ISYE 6644 Simulation, Math Bootcamp

Liuyi Ye

1. Calculus

First of all, let's suppose that $f(x)$ is a *function* that maps values of x from a certain *domain* X to a certain *range* Y, which we can denote by the shorthand $f: X \to Y$.

Example If $f(x) = x^2$, then the function takes x-values from the real line $\mathbb{R}$ to the nonnegative portion of the real line $\mathbb{R}^+$.

Definition We say that $f(x)$ is a *continuous* function if, for any x_0 and $x \in X$, we have $\lim_{x \to x_0} f(x) = f(x_0)$, where "lim" denotes a *limit* and $f(x)$ is assumed to exist for all $x \in X$.

Example The function $f(x) = 3x^2$ is continuous for all x. The function $f(x) = \lfloor x \rfloor$ (round down to the nearest integer, e.g., $\lfloor 3.4 \rfloor = 3$) has a "jump" discontinuity at any integer x. □

Definition If $f(x)$ is continuous, then the *derivative* (slope) is

$$\frac{d}{dx} f(x) \equiv f'(x) \equiv \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

assuming it exists and is well-defined for any given x.

Some useful formula and properties

$$\begin{aligned}
[x^k]' &= kx^{k-1}, \\
[e^x]' &= e^x, \\
[\sin(x)]' &= \cos(x), \\
[\cos(x)]' &= -\sin(x), \\
[\ell n(x)]' &= 1/x, \\
[\arctan(x)]' &= 1/(1+x^2).
\end{aligned}$$

Theorem Some well-known properties of derivatives are:

$$\begin{aligned}
[af(x) + b]' &= af'(x), \\
[f(x) + g(x)]' &= f'(x) + g'(x), \\
[f(x)g(x)]' &= f'(x)g(x) + f(x)g'(x) \quad \text{(product rule)}, \\
\left[\frac{f(x)}{g(x)}\right]' &= \frac{g(x)f'(x) - f(x)g'(x)}{g^2(x)} \quad \text{(quotient rule)}^1, \\
[f(g(x))]' &= f'(g(x))g'(x) \quad \text{(chain rule)}^2.
\end{aligned}$$

Second derivative

Remark The second derivative $f''(x) \equiv \frac{d}{dx} f'(x)$ and is the "slope of the slope." If $f(x)$ is "position," then $f'(x)$ can be regarded as "velocity," and as $f''(x)$ as "acceleration."

The minimum or maximum of $f(x)$ can only occur when the slope of $f(x)$ is zero, i.e., only when $f'(x) = 0$, say at $x = x_0$.

Then if $f''(x_0) < 0$, you get a max; if $f''(x_0) > 0$, you get a min; and if $f''(x_0) = 0$, you get a *point of inflection*.

Example Find x that minimizes $f(x) = e^{2x} + e^{-x}$. The min can only occur when $f'(x) = 2e^{2x} - e^{-x} = 0$. This occurs at $x_0 = -(1/3)\ell n(2)$. It's easy to show that $f''(x) > 0$ for all x; so x_0 is the min. □

2. Formal ways to search solutions to equations (finding a 0)

How might you find a 0 for a complicated nonlinear function, i.e., *x* such that *f(x)* = 0?

- Trial-and-error (not so great).
- Bisection (divide-and-conquer).
- Newton's method (or some variation)
- Fixed-point method (we'll do this later).

Bisection. By the Intermediate Value Theorem.
Newton's method,

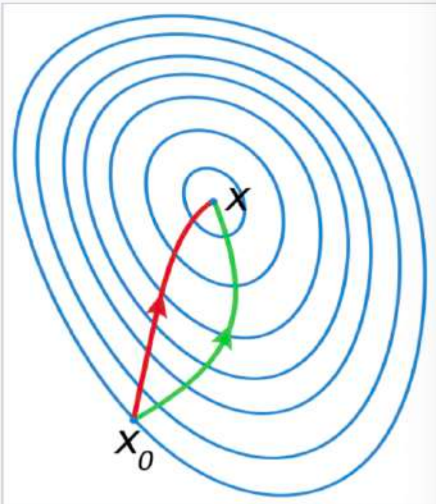
Newton's Method: Suppose you can find a reasonable first guess for the zero, say, x_i, where we start off at iteration $i = 0$. If $g(x)$ has a nice, well-behaved derivative (which doesn't happen to be too flat near the zero of $g(x)$), then iterate your guess as follows:

$$x_{i+1} = x_i - \frac{g(x_i)}{g'(x_i)}.$$

Keep going until things appear to converge.

This makes sense since for x_i and x_{i+1} close to each other and the zero $x^\star$, we have

$$g'(x_i) \approx \frac{g(x^\star) - g(x_i)}{x^\star - x_i}.$$



A comparison of gradient descent (green) and Newton's method (red) for minimizing a function (with small step sizes). Newton's method uses curvature information (i.e. the second derivative) to take a more direct route.

It is worth noting that we use gradient descent more in ML algorithms.

A comparison can be found here (Gatech ISYE 6416 Computational Statistics, Lecture 4, Newton's method and Gradient descent).

The basic idea is that:

Gradient descent tries to find a minimum x by using information from the first derivative of the function.

Newton's method tries to find a point x satisfying f'(x) = 0 by approximating f' with a linear function g and then solving for the root of that function explicitly.

3. Integration

Definition The function $F(x)$ having derivative $f(x)$ is called the *antiderivative*. The antiderivative is denoted $F(x) = \int f(x)\,dx$; and this is also called the *indefinite integral* of $f(x)$.

