}))^{\mathrm{t}} \mathbf{Y} \quad \text{quand } n \to \infty.$$

Idea of the proof: Taylor series approximation.

CI for a function of several means

Want a CI on $g(\boldsymbol{\mu})$ where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ and g is smooth but generally nonlinear. We estimate μ_j by Y_{jn}.

If $\mathbf{Y}_n = (Y_{1n}, \ldots, Y_{dn}) \to \boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ and $g : \mathbb{R}^d \to \mathbb{R}$ is continuous, then $g(\mathbf{Y}_n) \to g(\boldsymbol{\mu})$.

Delta theorem. Suppose $g : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable in a neighborhood of $\boldsymbol{\mu}$, and ∇g its gradient. If $r(n)(\mathbf{Y}_n - \boldsymbol{\mu}) \Rightarrow \mathbf{Y}$ when $n \to \infty$, then

$$r(n)(g(\mathbf{Y}_n) - g(\boldsymbol{\mu})) \Rightarrow (\nabla g(\boldsymbol{\mu}))^{\mathrm{t}}\mathbf{Y} \quad \text{quand } n \to \infty.$$

Idea of the proof: Taylor series approximation.

Corollary. If $\sqrt{n}(\mathbf{Y}_n - \boldsymbol{\mu}) \;\Rightarrow\; \mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_y)$ when $n \to \infty$ (e.g., if $\mathbf{Y}_n$ is an average and obeys the CLT), then we have the CLT:

$$\frac{\sqrt{n}(g(\mathbf{Y}_n) - g(\boldsymbol{\mu}))}{\sigma_g} \;\Rightarrow\; N(0,1) \quad \text{when } n \to \infty,$$

where $\sigma_g{}^2 = \mathrm{Var}[(\nabla g(\boldsymbol{\mu}))^{\mathrm{t}}\mathbf{Y}] = \mathbb{E}[((\nabla g(\boldsymbol{\mu}))^{\mathrm{t}}\mathbf{Y})^2]$
$= (\nabla g(\boldsymbol{\mu}))^{\mathrm{t}}\mathbb{E}[\mathbf{Y}\mathbf{Y}^{\mathrm{t}}]\nabla g(\boldsymbol{\mu}) = (\nabla g(\boldsymbol{\mu}))^{\mathrm{t}}\boldsymbol{\Sigma}_y \nabla g(\boldsymbol{\mu})$.

Corollary. If $\sqrt{n}(\mathbf{Y}_n - \boldsymbol{\mu}) \;\Rightarrow\; \mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_y)$ when $n \to \infty$ (e.g., if $\mathbf{Y}_n$ is an average and obeys the CLT), then we have the CLT:

$$\frac{\sqrt{n}(g(\mathbf{Y}_n) - g(\boldsymbol{\mu}))}{\sigma_g} \;\Rightarrow\; N(0,1) \quad \text{when } n \to \infty,$$

where $\sigma_g{}^2 = \mathrm{Var}[(\nabla g(\boldsymbol{\mu}))^{\mathrm{t}}\mathbf{Y}] = \mathbb{E}[((\nabla g(\boldsymbol{\mu}))^{\mathrm{t}}\mathbf{Y})^2]$
$= (\nabla g(\boldsymbol{\mu}))^{\mathrm{t}}\mathbb{E}[\mathbf{Y}\mathbf{Y}^{\mathrm{t}}]\nabla g(\boldsymbol{\mu}) = (\nabla g(\boldsymbol{\mu}))^{\mathrm{t}}\boldsymbol{\Sigma}_y \nabla g(\boldsymbol{\mu})$.

Examples:
(a) $g(\mu) = \ln \mu$ where $\mu = \mu_1 > 0$;
(b) $g(\mu_1, \mu_2) = \mu_2/\mu_1$;
(c) $\mathrm{Var}[X] = g(\mu_1, \mu_2) = \mu_2 - \mu_1^2$ where $\mu_k = \mathbb{E}[X^k]$,
(d) $\mathrm{Cov}[X, Y] = g(\mu_1, \mu_3, \mu_3)$ where $\mu_1 = \mathbb{E}[X]$, $\mu_2 = \mathbb{E}[Y]$, $\mu_3 = \mathbb{E}[XY]$.

