Calculus, Probability, and Statistics Primers

Dave Goldsman

Georgia Institute of Technology, Atlanta, GA, USA

12/30/18

Outline

Calculus Primer

Goal: This section provides a brief review of various calculus tidbits that we'll be using later on.

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. $\square$

Definition If $f(x)$ is continuous, then it is *differentiable* (has a *derivative*) if

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

exists and is well-defined for any given x. Think of the derivative as the slope of the function.

Example Some well-known derivatives are:

$$[x^k]' = kx^{k-1},$$

$$[e^x]' = e^x,$$

$$[\sin(x)]' = \cos(x),$$

$$[\cos(x)]' = -\sin(x),$$

$$[\ell\mathrm{n}(x)]' = \frac{1}{x},$$

$$[\arctan(x)]' = \frac{1}{1+x^2}. \quad \square$$

Theorem Some well-known properties of derivatives are:

$$[af(x) + b]' = af'(x),$$

$$[f(x) + g(x)]' = f'(x) + g'(x),$$

$$[f(x)g(x)]' = f'(x)g(x) + f(x)g'(x) \quad \text{(product rule)},$$

$$\left[\frac{f(x)}{g(x)}\right]' = \frac{g(x)f'(x) - f(x)g'(x)}{g^2(x)} \quad \text{(quotient rule)}^{1},$$

$$[f(g(x))]' = f'(g(x))g'(x) \quad \text{(chain rule)}^{2}.$$

[1] Ho dee Hi minus Hi dee Ho over Ho Ho.

[2] www.youtube.com/watch?v=gGAiW5dOnKo

Example Suppose that $f(x) = x^2$ and $g(x) = \ell\text{n}(x)$. Then

$$\begin{aligned}
[f(x)g(x)]' &= \frac{d}{dx}x^2\ell\text{n}(x) \;=\; 2x\ell\text{n}(x) + x, \\
\left[\frac{f(x)}{g(x)}\right]' &= \frac{d}{dx}\frac{x^2}{\ell\text{n}(x)} \;=\; \frac{2x\ell\text{n}(x) - x}{\ell\text{n}^2(x)}, \\
[f(g(x))]' &= 2g(x)g'(x) \;=\; \frac{2\ell\text{n}(x)}{x}. \quad \square
\end{aligned}$$

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 $x = x_0$. Exception: Check the endpoints of your interval of interest as well.

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 the value of x that minimizes $f(x) = e^{2x} + e^{-x}$. The minimum can only occur when $f'(x) = 2e^{2x} - e^{-x} = 0$. After a little algebra, we find that this occurs at $x_0 = -(1/3)\ell\text{n}(2) \approx -0.231$. It's also easy to show that $f''(x) > 0$ for all x; and so x_0 yields a minimum. $\Box$

Finding Zeroes: Speaking of solving for a 0, how might you do it if a continuous function $g(x)$ is a complicated nonlinear fellow?

- 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: Suppose you can find x_1 and x_2 such that $g(x_1) < 0$ and $g(x_2) > 0$. (We'll follow similar logic if the inequalities are both reversed.) By the Intermediate Value Theorem (which you may remember), there must be a zero in $[x_1, x_2]$, that is, $x^\star \in [x_1, x_2]$ such that $g(x^\star) = 0$.

Thus, take $x_3 = (x_1 + x_2)/2$. If $g(x_3) < 0$, then there must be a zero in $[x_3, x_2]$. Otherwise, if $g(x_3) > 0$, then there must be a zero in $[x_1, x_3]$. In either case, you've reduced the length of the search interval.

Continue in this same manner until the length of the search interval is as small as desired.

Exercise: Try this out for $g(x) = x^2 - 2$, and come up with an approximation for $\sqrt{2}$.

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

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

Exercise: Try Newton out for $g(x) = x^2 - 2$, noting that the iteration step is to set

$$x_{i+1} \ = \ x_i - \frac{x_i^2 - 2}{2x_i} \ = \ \frac{x_i}{2} + \frac{1}{x_i}.$$

Let's start with a bad guess of $x_1 = 1$. Then

$$\begin{aligned} x_2 \ &= \ \frac{x_1}{2} + \frac{1}{x_1} \ = \ \frac{1}{2} + 1 \ = \ 1.5 \\ x_3 \ &= \ \frac{x_2}{2} + \frac{1}{x_2} \ \approx \ \frac{1.5}{2} + \frac{1}{1.5} \ = \ 1.4167 \\ x_4 \ &= \ \frac{x_3}{2} + \frac{1}{x_3} \ \approx \ 1.4142 \quad \text{Wow!} \quad \square \end{aligned}$$

Integration

Definition The function $F(x)$ having derivative $f(x)$ is called the *antiderivative* (or *indefinite integral*). It is denoted by $F(x) = \int f(x)\,dx$.

