

MATH/STAT 455: Mathematical Statistics

Taylor Okonek

2023-12-07

Table of contents

Welcome to Mathematical Statistics!	4
1 Probability: A Brief Review	5
1.1 Learning Objectives	5
1.2 Reading Guide	5
1.2.1 Reading Questions	5
1.3 Definitions	6
1.3.1 Distributions Table	10
1.4 Theorems	11
1.4.1 Transforming Continuous Random Variables	13
1.5 Worked Examples	15
2 Maximum Likelihood Estimation	23
2.1 Learning Objectives	26
2.2 Reading Guide	26
2.2.1 Reading Questions	27
2.3 Definitions	27
2.4 Theorems	28
2.5 Worked Examples	29
3 Method of Moments	33
3.1 Learning Objectives	34
3.2 Reading Guide	34
3.2.1 Reading Questions	34
3.3 Definitions	35
3.4 Theorems	35
3.5 Worked Examples	36
4 Properties of Estimators	38
4.1 Learning Objectives	39
4.2 Reading Guide	39
4.2.1 Reading Questions	40
4.3 Definitions	40
4.4 Theorems	42
4.5 Worked Examples	45

5	Consistency	50
6	Asymptotics & the Central Limit Theorem	51
7	Computational Optimization	52
8	Bayesian Inference	53
9	Decision Theory	54
10	Hypothesis Testing	55
	References	56

Welcome to Mathematical Statistics!

This book contains the course notes for *MATH/STAT 455: Mathematical Statistics* at Macalester College, as taught by Prof. [Taylor Okonek](#). These notes draw from course notes created by Prof. [Kelsey Grinde](#), and heavily from the course textbook, *An Introduction to Mathematical Statistics and Its Applications* by Richard Larsen and Morris Marx (6th Edition). Each chapter will contain (at a minimum):

1. Learning Objectives
2. Reading Guide
3. Definitions
4. Theorems
5. Worked Examples

I will be editing and adding to these notes throughout Spring 2024, so please check consistently for updates!

If you find any typos or have other questions, please email tokonek@macalester.edu.

1 Probability: A Brief Review

MATH/STAT 455 builds directly on topics covered in *MATH/STAT 354: Probability*. You're not expected to perfectly remember everything from *Probability*, but you will need to have sufficient facility with the following topics covered in this review Chapter in order to grasp the majority of concepts covered in *MATH/STAT 455*.

1.1 Learning Objectives

By the end of this chapter, you should be able to...

- Distinguish between important probability models (e.g., Normal, Binomial)
- Derive the expectation and variance of a single random variable or a sum of random variables
- Define the moment generating function and use it to find moments or identify pdfs

1.2 Reading Guide

Associated Readings: Chapters 2-4 (pages 15-277)

1.2.1 Reading Questions

1. Which probability distributions are appropriate for *quantitative* (continuous) random variables?
2. Which probability distributions are appropriate for *categorical* random variables?
3. *Independently and Identically Distributed (iid)* random variables are an incredibly important assumption involved in many statistical methods. Why do you think it might be important/useful for random variables to have this property?

1.3 Definitions

You are expected to know the following definitions:

Random Variable

A random variable is a function that takes inputs from a sample space of all possible outcomes, and outputs real values or probabilities. As an example, consider a coin flip. The sample space of all possible outcomes consists of “heads” and “tails”, and each outcome is associated with a probability (50% each, for a fair coin). For our purposes, you should know that random variables have probability density (or mass) functions, and are either discrete or continuous based on the number of possible outcomes a random variable may take. Random variables are often denoted with capital Roman letters, like X , Y , Z , etc.

Probability density function (discrete, continuous)

- Note: I don’t care if you call a pmf a pdf... I will probably do this continuously throughout the semester. We don’t need to be picky about this in *MATH/STAT 455*.

There are many different accepted ways to write the notation for a pdf of a random variables. Any of the following are perfectly appropriate for this class: $f(x)$, $\pi(x)$, $p(x)$, $f_X(x)$. I typically use either π or p , but might mix it up occasionally.

Key things I want you to know about probability density functions:

- $\pi(x) \geq 0$, everywhere. This should make sense (hopefully) because probabilities cannot be negative!
- $\int_{-\infty}^{\infty} \pi(x) = 1$. This should also (hopefully) makes sense. Probabilities can’t be *greater* than one, and the probability of event occurring *at all (ever)* should be equal to one, if the event x is a random variable.

Cumulative distribution function (discrete, continuous)

Cumulative distribution functions we’ll typically write as $F_X(x)$. or $F(x)$, for short. It is important to know that

$$F_X(x) = \Pr(X \leq x),$$

or in words, “the cumulative distribution function is the probability that a random variable lies before x .” If you write $\Pr(X < x)$ instead of \leq , you’re fine. The probability that a random variable is exactly one number (for an RV with a continuous pdf) is zero anyway, so these are the same thing. Key things I want you to know about cumulative distribution functions:

- $F(x)$ is non-decreasing. This is in part where the “cumulative” piece comes in to play. Recall that probabilities are basically integrals or sums. If we’re integrating over something positive, and our upper bound for our integral *increases*, the area under the curve (cumulative probability) will increase as well.
- $0 \leq F(x) \leq 1$ (since probabilities have to be between zero and one!)
- $\Pr(a < X \leq b) = F(b) - F(a)$ (because algebra)

Joint probability density function

A joint probability density function is a probability distribution defined for more than one random variable at a time. For two random variables, X and Z , we could write their joint density function as $f_{X,Z}(x, z)$, or $f(x, z)$ for short. The joint density function encodes all sorts of fun information, including *marginal* distributions for X and Z , and conditional distributions (see next **bold** definition). We can think of the joint pdf as listing all possible pairs of outputs from the density function $f(x, z)$, for varying values of x and z . Key things I want you to know about joint pdfs:

- How to get a marginal pdf from a joint pdf:

Suppose I want to know $f_X(x)$, and I know $f_{X,Z}(x, z)$. Then I can integrate or “average over” Z to get

$$f_X(x) = \int f_{X,Z}(x, z) dz$$

- The relationship between conditional pdfs, marginal pdfs, joint pdfs, and Bayes’ theorem/rule
- How to obtain a joint pdf for *independent* random variables: just multiply their marginal pdfs together! This is how we will (typically) think about likelihoods!
- How to obtain a marginal pdf from a joint pdf when random variables are independent *without integrating* (think, “separability”)

Conditional probability density function

A conditional pdf denotes the probability distribution for a (set of) random variable(s), *given that* the value for another (set of) random variable(s) is known. For two random variables, X and Z , we could write the conditional distribution of X “given” Z as $f_{X|Z}(x | z)$, where the “conditioning” is denoted by a vertical bar (in LaTeX, this is typeset using “\mid”). Key things I want you to know about conditional pdfs:

- The relationship between conditional pdfs, marginal pdfs, joint pdfs, and Bayes’ theorem/rule
- How to obtain a conditional pdf from a joint pdf (again, think Bayes’ rule)

- Relationship between conditional pdfs and independence (see next **bold** definition)

Independence

Two random variables X and Z are *independent* if and only if:

- $f_{X,Z}(x, z) = f_X(x)f_Z(z)$ (their joint pdf is “separable”)
- $f_{X|Z}(x | z) = f_X(x)$ (the pdf for X does not depend on Z in any way)

Note that the “opposite” is also true: $f_{Z|X}(z | x) = f_Z(z)$

In notation, we denote that two variables are independent as $X \perp\!\!\!\perp Z$, or $X \perp Z$. In LaTeX, the *latter* is typeset as “\perp”, and the former is typeset as “\perp\!\!\!\perp”. As a matter of personal preference, I (Taylor) prefer $\perp\!\!\!\perp$, but I don’t like typing it out every time. Consider using the “\newcommand” functionality in LaTeX to create a shorthand for this for your documents!

Expected Value / Expectation

The expectation (or expected value) of a random variable is defined as:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

Expected value is a weighted average, where the average is over all possible values a random variable can take, weighted by the probability that those values occur. Key things I want you to know about expectation:

- The relationship between expectation, variance, and moments (specifically, that $E[X]$ is the 1st moment!)
- The “law of the unconscious statistician” (see the Theorems section of this chapter)
- Expectation of linear transformations of random variables (see **Theorems** section of this chapter)

Variance

The variance of a random variable is defined as:

$$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

In words, we can read this as “the expected value of the squared deviation from the mean” of a random variable X . Key things I want you to know about variance:

- The relationship between expectation, variance, and moments (hopefully clear, given the formula for variance)

- The relationship between variance and standard deviation: $Var(X) = sd(X)^2$
- The relationship between variance and covariance: $Var(X) = Cov(X, X)$
- $Var(X) \geq 0$. This should make sense, given that we're taking the expectation of something "squared" in order to calculate it!
- $Var(c) = 0$ for any constant, c .
- Variance of linear transformations of random variables (see **Theorems** section of this chapter)

r^{th} moment

The r^{th} moment of a probability distribution is given by $E[X^r]$. For example, when $r = 1$, the r^{th} moment is just the expectation of the random variable X . Key things I want you to know about moments:

- The relationship between moments, expectation, and variance
 - For example, if you know the first and second moments of a distribution, you should be able to calculate the variance of a random variable with that distribution!
- The relationship between moments and *moment generating functions* (see **Theorems** section of this chapter)

Covariance

The covariance of two random variables is a measure of their *joint* variability. We denote the covariance of two random variables X and Z as $Cov(X, Z)$, and

$$Cov(X, Z) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Some things I want you to know about covariance:

- $Cov(X, X) = Var(X)$
- $Cov(X, Y) = Cov(Y, X)$ (order doesn't matter)

Moment Generating Function (MGF)

The moment generating function of a random variable X is defined as

$$M_X(t) = E[e^{tX}]$$

A few things to note:

- $M_X(0) = 1$, always.