Example. (Ratio of expectations.)
Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. replicates of (X, Y) and suppose we want to estimate $\nu = \mathbb{E}[X]/\mathbb{E}[Y]$ by

$$\hat{\nu}_n = \frac{\bar{X}_n}{\bar{Y}_n} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n Y_i}.$$

Example. (Ratio of expectations.)
Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. replicates of (X, Y) and suppose we want to estimate $\nu = \mathbb{E}[X]/\mathbb{E}[Y]$ by

$$\hat{\nu}_n = \frac{\bar{X}_n}{\bar{Y}_n} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n Y_i}.$$

Biaised but strongly consistent estimator of ν.

Example. (Ratio of expectations.)
Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. replicates of (X, Y) and suppose we want to estimate $\nu = \mathbb{E}[X]/\mathbb{E}[Y]$ by

$$\hat{\nu}_n = \frac{\bar{X}_n}{\bar{Y}_n} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n Y_i}.$$

Biaised but strongly consistent estimator of ν.

Let $\mu_1 = \mathbb{E}[X]$, $\mu_2 = \mathbb{E}[Y]$, $g(\mu_1, \mu_2) = \mu_1/\mu_2$, $\sigma_1^2 = \mathrm{Var}[X]$, $\sigma_2^2 = \mathrm{Var}[Y]$, and $\sigma_{12} = \mathrm{Cov}[X, Y]$.

Example. (Ratio of expectations.)
Let $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d. replicates of (X, Y) and suppose we want to estimate $\nu = \mathbb{E}[X]/\mathbb{E}[Y]$ by

$$\hat{\nu}_n = \frac{\bar{X}_n}{\bar{Y}_n} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n Y_i}.$$

Biaised but strongly consistent estimator of ν.

Let $\mu_1 = \mathbb{E}[X]$, $\mu_2 = \mathbb{E}[Y]$, $g(\mu_1, \mu_2) = \mu_1/\mu_2$, $\sigma_1^2 = \mathrm{Var}[X]$, $\sigma_2^2 = \mathrm{Var}[Y]$, and $\sigma_{12} = \mathrm{Cov}[X, Y]$.

The gradient of g is $\nabla g(\mu_1, \mu_2) = (1/\mu_2, -\mu_1/\mu_2^2)^{\mathsf{t}}$.

By applying the Delta theorem, we obtain

$$\sqrt{n}(\hat{\nu}_n - \nu) \;\Rightarrow\; (W_1, W_2) \cdot \nabla g(\mu_1, \mu_2) = W_1/\mu_2 - W_2\mu_1/\mu_2^2 \sim N(0, \sigma_g^2)$$

where

$$\begin{aligned} \sigma_g^2 &= (\nabla g(\boldsymbol{\mu}))^{\mathrm{t}} \boldsymbol{\Sigma}_y \nabla g(\boldsymbol{\mu}) \\ &= (\sigma_1^2 + \sigma_2^2 \nu^2 - 2\sigma_{12}\nu)/\mu_2^2 \end{aligned}$$

can be replaced by its estimator

$$\hat{\sigma}_{g,n}^2 = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 \hat{\nu}_n^2 - 2\hat{\sigma}_{12}\hat{\nu}_n}{(\bar{Y}_n)^2},$$

in which

$$\begin{aligned} \hat{\sigma}_1^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2, \\ \hat{\sigma}_2^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y}_n)^2, \\ \hat{\sigma}_{12}^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)(Y_j - \bar{Y}_n). \end{aligned}$$

Higher order

To reduce the bias, one can also add another term to the Taylor series, if g is twice continuously differentiable:

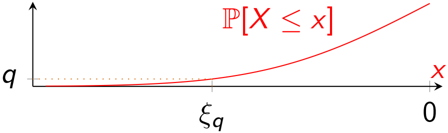
$$g(\mathbf{Y}_n) - g(\boldsymbol{\mu}) = (\nabla g(\boldsymbol{\mu}))^{\mathsf{t}}(\mathbf{Y}_n - \boldsymbol{\mu}) + (\mathbf{Y}_n - \boldsymbol{\mu})^{\mathsf{t}}\mathbf{H}(\boldsymbol{\mu})(\mathbf{Y}_n - \boldsymbol{\mu})/2 + o(\|\mathbf{Y}_n - \boldsymbol{\mu}\|^2),$$

where $\mathbf{H}$ is the Hessian matrix of g at $\boldsymbol{\mu}$. But this can increase the variance of the CI.

