# Computer Simulation

## Module 2: Calculus, Probability, and Statistics Primers


**Dave Goldsman, Ph.D.**

*Professor*

Stewart School of Industrial and Systems Engineering


## Covariance and Correlation

GTx

# Lesson Overview

Last Time: We kicked butt on conditional expectation and its applications.

This Time: We'll talk about independence, covariance, correlation, and related results.

Correlation shows up all over the place in simulation.

**"Definition"** (two-dimensional LOTUS): Suppose that $h(X, Y)$ is some function of the RV's $X$ and $Y$. Then

$$\mathrm{E}[h(X,Y)] = \begin{cases} \sum_x \sum_y h(x,y) f(x,y) & \text{if } (X,Y) \text{ is discrete} \\ \int_{\mathbb{R}} \int_{\mathbb{R}} h(x,y) f(x,y)\, dx\, dy & \text{if } (X,Y) \text{ is continuous} \end{cases}$$

**Theorem**: Whether or not $X$ and $Y$ are independent, we have $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$.

**Theorem**: If $X$ and $Y$ are *independent*, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

(Stay tuned for dependent case.)

**Definition**: $X_1, \ldots, X_n$ form a *random sample* from $f(x)$ if (i) $X_1, \ldots, X_n$ are independent, and (ii) each $X_i$ has the same pdf (or pmf) $f(x)$.

**Notation**: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$. (The term "iid" reads *independent and identically distributed.*)

**Example**: If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$ and the *sample mean* $\bar{X}_n \equiv \sum_{i=1}^n X_i/n$, then $\mathrm{E}[\bar{X}_n] = \mathrm{E}[X_i]$ and $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X_i)/n$. Thus, the variance *decreases* as $n$ increases. $\square$

But not all RV's are independent...

**Definition**: The *covariance* between $X$ and $Y$ is

$$\mathrm{Cov}(X,Y) \equiv \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])] = \mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y].$$

Note that $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$.

**Theorem**: If $X$ and $Y$ are independent RV's, then $\mathrm{Cov}(X,Y) = 0$.

**Remark**: $\mathrm{Cov}(X,Y) = 0$ doesn't mean $X$ and $Y$ are independent!

**Example**: Take $X \sim \mathrm{Unif}(-1,1)$ and $Y = X^2$. Dependent! But

$$\mathrm{Cov}(X,Y) = \mathrm{E}[X^3]-\mathrm{E}[X]\mathrm{E}[X^2] = 0 \ \text{(symmetry)}$$

**Theorem**: $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X, Y)$.

**Theorem**: Whether or not $X$ and $Y$ are independent,

$$\begin{aligned} \mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y) \\ \mathrm{Var}(X-Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y). \end{aligned}$$

**Definition**: The *correlation* between $X$ and $Y$ is

$$\rho \equiv \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

**Theorem**: $-1 \le \rho \le 1$.

**Example**: Consider the following joint pmf.

| $f(x,y)$ | $X=2$ | $X=3$ | $X=4$ | $f_Y(y)$ |
|---|---|---|---|---|
| $Y=40$ | 0.00 | 0.20 | 0.10 | 0.3 |
| $Y=50$ | 0.15 | 0.10 | 0.05 | 0.3 |
| $Y=60$ | 0.30 | 0.00 | 0.10 | 0.4 |
| $f_X(x)$ | 0.45 | 0.30 | 0.25 | 1 |

$\mathrm{E}[X] = 2.8$, $\mathrm{Var}(X) = 0.66$,

$\mathrm{E}[Y] = 51$, $\mathrm{Var}(Y) = 69$,

$$\mathrm{E}[XY] = \sum_x \sum_y xy f(x,y) = 140,$$

$$\rho = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = -0.415. \quad \square$$

**Portfolio Example:** Consider two assets, $S_1$ and $S_2$, with expected returns $\mathrm{E}[S_1] = \mu_1$ and $\mathrm{E}[S_2] = \mu_2$, and variabilities $\mathrm{Var}(S_1) = \sigma_1^2$, $\mathrm{Var}(S_2) = \sigma_2^2$, and $\mathrm{Cov}(S_1, S_2) = \sigma_{12}$.