- If two random variables have the same MGF, they have the same probability distribution!
- MGFs are sometimes useful for showing how different random variables are related to each other

1.3.1 Distributions Table

You are also expected to know the probability distributions contained in Table 1, below. Note that you *do not* need to memorize the pdfs for these distributions, but you *should* be familiar with what types of random variables (continuous/quantitative, categorical, integer-valued, etc.) may take on different distributions. The more familiar you are with the forms of the pdfs, the easier/faster it will be to work through problem sets and quizzes.

Table 1.1: *Table 1.* Table of main probability distributions we will work with for *MATH/STAT 455*.

Distribution	PDF/PMF	Parameters	Support
Uniform	$\pi(x) = \frac{1}{\beta - \alpha}$	$\alpha \in \mathbb{R}, \beta \in \mathbb{R}$	$x \in [\alpha, \beta]$
Normal	$\pi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x - \mu)^2)$	$\mu \in \mathbb{R}, \sigma > 0$	$x \in \mathbb{R}$
Multivariate Normal	$\pi(\mathbf{x}) = (2\pi)^{-k/2} \Sigma ^{-1/2} \exp(-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu))$	$\mu \in \mathbb{R}^k, \Sigma \in \mathbb{R}^{k \times k}$, positive semi-definite (in practice, almost always positive definite)	$x \in \mathbb{R}^k$
Gamma	$\pi(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	α (shape), β (rate) > 0	$x \in (0, \infty)$
Chi-squared	$\pi(x) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}$	$\nu > 0$	$x \in [0, \infty)$
Exponential	$\pi(x) = \beta e^{-\beta x}$	$\beta > 0$	$x \in [0, \infty)$
Student-\$t\$	$\pi(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}} (1 + \frac{x^2}{\nu})^{-(\nu+1)/2}$	$\nu > 0$	$x \in \mathbb{R}$
Beta	$\pi(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$\alpha, \beta > 0$	$x \in [0, 1]$
Poisson	$\pi(x) = \frac{\lambda^x e^{-\lambda}}{x!}$	$\lambda > 0$	$x \in \mathbb{N}$
Binomial	$\pi(x) = \binom{n}{x} p^x (1-p)^{n-x}$	$p \in [0, 1], n = \{0, 1, 2, \dots\}$	$x \in \{0, 1, \dots, n\}$
Multinomial	$\pi(\mathbf{x}) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$	$p_i > 0$, $p_1 + \dots + p_k = 1$, $n = \{0, 1, 2, \dots\}$	$\{x_1, \dots, x_k \mid \sum_{i=1}^k x_i = n, x_i \geq 0 (i = 1, \dots, k)\}$

Distribution	PDF/PMF	Parameters	Support
Negative Binomial	$\pi(x) = \binom{x+r-1}{x}(1-p)^x p^r$	$r > 0, p \in [0, 1]$	$x \in \{0, 1, \dots\}$

1.4 Theorems

- Law of Total Probability

$$P(A) = \sum_n P(A \cap B_n),$$

or

$$P(A) = \sum_n P(A | B_n)P(B_n)$$

- Bayes' Theorem

$$\pi(A | B) = \frac{\pi(B | A)\pi(A)}{\pi(B)}$$

- Relationship between pdf and cdf

$$F_Y(y) = \int_{-\infty}^y f_Y(t)dt$$

$$\frac{\partial}{\partial y} F_Y(y) = f_Y(y)$$

- Expectation of random variables

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx$$

– “Law of the Unconscious Statistician”

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

- Expectation and variance of linear transformations of random variables

$$E[cX + b] = cE[X] + b$$

$$Var[cX + b] = c^2 Var[X]$$

- Relationship between mean and variance

$$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

Also, recall that $Cov[X, X] = Var[X]$.

- Finding a marginal pdf from a joint pdf

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$$

- Independence of random variables and joint pdfs

If two random variables are independent, their joint pdf will be *separable*. For example, if X and Y are independent, we could write

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

- Expected value of a product of independent random variables

Suppose random variables X_1, \dots, X_n are independent. Then we can write,

$$E \left[\prod_{i=1}^n X_i \right] = \prod_{i=1}^n E[X_i]$$

- Covariance of independent random variables

If X and Y are independent, then $Cov(X, Y) = 0$. We can show this by noting that

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] \tag{1.1}$$

$$= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \tag{1.2}$$

$$= E[XY] - E[XE[Y]] - E[YE[X]] + E[X]E[Y] \tag{1.3}$$

$$= 2E[X]E[Y] - 2E[X]E[Y] \tag{1.4}$$

$$= 0 \tag{1.5}$$

- Using MGFs to find moments

Recall that the moment generating function of a random variable X , denoted by $M_X(t)$ is

$$M_X(t) = E[e^{tX}]$$

Then the n th moment of the probability distribution for X , $E[X^n]$, is given by

$$\left. \frac{\partial M_X}{\partial t^n} \right|_{t=0}$$

where the above reads as “the n th derivative of the moment generating function, evaluated at $t = 0$.”

- Using MGFs to identify pdfs

MGFs uniquely identify probability density functions. If X and Y are two random variables where for all values of t , $M_X(t) = M_Y(t)$, then $F_X(x) = F_Y(y)$.

- Central Limit Theorem

The classical CLT states that for independent and identically distributed (iid) random variables X_1, \dots, X_n , with expected value $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$, the sample average (centered and standardized) converges in distribution to a standard normal distribution at a root- n rate. Notationally, this is written as

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2)$$




A fun aside: this is only *one* CLT, often referred to as the Levy CLT. There are other CLTs, such as the Lyapunov CLT and Lindeberg-Feller CLT!

1.4.1 Transforming Continuous Random Variables

We will *often* take at face value previously proven *relationships* between random variables. What I mean by this, as an example, is that it is a nice (convenient) fact that a sum of two independent normal random variables is *still* normally distributed, with a nice form for the mean and variance. In particular, if $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\theta, \nu^2)$, then $X + Y \sim N(\mu + \theta, \sigma^2 + \nu^2)$. Most frequently used examples of these sorts of relationships can be found in the “Related Distributions” section of the Wikipedia page for a given probability distribution. Unless I explicitly ask you to derive/show how certain variables are related to each other, you can just state the known relationship, use it, and move on!

If I *do* ask you to derive/show these things, there are a few different ways we can go about this. For this course, I only expect you to know the “CDF method” for *one function of one random variable*, as we’ll demonstrate below.

Theorem. Let X be a continuous random variable with pdf $f_X(x)$. Define a new random variable $Y = g(X)$, for nice* functions g . Then $f_Y(y) = f_X(g^{-1}(y)) \times \frac{1}{g'(g^{-1}(y))}$.

 *By *nice* functions we mean functions that are strictly increasing and smooth *on the required range*. As an example, $\exp(x)$ is a smooth, strictly increasing function; $|x|$ is not on the *whole real line*, but *is* from $(0, \infty)$ (where a lot of useful pdfs are defined). For the purposes of this class, every function that you will need to do this for will be “nice.” Note that there are also considerations that need to be taken regarding the *range* of continuous random variables when considering transforming them. We will mostly ignore these considerations in this class, but a technically complete derivation (or proof) must consider them.

Proof. We can write

$$\begin{aligned} f_Y(y) &= \frac{\partial}{\partial y} F_Y(y) \\ &= \frac{\partial}{\partial y} \Pr(Y \leq y) \\ &= \frac{\partial}{\partial y} \Pr(g(X) \leq y) \\ &= \frac{\partial}{\partial y} \Pr(X \leq g^{-1}(y)) \\ &= \frac{\partial}{\partial y} F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \times \frac{\partial}{\partial y} g^{-1}(y) \end{aligned}$$

where to obtain the last equality we use chain rule! Now we require some statistical trickery to continue... (note that this method is called the “CDF method” because we go *through* the CDF to derive the distribution for Y)

You will *especially* see this in the Bayes chapter of our course notes, but it is often true that our lives are made easier as statisticians if we multiply things by one, or add zero. What exactly do I mean? Rearranging gross looking formulas into things we are familiar with (like pdfs, for example) often makes our lives easier and allows us to avoid dealing with such grossness. Here, the grossness is less obvious, but nonetheless relevant. Note that we can write

$$\begin{aligned}
y &= y \\
y &= g(g^{-1}(y)) \\
\frac{\partial}{\partial y} y &= \frac{\partial}{\partial y} g(g^{-1}(y)) \\
1 &= g'(g^{-1}(y)) \frac{\partial}{\partial y} g^{-1}(y) \quad (\text{chain rule again!}) \\
\frac{1}{g'(g^{-1}(y))} &= \frac{\partial}{\partial y} g^{-1}(y)
\end{aligned}$$

The right-hand side should look familiar: it is exactly what we needed to “deal with” in our proof! Returning to that proof, we have

$$\begin{aligned}
f_Y(y) &= f_X(g^{-1}(y)) \times \frac{\partial}{\partial y} g^{-1}(y) \\
&= f_X(g^{-1}(y)) \times \frac{1}{g'(g^{-1}(y))}
\end{aligned}$$

as desired.

1.5 Worked Examples

Problem 1: Suppose $X \sim \text{Exponential}(\lambda)$. Calculate $E[X]$ and $\text{Var}[X]$.