Fundamental Theorem of Calculus: If $f(x)$ is continuous, then the area under the curve for $x \in [a, b]$ is denoted and given by the *definite integral* [3]

$$\int_a^b f(x)\,dx \equiv F(x)\Big|_a^b \equiv F(b) - F(a).$$

$$\int x^k\,dx = \frac{x^{k+1}}{k+1} + C,\ k \neq -1,$$

$$\int \frac{dx}{x} = \ell n|x| + C,$$

$$\int e^x\,dx = e^x + C,$$

$$\int \cos(x)\,dx = \sin(x) + C,$$

$$\int \frac{dx}{1+x^2} = \arctan(x) + C.$$

Some useful integrals s to remember

Some useful properties of definite integrals

Theorem Some well-known properties of definite integrals are:

$$\int_a^a f(x)\,dx = 0,$$

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx,$$

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

Theorem Some other properties of general integrals are:

$$\int [f(x) + g(x)]\,dx = \int f(x)\,dx + \int g(x)\,dx,$$

$$\int f(x)g'(x)\,dx = f(x)g(x) - \int g(x)f'(x)\,dx \quad \text{(parts)}^4,$$

$$\int f(g(x))g'(x)\,dx = \int f(u)\,du \quad \text{(substitution rule)}^5.$$

Taylor Series Expansion about a point *a* of *f(x)*

Definition Derivatives of arbitrary order k can be written as $f^{(k)}(x)$ or $\frac{d^k}{dx^k}f(x)$. By convention, $f^{(0)}(x) = f(x)$.

The *Taylor series expansion* of $f(x)$ about a point a is given by

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)(x-a)^k}{k!}.$$

The *Maclaurin series* is simply Taylor expanded around $a = 0$.

Some Maclaurin series to remember

$$\sin(x) = \sum_{k=0}^{\infty} \frac{(-1)^{k+1}x^{2k+1}}{(2k+1)!},$$

$$\cos(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!},$$

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Some useful Series

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2},$$

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6},$$

$$\sum_{k=0}^{\infty} p^k = \frac{1}{1-p} \quad (\text{for } -1 < p < 1).$$

L' Hospital's Rule

Theorem Occasionally, we run into trouble when taking indeterminate ratios of the form $0/0$ or ∞/∞. In such cases, *L'Hôspital's Rule*[6] is useful: If the limits $\lim_{x\to a} f(x)$ and $\lim_{x\to a} g(x)$ both go to 0 or both go to ∞, then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

[6] This rule makes me sick.

Example L'Hôspital shows that

$$\lim_{x\to 0} \frac{\sin(x)}{x} = \lim_{x\to 0} \frac{\cos(x)}{1} = 1. \quad \square$$

4. Numerical Integration, Computer Exercises

Riemann sums.

Computer Exercise: Let's do some easy integration via *Riemann sums*. Simply approximate the area under the nice, continuous function $f(x)$ from a to b by adding up the areas of n adjacent rectangles of width $\Delta x = (b-a)/n$ and height $f(x_i)$, where $x_i = a + i\Delta x$ is the right-hand endpoint of the ith rectangle. Thus,

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{n} f(x_i)\Delta x = \frac{b-a}{n} \sum_{i=1}^{n} f\left(a + \frac{i(b-a)}{n}\right).$$

In fact, as $n \to \infty$, this result becomes an equality.

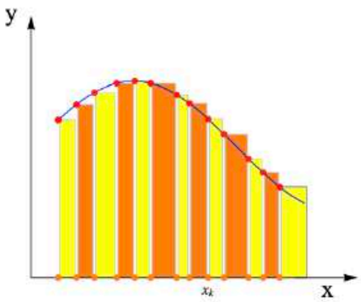
A more general Riemann sum is obtained by choosing n points in $[a, b]$ and defining

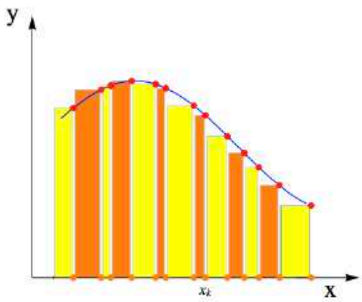
$$S_n = \sum f(y_j)(x_{j+1} - x_j) = \sum_{y_j} f(y_j)\Delta x_j$$

where y_j is in (x_j, x_{j+1}).

This generalization allows to use a small mesh size where the function fluctuates a lot.

The sum $\sum f(x_j)\Delta x_j$ is called the **left Riemann sum**, the sum $\sum f(x_{j+1})\Delta x_j$ the **right Riemann sum**.





If $x_0 = a, x_n = b$ and $\max_j \Delta x_j \to 0$ for $n \to \infty$ then S_n converges to $\int_a^b f(x)\, dx$.