Quantile estimation

If X had cdf F, the q-quantile of F is

$$\xi_q = F^{-1}(q) = \inf\{x : F(x) \geq q\}.$$



Let $X_{(1)}, \ldots, X_{(n)}$ an $\mathrm{i.i.d.}$ sample of X, sorted, and $\hat{F}_n$ the empirical cdf.

Let $X_{(1)}, \ldots, X_{(n)}$ an $\mathrm{i.i.d.}$ sample of X, sorted, and $\hat{F}_n$ the empirical cdf.

A simple estimator of ξ_q is the empirical quantile

$$\hat{\xi}_{q,n} = \hat{F}_n^{-1}(q) = \inf\{x : \hat{F}_n(x) \geq q\} = X_{(\lceil nq \rceil)}.$$

Let $X_{(1)}, \ldots, X_{(n)}$ an i.i.d. sample of X, sorted, and $\hat{F}_n$ the empirical cdf.

A simple estimator of ξ_q is the empirical quantile

$$\hat{\xi}_{q,n} = \hat{F}_n^{-1}(q) = \inf\{x : \hat{F}_n(x) \geq q\} = X_{(\lceil nq \rceil)}.$$

Theorem.
(i) For each q, $\hat{\xi}_{q,n} \stackrel{\text{w.p.1}}{\to} \xi_q$ when $n \to \infty$ (strongly consistant).
(ii) If X has a density $f > 0$ continuous in a neighborhood of ξ_q, then (CLT):

$$\frac{\sqrt{n}(\hat{\xi}_{q,n} - \xi_q) f(\xi_q)}{\sqrt{q(1-q)}} \Rightarrow N(0,1) \quad \text{when } n \to \infty.$$

Let $X_{(1)}, \ldots, X_{(n)}$ an $\mathrm{i.i.d.}$ sample of X, sorted, and $\hat{F}_n$ the empirical cdf.

A simple estimator of ξ_q is the empirical quantile

$$\hat{\xi}_{q,n} = \hat{F}_n^{-1}(q) = \inf\{x : \hat{F}_n(x) \geq q\} = X_{(\lceil nq \rceil)}.$$

Theorem.
(i) For each q, $\hat{\xi}_{q,n} \stackrel{\mathrm{w.p.1}}{\to} \xi_q$ when $n \to \infty$ (strongly consistant).
(ii) If X has a density $f > 0$ continuous in a neighborhood of ξ_q, then (CLT):

$$\frac{\sqrt{n}(\hat{\xi}_{q,n} - \xi_q) f(\xi_q)}{\sqrt{q(1-q)}} \Rightarrow N(0,1) \quad \text{when } n \to \infty.$$

This imples $\mathrm{Var}[\hat{\xi}_{q,n}] \approx \dfrac{q(1-q)}{nf^2(\xi_q)}$.

This CLT shows that this estimator has large variance if $f(\xi_q)$ is small.
To use this IC, we need an estimate of $f(\xi_q)$: difficult.

Non-asymptotic exact method for an IC on ξ_q:
Suppose F is continuous at ξ_q.

Non-asymptotic exact method for an IC on ξ_q:
Suppose F is continuous at ξ_q.
Let B be the number of observations $X_{(i)}$ smaller than ξ_q.
Since $\mathbb{P}[X < \xi_q] = q$, B is Binomial(n, q).