We know that $f(x) = \lambda e^{-\lambda x}$. If we can calculate $E[X]$ and $E[X^2]$, then we’re basically done! We can write

$$\begin{aligned}
E[X] &= \int_0^{\infty} x \lambda e^{-\lambda x} dx \\
&= \lambda \int_0^{\infty} x e^{-\lambda x} dx
\end{aligned}$$

And now we need integration by parts! Set $u = x$, $dv = e^{-\lambda x} dx$. Then $du = 1 dx$ and $v = \frac{-1}{\lambda} e^{-\lambda x}$. Since $\int u dv = uv - \int v du$, we can continue

$$\begin{aligned}
E[X] &= \lambda \int_0^{\infty} x e^{-\lambda x} dx \\
&= \lambda \left(-\frac{x}{\lambda} e^{-\lambda x} \Big|_0^{\infty} - \int_0^{\infty} \frac{-1}{\lambda} e^{-\lambda x} dx \right) \\
&= \lambda \left(- \int_0^{\infty} \frac{-1}{\lambda} e^{-\lambda x} dx \right) \\
&= \lambda \left(\frac{-1}{\lambda^2} e^{-\lambda x} \Big|_0^{\infty} \right) \\
&= \frac{-1}{\lambda} e^{-\lambda x} \Big|_0^{\infty} \\
&= \frac{1}{\lambda} e^{-0} \\
&= \frac{1}{\lambda}
\end{aligned}$$

We can follow a similar process to get $E[X^2]$ (using the law of the unconscious statistician!). We can write

$$\begin{aligned}
E[X^2] &= \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx \\
&= \lambda \int_0^{\infty} x^2 e^{-\lambda x} dx
\end{aligned}$$

And now we need integration by parts again! Set $u = x^2$, $dv = e^{-\lambda x} dx$. Then $du = 2x dx$ and $v = \frac{-1}{\lambda} e^{-\lambda x}$. Since $\int u dv = uv - \int v du$, we can continue

$$\begin{aligned}
E[X] &= \lambda \int_0^{\infty} x^2 e^{-\lambda x} dx \\
&= \lambda \left(-\frac{x^2}{\lambda} e^{-\lambda x} \Big|_0^{\infty} - \int_0^{\infty} \frac{-2}{\lambda} x e^{-\lambda x} dx \right) \\
&= \lambda \left(-\frac{x^2}{\lambda} e^{-\lambda x} \Big|_0^{\infty} + \frac{2}{\lambda} \int_0^{\infty} x e^{-\lambda x} dx \right) \\
&= \lambda \left(-\frac{x^2}{\lambda} e^{-\lambda x} \Big|_0^{\infty} + \frac{2}{\lambda^3} \right) \\
&= \lambda \left(0 + \frac{2}{\lambda^3} \right) \\
&= \frac{2}{\lambda^2}
\end{aligned}$$

Now we can calculate $Var[X] = E[X^2] - E[X]^2$ as

$$Var[X] = E[X^2] - E[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

And so we have $E[X] = \frac{1}{\lambda}$ and $Var[X] = \frac{1}{\lambda^2}$.

Problem 2: Show that an exponentially distributed random variable is “memoryless”, i.e. show that $\Pr(X > s + x \mid X > s) = \Pr(X > x)$, $\forall s$.

Recall that the CDF of an exponential distribution is given by $F(x) = 1 - e^{-\lambda x}$. Thanks to Bayes rule, we can write

$$\begin{aligned}
\Pr(X > s + x \mid X > s) &= \frac{\Pr(X > s + x, X > s)}{\Pr(X > s)} \\
&= \frac{\Pr(X > s + x)}{\Pr(X > s)} \\
&= \frac{1 - \Pr(X \leq s + x)}{1 - \Pr(X \leq s)} \\
&= \frac{1 - F(s + x)}{1 - F(s)}
\end{aligned}$$

where the second equality is true because $x > 0$. Then we can write

$$\begin{aligned}
\Pr(X > s + x \mid X > s) &= \frac{1 - F(s + x)}{1 - F(s)} \\
&= \frac{1 - (1 - e^{-\lambda(s+x)})}{1 - (1 - e^{-\lambda s})} \\
&= \frac{e^{-\lambda(s+x)}}{e^{-\lambda s}} \\
&= \frac{e^{-\lambda s - \lambda x}}{e^{-\lambda s}} \\
&= e^{-\lambda x} \\
&= 1 - F(x) \\
&= \Pr(X > x)
\end{aligned}$$

and we're done!

Problem 3: Suppose $X \sim \text{Exponential}(1/\lambda)$, and $Y \mid X \sim \text{Poisson}(X)$. Show that $Y \sim \text{Geometric}(1/(1 + \lambda))$.

Note that we can write $f(x, y) = f(y \mid x)f(x)$, and $f(y) = \int f(x, y)dx$. Then

$$f(x, y) = \left(\frac{1}{\lambda}e^{-x/\lambda}\right) \left(\frac{x^y e^{-x}}{y!}\right)$$

And so,

$$\begin{aligned}
f(y) &= \int f(x, y)dx \\
&= \int \left(\frac{1}{\lambda}e^{-x/\lambda}\right) \left(\frac{x^y e^{-x}}{y!}\right) dx \\
&= \frac{1}{\lambda y!} \int x^y e^{-x(1+\lambda)/\lambda} dx
\end{aligned}$$

And we can again use integration by parts! Let $u = x^y$ and $dv = e^{-x(1+\lambda)/\lambda} dx$. Then we have $du = yx^{y-1}dx$ and $v = -\frac{\lambda}{1+\lambda}e^{-x(1+\lambda)/\lambda}$, and we can write

$$\begin{aligned}
f(y) &= \frac{1}{\lambda y!} \int x^y e^{-x(1+\lambda)/\lambda} dx \\
&= \frac{1}{\lambda y!} \left(x^y \frac{\lambda}{1+\lambda} e^{-x(1+\lambda)/\lambda} \Big|_{x=0}^{x=\infty} + \int \frac{\lambda}{1+\lambda} e^{-x(1+\lambda)/\lambda} y x^{y-1} dx \right) \\
&= \frac{1}{\lambda y!} \left(\int \frac{\lambda}{1+\lambda} e^{-x(1+\lambda)/\lambda} y x^{y-1} dx \right) \\
&= \frac{1}{\lambda y!} \left(\frac{\lambda}{1+\lambda} \right) y \left(\int e^{-x(1+\lambda)/\lambda} x^{y-1} dx \right)
\end{aligned}$$

This *looks* gross, but it's actually not so bad. Note that, since Y is Poisson, it can only take integer values beginning at 1! Then we can *repeat* the process of integration by parts *y times* in order to get rid of x^{y-1} term on the inside of the integral. Specifically, each time we do this process we will pull out a $(\frac{\lambda}{1+\lambda})$, and a $y-i$ for the i th integration by parts step (try this one or two steps for yourself to see how it will simplify if you find this unintuitive!). We end up with,

$$\begin{aligned}
f(y) &= \frac{1}{\lambda y!} \left(\frac{\lambda}{1+\lambda} \right)^y y! \\
&= \frac{1}{\lambda} \left(\frac{\lambda}{1+\lambda} \right)^y
\end{aligned}$$

Now let $p = \frac{1}{1+\lambda}$. If we can show that $f(y) \sim \text{Geometric}(p)$ then we're done. Note that $1-p = \lambda/(1+\lambda)$. We have

$$\begin{aligned}
f(y) &= \frac{1}{\lambda} (1-p)^y \\
&= \frac{1}{\lambda} (1-p)^{y-1} (1-p) \\
&= (1-p)^{y-1} \frac{1}{\lambda} \left(\frac{\lambda}{1+\lambda} \right) \\
&= (1-p)^{y-1} \left(\frac{1}{1+\lambda} \right) \\
&= (1-p)^{y-1} p
\end{aligned}$$

which is exactly the pdf of a geometric random variable with parameter p and trials that begin at 1.

An *alternative* solution (which perhaps embodies the phrase “work smarter, not harder”) actually doesn't involve integration by parts at all! As statisticians, we typically like to avoid

actually integrating anything whenever possible, and this is often achieved by manipulating algebra enough to essentially “create” a pdf out of what we see (since pdfs integrate to 1!). Massive props to a student for solving this problem in a much “easier” way, answer below:

$$\begin{aligned}
f(y) &= \int_0^\infty f(y | x) f(x) dx \\
&= \int_0^\infty \left(\frac{1}{\lambda} e^{-\frac{x}{\lambda}}\right) \left(\frac{x^y}{y!} e^{-x}\right) dx \\
&= \frac{1}{\lambda y!} \int_0^\infty x^y e^{-\frac{x}{\lambda}(1+\lambda)} dx \\
&= \frac{1}{\lambda y!} \int_0^\infty \frac{(\frac{1+\lambda}{\lambda})^{y+1}}{(\frac{1+\lambda}{\lambda})^{y+1}} \frac{\Gamma(y+1)}{\Gamma(y+1)} x^{(y+1)-1} e^{-\frac{x}{\lambda}(1+\lambda)} dx \\
&= \frac{\Gamma(y+1)}{\lambda y! (\frac{1+\lambda}{\lambda})^{y+1}} \int_0^\infty \frac{(\frac{1+\lambda}{\lambda})^{y+1}}{\Gamma(y+1)} x^{(y+1)-1} e^{-\frac{x}{\lambda}(1+\lambda)} dx \\
&= \frac{\Gamma(y+1)}{\lambda y! (\frac{1+\lambda}{\lambda})^{y+1}} (1) \\
&= \frac{y!}{\lambda y! (\frac{1+\lambda}{\lambda})^{y+1}} \\
&= \frac{\lambda^{-1}}{(\frac{1+\lambda}{\lambda})^{y+1}} \\
&= \frac{\lambda^y}{(1+\lambda)^{y+1}} \\
&= \frac{1}{(1+\lambda)} \frac{\lambda^y}{(1+\lambda)^y} \\
&= \frac{1}{(1+\lambda)} \left(1 - \frac{1}{(1+\lambda)}\right)^y \\
&= p(1-p)^y \quad \left(\text{where } p = \frac{1}{1+\lambda}\right)
\end{aligned}$$

Note that we arrive at a slightly different answer with this approach. Specifically, we arrive at the pdf of a geometric random variable with parameter p and trials that begin at 0, as opposed to 1. There’s some subtlety here that we’re going to choose to ignore.