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]

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

[3] "I'm *really* an integral!"

Example Some well-known indefinite integrals are:

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

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

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

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

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

where C is an arbitrary constant. $\square$

Example It is easy to see that

$$\int \frac{d\,\text{cabin}}{\text{cabin}} \;=\; \ell\text{n}|\text{cabin}| + C \;=\; \text{houseboat.} \quad \square$$

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:

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

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

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

[4] www.youtube.com/watch?v=OTzLVIc-O5E
[5] www.youtube.com/watch?v=eswQl-hcvU0

Example Using integration by parts with $f(x) = x$ and $g'(x) = e^{2x}$ and the chain rule, we have

$$\int_0^1 xe^{2x}\,dx \;=\; \frac{xe^{2x}}{2}\Big|_0^1 - \int_0^1 \frac{e^{2x}}{2}\,dx \;=\; \frac{e^2}{2} - \frac{e^{2x}}{4}\Big|_0^1 = \frac{e^2+1}{4}. \quad \square$$

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 $a = 0$.

Example Here are some famous Maclaurin series.

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

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

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

Example And while we're at it, here are some miscellaneous sums that you should know.

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

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

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

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

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

Example L'Hôspital shows that

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

[6]This rule makes me sick.

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.

Try it out on $\int_0^1 \sin(\pi x/2)\,dx$ (which secretly equals $2/\pi$) for different values of n, and see for yourself.

Riemann (cont'd): Since I'm such a nice guy, I've made things easy for you. In this problem, I've thoughtfully taken $a = 0$ and $b = 1$, so that $\Delta x = 1/n$ and $x_i = i/n$, which simplifies the notation a bit. Then

$$\begin{aligned} \int_a^b f(x)\,dx &= \int_0^1 f(x)\,dx \\ &\approx \sum_{i=1}^{n} f(x_i)\Delta x \\ &= \frac{1}{n}\sum_{i=1}^{n} \sin\left(\frac{\pi i}{2n}\right). \end{aligned}$$

For $n = 100$, this calculates out to a value of 0.6416, which is pretty close to the true answer of $2/\pi \approx 0.6366$. $\square$

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

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

Again try it out on $\int_0^1 \sin(\pi x/2)\,dx$.

Computer Exercise, Monte Carlo version: You will soon learn a Monte Carlo method to accomplish approximate integration. Just take my word for it for now. Let $U_1, U_2, \ldots, U_n$ denote a sequence of Unif(0,1) random numbers, which can be obtained from Excel using `RAND()`. It can be shown that

$$\int_a^b f(x)\,dx \quad\approx\quad \frac{b-a}{n}\sum_{i=1}^{n} f(a+(b-a)U_i),$$

with the result becoming an equality as $n \to \infty$.

Yet again try it out on $\int_0^1 \sin(\pi x/2)\,dx$.

Outline

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: If $P(A \cap B) = P(A)P(B)$, then A and B are *independent* events.

Theorem: If A and B are independent, then $P(A|B) = P(A)$.

Example: Toss two dice. Let $A =$ "Sum is 7" and $B =$ "First die is 4". Then

$$P(A) = 1/6, \quad P(B) = 1/6, \quad \text{and}$$

$$P(A \cap B) = P((4,3)) = 1/36 = P(A)P(B).$$

So A and B are independent. $\square$

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,

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

Definition: If the set 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 $\sum_x f(x) = 1$.

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

$$f(x) = \begin{cases} 1/4 & \text{if } x = 0 \text{ or } 2 \\ 1/2 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \square$$

Examples: Here are some well-known discrete RV's that you may know: Bernoulli(p), Binomial(n, p), Geometric(p), Negative Binomial, Poisson(λ), etc.

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

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

Examples: Here are some well-known continuous RV's: Uniform(a, b), Exponential(λ), Normal(μ, σ^2), etc.

Notation: "$\sim$" means "is distributed as." For instance, $X \sim \mathrm{Unif}(0,1)$ means that X has the uniform distribution on [0,1].

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

Simulating Random Variables

We'll make a brief aside here to show how to simulate some very simple random variables.

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 $X = \lceil 7.3 \rceil = 8$. $\Box$

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	$f(x)$	$P(X \le x)$	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 $X = 4.2$. $\square$

Now we'll use the inverse transform method to generate a continuous random variable. We'll talk about the following result a little later…

Theorem: If X is a continuous random variable with cdf $F(x)$, then the random variable $F(X) \sim \text{Unif}(0,1)$.