Define a *portfolio* $P = wS_1 + (1-w)S_2$, where $w \in [0,1]$. Then

$$\mathrm{E}[P] = w\mu_1 + (1-w)\mu_2$$

$$\mathrm{Var}(P) = w^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2w(1-w)\sigma_{12}.$$

Setting $\frac{d}{dw}\mathrm{Var}(P) = 0$, we obtain the critical point that (hopefully) minimizes the variance of the portfolio,

$$w = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}. \quad \square$$

**Portfolio Exercise:** Suppose $\mathrm{E}[S_1] = 0.2$, $\mathrm{E}[S_2] = 0.1$, $\mathrm{Var}(S_1) = 0.2$, $\mathrm{Var}(S_2) = 0.4$, and $\mathrm{Cov}(S_1, S_2) = -0.1$.

What value of $w$ maximizes the expected return of the portfolio?

What value of $w$ minimizes the variance? (Note the negative covariance I've introduced into the picture.)

Let's talk trade-offs.

# Summary

Defined covariance and correlation and gave a couple of practical examples.

Next Time: Quickie review of some favorite distributions.

Poisson arrivals are just around the corner!

# Computer Simulation

## Module 2: Calculus, Probability, and Statistics Primers


**Dave Goldsman, Ph.D.**

*Professor*

Stewart School of Industrial and Systems Engineering


## Probability Distributions

# Lesson Overview

Last Time: Defined covariance and correlation and gave a couple of practical examples.

This Time: These are a few of my favorite distributions.

We'll go through a little list of important distributions, both discrete and continuous.

GTx

## Some Discrete Distributions...

$X \sim \text{Bernoulli}(p)$.

$$f(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p \ (= q) & \text{if } x = 0 \end{cases}$$

$\text{E}[X] = p$, $\text{Var}(X) = pq$, $M_X(t) = pe^t + q$.

$Y \sim \text{Binomial}(n, p)$. If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$ (i.e., *Bernoulli(p) trials*), then $Y = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$.

$$f(y) = \binom{n}{y} p^y q^{n-y}, \quad y = 0, 1, \ldots, n.$$

$\text{E}[Y] = np$, $\text{Var}(Y) = npq$, $M_Y(t) = (pe^t + q)^n$.

$X \sim \text{Geometric}(p)$ is the number of Bern($p$) trials until a success occurs. For example, "FFFS" implies that $X = 4$.

$$f(x) = q^{x-1}p, \quad x = 1, 2, \ldots.$$

$\text{E}[X] = 1/p$, $\text{Var}(X) = q/p^2$, $M_X(t) = pe^t/(1 - qe^t)$.

$Y \sim \text{NegBin}(r, p)$ is the sum of $r$ iid Geom($p$) RV's, i.e., the time until the $r$th success occurs. For example, "FFFSSFS" implies that $\text{NegBin}(3, p) = 7$.

$$f(y) = \binom{y-1}{r-1} q^{y-r} p^r, \quad y = r, r+1, \ldots.$$

$E[Y] = r/p$, $\text{Var}(Y) = qr/p^2$.

$X \sim \text{Poisson}(\lambda)$.

**Definition:** A *counting process* $N(t)$ tallies the number of "arrivals" observed in $[0, t]$. A *Poisson process* is a counting process satisfying the following.

i. Arrivals occur one-at-a-time at rate $\lambda$ (e.g., $\lambda = 4$ customers/hr)

ii. Independent increments, i.e., the numbers of arrivals in disjoint time intervals are independent.

iii. Stationary increments, i.e., the distribution of the number of arrivals in $[s, s+t]$ only depends on $t$.

GTx

$X \sim \text{Pois}(\lambda)$ is the number of arrivals that a Poisson process experiences in one time unit, i.e., $N(1)$.