Problem 4: Suppose that $X \sim N(\mu, \sigma^2)$, and let $Y = \frac{X-\mu}{\sigma}$. Find the distribution of Y (simplifying all of your math will be useful for this problem).

To solve this problem, we can use the theorem on transforming continuous random variables. We must first define our function g that relates X and Y . In this case, we have $g(a) = \frac{a-\mu}{\sigma}$. Now all we need to do is collect the mathematical “pieces” we need to use theorem: $g^{-1}(\frac{a}{\sigma})$, and $g'(a)$, and finally, the pdf of a normal random variable. We have

$$\begin{aligned}
f_X(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \\
g^{-1}(a) &= \sigma a + \mu \\
g'(a) &= \frac{\partial}{\partial a} \left(\frac{a - \mu}{\sigma} \right) = \frac{1}{\sigma}
\end{aligned}$$

Putting it all together, we have

$$\begin{aligned}
f_Y(y) &= f_X(g^{-1}(y)) \times \frac{1}{g'(g^{-1}(y))} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\sigma y + \mu - \mu)^2\right) \times \sigma \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(\sigma y)^2\right) \times \sigma \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\sigma^2 y^2\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)
\end{aligned}$$

and note that this is the pdf of a normally distributed random variable with mean 0 and variance 1! Thus, we have shown that $\frac{X-\mu}{\sigma} \sim N(0,1)$. **Fun Fact:** If this random variable reminds you of a Z-score, *it should!*

Problem 5: Suppose the joint pdf of two random variables X and Y is given by $f_{X,Y}(x,y) = \lambda\beta e^{-x\lambda-y\beta}$. Determine if X and Y are independent, showing why or why not.

To determine whether X and Y are independent (or not), we need to determine if their joint pdf is “separable.” Doing some algebra, we can see that

$$\begin{aligned}
f_{X,Y}(x,y) &= \lambda\beta e^{-x\lambda-y\beta} \\
&= \lambda\beta e^{-x\lambda} e^{-y\beta} \\
&= (\lambda e^{-x\lambda}) (\beta e^{-y\beta})
\end{aligned}$$

and so since we can write the joint distribution as a function of X multiplied by a function of Y , X and Y are independent (and in this case, both have exponential distributions).

Problem 6: Suppose the joint pdf of two random variables X and Y is given by $f_{X,Y}(x,y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{n}{y} x^{y+\alpha-1} (1-x)^{n-y+\beta-1}$. Determine if X and Y are independent, showing why or why not.

To determine whether X and Y are independent (or not), we need to determine if their joint pdf is “separable.” Right away, we should note that a piece of the pdf contains x^y , and therefore we are *never* going to be able to fully separate out this joint pdf into a function of x times a function y . Therefore, X and Y are *not* independent. In this case, we actually have $X \sim \text{Beta}(\alpha, \beta)$, and $Y | X \sim \text{Binomial}(n, y)$ (we’ll return to this example in the Bayes chapter!).

2 Maximum Likelihood Estimation

In *Probability*, you calculated probabilities of events by assuming a probability model for data and then *assuming you knew the value of the parameters* in that model. In *Mathematical Statistics*, we will similarly write down a probability model but then we will use observed data to *estimate the value of the parameters* in that model.

There is more than one technique that you can use to estimate the value of an unknown parameter. You're already familiar with one technique—**least squares estimation**—from *STAT 155*. We'll review the ideas behind that approach later in the course. To start, we'll explore two other widely used estimation techniques: **maximum likelihood estimation** (this chapter) and the **method of moments** (next chapter).

Introduction to MLE

To understand maximum likelihood estimation, we can first break down each individual word in that phrase: (1) maximum, (2) likelihood, (3) estimation. We'll start in reverse order.

Recall from your introductory statistics course that we are (often) interested in estimating *true, unknown parameters* in statistics, using some data. Our best guess at the truth, based on the data we observe / sample that we have, is an *estimate* of the truth (given some modeling assumptions). This is all the “estimation” piece is getting at here. We're going to be learning about a method that produces estimates!

The likelihood piece may be less familiar to you. A likelihood is essentially a fancy form of a function (see the Definitions section for an *exact* definition), that combines an assumed probability distribution for your data, with some unknown parameters.* The key here is that a likelihood is a *function*. It may *look* more complicated than a function like $y = mx + b$, but we can often manipulate them in a similar fashion, which comes in handy when trying to find the...

Maximum! We've maximized functions before, and we can do it again! There are ways to maximize functions numerically (using certain algorithms, such as Newton-Raphson for example, which we'll cover in a later chapter), but we will primarily focus on maximizing likelihoods *analytically* in this course to help us build intuition.

Recall from calculus: To maximize a function we...

1. Take the derivative of the function

2. Set the derivative equal to zero
3. Solve!
4. (double check that the second derivative is negative, so that it's actually a maximum as opposed to a minimum)
5. (also check the endpoints)

The last two steps we'll often skip in this class, since things have a tendency to work out nicely with most likelihood functions. If we are trying to maximize a likelihood with *multiple* parameters, there are a few different ways we can go about this. One way (which is nice for distributions like the multivariate normal) is to place all of the parameters in a vector, write the distribution in terms of matrices and vectors, and then use matrix algebra to obtain all of the MLEs for each parameter at once! An alternative way is to take *partial* derivatives of the likelihood function with respect to each parameter, and solve a *system* of equations to obtain MLEs for each parameter. We'll see an example of this in Problem Set 1 as well as Worked Example 2!

One final thing to note (before checking out worked examples and making sure you have a grasp on definitions and theorems) is that it is often *easier* to maximize the *log-likelihood* as opposed to the likelihood... un-logged. This is for a variety of reasons, one of which is that many common probability density functions contain some sort of e^x term, and logging (*natural* logging) simplifies that for us. Another one is that log rules sometimes make taking derivatives easier. The value of a parameter that maximizes the log-likelihood is the same value that maximizes the likelihood, un-logged (since log is a monotone, increasing function). This is truly just a convenience thing!

When maximizing the “usual” way doesn’t work...

To maximize a function what I'm calling the “usual” way involves the five steps listed above. Unfortunately, sometimes this doesn't work. We typically recognize that the process won't work once we get to step 3, and realize that “solving” ends up giving us an MLE that doesn't depend at all on our data. When this happens, it's usually because the MLE is an *order statistic* (see Definitions section of this chapter), and usually because the distribution of our random variable has a range that depends on our unknown parameter. An example of this (that will appear on your homework) occurs when $X_1, \dots, X_n \sim \text{Uniform}(0, \theta)$. In this case, the range of X_i depends directly on θ , since it cannot be any *greater than* θ .

In these cases, the process of finding the MLE for our unknown parameter usually involves plotting the likelihood as a function of the unknown parameter. We then look at where that function achieves its maximum (usually at one of the endpoints), and determine which observation (again, typically the minimum or maximum) will maximize our likelihood. An example of this can be found in Example 5.2.4 in our course textbook.

Maximum Likelihood: Does it make sense? Is it even “good”?

Let’s think for a minute about why maximum likelihood, as a procedure for producing estimates of parameters, might make sense. Given a distributional assumption* (a probability density function) for *independent* random variables, we define a “likelihood” as a product of their densities. We can think of this intuitively as just the “likelihood” or “chance” that our data occurs, given a specific distribution. Maximum likelihood estimators then tell us, given that assumed likelihood, **what parameter values make our observed data *most likely*** to have occurred.

So. Does it make sense? I would argue, intuitively, yes! Yes, it does. Is it good? That’s perhaps a different question with a more complicated answer. It’s a good baseline, certainly, and foundational to *much* of statistical theory. We’ll see in a later chapter that maximum likelihood estimates have good properties related to having minimal variance among a larger class of estimators (yay!), but the maximum likelihood estimators we will consider in this course rely on *parametric* assumptions (i.e. we assume that the data follows a specific probability distribution in order to calculate MLEs). There are ways around these assumptions, but they are outside the scope of our course.

*🌶️ Note that distributions are only involved in *parametric* methods, as opposed to non-parametric and semi-parametric methods, the latter of which are for independent study or a graduate course in statistics!

Relation to Least-Squares

Recall that we typically write a simple linear regression model in one of two ways. For n observations X_1, \dots, X_n with outcomes Y_1, \dots, Y_n , we can write

$$E[Y_i | X_i] = \beta_0 + \beta_1 X_i$$

or we can write

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $E[\epsilon_i] = 0$. The latter equation makes it more clear where residuals come into play (they are just given by ϵ_i), and the former perhaps makes it more clear why the word “average” usually finds its way into our interpretations of regression coefficients. The second form, however, allows us to make it more clear how we would write up a “least-squares” equation.

Recall that the least-squares line (or, line of “best” fit) is the line that minimizes the *sum of squared residuals*. Parsing these words out, note that our residuals can be written as

$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i.$$

Squared residuals are then written as

$$\epsilon_i^2 = (Y_i - \beta_0 - \beta_1 X_i)^2,$$

and finally, the *sum* of squared residuals is given by

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

We can find what values of β_0 and β_1 minimize this sum by taking partial derivatives, setting equations equal to zero, and solving. It turns out that if let $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$ where σ^2 is *known*, then the MLE for β_0 and β_1 are equivalent to the values of β_0 and β_1 that minimize the sum of squared residuals!