Non-asymptotic exact method for an IC on ξ_q:
Suppose F is continuous at ξ_q.
Let B be the number of observations $X_{(i)}$ smaller than ξ_q.
Since $\mathbb{P}[X < \xi_q] = q$, B is Binomial(n, q).
If $1 \leq j < k \leq n$, we have

$$\mathbb{P}[X_{(j)} < \xi_q \leq X_{(k)}] = \mathbb{P}[j \leq B < k] = \sum_{i=j}^{k-1} \binom{n}{i} q^i (1-q)^{n-i}.$$

Non-asymptotic exact method for an IC on ξ_q:
Suppose F is continuous at ξ_q.
Let B be the number of observations $X_{(i)}$ smaller than ξ_q.
Since $\mathbb{P}[X < \xi_q] = q$, B is Binomial(n, q).
If $1 \leq j < k \leq n$, we have

$$\mathbb{P}[X_{(j)} < \xi_q \leq X_{(k)}] = \mathbb{P}[j \leq B < k] = \sum_{i=j}^{k-1} \binom{n}{i} q^i (1-q)^{n-i}.$$

We choose j and k for this sum to be $\geq 1 - \alpha$.
(Can be a unilateral or bilateral interval.)

Non-asymptotic exact method for an IC on ξ_q:
Suppose F is continuous at ξ_q.
Let B be the number of observations $X_{(i)}$ smaller than ξ_q.
Since $\mathbb{P}[X < \xi_q] = q$, B is Binomial(n, q).
If $1 \leq j < k \leq n$, we have

$$\mathbb{P}[X_{(j)} < \xi_q \leq X_{(k)}] = \mathbb{P}[j \leq B < k] = \sum_{i=j}^{k-1} \binom{n}{i} q^i (1-q)^{n-i}.$$

We choose j and k for this sum to be $\geq 1 - \alpha$.
(Can be a unilateral or bilateral interval.)

If n is large and q not too close to 0 or 1, we can approximate the binomial by a normal: $(B - nq)/\sqrt{nq(1-q)} \approx N(0, 1)$.

Non-asymptotic exact method for an IC on ξ_q:
Suppose F is continuous at ξ_q.
Let B be the number of observations $X_{(i)}$ smaller than ξ_q.
Since $\mathbb{P}[X < \xi_q] = q$, B is Binomial(n, q).
If $1 \le j < k \le n$, we have

$$\mathbb{P}[X_{(j)} < \xi_q \le X_{(k)}] = \mathbb{P}[j \le B < k] = \sum_{i=j}^{k-1} \binom{n}{i} q^i (1-q)^{n-i}.$$

We choose j and k for this sum to be $\ge 1 - \alpha$.
(Can be a unilateral or bilateral interval.)

If n is large and q not too close to 0 or 1, we can approximate the binomial by a normal: $(B - nq)/\sqrt{nq(1-q)} \approx N(0,1)$.
This gives $j = \lfloor nq + 1 - \delta \rfloor$ and $k = \lfloor nq + 1 + \delta \rfloor$, where $\delta = \sqrt{nq(1-q)}\Phi^{-1}(1 - \alpha/2)$.

Non-asymptotic exact method for an IC on ξ_q:
Suppose F is continuous at ξ_q.
Let B be the number of observations $X_{(i)}$ smaller than ξ_q.
Since $\mathbb{P}[X < \xi_q] = q$, B is Binomial(n, q).
If $1 \le j < k \le n$, we have

$$\mathbb{P}[X_{(j)} < \xi_q \le X_{(k)}] = \mathbb{P}[j \le B < k] = \sum_{i=j}^{k-1} \binom{n}{i} q^i (1-q)^{n-i}.$$

We choose j and k for this sum to be $\ge 1 - \alpha$.
(Can be a unilateral or bilateral interval.)

If n is large and q not too close to 0 or 1, we can approximate the binomial by a normal: $(B - nq)/\sqrt{nq(1-q)} \approx N(0, 1)$.
This gives $j = \lfloor nq + 1 - \delta \rfloor$ and $k = \lfloor nq + 1 + \delta \rfloor$, where $\delta = \sqrt{nq(1-q)}\Phi^{-1}(1 - \alpha/2)$.

To improve the estimator, one can replace $\hat{F}_n$ by a smoother fonction (e.g., an interpolation, a kernel density estimator, etc.).
In all cases, the estimator is biased, but this can reduce the bias.