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots.$$

$\text{E}[X] = \lambda = \text{Var}(X)$, $M_X(t) = e^{\lambda(e^t - 1)}$.

## Some Continuous Distributions...

$X \sim \text{Uniform}(a, b)$. $f(x) = \frac{1}{b-a}$ for $a \le x \le b$, $\text{E}[X] = \frac{a+b}{2}$, $\text{Var}(X) = \frac{(b-a)^2}{12}$, $M_X(t) = (e^{tb} - e^{ta})/(tb - ta)$.

$X \sim \text{Exponential}(\lambda)$. $f(x) = \lambda e^{-\lambda x}$ for $x \ge 0$, $\text{E}[X] = 1/\lambda$, $\text{Var}(X) = 1/\lambda^2$, $M_X(t) = \lambda/(\lambda - t)$ for $t < \lambda$.

**Theorem**: The Exp($\lambda$) has the *memoryless property*, i.e., for $s, t > 0$, $P(X > s + t | X > s) = P(X > t)$.

**Example**: If $X \sim \text{Exp}(1/100)$, then

$$P(X > 200 | X > 50) \;=\; P(X > 150) \;=\; e^{-\lambda t} \;=\; e^{-150/100}.$$

$X \sim \text{Gamma}(\alpha, \lambda)$. Recall the gamma fn $\Gamma(\alpha) \equiv \int_0^\infty t^{\alpha-1} e^{-t}\, dt$.

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad x \geq 0.$$

$\text{E}[X] = \alpha/\lambda$, $\text{Var}(X) = \alpha/\lambda^2$, $M_X(t) = \left[\lambda/(\lambda - t)\right]^\alpha$ for $t < \lambda$.

If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$, then $Y \equiv \sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$.
The $\text{Gamma}(n, \lambda)$ is also called the $\text{Erlang}_n(\lambda)$. It has cdf

$$F_Y(y) = 1 - e^{-\lambda y} \sum_{j=0}^{n-1} \frac{(\lambda y)^j}{j!}, \quad y \geq 0.$$

$X \sim \text{Triangular}(a, b, c)$. Good for models with limited data — $a$ is the smallest possible value, $b$ is the "most likely," and $c$ is the largest.

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{if } a < x \le b \\ \frac{2(c-x)}{(c-b)(c-a)} & \text{if } b < x \le c \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \mathrm{E}[X] = \frac{a+b+c}{3}.$$

$X \sim \text{Beta}(a, b)$. $f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$, $0 \le x \le 1$, $a, b > 0$.

$$\mathrm{E}[X] = \frac{a}{a+b} \quad \text{and} \quad \mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

$X \sim \text{Normal}(\mu, \sigma^2)$. Most important distribution.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], \quad x \in \mathbb{R}.$$

$\text{E}[X] = \mu$, $\text{Var}(X) = \sigma^2$, and $M_X(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$.

**Theorem:** If $X \sim \text{Nor}(\mu, \sigma^2)$, then $aX + b \sim \text{Nor}(a\mu + b, a^2\sigma^2)$.

**Corollary:** If $X \sim \text{Nor}(\mu, \sigma^2)$, then $Z \equiv \frac{X-\mu}{\sigma} \sim \text{Nor}(0, 1)$, the *standard normal distribution*, with pdf $\phi(z) \equiv \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ and cdf $\Phi(z)$, which is tabled. E.g., $\Phi(1.96) \doteq 0.975$.

**Theorem:** If $X_1$ and $X_2$ are *independent* with $X_i \sim \text{Nor}(\mu_i, \sigma_i^2)$, $i = 1, 2$, then $X_1 + X_2 \sim \text{Nor}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

**Example:** Suppose $X \sim \text{Nor}(3, 4)$, $Y \sim \text{Nor}(4, 6)$, and $X$ and $Y$ are independent. Then $2X - 3Y + 1 \sim \text{Nor}(-5, 70)$. $\Box$

**Corollary** (of a previous theorem): If $X_1, \ldots, X_n$ are iid $\text{Nor}(\mu, \sigma^2)$, then the sample mean $\bar{X}_n \sim \text{Nor}(\mu, \sigma^2/n)$.