2.1 Learning Objectives

By the end of this chapter, you should be able to...

- Derive maximum likelihood estimators for parameters of common probability density functions
- Calculate maximum likelihood estimators “by hand” for common probability density functions
- Explain (in plain English) why maximum likelihood estimation is an intuitive approach to estimating unknown parameters using a combination of (1) observed data, and (2) a distributional assumption

2.2 Reading Guide

Associated Readings: Chapter 5 (Introduction through Example 5.2.5)

2.2.1 Reading Questions

1. What is the intuition behind the maximum likelihood estimation (MLE) approach?
2. What are the typical steps to find a MLE? (see Ex 5.2.1, 5.2.2, and Case Study 5.2.1; work through at least one of these examples in detail, filling in any steps that the textbook left out)
3. Are there ever situations when the typical steps to finding a MLE don't work? If so, what can we do instead to find the MLE? (see Ex 5.2.3, 5.2.4)
4. How do the steps to finding a MLE change when we have more than one unknown parameter? (see Ex 5.2.5)

2.3 Definitions

You are expected to know the following definitions:

Parameter

In a frequentist* framework, a parameter is a *fixed*, unknown truth (very philosophical). By fixed, I mean “not random”. We assume that there is some true unknown value, governing the generation of all possible random observations of all possible people and things *in the whole world*. We sometimes call this unknown governing process the “superpopulation” (think: all who ever have been, all who are, and all who ever will be).

Practically speaking, parameters are things that we want to estimate, and we will estimate them using observed data!

*Two main schools of thought in statistics are: (1) Frequentist (everything you've ever learned so far in statistics, realistically), and (2) Bayesian. We'll cover the latter, and differences between the two, in a later chapter. There's also technically Fiducial inference as a third school of thought, but that one's never been widely accepted.

Statistic/Estimator

A statistic (or “estimator”) is a function of your data, used to “estimate” an unknown parameter. Often, statistics/estimators will be functions of *means* or averages, as we'll see in the worked examples for this chapter!

Likelihood Function

Let x_1, \dots, x_n be a sample of size n of independent observations from the probability density function $f_X(x \mid \theta)$, where θ is a set of unknown parameters that define the pdf. Then the likelihood function $L(\theta)$ is the product of the pdf evaluated at each x_i ,

$$L(\theta) = \prod_{i=1}^n f_X(x_i | \theta).$$

Note that this *looks* exactly like the joint pdf for n independent random variables, but it is *interpreted* differently. A likelihood is a function of *parameters*, given a set of observations (random variables). A joint pdf is a function of random variables.

Note: The likelihood function is one of the reasons why we like independent observations so much! If observations aren't independent, we can't simply multiply all of their pdfs together to get a likelihood function.

Maximum Likelihood Estimate (MLE)

Let $L(\theta) = \prod_{i=1}^n f_X(x_i | \theta)$ be the likelihood function corresponding to a random sample of observations x_1, \dots, x_n . If θ_e is such that $L(\theta_e) \geq L(\theta)$ for all possible values θ , then θ_e is called a *maximum likelihood estimate* for θ .

Log-likelihood

In statistics, when we say “log,” we essentially always mean “ln” (or, natural log). The log-likelihood is then, hopefully unsurprisingly, given by $\log(L(\theta))$. One thing that's useful to note (and will come in handy when calculating MLEs, is that the log of a product is equal to a sum of logs. For likelihoods, that means

$$\log(L(\theta)) = \log\left(\prod_{i=1}^n f_X(x_i | \theta)\right) = \sum_{i=1}^n \log(f_X(x_i | \theta))$$

This will end up making it *much* easier to take derivatives than needing to deal with products!

Order Statistic

The k th order statistic is equal to a sample's k th smallest value. Practically speaking, there are essentially three order statistics we typically care about: the minimum, the median, and the maximum. We denote the minimum (or, first order statistic) in a sample of random variables X_1, \dots, X_n as $X_{(1)}$, the maximum as $X_{(n)}$, and the median $X_{(m+1)}$ where $n = 2m + 1$ *when n is odd*. Note that median is in fact not an order statistic if n is even (since the median is an average of two values, $X_{(m)}$ and $X_{(m+1)}$, in this case.

See Example 5.2.4 in the Textbook for an example of where order statistics occasionally come into play when calculating maximum likelihood estimates.

2.4 Theorems

None for this chapter!

2.5 Worked Examples

Problem 1: Suppose we observe n independent observations $X_1, \dots, X_n \sim \text{Bernoulli}(p)$, where $f_X(x) = p^x(1-p)^{1-x}$. Find the MLE of p .

We can write the likelihood function as

$$L(p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$$

Then the log-likelihood is given by

$$\begin{aligned} \log(L(p)) &= \log \left[\prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} \right] \\ &= \sum_{i=1}^n \log [p^{x_i} (1-p)^{1-x_i}] \\ &= \sum_{i=1}^n [\log(p^{x_i}) + \log((1-p)^{1-x_i})] \\ &= \sum_{i=1}^n [x_i \log(p) + (1-x_i) \log(1-p)] \\ &= \log(p) \sum_{i=1}^n x_i + \log(1-p) \sum_{i=1}^n (1-x_i) \\ &= \log(p) \sum_{i=1}^n x_i + \log(1-p) (n - \sum_{i=1}^n x_i) \end{aligned}$$

We can take the derivative of the log-likelihood with respect to p , and set it equal to zero...

$$\begin{aligned}
\frac{\partial}{\partial p} \log(L(p)) &= \frac{\partial}{\partial p} \left[\log(p) \sum_{i=1}^n x_i + \log(1-p)(n - \sum_{i=1}^n x_i) \right] \\
&= \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} \\
0 &\equiv \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} \\
\frac{\sum_{i=1}^n x_i}{p} &= \frac{n - \sum_{i=1}^n x_i}{1-p} \\
(1-p) \sum_{i=1}^n x_i &= p(n - \sum_{i=1}^n x_i) \\
\sum_{i=1}^n x_i - p \sum_{i=1}^n x_i &= pn - p \sum_{i=1}^n x_i \\
\sum_{i=1}^n x_i &= pn \\
\frac{1}{n} \sum_{i=1}^n x_i &= p
\end{aligned}$$

and by solving for p , we get that the MLE of p is equal to $\frac{1}{n} \sum_{i=1}^n x_i$. We will *often* see that the MLEs of parameters are functions of sample averages (in this case, just the identity function!).

Problem 2: Suppose X_1, X_2, \dots, X_n are a random sample from the Normal pdf with parameters μ and σ^2 :

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2},$$

for $-\infty < x < \infty$, $-\infty < \mu < \infty$, and $\sigma^2 > 0$. Find the MLEs of μ and σ^2 . (Note that this is Question 5 on the MLE section of Problem Set 1! For your HW, try your best to do this problem from scratch, without looking at the course notes!)

Since we are dealing with a likelihood with two parameters, we'll need to solve a *system* of equations to obtain the MLEs for μ and σ^2 .

$$\begin{aligned}
\log(L(\mu, \sigma^2)) &= \log\left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right)\right) \\
&= \sum_{i=1}^n \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \right] \\
&= \sum_{i=1}^n \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \right] \\
&= \frac{-n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2
\end{aligned}$$

Now we need to find $\frac{\partial}{\partial \sigma^2} \log(L(\mu, \sigma^2))$ and $\frac{\partial}{\partial \mu} \log(L(\mu, \sigma^2))$. Let's make our lives a little bit easier by setting $\sigma^2 \equiv \theta$ (so we don't trip ourselves up with the exponent). We get

$$\begin{aligned}
\frac{\partial}{\partial \theta} \log(L(\mu, \theta)) &= \frac{\partial}{\partial \theta} \left(\frac{-n}{2} \log(2\pi\theta) - \frac{1}{2\theta} \sum_{i=1}^n (x_i - \mu)^2 \right) \\
&= \frac{-2\pi n}{4\pi\theta} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\theta^2} \\
&= \frac{-n}{2\theta} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\theta^2}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial}{\partial \mu} \log(L(\mu, \theta)) &= \frac{\partial}{\partial \mu} \left(\frac{-n}{2} \log(2\pi\theta) - \frac{1}{2\theta} \sum_{i=1}^n (x_i - \mu)^2 \right) \\
&= \frac{\partial}{\partial \mu} \left(-\frac{1}{2\theta} \sum_{i=1}^n (x_i^2 - 2\mu x_i + \mu^2) \right) \\
&= \frac{\partial}{\partial \mu} \left(-\frac{1}{2\theta} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \right) \right) \\
&= \frac{\partial}{\partial \mu} \left(-\frac{1}{2\theta} (-2\mu \sum_{i=1}^n x_i + n\mu^2) \right) \\
&= \frac{\partial}{\partial \mu} \left(\frac{\sum_{i=1}^n x_i}{\theta} \mu - \frac{n}{2\theta} \mu^2 \right) \\
&= \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n}{\theta} \mu
\end{aligned}$$

We now have the following system of equations to solve:

$$\begin{aligned} 0 &\equiv \frac{-n}{2\theta} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\theta^2} \\ 0 &\equiv \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n}{\theta}\mu \end{aligned}$$

Typically, we solve one of the equations for *one* of the parameters, plug that into the other equation, and then go from there. We'll start by solving the second equation for μ .

$$\begin{aligned} 0 &= \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n}{\theta}\mu \\ \frac{n}{\theta}\mu &= \frac{\sum_{i=1}^n x_i}{\theta} \\ \mu &= \frac{1}{n} \sum_{i=1}^n x_i \end{aligned}$$

Well that's convenient! We already have the MLE for μ as being just the sample average. Plugging this into the first equation in our system we obtain

$$\begin{aligned} 0 &= \frac{-n}{2\theta} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\theta^2} \\ 0 &= \frac{-n}{2\theta} + \frac{\sum_{i=1}^n (x_i - \frac{1}{n} \sum_{i=1}^n x_i)^2}{2\theta^2} \\ \frac{n}{2\theta} &= \frac{\sum_{i=1}^n (x_i - \frac{1}{n} \sum_{i=1}^n x_i)^2}{2\theta^2} \\ n &= \frac{\sum_{i=1}^n (x_i - \frac{1}{n} \sum_{i=1}^n x_i)^2}{\theta} \\ \theta &= \frac{1}{n} \sum_{i=1}^n (x_i - \frac{1}{n} \sum_{i=1}^n x_i)^2 \\ \theta &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. And so finally, we have that the MLE for σ^2 is given by $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$, and the MLE for μ is given by \bar{x} !

