Stat 4352.002: Mathematical Statistics, Spring 2026 **Topic 1**

Statistical Inference

Lecturer: Yun Wei *Scribe: Students in Spring 2026*

Contents

Chapter 1

Transformation of random variables

1.1 Distributions of Functions of Random Variables

If X is a random variable with cdf $F_X(x)$, then any function of X, say $g(X)$, is also a random variable. Often $g(X)$ is of interest itself and we write $Y = g(X)$ to denote the new random variable $g(X)$. Since Y is a function of X, we can describe the probabilistic behavior of Y in terms of that of X. That is, for any set A,

$$P(Y \in A) = P\big(g(X) \in A\big),$$

showing that the distribution of Y depends on the functions F_X and g. Depending on the choice of g, it is sometimes possible to obtain a tractable expression for this probability.

Formally, if we write $y = g(x)$, the function $g(x)$ defines a mapping from the original sample space of X, $\mathcal{X}$, to a new sample space, $\mathcal{Y}$, the sample space of the random variable Y. That is,

$$g(x) : \mathcal{X} \to \mathcal{Y}.$$

We associate with g an inverse mapping, denoted by g^{-1}, which is a mapping from subsets of $\mathcal{Y}$ to subsets of $\mathcal{X}$, and is defined by

$$g^{-1}(A) = \{x \in \mathcal{X} : g(x) \in A\}.$$

Note that the mapping g^{-1} takes sets into sets; that is, $g^{-1}(A)$ is the set of points in $\mathcal{X}$ that $g(x)$ takes into the set A. It is possible for A to be a point set, say $A = \{y\}$. Then

$$g^{-1}(\{y\}) = \{x \in \mathcal{X} : g(x) = y\}.$$

In this case we often write $g^{-1}(y)$ instead of $g^{-1}(\{y\})$. The quantity $g^{-1}(y)$ can still be a set, however, if there is more than one x for which $g(x) = y$. If there is only one x for which $g(x) = y$, then $g^{-1}(y)$ is the point set $\{x\}$, and we will write $g^{-1}(y) = x$.

Remark 1.1.1. In particular, if g is an invertible function, then the inverse map matches the definition of inverse function.

If the random variable Y is now defined by $Y = g(X)$, we can write for any set $A \subset \mathcal{Y}$,

$$\begin{aligned} P(Y \in A) &= P\big(g(X) \in A\big) \\ &= P\big(\{x \in \mathcal{X} : g(x) \in A\}\big) \\ &= P\big(X \in g^{-1}(A)\big). \end{aligned}$$

This defines the probability distribution of Y.

1.1.1 Discrete distributions

If X is a discrete random variable, then $\mathcal{X}$ is countable. The sample space for $Y = g(X)$ is $\mathcal{Y} = \{y : y = g(x),\ x \in \mathcal{X}\}$, which is also a countable set. Thus, Y is also a discrete random variable. From (2.1.2), the pmf for Y is

$$f_Y(y) = P(Y = y) = \sum_{x \in g^{-1}(y)} P(X = x) = \sum_{x \in g^{-1}(y)} f_X(x), \qquad \text{for } y \in \mathcal{Y},$$

and $f_Y(y) = 0$ for $y \notin \mathcal{Y}$. In this case, finding the pmf of Y involves simply identifying $g^{-1}(y)$, for each $y \in \mathcal{Y}$, and summing the appropriate probabilities.

Example 1.1.2 (Binomial transformation)**.** A discrete random variable X has a *binomial* distribution if its pmf is of the form

$$f_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n,$$

where n is a positive integer and $0 \le p \le 1$. Values such as n and p that can be set to different values, producing different probability distributions, are called *parameters*. Consider the random variable $Y = g(X)$, where $g(x) = n - x$; that is, $Y = n - X$. Here $\mathcal{X} = \{0, 1, \ldots, n\}$ and $\mathcal{Y} = \{y : y = g(x),\ x \in \mathcal{X}\} = \{0, 1, \ldots, n\}$. For any $y \in \mathcal{Y}$, $n - x = g(x) = y$ if and only if $x = n - y$. Thus, $g^{-1}(y)$ is the single point $x = n - y$, and

$$\begin{aligned} f_Y(y) &= \sum_{x \in g^{-1}(y)} f_X(x) \\ &= f_X(n - y) \\ &= \binom{n}{n-y} p^{n-y} (1-p)^{n-(n-y)} \\ &= \binom{n}{y} (1-p)^y p^{n-y}. \qquad \left(\binom{n}{y} = \binom{n}{n-y}\right) \end{aligned}$$

Thus, we see that Y also has a binomial distribution, but with parameters n and $1 - p$. Intuitively, X is the number of heads, when we toss a coin with probability p getting a head n times; so $Y = n - X$ is the number of tails, which then is also a binomial distribution, but with parameters n and $1 - p$.