This is a special case of the *Law of Large Numbers*, which says that $\bar{X}_n$ approximates $\mu$ well as $n$ becomes large.

# Summary

Went over our favorite discrete and continuous distributions. Ended on a high note by ringing the bell with the normal.

Next Time: The normal figures prominently in The Law of Large Numbers and The Central Limit Theorem.

They’re not just good ideas – They’re the law!

# Computer Simulation

## Module 2: Calculus, Probability, and Statistics Primers

**Dave Goldsman, Ph.D.**
*Professor*
Stewart School of Industrial and Systems Engineering

## Limit Theorems

GTx

# Lesson Overview

Last Time: Went over a dictionary of distributions.

This Time: Take it to the limit one more time! What happens when the sample size gets big?

Normality happens! The Central Limit Theorem is the most-important result ever!

**Corollary** (of a previous theorem): If $X_1, \ldots, X_n$ are iid $\text{Nor}(\mu, \sigma^2)$, then the sample mean $\bar{X}_n \sim \text{Nor}(\mu, \sigma^2/n)$.

This is a special case of the *Law of Large Numbers*, which says that $\bar{X}_n$ approximates $\mu$ well as $n$ becomes large.

**Definition:** The sequence of RV's $Y_1, Y_2, \ldots$ with respective cdf's $F_{Y_1}(y), F_{Y_2}(y), \ldots$ *converges in distribution* to the RV $Y$ having cdf $F_Y(y)$ if $\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$ for all $y$ belonging to the continuity set of $Y$. Notation: $Y_n \xrightarrow{d} Y$.

**Idea:** If $Y_n \xrightarrow{d} Y$ and $n$ is large, then you ought to be able to approximate the distribution of $Y_n$ by the limit distribution of $Y$.

**Central Limit Theorem:** If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$ with mean $\mu$ and variance $\sigma^2$, then

$$Z_n \equiv \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \text{Nor}(0,1).$$

Thus, the cdf of $Z_n$ approaches $\Phi(z)$ as $n$ increases.

The CLT is the most-important theorem in the universe.

Usually works well if the pdf/pmf is fairly symmetric and $n \geq 15$.

We will eventually look at more-general versions of the CLT.

**Example:** If $X_1, X_2, \ldots, X_{100} \overset{\text{iid}}{\sim} \text{Exp}(1)$ (so $\mu = \sigma^2 = 1$), then

$$
\begin{aligned}
& P\left(90 \le \sum_{i=1}^{100} X_i \le 110\right) \\
&= \; P\left(\frac{90-100}{\sqrt{100}} \le Z_{100} \le \frac{110-100}{\sqrt{100}}\right) \\
&\approx \; P(-1 \le \text{Nor}(0,1) \le 1) \; = \; 0.6827.
\end{aligned}
$$

BTW, since $\sum_{i=1}^{100} X_i \sim \text{Erlang}_{k=100}(\lambda = 1)$, we can use the cdf (which may be tedious) or software such as Minitab to obtain the *exact* value of $P(90 \le \sum_{i=1}^{100} X_i \le 110) = 0.6835$.

Wow! The CLT and exact answers match nicely! $\square$

**Exercise:** Demonstrate that the CLT actually works.

1. Pick your favorite RV $X_1$. Simulate it and make a histogram.
2. Now suppose $X_1$ and $X_2$ are iid from your favorite distribution. Make a histogram of $X_1 + X_2$.
3. Now $X_1 + X_2 + X_3$.
4. ... Now $X_1 + X_2 + \cdots + X_n$ for some reasonably large $n$.
5. Does the CLT work for the Cauchy distribution, i.e., $X = \tan(2\pi U)$, where $U \sim \text{Unif}(0, 1)$?

# Summary

Went over some limit theorem results and applications.