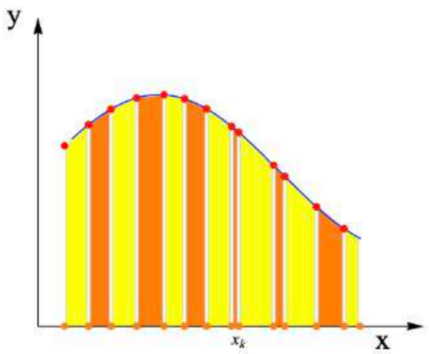
Trapezoid Version

Computer Exercise, Trapezoid version: Same numerical integration via the Trapezoid Rule (which usually works a little better than Riemann). Now we have

$$\begin{aligned}\int_a^b f(x)\, dx &\approx \left[\frac{f(x_0)}{2} + \sum_{i=1}^{n-1} f(x_i) + \frac{f(x_n)}{2}\right]\Delta x \\ &= \frac{b-a}{n}\left[\frac{f(a)}{2} + \sum_{i=1}^{n-1} f\left(a + \frac{i(b-a)}{n}\right) + \frac{f(b)}{2}\right].\end{aligned}$$

Trapezoid rule

The average between the left and right hand Riemann sum is called the **Trapezoid rule**. Geometrically, it sums up areas of trapezoids instead of rectangles.



The Trapezoid rule does not change things much in the case of equal spacing $x_k = a + (b-a)k/n$.

$$\frac{1}{2n}[f(x_0) + f(x_n)] + \frac{1}{n}\sum_{k=1}^{n-1} f(x_k) \ .$$

Monte Carlo Simulation

Monte Carlo Method

A powerful integration method is to chose n random points x_k in $[a,b]$ and look at the sum divided by n. Because it uses randomness, it is called **Monte Carlo method**.

The **Monte Carlo** integral is the limit S_n to infinity

$$S_n = \frac{1}{n}\sum_{k=1}^{n} f(x_k) \ ,$$

where x_k are n random values in $[a,b]$.

The law of large numbers in probability shows that the **Monte Carlo integral** is equivalent to the **Lebesgue integral** which is more powerful than the Riemann integral. Monte Carlo integration is interesting especially if the function is complicated.

3 Lets look at the **salt and pepper** function

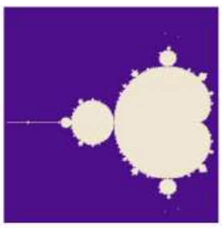
$$f(x) = \begin{cases} 1 & x \text{ rational} \\ 0 & x \text{ irrational} \end{cases}$$

The Riemann integral with equal spacing k/n is equal to 1 for every n. But this is only because we have evaluated the function at rational points, where it is 1.
The Monte Carlo integral gives zero because if we chose a random number in $[0,1]$ we hit an irrational number with probability 1.

Monte Carlo Intuition

The Salt and Pepper function and the Boston Salt and Pepper bridge (Anne Heywood).

1 The following two lines evaluate the **area of the Mandelbrot fractal** using Monte Carlo integration. The function F is equal to 1, if the parameter value c of the quadratic map $z \to z^2 + c$ is in the Mandelbrot set and 0 else. It shoots $100'000$ random points and counts what fraction of the square of area 9 is covered by the set. Numerical experiments give values close to the actual value around 1.51.... One could use more points to get more accurate estimates.



5. Probability

Basics

Will assume that you know about sample spaces, events, and the definition of probability.

Definition: $P(A|B) \equiv P(A \cap B)/P(B)$ is the *conditional probability of A given B*.

Example: Toss a fair die. Let $A = \{1, 2, 3\}$ and $B = \{3, 4, 5, 6\}$. Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{4/6} = 1/4. \quad \square$$

Definition: A *random variable* (RV) X is a function from the sample space Ω to the real line, i.e., $X : \Omega \to \mathbb{R}$.

Example: Let X be the sum of two dice rolls. Then $X((4,6)) = 10$. In addition,

$$P(X = x) = \begin{cases} 1/36 & \text{if } x = 2 \\ 2/36 & \text{if } x = 3 \\ \vdots & \\ 1/36 & \text{if } x = 12 \\ 0 & \text{otherwise} \end{cases}$$

Definition: If the number of possible values of a RV X is finite or countably infinite, then X is a *discrete* RV. Its *probability mass function* (pmf) is $f(x) \equiv P(X = x)$. Note that $\sum_x f(x) = 1$.

Definition: A *continuous* RV is one with probability zero at every individual point, and for which there exists a *probability density function* (pdf) $f(x)$ such that $P(X \in A) = \int_A f(x)\,dx$ for every set A. Note that $\int_{\mathbb{R}} f(x)\,dx = 1$.

Example: Pick a random number between 3 and 7. Then

$$f(x) = \begin{cases} 1/4 & \text{if } 3 \le x \le 7 \\ 0 & \text{otherwise} \end{cases} \qquad \square$$

Some continuous RVs are Uniform (a, b), Exponential(Lambda), Normal (mu, sigma^2)

Definition: For any RV X (discrete or continuous), the *cumulative distribution function* (cdf) is

$$F(x) \equiv P(X \leq x) = \begin{cases} \sum_{y \leq x} f(y) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{x} f(y)\,dy & \text{if } X \text{ is continuous} \end{cases}$$

Note that $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$. In addition, if X is continuous, then $\frac{d}{dx}F(x) = f(x)$.

Example: Flip 2 coins. Let X be the number of heads.

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/4 & \text{if } 0 \leq x < 1 \\ 3/4 & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2 \end{cases} \qquad \square$$

Example: if $X \sim \text{Exp}(\lambda)$ (i.e., X is exponential with parameter λ), then $f(x) = \lambda e^{-\lambda x}$ and $F(x) = 1 - e^{-\lambda x}$, $x \geq 0$. $\square$

6. Simulating Random Variable

Example (Discrete Uniform): Consider a D.U. on $\{1, 2, \ldots, n\}$, i.e., $X = i$ with probability $1/n$ for $i = 1, 2, \ldots, n$. (Think of this as an n-sided dice toss for you Dungeons and Dragons fans.)