Example: Value at Risk (VaR).

Let L be the net loss of an investment portfolio over a given time period.
The VaR is the value of x_q such that $\mathbb{P}[L > x_q] = q$.
It is $(1 - q)$-quantile of L.

Example: Value at Risk (VaR).

Let L be the net loss of an investment portfolio over a given time period.
The VaR is the value of x_q such that $\mathbb{P}[L > x_q] = q$.
It is $(1-q)$-quantile of L.

Common values: $q = 0.01$, period $= 2$ weeks (banks), or months or years (insurance, pension plans).

Use of the VaR can be criticized, since it provides a limited information. For example, if $x_{0.01} = 10^7$ dollars, what do we know?

Example: Value at Risk (VaR).

Let L be the net loss of an investment portfolio over a given time period.
The VaR is the value of x_q such that $\mathbb{P}[L > x_q] = q$.
It is $(1-q)$-quantile of L.

Common values: $q = 0.01$, period $= 2$ weeks (banks), or months or years (insurance, pension plans).

Use of the VaR can be criticized, since it provides a limited information.
For example, if $x_{0.01} = 10^7$ dollars, what do we know?

A complementary measure is the conditional VaR (CVaR):

$$\mathbb{E}[L \mid L > x_q].$$

Better: find and display distribution of L conditional on $L > x_q$.

Example: Value at Risk (VaR).

Let L be the net loss of an investment portfolio over a given time period.
The VaR is the value of x_q such that $\mathbb{P}[L > x_q] = q$.
It is $(1 - q)$-quantile of L.

Common values: $q = 0.01$, period $=$ 2 weeks (banks), or months or years (insurance, pension plans).

Use of the VaR can be criticized, since it provides a limited information.
For example, if $x_{0.01} = 10^7$ dollars, what do we know?

A complementary measure is the conditional VaR (CVaR):

$$\mathbb{E}[L \mid L > x_q].$$

Better: find and display distribution of L conditional on $L > x_q$.

Models for VaR estimation: often thousands of dependent assets. Asset prices (or log) can be defined as linear combination of factors, whose changes are modeled via a copula (e.g., Student copula) and normal or lognormal marginals.

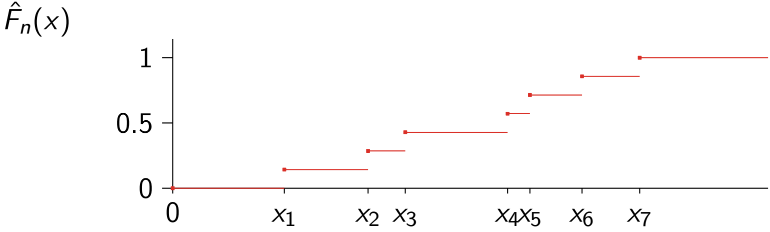
Except for very simple cases, the VaR cannot be computed exactly. It can be estimated by simulation. For very small p: rare-event simulation.

Estimating the cdf by the empirical cdf

With simulation, we can estimate not only $\mathbb{E}[X]$, but the entire distribution of X. It is characterized by its cdf, defined by $F(x) = \mathbb{P}[X \le x]$, for all $x \in \mathbb{R}$. The empirical cdf

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[x_i \le x].$$

jumps by $1/n$ at each observation x_i. Easy to compute and plot after sorting the observations $x_1, \ldots, x_n$ into $x_{(1)}, \ldots, x_{(n)}$.



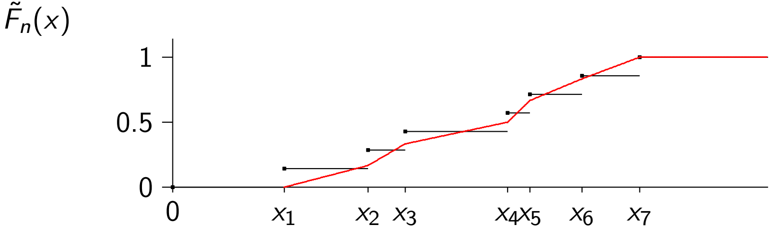
If $x_1, \ldots, x_n$ are $\mathrm{i.i.d.}$ random variables with cdf F, then

$$D_n = \sup_{-\infty<x<\infty} |\hat{F}_n(x) - F(x)| \to 0 \quad \text{w.p.1 when } n \to \infty,$$

and $D_n = O_p(n^{-1/2})$.