Next Time: Stats Attack!  We'll be starting our Statistics Primer.

# Computer Simulation

## Module 2: Calculus, Probability, and Statistics Primers


**Dave Goldsman, Ph.D.**

*Professor*

Stewart School of Industrial and Systems Engineering


## Introduction to Estimation

# Lesson Overview

Last Time: We completed our Probability Boot Camp!

This Time: We'll start our review of Statistics with some basics, including unbiased estimation and mean squared error.

If you are not yet (or no longer) a stats maven, simply go through the notes at your leisure.

**Definition**: A *statistic* is a function of the observations $X_1, \ldots, X_n$, and not explicitly dependent on any unknown parameters.

Examples of statistics: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Statistics are *random variables*. If we take two different samples, we'd expect to get two different values of a statistic.

A statistic is usually used to estimate some unknown *parameter* from the underlying probability distribution of the $X_i$'s.

Examples of parameters: $\mu$, $\sigma^2$.

Let $X_1, \ldots, X_n$ be iid RV's and let $T(\mathbf{X}) \equiv T(X_1, \ldots, X_n)$ be a statistic based on the $X_i$'s. Suppose we use $T(\mathbf{X})$ to estimate some unknown parameter $\theta$. Then $T(\mathbf{X})$ is called a *point estimator* for $\theta$.

**Examples**: $\bar{X}$ is usually a point estimator for the mean $\mu = \mathrm{E}[X_i]$, and $S^2$ is often a point estimator for the variance $\sigma^2 = \mathrm{Var}(X_i)$.

It would be nice if $T(\mathbf{X})$ had certain properties:

* Its expected value should equal the parameter it's trying to estimate.

* It should have low variance.

**Definition**: $T(\mathbf{X})$ is *unbiased* for $\theta$ if $\mathrm{E}[T(\mathbf{X})] = \theta$.

**Example/Theorem**: Suppose $X_1, \ldots, X_n$ are iid anything with mean $\mu$. Then

$$\mathrm{E}[\bar{X}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \mathrm{E}[X_i] = \mu.$$

So $\bar{X}$ is always unbiased for $\mu$. That's why $\bar{X}$ is the *sample mean*.

**Baby Example**: In particular, suppose $X_1, \ldots, X_n$ are iid Exp($\lambda$). Then $\bar{X}$ is unbiased for $\mu = \mathrm{E}[X_i] = 1/\lambda$.

But be careful.... $1/\bar{X}$ is *biased* for $\lambda$ in this exponential case, i.e., $\mathrm{E}[1/\bar{X}] \neq 1/\mathrm{E}[\bar{X}] = \lambda$.

**Example/Theorem**: Suppose $X_1, \ldots, X_n$ are iid anything with mean $\mu$ and variance $\sigma^2$. Then

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right] = \mathrm{Var}(X_i) = \sigma^2.$$

Thus, $S^2$ is always unbiased for $\sigma^2$. This is why $S^2$ is called the *sample variance*.

**Baby Example**: Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Exp}(\lambda)$. Then $S^2$ is unbiased for $\mathrm{Var}(X_i) = 1/\lambda^2$.

**Proof** (of general result): First, some algebra gives

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

Since $\mathrm{E}[X_1] = \mathrm{E}[\bar{X}]$ and $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X_1)/n = \sigma^2/n$, we have

$$\begin{aligned}
\mathrm{E}[S^2] &= \frac{\sum_{i=1}^{n} \mathrm{E}[X_i^2] - n\mathrm{E}[\bar{X}^2]}{n-1} = \frac{n}{n-1}\Big(\mathrm{E}[X_1^2] - \mathrm{E}[\bar{X}^2]\Big) \\
&= \frac{n}{n-1}\Big(\mathrm{Var}(X_1) + (\mathrm{E}[X_1])^2 - \mathrm{Var}(\bar{X}) - (\mathrm{E}[\bar{X}])^2\Big) \\
&= \frac{n}{n-1}(\sigma^2 - \sigma^2/n) = \sigma^2. \quad \square
\end{aligned}$$

**Remark**: $S$ is *biased* for the standard deviation $\sigma$.

**Big Example:** Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$, i.e., the p.d.f. is $f(x) = 1/\theta$, $0 < x < \theta$.