1.1.2 Continuous distributions

If X and Y are continuous random variables, then in some cases it is possible to find simple formulas for the cdf and pdf of Y in terms of the cdf and pdf of X and the function g. In the remainder of this section, we consider some of these cases.

The cdf of $Y = g(X)$ is

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P\big(g(X) \le y\big) \\ &= P\big(\{x \in \mathcal{X} : g(x) \le y\}\big). \end{aligned} \tag{1.1}$$

When transformations are made, it is important to keep track of the sample spaces of the random variables; otherwise, much confusion can arise. When the transformation is from X to $Y = g(X)$, it is most convenient to use

$$\mathcal{X} = \{x : f_X(x) > 0\} \qquad \text{and} \qquad \mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}.$$

The pdf of the random variable X is positive only on the set $\mathcal{X}$ and is 0 elsewhere. Such a set is called the *support* set of a distribution or, more informally, the *support* of a distribution. This terminology can also apply to a pmf or, in general, to any nonnegative function.

It is easiest to deal with functions $g(x)$ that are *monotone*, that is, those that satisfy either

$$u > v \Rightarrow g(u) > g(v) \quad \text{(increasing)} \qquad \text{or} \qquad u < v \Rightarrow g(u) > g(v) \quad \text{(decreasing)}.$$

If the transformation $x \to g(x)$ is monotone, then it is *one-to-one and onto* from $\mathcal{X} \to \mathcal{Y}$. That is, each x goes to only one y and each y comes from at most one x (one-to-one). Also, for $\mathcal{Y}$ defined as above, for each $y \in \mathcal{Y}$ there is an $x \in \mathcal{X}$ such that $g(x) = y$ (onto). Thus, the transformation g uniquely pairs xs and ys. If g is monotone, then g^{-1} is single-valued; that is, $g^{-1}(y) = x$ if and only if $y = g(x)$. If g is increasing, this implies that

$$\begin{aligned}\{x \in \mathcal{X} : g(x) \le y\} &= \{x \in \mathcal{X} : g^{-1}(g(x)) \le g^{-1}(y)\} \\ &= \{x \in \mathcal{X} : x \le g^{-1}(y)\}.\end{aligned}$$

If $g(x)$ is an increasing function, then using (2.1.4), we can write

$$F_Y(y) = \int_{\{x \in \mathcal{X} : x \le g^{-1}(y)\}} f_X(x)\,dx = \int_{-\infty}^{g^{-1}(y)} f_X(x)\,dx = F_X\big(g^{-1}(y)\big).$$

If the pdf of Y is continuous, it can be obtained by differentiating the cdf. Taking derivatives,

$$f_Y(y) = f_X\big(g^{-1}(y)\big)\frac{d}{dy}g^{-1}(y) = f_X\big(g^{-1}(y)\big)\left|\frac{d}{dy}g^{-1}(y)\right|,$$

where the last step follows since $g^{-1}(y)$ is increasing and thus has nonnegative derivatives.

If g is decreasing, this implies that

$$\begin{aligned}\{x \in \mathcal{X} : g(x) \le y\} &= \{x \in \mathcal{X} : g^{-1}(g(x)) \ge g^{-1}(y)\} \\ &= \{x \in \mathcal{X} : x \ge g^{-1}(y)\}. \qquad (2.1.9)\end{aligned}$$

If $g(x)$ is decreasing, we have

$$F_Y(y) = \int_{g^{-1}(y)}^{\infty} f_X(x)\,dx = 1 - F_X\big(g^{-1}(y)\big).$$

The continuity of X is used to obtain the second equality. Taking derivatives,

$$f_Y(y) = -f_X\big(g^{-1}(y)\big)\frac{d}{dy}g^{-1}(y) = f_X\big(g^{-1}(y)\big)\left|\frac{d}{dy}g^{-1}(y)\right|,$$

where the last step follows since $g^{-1}(y)$ is decreasing and thus has nonpositive derivatives.

Define the indicator function $1_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$.