If $U \sim \text{Unif}(0,1)$, we can obtain a D.U. random variate simply by setting $X = \lceil nU \rceil$, where $\lceil \cdot \rceil$ is the "ceiling" (or "round up") function.

For example, if $n = 10$ and we sample a Unif(0,1) random variable $U = 0.73$, then $X = \lceil 7.3 \rceil = 8$. $\square$

Example (Another Discrete Random Variable):

$$P(X = x) = \begin{cases} 0.25 & \text{if } x = -2 \\ 0.10 & \text{if } x = 3 \\ 0.65 & \text{if } x = 4.2 \\ 0 & \text{otherwise} \end{cases}$$

Can't use a die toss to simulate this random variable. Instead, use what's called the *inverse transform method*.

x	$f(x)$	$P(X \leq x)$	Unif(0,1)'s
-2	0.25	0.25	$[0.00, 0.25]$
3	0.10	0.35	$(0.25, 0.35]$
4.2	0.65	1.00	$(0.35, 1.00)$

Sample $U \sim \text{Unif}(0,1)$. Choose the corresponding x-value, i.e., $X = F^{-1}(U)$. For example, $U = 0.46$ means that $X = 4.2$. $\square$

Inverse Transform Method

Inverse Transform Theorem: Let X be a continuous random variable with c.d.f. $F(x)$. Then $F(X) \sim \mathcal{U}(0,1)$.

Proof: Let $Y = F(X)$ and suppose that Y has c.d.f. $G(y)$. Then

$$\begin{aligned} G(y) &= \mathrm{P}(Y \leq y) = \mathrm{P}(F(X) \leq y) \\ &= \mathrm{P}(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y. \quad \Box \end{aligned}$$

In the above, we can define the inverse c.d.f. by

$$F^{-1}(y) = \min[x : F(x) \geq y] \quad y \in [0,1].$$

This representation can be applied to continuous or *discrete* or mixed distributions (see figure).

Some random variable generation related stuff by Dave Goldsman can be found here

This suggests a way to generate realizations of the RV X. Simply set $F(X) = U \sim \mathrm{Unif}(0,1)$ and solve for $X = F^{-1}(U)$.

Example: Suppose $X \sim \mathrm{Exp}(\lambda)$. Then $F(x) = 1 - e^{-\lambda x}$ for $x > 0$. Set $F(X) = 1 - e^{-\lambda X} = U$. Solve for X,

$$X = \frac{-1}{\lambda}\ell\mathrm{n}(1-U) \sim \mathrm{Exp}(\lambda). \quad \Box$$

Example (Generating Uniforms): The above RV generation examples required us to generate "practically" independent and identically distributed (iid) Unif(0,1) RV's.

If you don't like programming, you can use Excel function `RAND()` or something similar to generate Unif(0,1)'s.

Here's an algorithm to generate *pseudo-random numbers (PRN's)*, i.e., a series $R_1, R_2, \ldots$ of *deterministic* numbers that *appear* to be iid Unif(0,1). Pick a *seed* integer X_0, and calculate

$$X_i = 16807 X_{i-1} \mathrm{mod}(2^{31}-1), \quad i = 1,2,\ldots.$$

Then set $R_i = X_i/(2^{31}-1)$, $i = 1,2,\ldots$.

2^32 -1 is prime. A comprehensive review of Mersenne Primes are here.

7. Expected Values of Random Variables

Definition: The *expected value* (or *mean*) of a RV X is

$$\mathrm{E}[X] \equiv \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x f(x)\, dx & \text{if } X \text{ is continuous} \end{cases} = \int_{\mathbb{R}} x\, dF(x).$$

Example: Suppose that $X \sim \text{Bernoulli}(p)$. Then

$$X = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob. } 1-p\ (= q) \end{cases}$$

and we have $\mathrm{E}[X] = \sum_x x f(x) = p$. □

Example: Suppose that $X \sim \text{Uniform}(a, b)$. Then

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

and we have $\mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\, dx = (a+b)/2$. □

Example: Suppose that $X \sim \text{Exponential}(\lambda)$. Then

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

and we have (after integration by parts and L'Hôspital's Rule)

$$\mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\, dx = \int_0^\infty x \lambda e^{-\lambda x}\, dx = \frac{1}{\lambda}. \quad \square$$

The Law of Unconscious Statistician 😂

Def/Thm (the "Law of the Unconscious Statistician" or "LOTUS"): Suppose that $h(X)$ is some function of the RV X. Then

$$\mathrm{E}[h(X)] = \begin{cases} \sum_x h(x) f(x) & \text{if } X \text{ is disc} \\ \int_{\mathbb{R}} h(x) f(x)\, dx & \text{if } X \text{ is cts} \end{cases} = \int_{\mathbb{R}} h(x)\, dF(x).$$

The function $h(X)$ can be anything "nice", e.g., $h(X) = X^2$ or $1/X$ or $\sin(X)$ or $\ell n(X)$.

Example: Suppose X is the following discrete RV:

x	2	3	4
$f(x)$	0.3	0.6	0.1

Then $\mathrm{E}[X^3] = \sum_x x^3 f(x) = 8(0.3) + 27(0.6) + 64(0.1) = 25.$ $\square$

Example: Suppose $X \sim \mathrm{Unif}(0, 2)$. Then

$$\mathrm{E}[X^n] = \int_{\mathbb{R}} x^n f(x)\, dx = 2^n/(n+1). \quad \square$$

Moments, Calculating Variance

Definitions: $\mathrm{E}[X^n]$ is the nth *moment* of X.