Consider two estimators: $Y_1 \equiv 2\bar{X}$ and $Y_2 \equiv \frac{n+1}{n} \max_{1 \le i \le n} X_i$

Since $\mathrm{E}[Y_1] = 2\mathrm{E}[\bar{X}] = 2\mathrm{E}[X_i] = \theta$, we see that $Y_1$ is unbiased for $\theta$.

It's also the case that $Y_2$ is unbiased, but it takes a little more work to show this. As a first step, let's get the cdf of $M \equiv \max_i X_i \ldots$

$$
\begin{aligned}
P(M \le y) &= P(X_1 \le y \text{ and } X_2 \le y \text{ and } \cdots \text{ and } X_n \le y) \\
&= \prod_{i=1}^{n} P(X_i \le y) = [P(X_1 \le y)]^n \quad (X_i\text{'s are iid}) \\
&= \left[\int_0^y f_{X_1}(x)\,dx\right]^n = \left[\int_0^y 1/\theta\,dx\right]^n = (y/\theta)^n.
\end{aligned}
$$

This implies that the p.d.f. of $M$ is

$$
f_M(y) \equiv \frac{d}{dy}(y/\theta)^n = \frac{ny^{n-1}}{\theta^n}.
$$

GTx

Then

$$\mathrm{E}[M] = \int_0^\theta y f_M(y)\,dy = \int_0^\theta \frac{ny^n}{\theta^n} = \frac{n\theta}{n+1}.$$

Whew! So we see that $Y_2 = \frac{n+1}{n} \max_{1 \le i \le n} X_i$ is unbiased for $\theta$.

So both $Y_1$ and $Y_2$ are unbiased for $\theta$, but which is better?

Let's now compare *variances*. After similar algebra, we have

$$\mathrm{Var}(Y_1) = \frac{\theta^2}{3n} \quad \text{and} \quad \mathrm{Var}(Y_2) = \frac{\theta^2}{n(n+2)}.$$

Thus, $Y_2$ has *much lower variance* than $Y_1$. $\square$

**Definition**: The *bias* of an estimator $T(\mathbf{X})$ is $\mathrm{Bias}(T) \equiv \mathrm{E}[T] - \theta$.

The *mean squared error* of $T(\mathbf{X})$ is $\mathrm{MSE}(T) \equiv \mathrm{E}[(T - \theta)^2]$.

**Remark**: After some algebra, we get an easier expression for MSE that combines the bias and variance of an estimator

$$\mathrm{MSE}(T) = \mathrm{Var}(T) + (\underbrace{\mathrm{E}[T] - \theta}_{\mathrm{Bias}})^2.$$

Lower MSE is better — even if there's a little bias.

**Definition**: The *relative efficiency* of $T_2$ to $T_1$ is $\mathrm{MSE}(T_1)/\mathrm{MSE}(T_2)$. If this quantity is < 1, then we'd want $T_1$.

**Example**: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Unif}(0, \theta)$.

Two estimators: $Y_1 = 2\bar{X}$ and $Y_2 = \frac{n+1}{n} \max_i X_i$.

Showed before $\mathrm{E}[Y_1] = \mathrm{E}[Y_2] = \theta$ (so both are unbiased).

Also, $\mathrm{Var}(Y_1) = \frac{\theta^2}{3n}$ and $\mathrm{Var}(Y_2) = \frac{\theta^2}{n(n+2)}$.

Thus, $\mathrm{MSE}(Y_1) = \frac{\theta^2}{3n}$ and $\mathrm{MSE}(Y_2) = \frac{\theta^2}{n(n+2)}$, so $Y_2$ is better.

# Summary

We began our Stats Attack with an intro to point estimation, including unbiasedness and MSE.