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

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

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

Example (Generating Uniforms): All of the above RV generation examples relied on our ability to generate a Unif(0,1) RV. For now, let's assume that we can generate numbers that are "practically" iid Unif(0,1).

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

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

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

Here's an easy FORTRAN implementation of the above algorithm (from Bratley, Fox, and Schrage).

```
FUNCTION UNIF(IX)
K1 = IX/127773     (this division truncates, e.g., 5/3 = 1.)
IX = 16807*(IX - K1*127773) - K1*2836     (update seed)
IF(IX.LT.0)IX = IX + 2147483647
UNIF = IX * 4.656612875E-10
RETURN
END
```

In the above function, we input a positive integer IX and the function returns the PRN UNIF, as well as an updated IX that we can use again. □

Some Exercises: In the following, I'll assume that you can use Excel (or whatever) to simulate independent Unif(0,1) RV's. (We'll review independence in a little while.)

1. Make a histogram of $X_i = -\ell\text{n}(U_i)$, for $i = 1, 2, \ldots, 10000$, where the U_i's are independent Unif(0,1) RV's. What kind of distribution does it look like?
2. Suppose X_i and Y_i are independent Unif(0,1) RV's, $i = 1, 2, \ldots, 10000$. Let $Z_i = \sqrt{-2\ell\text{n}(X_i)}\,\sin(2\pi Y_i)$, and make a histogram of the Z_i's based on the 10000 replications.
3. Suppose X_i and Y_i are independent Unif(0,1) RV's, $i = 1, 2, \ldots, 10000$. Let $Z_i = X_i/(X_i - Y_i)$, and make a histogram of the Z_i's based on the 10000 replications. This may be somewhat interesting. It's possible to derive the distribution analytically, but it takes a lot of work.

Great Expectations

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 $\mathrm{E}[X] = \sum_x x f(x) = p$. $\square$

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

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)

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

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 $\ell\mathrm{n}(X)$.

Example: Suppose X is the following discrete RV:

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

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

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

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

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 $X \sim \text{Bern}(p)$. Recall that $\text{E}[X] = p$. Then

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

$$\text{Var}(X) \;=\; \text{E}[X^2] - (\text{E}[X])^2 \;=\; p(1-p). \quad \Box$$

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

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

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

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

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

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

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

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

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

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

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

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

Further,

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

Thus,

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

Moment generating functions have many other important uses, some of which we'll talk about in this course.

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

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,

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

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

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

Thus, the pdf of Y is $g(y) = G'(y) = 1$, $0 < y < 1$, indicating 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

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

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

This result is of fundamental importance when it comes to generating random variates during a simulation.

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

So the Inverse Transform Theorem implies that

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

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

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

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

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

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

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

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!

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

Jointly Distributed Random Variables

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

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

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

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

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

Then the marginal pdf's are:

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

and

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

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

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, $\int_{\mathbf{R}} f(y|x)\,dy = 1$, for any x.

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

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

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,

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

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 person who is 7 feet tall ($\mathrm{E}[Y|X=7]$) will probably be greater than that of a random person from the entire population ($\mathrm{E}[Y]$).

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

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

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

Proof (cts case): By the Unconscious Statistician,

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

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 $\mathrm{E}[Y|x]$. 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):

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

Notice that both answers are the same (good)! $\square$

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

Let $Y \sim \text{Geom}(p)$, e.g., Y could be the number of coin flips before H appears, where $P(\text{H}) = p$. 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:

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

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

But if you are not quite willing to believe me,. . .

…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 $A = \{Y < X\}$.
$\square$

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

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

Theorem (variance decomposition):

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

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

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

Similarly,

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

Thus,

$$\mathrm{E}\left[\mathrm{Var}(Y|X)\right] + \mathrm{Var}\left[\mathrm{E}(Y|X)\right] \;=\; \mathrm{E}(Y^2) - \{\mathrm{E}(Y)\}^2 \;=\; \mathrm{Var}(Y). \qquad \square$$

"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...

Covariance and Correlation

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: Suppose $X \sim \mathrm{Unif}(-1,1)$ and $Y = X^2$. Then X and Y are clearly dependent. However,

$$\mathrm{Cov}(X,Y) = \mathrm{E}[X^3]-\mathrm{E}[X]\mathrm{E}[X^2] = \mathrm{E}[X^3] = \int_{-1}^{1} \frac{x^3}{2}\,dx = 0. \quad \square$$

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

Theorem: Whether or not X and Y are independent,

$$\mathrm{Var}(X+Y) \;=\; \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$$

and

$$\mathrm{Var}(X-Y) \;=\; \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y).$$

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

$\mathrm{E}[X] = 2.8$, $\mathrm{Var}(X) = 0.66$, $\mathrm{E}[Y] = 51$, $\mathrm{Var}(Y) = 69$,

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

and

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

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

$$\begin{aligned} \mathrm{E}[P] &= w\mu_1 + (1-w)\mu_2 \\ \mathrm{Var}(P) &= w^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2w(1-w)\sigma_{12}. \end{aligned}$$

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.