$\mathrm{E}[(X - \mathrm{E}[X])^n]$ is the nth *central moment* of X.

$\mathrm{Var}(X) \equiv \mathrm{E}[(X - \mathrm{E}[X])^2]$ is the *variance* of X.

The *standard deviation* of X is $\sqrt{\mathrm{Var}(X)}$.

Theorem: $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$ (sometimes easier to calculate this way).

Example: Suppose $X \sim \mathrm{Bern}(p)$. Recall that $\mathrm{E}[X] = p$. Then

$$\mathrm{E}[X^2] = \sum_x x^2 f(x) = p \quad \text{and}$$

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = p(1-p). \quad \square$$

Example: Suppose $X \sim \mathrm{Exp}(\lambda)$. By LOTUS,

$$\mathrm{E}[X^n] = \int_0^\infty x^n \lambda e^{-\lambda x}\, dx = n!/\lambda^n.$$

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = 1/\lambda^2. \quad \square$$

Theorem: $\mathrm{E}[aX+b] = a\mathrm{E}[X]+b$ and $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$.

Example: If $X \sim \mathrm{Exp}(3)$, then

$$\mathrm{E}[-2X+7] = -2\mathrm{E}[X]+7 = -\frac{2}{3}+7.$$

$$\mathrm{Var}(-2X+7) = (-2)^2\mathrm{Var}(X) = \frac{4}{9}. \quad \square$$

8. Moment generating function

Definition: $M_X(t) \equiv \mathrm{E}[e^{tX}]$ is the *moment generating function* (mgf) of the RV X. ($M_X(t)$ is a function of t, *not* of X!)

Example: $X \sim \mathrm{Bern}(p)$. Then

$$M_X(t) = \mathrm{E}[e^{tX}] = \sum_x e^{tx} f(x) = e^{t\cdot 1}p + e^{t\cdot 0}q = pe^t + q.$$

Example: $X \sim \mathrm{Exp}(\lambda)$. Then

$$M_X(t) = \int_{\mathbb{R}} e^{tx} f(x)\,dx = \lambda \int_0^\infty e^{(t-\lambda)x}\,dx = \frac{\lambda}{\lambda - t} \quad \text{if } \lambda > t.$$

Theorem: Under certain technical conditions,

$$\mathrm{E}[X^k] = \frac{d^k}{dt^k} M_X(t)\Big|_{t=0}, \quad k = 1, 2, \ldots.$$

Thus, you can *generate* the moments of X from the mgf.

Example: $X \sim \mathrm{Exp}(\lambda)$. Then $M_X(t) = \frac{\lambda}{\lambda - t}$ for $\lambda > t$. So

$$\mathrm{E}[X] = \frac{d}{dt} M_X(t)\Big|_{t=0} = \frac{\lambda}{(\lambda - t)^2}\Big|_{t=0} = 1/\lambda.$$

Further,

$$\mathrm{E}[X^2] = \frac{d^2}{dt^2} M_X(t)\Big|_{t=0} = \frac{2\lambda}{(\lambda - t)^3}\Big|_{t=0} = 2/\lambda^2.$$

Thus,

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = 1/\lambda^2. \quad \square$$

9. Functions of a Random Variable

Problem: Suppose we have a RV X with pmf/pdf $f(x)$. Let $Y = h(X)$. Find $g(y)$, the pmf/pdf of Y.

Examples (take my word for it for now):

If $X \sim \text{Nor}(0,1)$, then $Y = X^2 \sim \chi^2(1)$.

If $U \sim \text{Unif}(0,1)$, then $Y = -\frac{1}{\lambda}\ell n(U) \sim \text{Exp}(\lambda)$.

Discrete Example: Let X denote the number of H's from two coin tosses. We want the pmf for $Y = X^3 - X$.

x	0	1	2
$f(x)$	1/4	1/2	1/4
$y = x^3 - x$	0	0	6

This implies that $g(0) = P(Y = 0) = P(X = 0 \text{ or } 1) = 3/4$ and $g(6) = P(Y = 6) = 1/4$. In other words,

$$g(y) = \begin{cases} 3/4 & \text{if } y = 0 \\ 1/4 & \text{if } y = 6 \end{cases} . \quad \square$$

Continuous Example: Suppose X has pdf $f(x) = |x|$, $-1 \le x \le 1$. Find the pdf of $Y = X^2$.

First of all, the cdf of Y is

$$\begin{aligned} G(y) &= P(Y \le y) \\ &= P(X^2 \le y) \\ &= P(-\sqrt{y} \le X \le \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} |x|\,dx \;=\; y, \quad 0 < y < 1. \end{aligned}$$

The pdf of Y is $g(y) = G'(y) = 1$, so that $Y \sim \text{Unif}(0,1)$. $\square$

Inverse Transform Theorem: Suppose X is a continuous random variable having cdf $F(x)$. Then, amazingly, $F(X) \sim \text{Unif}(0,1)$.

Proof: Let $Y = F(X)$. Then the cdf of Y is

$$\begin{aligned} P(Y \le y) &= P(F(X) \le y) \\ &= P(X \le F^{-1}(y)) \\ &= F(F^{-1}(y)) \;=\; y, \end{aligned}$$

which is the cdf of the Unif(0,1). $\square$

This result is of great importance when it comes to generating RV's during a simulation.