Next Time: We'll study maximum likelihood point estimators, which are often quite flexible, even if they're occasionally a teensy biased.

GTx

# Computer Simulation

## Module 2: Calculus, Probability, and Statistics Primers


**Dave Goldsman, Ph.D.**

*Professor*

Stewart School of Industrial and Systems Engineering


## Maximum Likelihood Estimation

# Lesson Overview

Last Time: Started our Stats Bootcamp with review of unbiased estimators and MSE.

This Time: Maximum likelihood estimation – perhaps the most popular point estimation method.

Very flexible technique that many software packages use to help estimate distributions.

**Definition**: Consider an iid random sample $X_1, \ldots, X_n$, where each $X_i$ has pdf/pmf $f(x)$. Further, suppose that $\theta$ is some unknown parameter from $X_i$. The *likelihood function* is $L(\theta) \equiv \prod_{i=1}^{n} f(x_i)$.

**Definition**: The *maximum likelihood estimator* (MLE) of $\theta$ is the value of $\theta$ that maximizes $L(\theta)$. The MLE is a function of the $X_i$'s and is a RV.

**Example**: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. Find the MLE for $\lambda$.

$$L(\lambda) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} x_i\Big).$$

Now maximize $L(\lambda)$ with respect to $\lambda$.

Could take the derivative and plow through all of the horrible algebra.

Useful Trick: Since the natural log function is one-to-one, it's easy to see that the $\lambda$ that maximizes $L(\lambda)$ also maximizes $\ell\mathrm{n}(L(\lambda))$!

$$L(\lambda) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} x_i\Big).$$

$$\ell\mathrm{n}(L(\lambda)) = \ell\mathrm{n}\left(\lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} x_i\Big)\right) = n\ell\mathrm{n}(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

This makes our job less horrible.

$$\frac{d}{d\lambda}\ell\mathrm{n}(L(\lambda)) = \frac{d}{d\lambda}\Big(n\ell\mathrm{n}(\lambda) - \lambda \sum_{i=1}^{n} x_i\Big) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \equiv 0.$$

This implies that the MLE is $\hat{\lambda} = 1/\bar{X}$. $\square$

**Remarks:** (1) $\hat{\lambda} = 1/\bar{X}$ makes sense since $\mathrm{E}[X] = 1/\lambda$.

(2) At the end, we put a little $\widehat{\text{hat}}$ over $\lambda$ to indicate that this is the MLE.

(3) At the end, we make all of the little $x_i$'s into big $X_i$'s to indicate that this is a RV.

(4) Just to be careful, you probably ought to perform a second-derivative test, but I won't blame you if you don't.

**Theorem** (Invariance Property of MLE's): If $\hat{\theta}$ is the MLE of some parameter $\theta$ and $h(\cdot)$ is a 1:1 function, then $h(\hat{\theta})$ is the MLE of $h(\theta)$.

**Example**: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. The *survival function* is

$$\bar{F}(x) = P(X > x) = 1 - F(x) = e^{-\lambda x}.$$

In addition, we saw that the MLE for $\lambda$ is $\hat{\lambda} = 1/\bar{X}$.

Then the invariance property says that the MLE of $\bar{F}(x)$ is

$$\widehat{\bar{F}}(x) = e^{-\hat{\lambda} x} = e^{-x/\bar{X}}.$$

This kind of thing is used all of the time the actuarial sciences.

# Summary

Went over some basics on Maximum Likelihood Estimators. These will be useful later in the course, e.g., when we do simulation input analysis.

Next Time: Point estimators are fine, but we can do better. The next lesson will deal with confidence intervals. I'm confident that you'll enjoy it.

GTx

# Computer Simulation

## Module 2: Calculus, Probability, and Statistics Primers

**Dave Goldsman, Ph.D.**
*Professor*
Stewart School of Industrial and Systems Engineering

## Confidence Intervals

# Lesson Overview

Last Time: Reviewed maximum likelihood estimation for an unknown parameter θ.

This Time: we'll do even better now… confidence intervals for θ!