Estimating the density of X

To estimate the density f of X, a first idea could be to take the derivative (the slope) of a piecewise-linear interpolation of $\hat{F}_n$.

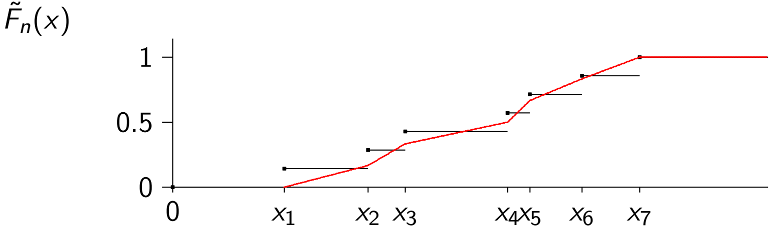


This gives the piecewise-constant empirical density:

$$\tilde{f}_n(x) = \frac{1}{(n-1)(x_{(i+1)} - x_{(i)})} \qquad \text{for } x_{(i)} \leq x \leq x_{(i+1)}.$$

Estimating the density of X

To estimate the density f of X, a first idea could be to take the derivative (the slope) of a piecewise-linear interpolation of $\hat{F}_n$.



This gives the piecewise-constant empirical density:

$$\tilde{f}_n(x) = \frac{1}{(n-1)(x_{(i+1)} - x_{(i)})} \qquad \text{for } x_{(i)} \le x \le x_{(i+1)}.$$

Between two observations very close to each other, this density is very large! It is irregular and does not converge to f when $n \to \infty$.

Example. Let f be the $U(0,1)$ density and let

$$d_n = \min_{1 \le i < n} (x_{(i+1)} - x_{(i)}).$$

One can show that $n(n-1)d_n \;\Rightarrow\; \text{Expon}(1)$ when $n \to \infty$.

Example. Let f be the $U(0,1)$ density and let

$$d_n = \min_{1\leq i<n}(x_{(i+1)} - x_{(i)}).$$

One can show that $n(n-1)d_n \;\Rightarrow\;$ Expon(1) when $n\to\infty$.

$$\mathbb{P}\left[\max_{0\leq x\leq 1}\tilde{f}_n(x) > y\right] = \mathbb{P}\left[\frac{1}{(n-1)d_n} > y\right] = \mathbb{P}[n(n-1)d_n < n/y] \approx 1-e^{-n/y}$$

For fixed y, converges to 1 exponentially fast when $n\to\infty$.
Thus, the density $\tilde{f}_n$ will have very large peaks when n is large!

Example. Let f be the $U(0,1)$ density and let

$$d_n = \min_{1 \le i < n} (x_{(i+1)} - x_{(i)}).$$

One can show that $n(n-1)d_n \;\Rightarrow\; \text{Expon}(1)$ when $n \to \infty$.

$$\mathbb{P}\left[\max_{0 \le x \le 1} \tilde{f}_n(x) > y\right] = \mathbb{P}\left[\frac{1}{(n-1)d_n} > y\right] = \mathbb{P}[n(n-1)d_n < n/y] \approx 1 - e^{-n/y}$$

For fixed y, converges to 1 exponentially fast when $n \to \infty$.
Thus, the density $\tilde{f}_n$ will have very large peaks when n is large!

Sometimes plotting F may be sufficient, but often a good density estimator gives better visual information. How can we estimate the density?

Histograms

We have n observations over $[a, b]$.
We partition the interval in m pieces of length $h = (b - a)/m$.
The histogram density (height) $f_{\mathrm{h},n}$ is constant over each interval and proportional to the number of observations in the interval.

Histograms

We have n observations over $[a, b]$.
We partition the interval in m pieces of length $h = (b - a)/m$.
The histogram density (height) $f_{\mathrm{h},n}$ is constant over each interval and proportional to the number of observations in the interval.