Example (how to generate exponential RV's): Suppose $X \sim \text{Exp}(\lambda)$, with cdf $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$.

So the Inverse Transform Theorem implies that

$$F(X) = 1 - e^{-\lambda X} \sim \text{Unif}(0,1).$$

Let $U \sim \text{Unif}(0,1)$ and set $F(X) = U$. Then we have

$$X = \frac{-1}{\lambda}\,\ell\text{n}(1-U) \sim \text{Exp}(\lambda).$$

For instance, if $\lambda = 2$ and $U = 0.27$, then $X = 0.157$ is an Exp(2) realization. $\Box$

Exercise: Suppose that X has the Weibull distribution with cdf

$$F(x) = 1 - e^{-(\lambda x)^\beta}, x > 0.$$

If you set $F(X) = U$ and solve for X, show that you get

$$X = \frac{1}{\lambda}\left[-\ell\text{n}(1-U)\right]^{1/\beta}.$$

Now pick your favorite λ and β, and use this result to generate values of X. In fact, make a histogram of your X values. Are there any interesting values of λ and β you could've chosen?

Bonus Theorem: Here's another way to get the pdf of $Y = h(X)$ for some nice continuous function $h(\cdot)$. The cdf of Y is

$$F_Y(y) = P(Y \leq y) = P(h(X) \leq y) = P(X \leq h^{-1}(y)).$$

By the chain rule (and since a pdf must be ≥ 0), the pdf of Y is

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X(h^{-1}(y))\left|\frac{d}{dy}h^{-1}(y)\right|.$$

And now, here's how to prove LOTUS!

$$\begin{aligned} \text{E}[Y] &= \int_{\mathbb{R}} y f_Y(y)\,dy = \int_{\mathbb{R}} y f_X(h^{-1}(y))\left|\frac{d}{dy}h^{-1}(y)\right| dy \\ &\text{“=”} \int_{\mathbb{R}} y f_X(h^{-1}(y))dh^{-1}(y) = \int_{\mathbb{R}} h(x) f_X(x)\,dx. \quad \Box \end{aligned}$$

10. Jointly Distributed Random Variables

Idea: Consider two random variables interacting together — think height and weight.

Definition: The *joint cdf* of X and Y is

$$F(x,y) \equiv P(X \le x, Y \le y), \quad \text{for all } x, y.$$

Remark: The *marginal cdf* of X is $F_X(x) = F(x,\infty)$. (We use the X subscript to remind us that it's just the cdf of X all by itself.) Similarly, the *marginal cdf* of Y is $F_Y(y) = F(\infty, y)$.

Definition: If X and Y are discrete, then the *joint pmf* of X and Y is $f(x,y) \equiv P(X = x, Y = y)$. Note that $\sum_x \sum_y f(x,y) = 1$.

Remark: The *marginal pmf* of X is

$$f_X(x) = P(X = x) = \sum_y f(x,y).$$

The *marginal pmf* of Y is

$$f_Y(y) = P(Y = y) = \sum_x f(x,y).$$

Example: The following table gives the joint pmf $f(x,y)$, along with the accompanying marginals.

$f(x,y)$	$X=2$	$X=3$	$X=4$	$f_Y(y)$
$Y=4$	0.3	0.2	0.1	0.6
$Y=6$	0.1	0.2	0.1	0.4
$f_X(x)$	0.4	0.4	0.2	1

Definition: If X and Y are continuous, then the *joint pdf* of X and Y is $f(x,y) \equiv \frac{\partial^2}{\partial x \partial y} F(x,y)$. Note that $\int_{\mathbb{R}} \int_{\mathbb{R}} f(x,y)\, dx\, dy = 1$.

Remark: The *marginal pdf's* of X and Y are

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\, dy \quad \text{and} \quad f_Y(y) = \int_{\mathbb{R}} f(x,y)\, dx.$$

Example: Suppose the joint pdf is

$$f(x,y) = \frac{21}{4}x^2y, \quad x^2 \le y \le 1.$$

Then the marginal pdf's are:

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,dy = \int_{x^2}^{1} \frac{21}{4}x^2y\,dy = \frac{21}{8}x^2(1-x^4),\ -1 \le x \le 1$$

and

$$f_Y(y) = \int_{\mathbb{R}} f(x,y)\,dx = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{21}{4}x^2y\,dx = \frac{7}{2}y^{5/2}, \quad 0 \le y \le 1.$$

Independent Random Variable

Definition: X and Y are *independent* RV's if

$$f(x,y) = f_X(x)f_Y(y) \quad \text{for all } x, y.$$

Theorem: X and Y are indep if you can write their joint pdf as $f(x,y) = a(x)b(y)$ for some functions $a(x)$ and $b(y)$, and x and y don't have funny limits (their domains do not depend on each other).

Examples: If $f(x,y) = cxy$ for $0 \le x \le 2$, $0 \le y \le 3$, then X and Y are independent.

If $f(x,y) = \frac{21}{4}x^2y$ for $x^2 \le y \le 1$, then X and Y are *not* independent.

If $f(x,y) = c/(x+y)$ for $1 \le x \le 2$, $1 \le y \le 3$, then X and Y are *not* independent. $\square$

Conditional pdf

Definition: The *conditional pdf* (or *pmf*) of Y given $X = x$ is $f(y|x) \equiv f(x,y)/f_X(x)$ (assuming $f_X(x) > 0$).

This is a legit pmf/pdf. For example, in the continuous case, $\int_{\mathbb{R}} f(y|x)\,dy = 1$, for any x.