3 Method of Moments

At this point in the course, we've now seen one (hopefully intuitive) way to obtain estimators for unknown parameters in probability distributions: maximum likelihood estimation. An alternative approach to producing a “reasonable” estimator for an unknown parameter is called the “Method of Moments.” As the name implies, this method uses moments to derive estimators! Recall from probability theory that the r th moment of a probability distribution for X is given by $E[X^r]$. We can make use of relationships that between *theoretical* moments and *sample* moments to derive reasonable estimators!

In general, the steps involved in obtaining a MOM estimator are:

1. Determine how many equations are in the system we need to solve
2. Find the theoretical moments
3. Set theoretical moments equal to sample moments
4. Solve!

Why do we need more than one approach to obtain estimators?

We already have maximum likelihood estimation, and it seems reasonable, so why might we want another approach to obtaining estimates? A few reasons!

One is that estimators vary with regards their theoretical “properties” (as we'll see in the following chapters). These properties are one way to define how “good” an estimator is, and we ideally want our estimators to be the best of the best.

Another reason why we might sometimes want another approach to obtaining estimators, quite frankly, is that maximum likelihood estimators are sometimes a pain to calculate. In some cases, there isn't even a closed form solution for the parameter we're trying to estimate. In these scenarios, we need numerical optimization in order to obtain maximum likelihood estimators. While numerical optimization isn't the end of the world (it's actually often quite easy to implement), it can be *very* computationally intensive for more complex likelihoods. In general, if we can obtain a closed form estimator *analytically* (via calculus/algebra, for example), we'll be better off in the long run.* With the method of moments approach, it is often much easier to obtain a closed form estimator analytically. An example of this can be found in Example 5.2.7 in the course textbook.

* This is mostly a function the fact that much statistics research focuses on developing new methods for solving problems and analyzing data (think: linear regression but fancier, linear regression but *new* somehow, etc.). Statistics is inherently practical. You (probably) want any methods that you develop to be practically usable by people who are perhaps not statisticians. No one is going to use your method if it takes an unreasonably long time to compute an estimator. Imagine how irritating it would be if it took your machine two days to compute linear regression coefficients in R, for example.

3.1 Learning Objectives

By the end of this chapter, you should be able to...

- Derive method of moments estimators for parameters of common probability density functions
- Explain (in plain English) why method of moments estimation is an intuitive approach to estimating unknown parameters

3.2 Reading Guide

Associated Readings: Chapter 5 (“The Method of Moments” through Example 5.2.7)

3.2.1 Reading Questions

1. What is the intuition behind the method of moments (MOM) procedure for estimating unknown parameters?
2. What are the typical steps to find a MOM estimator? (see Ex 5.2.6, 5.2.7, and Case Study 5.2.2; work through at least one of these examples in detail, filling in any steps that the textbook left out)
3. What advantages does the MOM approach offer compared to MLE?
4. Do the MOM and MLE approaches always yield the same estimate? (look through the examples in Section 5.2 and try using the other approach — do you always get the same answer?)

3.3 Definitions

You are expected to know the following definitions:

Theoretical Moment

The r^{th} *theoretical* moment of a probability distribution is given by $E[X^r]$. For example, when $r = 1$, the r^{th} moment is just the expectation of the random variable X .

Sample Moment

The r^{th} *sample* moment of a probability distribution is given by $\frac{1}{n} \sum_{i=1}^n x_i^r$, for a random sample of observations x_1, \dots, x_n .

Method of Moments Estimates

Let x_1, \dots, x_n be a random sample of observations from the pdf $f_X(x | \theta)$. The method of moments estimates are then the solutions to the set of s equations given by

$$\begin{aligned} E[X] &= \frac{1}{n} \sum_{i=1}^n x_i \\ E[X^2] &= \frac{1}{n} \sum_{i=1}^n x_i^2 \\ &\vdots \\ E[X^s] &= \frac{1}{n} \sum_{i=1}^n x_i^s \end{aligned}$$

If our pdf depends on only a single unknown parameter, we only need to solve the first equation. If we have two unknown parameters, we need to solve the system of the first two equations. So on and so forth.

3.4 Theorems

None for this chapter!

3.5 Worked Examples

In general (for these worked examples *as well as the problem sets*), I do not expect you to calculate theoretical moments by hand. We practiced that in the probability review chapter, and now we can use those known theoretical moments to make our lives easier.

Problem 1: Suppose $X_1, \dots, X_n \sim \text{Poisson}(\lambda)$. Find the MLE of λ *and* the MOM estimator of λ .

To obtain the MLE, note that we can write the likelihood as

$$L(\lambda) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

and the log-likelihood as

$$\log(L(\lambda)) = \sum_{i=1}^n [x_i \log(\lambda) - \lambda - \log(x_i!)]$$

where I've used one “log rule” in the above to simplify: $\log(a^b) = b \times \log(a)$. Taking the derivative of the log-likelihood and setting it equal to zero, we obtain

$$\begin{aligned} \frac{\partial}{\partial \lambda} \log(L(\lambda)) &= \frac{1}{\lambda} \sum_{i=1}^n x_i - n \\ 0 &\equiv \frac{1}{\lambda} \sum_{i=1}^n x_i - n \\ n &= \frac{1}{\lambda} \sum_{i=1}^n x_i \\ \lambda &= \frac{1}{n} \sum_{i=1}^n x_i \end{aligned}$$

and so the MLE for λ is the sample average. To obtain the MOM estimator for λ , first note that the pdf contains only one parameter. Therefore, we only need to set the first theoretical moment equal to the first sample moment, and solve. Note that the first theoretical moment of a Poisson distribution is $E[X] = \lambda$, and so equating this to the first sample moment, we obtain that the MOM estimator for λ is again, just the sample average! Much “easier” to compute than the MLE, in this case.

Problem 2: Suppose $X_1, \dots, X_n \sim \text{Bernoulli}(\theta)$. Find the MOM estimator for θ .

Note that our pdf contains only one parameter. Then we only need to solve a “system” of one equation. We have

$$E[X] = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\theta = \frac{1}{n} \sum_{i=1}^n x_i$$

and we’re done! The system is pretty easy to “solve” when the theoretical moment is exactly the parameter we’re interested in.

Problem 3: Suppose $Y_1, \dots, Y_n \sim \text{Uniform}(0, \theta)$. Find the MOM estimator for θ .

Note that our pdf contains only one parameter. Then we only need to solve a “system” of one equation. We have

$$E[Y] = \frac{1}{n} \sum_{i=1}^n y_i$$

$$\frac{\theta}{2} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$\theta = 2\bar{y}$$

And so the MOM estimator for θ is 2 times the sample mean. Note that this is an example where the MOM estimator and MLE are not the same (you derived the MLE on your first problem set)!

4 Properties of Estimators

Now that we've developed the tools for deriving estimators of unknown parameters, we can start thinking about different metrics for determining how “good” our estimators actually are. In general, we like our estimators to be:

- **Unbiased:** Our estimate should be estimating *what it's supposed to be estimating*, for lack of a better phrase. Bias (or, unbiased-ness, in this case) is related to accuracy. In introductory statistics, you likely discussed sample bias (or, whether or not the data you collect is representative of the population you are trying to make inference on) and information bias (or, whether the values of the data you collect are representative of the people who report them). If you have a biased sample or biased information, your estimates (think, regression coefficients) are likely going to misrepresent true relationships in the population. Bias of *estimates* has a very specific definition in statistical theory that is *distinct* from sample bias and information bias. Questions of sample bias and information bias are important to consider when collecting and analyzing data, and questions of whether or not our estimates are biased are important to consider *prior* to analyzing data.
- **Precise:** In short, if our estimates are wildly uncertain (think, gigantic confidence intervals), they'll essentially be of no use to us from a practical perspective. As an extreme example, consider how you would feel if a new cancer drug was released with *very* severe side-effects, but scientists noted that the drug would increase cancer patients expected survival time by somewhere between 1 and 700 days. Are we really certain enough, in this case, that the benefits of the drug outweigh the potential costs? What if instead, the expected survival time would increase between 650 and 700 days? Would that change your answer? These types of questions are precisely (ha!) why precision is important. Again, you've likely discussed precision (colloquially) in an introductory statistics course. In statistical theory, precision (similar to bias) has a very specific definition. So long as our estimates are unbiased, we want to minimize variance (and therefore increase precision) as much as we possibly can. Even at the same sample size, some estimates are more precise than others, which we'll explore in this chapter.

The Bias-Variance Trade-off

If you are familiar with machine learning techniques or models for prediction purposes more generally (as opposed to inference), you may have stumbled upon the phrase “bias-variance

trade-off.” In scenarios where we want to make good predictions for new observations using a statistical model, one way to measure how “well” our model is predicting new observations is through minimizing **mean squared error**. Intuitively, this is something we should *want* to minimize: “errors” (the difference between a predicted value and an observed value) are bad, we square them because the direction of the error (above or below) shouldn’t matter too much, and average over them because we need a summary measure of all our errors combined, and an average seems reasonable. In statistical terms, mean squared error has a very specific definition (see below) as the expected value of what is sometimes called a *loss function* (where in this case, loss is defined as squared error loss). We’ll return to this in the decision theory chapter of our course notes.