To minimize the mean integrated square error (MISE) $\mathbb{E} \int_a^b [f_{\mathrm{h},n}(x) - f(x)]^2 dx$, we must choose h such that $h^3 n \int_a^b (f'(x))^2 dx \approx 6$. We then have $m = (b - a)/h = \mathcal{O}(n^{1/3})$ and MISE $= \mathcal{O}(n^{-2/3})$. We should double m when n is multiplied by 8.

Histograms

We have n observations over $[a, b]$.
We partition the interval in m pieces of length $h = (b - a)/m$.
The histogram density (height) $f_{\mathrm{h},n}$ is constant over each interval and proportional to the number of observations in the interval.

To minimize the mean integrated square error (MISE) $\mathbb{E} \int_a^b [f_{\mathrm{h},n}(x) - f(x)]^2 dx$, we must choose h such that $h^3 n \int_a^b (f'(x))^2 dx \approx 6$. We then have $m = (b - a)/h = \mathcal{O}(n^{1/3})$ and MISE $= \mathcal{O}(n^{-2/3})$. We should double m when n is multiplied by 8.

Better: polygonal interpolation of the histogram.
Gives MISE $= \mathcal{O}(n^{-4/5})$ with $m = \mathcal{O}(n^{1/5})$. In this case, we double m when n is multiplied by 32.

Kernel density estimation

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} k((x - x_i)/h),$$

where k is a base density named the kernel and
$h > 0$ is a constant named the smoothing factor.

Kernel density estimation

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k((x - x_i)/h),$$

where k is a base density named the kernel and
$h > 0$ is a constant named the smoothing factor.
The change of variable $y = (x - x_i)/h$ gives

$$\frac{1}{h} \int_{-\infty}^{\infty} k((x - x_i)/h)dx = \int_{-\infty}^{\infty} k(y)dy = 1.$$

This generalizes to multivariate distributions (d dimensions).

Kernel density estimation

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k((x - x_i)/h),$$

where k is a base density named the kernel and
$h > 0$ is a constant named the smoothing factor.
The change of variable $y = (x - x_i)/h$ gives

$$\frac{1}{h} \int_{-\infty}^{\infty} k((x - x_i)/h)dx = \int_{-\infty}^{\infty} k(y)dy = 1.$$

This generalizes to multivariate distributions (d dimensions).

Main difficulty: choices of k and (mostly) h.

Kernel density estimation

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k((x - x_i)/h),$$

where k is a base density named the kernel and
$h > 0$ is a constant named the smoothing factor.
The change of variable $y = (x - x_i)/h$ gives

$$\frac{1}{h} \int_{-\infty}^{\infty} k((x - x_i)/h) dx = \int_{-\infty}^{\infty} k(y) dy = 1.$$

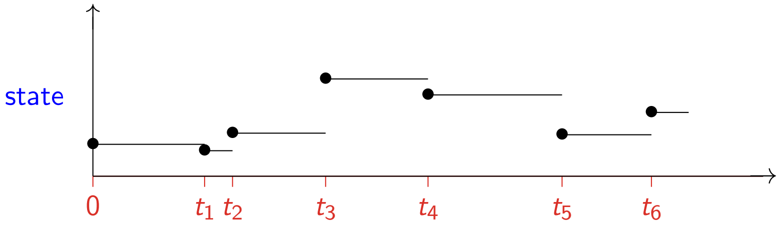
This generalizes to multivariate distributions (d dimensions).

Main difficulty: choices of k and (mostly) h.

Simpler and much more common: just plot a histogram.

Discrete-event Simulation

Events $e_0, e_1, e_2, \ldots$ occur at times $0 = t_0 \leq t_1 \leq t_2 \leq \cdots$.
Let $\mathcal{S}_i$ be the state of the system just after event e_i.
Simulation time: current value of t_i.



$(t_i, \mathcal{S}_i)$ must contain enough information so we can continue running the simulation with it and also the random variables generated afterward, at events e_j for $j > i$.

A discrete-event simulation program requires:

- A simulation clock.
- A list of the planned future events, by chronological order (different possible implementations).
- A "procedure" or "method" to execute for each type of event.