Example: Suppose $f(x,y) = \frac{21}{4}x^2y$ for $x^2 \le y \le 1$. Then

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{21}{4}x^2y}{\frac{21}{8}x^2(1-x^4)} = \frac{2y}{1-x^4}, \quad x^2 \le y \le 1.$$

Theorem: If X and Y are indep, then $f(y|x) = f_Y(y)$ for all x, y.

Proof: By definition of conditional and independence,

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)}. \quad \square$$

Definition: The *conditional expectation* of Y given $X = x$ is

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

Example: The expected weight of a 7' tall guy ($\mathrm{E}[Y|X=7]$) is > the expected weight of a totally random guy ($\mathrm{E}[Y]$).

Old Cts Example: $f(x,y) = \frac{21}{4}x^2 y$, if $x^2 \le y \le 1$. Then

$$\mathrm{E}[Y|x] = \int_{\mathbb{R}} y f(y|x)\, dy = \int_{x^2}^{1} \frac{2y^2}{1-x^4}\, dy = \frac{2}{3} \cdot \frac{1-x^6}{1-x^4}. \quad \square$$

11.Conditional Expectation (Cont'd) and Applications

Double Expectation

Old Example: Suppose $f(x,y) = \frac{21}{4}x^2 y$, if $x^2 \le y \le 1$. By previous examples, we know $f_X(x)$, $f_Y(y)$, and $\mathrm{E}[Y|x]$. Let's find $\mathrm{E}[Y]$.

Solution #1 (old, boring way):

$$\mathrm{E}[Y] = \int_{\mathbb{R}} y f_Y(y)\, dy = \int_0^1 \frac{7}{2} y^{7/2}\, dy = \frac{7}{9}.$$

Solution #2 (new, exciting way):

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[\mathrm{E}(Y|X)] = \int_{\mathbb{R}} \mathrm{E}(Y|x) f_X(x)\, dx \\ &= \int_{-1}^{1} \left(\frac{2}{3} \cdot \frac{1-x^6}{1-x^4}\right)\left(\frac{21}{8}x^2(1-x^4)\right) dx = \frac{7}{9}. \end{aligned}$$

Theorem (double expectations): $\mathrm{E}[\mathrm{E}(Y|X)] = \mathrm{E}[Y]$.

Proof (cts case): By the Unconscious Statistician,

$$\begin{aligned}
\mathrm{E}[\mathrm{E}(Y|X)] &= \int_{\mathbb{R}} \mathrm{E}(Y|x) f_X(x)\, dx \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y f(y|x)\, dy \right) f_X(x)\, dx \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} y f(y|x) f_X(x)\, dx\, dy \\
&= \int_{\mathbb{R}} y \int_{\mathbb{R}} f(x,y)\, dx\, dy \\
&= \int_{\mathbb{R}} y f_Y(y)\, dy = \mathrm{E}[Y]. \quad \square
\end{aligned}$$

Example: A cutesy way to calculate the mean of the Geometric distribution.

Let $Y \sim \mathrm{Geom}(p)$, e.g., Y could be the number of coin flips before H appears, where $P(\mathrm{H}) = p$. From Baby Probability class, we know that the pmf of Y is $f_Y(y) = P(Y = y) = q^{y-1}p$, for $y = 1, 2, \ldots$.

Then the old-fashioned way to calculate the mean is:

$$\mathrm{E}[Y] = \sum_y y f_Y(y) = \sum_{y=1}^{\infty} y q^{y-1} p = 1/p,$$

where the last step follows because I tell you so. $\square$

...Let's use double expectation to do what's called a "standard one-step conditioning argument". Define $X = 1$ if the first flip is H; and $X = 0$ otherwise.

Based on the result X of the first step, we have

$$\begin{aligned}
\mathrm{E}[Y] &= \mathrm{E}[\mathrm{E}(Y|X)] = \sum_x \mathrm{E}(Y|x) f_X(x) \\
&= \mathrm{E}(Y|X=0)P(X=0) + \mathrm{E}(Y|X=1)P(X=1) \\
&= (1 + \mathrm{E}[Y])(1-p) + 1(p). \quad \text{(why?)}
\end{aligned}$$

Solving, we get $\mathrm{E}[Y] = 1/p$ again! $\square$

Computing Probabilities by Conditioning

Let A be some event, and define the RV $Y = 1$ if A occurs; and $Y = 0$ otherwise. Then

$$\mathrm{E}[Y] = \sum_y y f_Y(y) = P(Y = 1) = P(A).$$

Similarly, for any RV X, we have

$$\mathrm{E}[Y|X = x] = \sum_y y f_Y(y|x) = P(Y = 1|X = x) = P(A|X = x).$$

Thus,

$$\begin{aligned} P(A) &= \mathrm{E}[Y] = \mathrm{E}[\mathrm{E}(Y|X)] \\ &= \int_{\mathbb{R}} \mathrm{E}[Y|X = x] dF_X(x) \\ &= \int_{\mathbb{R}} P(A|X = x) dF_X(x). \end{aligned}$$

Example/Theorem: If X and Y are independent cts RV's, then

$$P(Y < X) = \int_{\mathbb{R}} P(Y < x) f_X(x)\, dx.$$

Proof: Follows from above result if we let the event $A = \{Y < X\}$.
$\Box$

Example: If $X \sim \mathrm{Exp}(\mu)$ and $Y \sim \mathrm{Exp}(\lambda)$ are indep RV's, then

$$\begin{aligned} P(Y < X) &= \int_{\mathbb{R}} P(Y < x) f_X(x)\, dx \\ &= \int_0^\infty (1 - e^{-\lambda x}) \mu e^{-\mu x}\, dx \\ &= \frac{\lambda}{\lambda + \mu}. \quad \Box \end{aligned}$$