Some Probability Distributions

First, some discrete distributions…

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

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

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

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

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

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

$X \sim \text{Geometric}(p)$ 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 rth success occurs. For example, "FFFSSFS" implies that $\text{NegBin}(3, p) = 7$.

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

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

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

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 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., $N(1)$.

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

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

Now, 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 exponential distribution has the *memoryless property* (and is the only continuous distribution with this property), i.e., for $s, t > 0$, $P(X > s + t | X > s) = P(X > t)$.

Example: Suppose $X \sim \text{Exp}(\lambda = 1/100)$. Then

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

$X \sim \text{Gamma}(\alpha, \lambda)$.

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

where the gamma function is

$$\Gamma(\alpha) \;\equiv\; \int_0^\infty t^{\alpha-1} e^{-t}\, dt.$$

$\text{E}[X] = \alpha/\lambda$, $\text{Var}(X) = \alpha/\lambda^2$, $M_X(t) = \Big[\lambda/(\lambda - t)\Big]^\alpha$ 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 modeling things with limited data — a is the smallest possible value, b is the "most likely," and c is the largest.

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{if } a < x \leq b \\ \frac{2(c-x)}{(c-b)(c-a)} & \text{if } b < x \leq c \\ 0 & \text{otherwise} \end{cases}.$$

$\mathrm{E}[X] = (a + b + c)/3$.

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

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

$X \sim \text{Normal}(\mu, \sigma^2)$. 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 $2X - 3Y + 1 \sim \text{Nor}(-5, 70)$. $\Box$

Limit Theorems

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.

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 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.

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

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

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

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

By the way, 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 $P(90 \leq \sum_{i=1}^{100} X_i \leq 110) = 0.6835$.

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 $U \sim \mathrm{Unif}(0, 1)$?

Outline

Intro to Estimation

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.

Unbiased Estimators

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 called the *sample mean*.

Baby Example: In particular, suppose $X_1, \ldots, X_n$ are iid Exp(λ). 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., $\mathrm{E}[1/\bar{X}] \neq 1/\mathrm{E}[\bar{X}] = \lambda$.

Example/Theorem: Suppose $X_1, \ldots, X_n$ are iid anything with mean μ and variance σ^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 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 $\mathrm{E}[X_1] = \mathrm{E}[\bar{X}]$ and $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X_1)/n = \sigma^2/n$, we have

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

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

Big Example: Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$, i.e., the pdf 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 $\text{E}[Y_1] = 2\text{E}[\bar{X}] = 2\text{E}[X_i] = \theta$, we see that Y_1 is unbiased for θ.

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 $M \equiv \max_i X_i$,

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

This implies that the pdf of M is

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

and this implies that

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

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

So both Y_1 and Y_2 are unbiased for θ, 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. $\square$

Mean Squared Error

Definition: The *bias* of $T(\mathbf{X})$ as an estimator of θ 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 $\text{MSE}(T_1)/\text{MSE}(T_2)$. 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.

Maximum Likelihood Estimators

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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. 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. Too tedious. Need a trick. . . .

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 $\ell\text{n}(L(\lambda))$!

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

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 $\mathrm{E}[X] = 1/\lambda$.

(2) At the end, we put a little $\widehat{\mathrm{hat}}$ 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's

Theorem (Invariance Property): If $\hat{\theta}$ is the MLE of some parameter θ and $h(\cdot)$ is a one-to-one function, then $h(\hat{\theta})$ is the MLE of $h(\theta)$.

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. We define the *survival function* as

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

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. $\square$

Distributional Results and Confidence Intervals

There are a number of distributions (including the normal) that come up in statistical sampling problems. Here are a few:

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 $\mathrm{E}[Y] = k$ 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 μ and σ^2 under a variety of assumptions.

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

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

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

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

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

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

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

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

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

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

Exercise: 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

$$\bar{X} = 9.8525 \quad \text{and} \quad S = 0.0965.$$

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

$$H = t_{\alpha/2,n-1}\sqrt{S^2/n} = \frac{(2.093)(0.0965)}{\sqrt{20}} = 0.0451.$$

Thus, the CI is $\mu \in \bar{X} \pm H$, or $9.8074 \le \mu \le 9.8976$. $\square$