We'll use CIs throughout the course, especially when we do output analysis.

Some distributions that arise in confidence intervals...

**Definitions:** If $Z_1, Z_2, \ldots, Z_k$ are iid Nor(0,1), then $Y = \sum_{i=1}^{k} Z_i^2$ has the $\chi^2$ *distribution with $k$ degrees of freedom (df)*. Notation: $Y \sim \chi^2(k)$. Note that $\mathrm{E}[Y] = k$ and $\mathrm{Var}(Y) = 2k$.

If $Z \sim \mathrm{Nor}(0,1)$, $Y \sim \chi^2(k)$, and $Z$ and $Y$ are independent, then $T = Z/\sqrt{Y/k}$ has the *Student t distribution with $k$ df*. Notation: $T \sim t(k)$. Note that the $t(1)$ is the *Cauchy* distribution.

If $Y_1 \sim \chi^2(m)$, $Y_2 \sim \chi^2(n)$, and $Y_1$ and $Y_2$ are independent, then $F = (Y_1/m)/(Y_2/n)$ has the *F distribution with $m$ and $n$ df*. Notation: $F \sim F(m,n)$.

How (and why) would one use the above facts? Because they can be used to construct *confidence intervals* (CIs) for $\mu$ and $\sigma^2$ under a variety of assumptions.

A $100(1-\alpha)\%$ two-sided CI for an unknown parameter $\theta$ is a random interval $[L, U]$ such that $P(L \le \theta \le U) = 1 - \alpha$.

Here are some examples / theorems, all of which assume that the $X_i$'s are iid normal...

**Example:** If $\sigma^2$ is *known*, then a $100(1-\alpha)\%$ CI for $\mu$ is

$$\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \le \mu \le \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}},$$

where $z_\gamma$ is the $1-\gamma$ quantile of the standard normal distribution, i.e., $z_\gamma \equiv \Phi^{-1}(1-\gamma)$.

**Example:** If $\sigma^2$ is *unknown*, then a $100(1-\alpha)\%$ CI for $\mu$ is

$$\bar{X}_n - t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}} \le \mu \le \bar{X}_n + t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}},$$

where $t_{\gamma,\nu}$ is the $1-\gamma$ quantile of the $t(\nu)$ distribution.

**Example:** A $100(1-\alpha)\%$ CI for $\sigma^2$ is

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}},$$

where $\chi^2_{\gamma,\nu}$ is the $1-\gamma$ quantile of the $\chi^2(\nu)$ distribution.

**Example:** Here are 20 residual flame times (in sec.) of treated specimens of children's nightwear.

(Don't worry — children were not in the nightwear when the clothing was set on fire.)

| | | | | |
|---|---|---|---|---|
| 9.85 | 9.93 | 9.75 | 9.77 | 9.67 |
| 9.87 | 9.67 | 9.94 | 9.85 | 9.75 |
| 9.83 | 9.92 | 9.74 | 9.99 | 9.88 |
| 9.95 | 9.95 | 9.93 | 9.92 | 9.89 |

Let's get a 95% CI for the mean residual flame time.

After a little algebra, we get

$$\bar{X} = 9.8475 \quad \text{and} \quad S = 0.0954.$$

Further, you can use the Excel function `t.inv(0.975,19)` to get $t_{\alpha/2,n-1} = t_{0.025,19} = 2.093$.

Then the half-length of the CI is

$$H = t_{\alpha/2,n-1}\sqrt{S^2/n} = \frac{(2.093)(0.0954)}{\sqrt{20}} = 0.0446.$$

Thus, the CI is $\mu \in \bar{X} \pm H$, or $9.8029 \le \mu \le 9.8921$. $\square$

# Summary

Reviewed a few CIs for the mean and variance of a normal distribution.  More coming up later when we need them.

This completes Module 2, where we went to Calc, Prob, and Stats Bootcamps, sneaking in some simulation here and there.

Module 3 concerns hand and spreadsheet simulations.