Theorem (variance decomposition):

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}[\mathrm{E}(Y|X)]$$

Proof (from Ross): By definition of variance and double expectation,

$$\begin{aligned} \mathrm{E}[\mathrm{Var}(Y|X)] &= \mathrm{E}\left[\mathrm{E}(Y^2|X) - \{\mathrm{E}(Y|X)\}^2\right] \\ &= \mathrm{E}(Y^2) - \mathrm{E}\left[\{\mathrm{E}(Y|X)\}^2\right]. \end{aligned}$$

Similarly,

$$\begin{aligned}\operatorname{Var}[\mathrm{E}(Y|X)] &= \mathrm{E}\left[\{\mathrm{E}(Y|X)\}^2\right] - \{\mathrm{E}[\mathrm{E}(Y|X)]\}^2 \\ &= \mathrm{E}\left[\{\mathrm{E}(Y|X)\}^2\right] - \{\mathrm{E}(Y)\}^2.\end{aligned}$$

Thus, putting the last two results together,

$$\mathrm{E}[\operatorname{Var}(Y|X)] + \operatorname{Var}[\mathrm{E}(Y|X)] = \mathrm{E}(Y^2) - \{\mathrm{E}(Y)\}^2 = \operatorname{Var}(Y). \quad \square$$

12.Covariance and Correlation

"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 $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$.

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 $\operatorname{Var}(\bar{X}_n) = \operatorname{Var}(X_i)/n$. Thus, the variance *decreases* as n increases. $\square$

But not all RV's are independent...

Definition: The *covariance* between X and Y is

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

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

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

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

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

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

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

$\mathrm{E}[X] = 2.8$, $\mathrm{Var}(X) = 0.66$,
$\mathrm{E}[Y] = 51$, $\mathrm{Var}(Y) = 69$,
$\mathrm{E}[XY] = \sum_x \sum_y xyf(x,y) = 140$,

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

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,

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

13.Probability Distribution

- Discreet Distribution
  - Bernoulli

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

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

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

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

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

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

  - Geometric

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

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

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

  - Neg Bernoulli

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

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

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

- Poisson

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

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 λ (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.

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

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

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

- Continuous Distribution

$X \sim \text{Uniform}(a, b)$. $f(x) = \frac{1}{b-a}$ 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 $\text{Exp}(\lambda)$ 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 $\Gamma(\alpha) \equiv \int_0^\infty t^{\alpha-1}e^{-t}\,dt$.

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

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

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

$X \sim \text{Triangular}(a, b, c)$. 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.

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

$\mathrm{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$, 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 $2X - 3Y + 1 \sim \text{Nor}(-5, 70)$. $\square$

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 μ well as n becomes large.

14.Limit Theorems

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

This is a special case of the *Law of Large Numbers*, which says that $\bar{X}_n$ approximates μ 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.

'Converges in distribution '

- **Central Limit Theorem**

Central Limit Theorem: If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$ with mean μ and variance σ^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.

15. Introduction to Estimation, Unbiasedness and MSE

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: μ, σ^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 θ. Then $T(\mathbf{X})$ is called a *point estimator* for θ.

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.

- Unbiasedness of an estimator

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

Example/Theorem: Suppose $X_1, \ldots, X_n$ are iid anything with mean μ. 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 μ. That's why $\bar{X}$ is the *sample mean*.

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

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

- Sample variance is an unbiased estimator of population variance

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

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

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

Baby Example: Suppose $X_1, \ldots, X_n$ are iid Exp(λ). 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 $E[X_1] = E[\bar{X}]$ and $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X_1)/n = \sigma^2/n$, we have

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

Remark: S is *biased* for the standard deviation σ.

For two unbiased estimators, we compare their variances.

-MSE and Relative Efficiency of estimators

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

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

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

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

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

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

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

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

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

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

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

16. Maximum Likelihood Estimation

Definition: Consider an iid random sample $X_1, \ldots, X_n$, where each X_i has pdf/pmf $f(x)$. Further, suppose that θ 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 θ is the value of θ that maximizes $L(\theta)$. The MLE is a function of the X_i's and is a RV.

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

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

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

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 λ that maximizes $L(\lambda)$ also maximizes $\ell\text{n}(L(\lambda))$!

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

$$\ell\text{n}(L(\lambda)) = \ell\text{n}\bigg(\lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} x_i\Big)\bigg) = n\ell\text{n}(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

This makes our job less horrible.

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

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

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

(2) At the end, we put a little $\widehat{\text{hat}}$ over λ 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.

- **Invariance Property of MLE**

Theorem (Invariance Property of MLE's): If $\hat{\theta}$ is the MLE of some parameter θ 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 λ 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.

17.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 χ^2 *distribution with k degrees of freedom (df)*. Notation: $Y \sim \chi^2(k)$. Note that $\text{E}[Y] = k$ and $\text{Var}(Y) = 2k$.

If $Z \sim \text{Nor}(0,1)$, $Y \sim \chi^2(k)$, 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: $F \sim F(m, n)$.

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

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

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

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

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

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

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

Example: A $100(1-\alpha)\%$ CI for σ^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.