It just so happens that we can decompose mean squared error into a sum of two terms: the variance of our estimator + the bias of our estimator (squared). What this means for us is that two estimators may have the *exact same* MSE, but *very* different variances or biases (potentially). In general, if we hold MSE constant and imagine *increasing* the variance of our estimator, the bias would need to decrease accordingly to maintain the same MSE. This is where the “trade-off” comes from. MSE is an *incredibly* commonly used metric for assessing prediction models, but as we will see, doesn’t necessarily paint a full picture in terms of how “good” an estimator is. Smaller MSE does not automatically imply “better estimator,” just as smaller bias (in some cases) does not automatically imply “better estimator.”

4.1 Learning Objectives

By the end of this chapter, you should be able to...

- Calculate bias and variance of various estimators for unknown parameters
- Explain the distinction between bias and variance colloquially in terms of precision and accuracy, and why these properties are important
- Compare estimators in terms of their relative efficiency
- Justify why there exists a bias-variance trade-off, and explain what consequences this may have when comparing estimators

4.2 Reading Guide

Associated Readings: Chapter 5, Section 5.4 (“Properties of Estimators”) & 5.5 (“Minimum-Variance Estimators: The Cramér-Rao Lower Bound”)

4.2.1 Reading Questions

1. Intuitively, what is the difference between bias and precision?
2. What are the typical steps to checking if an estimator is unbiased? (see Examples 5.4.2, 5.4.3, and 5.4.4 in the textbook)
3. How can we construct unbiased estimators? (see comment in Example 5.4.2 and 5.4.4)
4. If an estimator is unbiased, is it also *asymptotically* unbiased? If an estimator is asymptotically unbiased, is it necessarily unbiased?
5. When comparing estimators, how can we determine which estimator is more efficient? (see Examples 5.4.5 and 5.4.6)
6. Describe, in your own words, what the Cramér-Rao inequality tells us.
7. What is the difference between a UMVUE and an efficient estimator? Does one imply the other? (see the Comment below Definition 5.5.2)

4.3 Definitions

You are expected to know the following definitions:

Unbiased

An estimator $\hat{\theta} = g(X_1, \dots, X_n)$ is an unbiased estimator for θ if $E[\hat{\theta}] = \theta$, for all θ .

Asymptotically Unbiased

An estimator $\hat{\theta} = g(X_1, \dots, X_n)$ is an *asymptotically* unbiased estimator for θ if $\lim_{n \rightarrow \infty} E[\hat{\theta}] = \theta$.

Precision

The precision of a random variable X is given by $\frac{1}{\text{Var}(X)}$.

Mean Squared Error (MSE)

The mean squared error of an estimator is given by

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$$

Relative Efficiency

The relative efficiency of an estimator $\hat{\theta}_1$ with respect to an estimator $\hat{\theta}_2$ is the ratio $\text{Var}(\hat{\theta}_2)/\text{Var}(\hat{\theta}_1)$.

Uniformly Minimum-Variance Unbiased Estimator (UMVUE)

An estimator $\hat{\theta}^*$ is the UMVUE if, for all estimators $\hat{\theta}$ in the class of unbiased estimators Θ ,

$$Var(\hat{\theta}^*) \leq Var(\hat{\theta})$$

Score

The score is defined as the first partial derivative with respect to θ of the log-likelihood function, given by

$$\frac{\partial}{\partial \theta} \log L(\theta | x)$$

Information Matrix

The information matrix* $I(\theta)$ for a collection of iid random variables X_1, \dots, X_n is the variance of the score, given by

$$I(\theta) = E \left[\left(\frac{\partial}{\partial \theta} \log L(\theta | x) \right)^2 \right] = -E \left[\frac{\partial^2}{\partial \theta^2} \log L(\theta | x) \right]$$

Note that the above formula *is* in fact the variance of the score, since we can show that the *expectation* of the score is 0 (under some regularity conditions). This is shown as part of the proof of the C-R lower bound in the Theorems section of this chapter.

The information matrix is sometimes written in terms of a pdf for a single random variable as opposed to a likelihood (this is what our textbook does, for example). In this case, we have $I(\theta) = nI_1(\theta)$, where the $I_1(\theta)$ on the right-hand side is defined as $E \left[\left(\frac{\partial}{\partial \theta} \log f_X(x | \theta) \right)^2 \right]$. Sometimes (as in the textbook) $I_1(\theta)$ is written without the subscript 1 which is a slight abuse of notation that is endlessly confusing (to me, at least). For this set of course notes, we'll always specify the information matrix in terms of a pdf for a single random variable with the subscript 1, for clarity.

*The information matrix is often referred to as the Fisher Information matrix, as it was developed by Sir Ronald Fisher. Fisher developed *much* of the core, statistical theory that we use today. He was also the founding chairman of the University of Cambridge Eugenics Society, and contributed to a large body of scientific work and public policy that promoted racist and classist ideals.

4.4 Theorems

Covariance Inequality (based on the Cauchy-Schwarz inequality)

Let X and Y be random variables. Then,

$$Var(X) \geq \frac{Cov(X, Y)^2}{Var(Y)}$$

The proof is quite clear on [Wikipedia](#).

Cramér-Rao Lower Bound

Let $f_Y(y | \theta)$ be a pdf with nice* conditions, and let Y_1, \dots, Y_n be a random sample from $f_Y(y | \theta)$. Let $\hat{\theta}$ be *any* unbiased estimator of θ . Then

$$\begin{aligned} Var(\hat{\theta}) &\geq \left\{ E \left[\left(\frac{\partial \log(L(\theta | y))}{\partial \theta} \right)^2 \right] \right\}^{-1} \\ &= - \left\{ E \left[\frac{\partial^2 \log(L(\theta | y))}{\partial \theta^2} \right] \right\}^{-1} \\ &= \frac{1}{I(\theta)} \end{aligned}$$

*our nice conditions that we need are that $f_Y(y | \theta)$ has continuous first- and second-order derivatives, which would quickly discover we need by looking at the form for the C-R lower bound, and that the set of values y where $f_Y(y | \theta) \neq 0$ does not depend on θ . If you are familiar with the concept of the “support” of a function, that is where this second condition comes from. The key here is that this condition allows to interchange derivatives and integrals, in particular, $\frac{\partial}{\partial \theta} \int f(x) dx = \int \frac{\partial}{\partial \theta} f(x) dx$, which we’ll need to complete the proof.

Proof.

Let $X = \frac{\partial \log L(\theta | \mathbf{y})}{\partial \theta}$. By the Covariance Inequality,

$$Var(\hat{\theta}) \geq \frac{Cov(\hat{\theta}, X)^2}{Var(X)}$$

and so if we can show

$$\begin{aligned}\frac{Cov(\hat{\theta}, X)^2}{Var(X)} &= \left\{ E \left[\left(\frac{\partial \log(L(\theta | \mathbf{y}))}{\partial \theta} \right)^2 \right] \right\}^{-1} \\ &= \frac{1}{I(\theta)}\end{aligned}$$

then we're done, as this is the C-R lower bound. Note first that

$$\begin{aligned}E[X] &= \int x f_Y(\mathbf{y} | \theta) d\mathbf{y} \\ &= \int \left(\frac{\partial \log L(\theta | \mathbf{y})}{\partial \theta} \right) f_Y(\mathbf{y} | \theta) d\mathbf{y} \\ &= \int \left(\frac{\partial \log f_Y(\mathbf{y} | \theta)}{\partial \theta} \right) f_Y(\mathbf{y} | \theta) d\mathbf{y} \\ &= \int \frac{\frac{\partial}{\partial \theta} f_Y(\mathbf{y} | \theta)}{f_Y(\mathbf{y} | \theta)} f_Y(\mathbf{y} | \theta) d\mathbf{y} \\ &= \int \frac{\partial}{\partial \theta} f_Y(\mathbf{y} | \theta) d\mathbf{y} \\ &= \frac{\partial}{\partial \theta} \int f_Y(\mathbf{y} | \theta) d\mathbf{y} \\ &= \frac{\partial}{\partial \theta} 1 \\ &= 0\end{aligned}$$

This means that

$$\begin{aligned}Var[X] &= E[X^2] - E[X]^2 \\ &= E[X^2] \\ &= E \left[\left(\frac{\partial \log L(\theta | \mathbf{y})}{\partial \theta} \right)^2 \right]\end{aligned}$$

and

$$\begin{aligned}
Cov(\hat{\theta}, X) &= E[\hat{\theta}X] - E[\hat{\theta}]E[X] \\
&= E[\hat{\theta}X] \\
&= \int \hat{\theta}x f_Y(\mathbf{y} | \theta) d\mathbf{y} \\
&= \int \hat{\theta} \left(\frac{\partial \log L(\theta | \mathbf{y})}{\partial \theta} \right) f_Y(\mathbf{y} | \theta) d\mathbf{y} \\
&= \int \hat{\theta} \left(\frac{\partial \log f_Y(\mathbf{y} | \theta)}{\partial \theta} \right) f_Y(\mathbf{y} | \theta) d\mathbf{y} \\
&= \int \hat{\theta} \frac{\frac{\partial}{\partial \theta} f_Y(\mathbf{y} | \theta)}{f_Y(\mathbf{y} | \theta)} f_Y(\mathbf{y} | \theta) d\mathbf{y} \\
&= \int \hat{\theta} \frac{\partial}{\partial \theta} f_Y(\mathbf{y} | \theta) d\mathbf{y} \\
&= \frac{\partial}{\partial \theta} \int \hat{\theta} f_Y(\mathbf{y} | \theta) d\mathbf{y} \\
&= \frac{\partial}{\partial \theta} E[\hat{\theta}] \\
&= \frac{\partial}{\partial \theta} \theta \\
&= 1
\end{aligned}$$

where $E[\hat{\theta}] = \theta$ since our estimator is unbiased. Putting this all together, we have

$$\begin{aligned}
Var[\hat{\theta}] &\geq \frac{Cov(\hat{\theta}, X)^2}{Var(X)} \\
&= \frac{1^2}{E \left[\left(\frac{\partial \log L(\theta | \mathbf{y})}{\partial \theta} \right)^2 \right]} \\
&= \frac{1}{I(\theta)}
\end{aligned}$$

as desired.

Comment: Note that what the Cramér-Rao lower bound tells us is that, if the variance of an unbiased estimator is *equal* to the Cramér-Rao lower bound, then that estimator has the *minimum possible variance* among all unbiased estimators there could possibly be. This allows us to *prove*, for example, whether or not an unbiased estimator is the UMVUE: If an unbiased estimator's variance achieves the C-R lower bound, then it is *optimal* according to the UMVUE criterion.

4.5 Worked Examples

Problem 1: Suppose $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Exponential}(1/\theta)$. Compute the MLE of θ , and show that it is an unbiased estimator of θ .

Note that we can write

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \frac{1}{\theta} e^{-x_i/\theta} \\ \log L(\theta) &= \sum_{i=1}^n \log\left(\frac{1}{\theta} e^{-x_i/\theta}\right) \\ &= \sum_{i=1}^n \log\left(\frac{1}{\theta}\right) - \sum_{i=1}^n x_i/\theta \\ &= -n \log(\theta) - \frac{1}{\theta} \sum_{i=1}^n x_i \\ \frac{\partial}{\partial \theta} \log L(\theta) &= \frac{\partial}{\partial \theta} \left(-n \log(\theta) - \frac{1}{\theta} \sum_{i=1}^n x_i \right) \\ &= -\frac{n}{\theta} + \frac{\sum_{i=1}^n x_i}{\theta^2} \end{aligned}$$

Setting this equal to 0 and solving for θ we obtain

$$\begin{aligned} 0 &\equiv -\frac{n}{\theta} + \frac{\sum_{i=1}^n x_i}{\theta^2} \\ \frac{n}{\theta} &= \frac{\sum_{i=1}^n x_i}{\theta^2} \\ n &= \frac{\sum_{i=1}^n x_i}{\theta} \\ \theta &= \frac{1}{n} \sum_{i=1}^n x_i \end{aligned}$$

and so the MLE for θ is the sample mean. To show that the MLE is unbiased, we note that

$$E \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} \sum_{i=1}^n \theta = \theta$$

as desired.

Problem 2: Suppose again that $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Exponential}(1/\theta)$. Let $\hat{\theta}_2 = Y_1$, and $\hat{\theta}_3 = nY_{(1)}$. Show that $\hat{\theta}_2$ and $\hat{\theta}_3$ are unbiased estimators of θ . Hint: use the fact that $Y_{(1)} \sim \text{Exponential}(n/\theta)$

Note that the mean of a random variable $Y \sim \text{Exponential}(\lambda)$ is given by $1/\lambda$. Then we can write

$$E[\hat{\theta}_2] = E[Y_1] = \frac{1}{1/\theta} = \theta$$

and

$$E[\hat{\theta}_3] = E[nY_{(1)}] = \frac{n}{n/\theta} = \theta$$

as desired.

Problem 3: Compare the variance of the estimators from Problems 1 and 2. Which is most efficient?

Recall that the variance of a random variable $Y \sim \text{Exponential}(\lambda)$ is given by $1/\lambda^2$. Let the MLE from Problem 1 be denoted $\hat{\theta}_1 = \bar{X}$. Then we can write

$$\text{Var}[\hat{\theta}_1] = \text{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[X_i] = \frac{1}{n^2} \left(\frac{n}{(1/\theta)^2}\right) = \frac{\theta^2}{n}$$

and

$$\text{Var}[\hat{\theta}_2] = \text{Var}[Y_1] = \frac{1}{(1/\theta)^2} = \theta^2$$

and

$$\text{Var}[\hat{\theta}_3] = \text{Var}[nY_{(1)}] = n^2 \text{Var}[Y_{(1)}] = \frac{n^2}{(n/\theta)^2} = \theta^2$$

Thus, the variance of the MLE, $\hat{\theta}_1$, is most efficient, and is n times smaller than the variance of both $\hat{\theta}_2$ and $\hat{\theta}_3$.

Problem 4: Suppose $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$. Show that the estimator $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$ and the estimator $\hat{\mu}_w = \sum_{i=1}^n w_i X_i$ are both unbiased estimators of μ , where $\sum_{i=1}^n w_i = 1$.

We can write

$$E[\hat{\mu}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

and

$$E[\hat{\mu}_w] = E\left[\sum_{i=1}^n w_i X_i\right] = \sum_{i=1}^n w_i E[X_i] = \sum_{i=1}^n w_i \mu = \mu \sum_{i=1}^n w_i = \mu$$

as desired.

Problem 5: Determine whether the estimator $\hat{\mu}$ or $\hat{\mu}_w$ is more efficient, in Problem 5, if we additionally impose the constraint $w_i \geq 0 \ \forall i$. (Note that this is a more “general” example based on Example 5.4.5 in the course textbook) (Hint: use the Cauchy-Schwarz inequality)

To determine relative efficiency, we must compute the variance of each estimator. We have

$$Var[\hat{\mu}] = Var\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \sum_{i=1}^n Var[X_i] = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \sigma^2/n$$

and

$$\begin{aligned} Var[\hat{\mu}_w] &= Var\left[\sum_{i=1}^n w_i X_i\right] \\ &= \sum_{i=1}^n Var[w_i X_i] \\ &= \sum_{i=1}^n w_i^2 Var[X_i] \\ &= \sum_{i=1}^n w_i^2 \sigma^2 \\ &= \sigma^2 \sum_{i=1}^n w_i^2 \end{aligned}$$

And so to determine which estimator is more efficient, we need to determine if $\frac{1}{n}$ is less than $\sum_{i=1}^n w_i^2$ (or not). The Cauchy-Schwarz inequality tells us that

$$\begin{aligned}
\left(\sum_{i=1}^n w_i \cdot 1\right)^2 &\leq \left(\sum_{i=1}^n w_i^2\right) \left(\sum_{i=1}^n 1^2\right) \\
\left(\sum_{i=1}^n w_i\right)^2 &\leq \left(\sum_{i=1}^n w_i^2\right) n \\
1 &\leq \left(\sum_{i=1}^n w_i^2\right) n \\
\frac{1}{n} &\leq \sum_{i=1}^n w_i^2
\end{aligned}$$

and therefore, $\hat{\mu}$ is more efficient than $\hat{\mu}_w$.

Problem 6: Suppose $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$. Show that the MLE for σ^2 is *biased*, and suggest a modified variance estimator for σ^2 that is *unbiased*. (Note that this is example 5.4.4 in our course textbook)

Recall that the MLE for σ^2 is given by $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$. Then

$$\begin{aligned}
E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right] &= \frac{1}{n} \sum_{i=1}^n E[(X_i - \bar{X})^2] \\
&= \frac{1}{n} \sum_{i=1}^n E[X_i^2 - 2X_i\bar{X} + \bar{X}^2] \\
&= \frac{1}{n} \sum_{i=1}^n E[X_i^2] - 2E \left[\frac{1}{n} \sum_{i=1}^n X_i \bar{X} \right] + E[\bar{X}^2] \\
&= \frac{1}{n} \sum_{i=1}^n E[X_i^2] - 2E \left[\bar{X} \frac{1}{n} \sum_{i=1}^n X_i \right] + E[\bar{X}^2] \\
&= \frac{1}{n} \sum_{i=1}^n E[X_i^2] - 2E[\bar{X}^2] + E[\bar{X}^2] \\
&= \frac{1}{n} \sum_{i=1}^n E[X_i^2] - E[\bar{X}^2]
\end{aligned}$$

Recall that since $X_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$, $\bar{X} \sim N(\mu, \sigma^2/n)$, and that we can write $Var[Y] + E[Y]^2 = E[Y^2]$ (definition of variance). Then we can write

$$\begin{aligned}
E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right] &= \frac{1}{n} \sum_{i=1}^n E[X_i^2] - E[\bar{X}^2] \\
&= \frac{1}{n} \sum_{i=1}^n (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2 \right) \\
&= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 \\
&= \sigma^2 - \frac{\sigma^2}{n} \\
&= \sigma^2 \left(1 - \frac{1}{n} \right) \\
&= \sigma^2 \left(\frac{n-1}{n} \right)
\end{aligned}$$

Therefore, since $E[\hat{\sigma}_{MLE}^2] \neq \sigma^2$, the MLE is unbiased. Note that

$$\begin{aligned}
E \left[\left(\frac{n}{n-1} \right) \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right] &= \left(\frac{n}{n-1} \right) \left(\frac{n-1}{n} \right) \sigma^2 \\
&= \sigma^2
\end{aligned}$$

and so the estimator $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is an unbiased estimator for σ^2 . This estimator is often called the “sample variance”, and is denoted by S^2 .

5 Consistency

Under development...

6 Asymptotics & the Central Limit Theorem

Under development...

7 Computational Optimization

Under development...

8 Bayesian Inference

Under development...

9 Decision Theory

Under development...

10 Hypothesis Testing

Under development...

References