Example 1.1.3 (Uniform-exponential relationship–I)**.** Suppose $X \sim f_X(x) = 1_{(0,1)}(x)$, the *uniform*$(0,1)$ distribution. It is straightforward to check that $F_X(x) = \begin{cases} 0 & x \le 0 \\ x & 0 < x < 1 \\ 1 & x \ge 1 \end{cases}$. We now make the transformation $Y = g(X) = -\log X$. Since

$$\frac{d}{dx}g(x) = \frac{d}{dx}(-\log x) = \frac{-1}{x} < 0, \qquad \text{for } 0 < x < 1,$$

$g(x)$ is a decreasing function. As X ranges between 0 and 1, $-\log x$ ranges between 0 and ∞, that is, $\mathcal{Y} = (0, \infty)$. For $y > 0$, $y = -\log x$ implies $x = e^{-y}$, so $g^{-1}(y) = e^{-y}$. Therefore, for $y > 0$,

$$F_Y(y) = 1 - F_X\big(g^{-1}(y)\big) = 1 - F_X(e^{-y}) = 1 - e^{-y}. \qquad (F_X(x) = x \text{ when } x \in (0,1) \text{ and } e^{-y} \in (0,1))$$

Of course, $F_Y(y) = 0$ for $y \leq 0$. Taking derivative, for $y > 0$,

$$f_Y(y) = e^{-y}.$$

Of course, $f_Y(y) = 0$ for $y \leq 0$. Using indicator notation, $F_Y(y) = (1 - e^{-y})1_{(0,\infty)}(y)$ and $f_Y(y) = e^{-y}1_{(0,\infty)}(y)$. **Note that it was necessary only to verify that $g(x) = -\log x$ is monotone on $(0,1)$, the support of X.**

1.1.3 Distributions of Non-Monotone Continuous Functions

Recall that if X is a continuous random variable and g is a monotone function, we can derive the cumulative distribution function (CDF) of $Y = g(X)$ from the CDF of X using the following:

$$g \text{ increasing } \Rightarrow F_Y(y) = F_X(g^{-1}(y)),$$
$$g \text{ decreasing } \Rightarrow F_Y(y) = 1 - F_X(g^{-1}(y)).$$

In both cases, the probability density (PDF) of Y is

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

But what if g is not monotone? In such case, it is often true that g will be monotone over certain intervals, which allows us to get an expression for $Y = g(X)$.

Theorem 1.1.4. *Suppose X has PDF $f_X(x)$ and $Y = g(X)$, and $\exists$ a partition of $\mathcal{X}$, $A_0, A_1, \ldots, A_k$ such that $\mathbb{P}(x \in A_0) = 0$ and functions $g_1(x), \ldots, g_k(x)$ defined on $A_1, \ldots, A_k$ satisfy the following:*

(i) $g(x) = g_i(x)$ *for* $x \in A_i$ *for* $i = 1, 2, \ldots, k$.

(ii) $g_i(x)$ *is monotone on* A_i *for* $i = 1, 2, \ldots, k$.

(iii) the set $\mathcal{Y} = \{y | y = g_i(x) \text{ for some } x \in A_i\}$ *is the same for* $i = 1, 2, \ldots, k$.

(iv) $g^{-1}(y)$ *has a continuous derivative on* $\mathcal{Y}$ *for* $i = 1, 2, \ldots, k$.

Then

$$f_Y(y) = \begin{cases} \sum_{i=1}^k f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right|, & \forall y \in \mathcal{Y} \\ 0, & \text{otherwise} \end{cases} = \sum_{i=1}^k f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right| 1_{\mathcal{Y}}(y)$$

.

One of the most common examples of this is the normal chi-square 1 distribution, also referred to as a chi-squared random variable with 1 degree of freedom.

Example 1.1.5 (Normal Chi-square 1 distribution)**.** . Suppose $X = N(0,1)$ is the standard normal distribution

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

and $Y = X^2$. In this case, let $A_0 = 0, A_1 = (-\infty, 0)$, and $A_2 = (0, \infty)$. Then $\mathbb{P}(x \in A_0) = 0$, $g_1(x) = x^2$ with domain A_1 is decreasing, and $g_2(x) = x^2$ with domain A_2 is increasing. Furthermore, $\mathcal{Y} = (0, \infty)$ for

$i = 1, 2$, and $\{g_1^{-1}(y) = -\sqrt{y},\ g_2^{-1}(y) = \sqrt{y}\}$ are continuous on $\mathcal{Y}$. Hence conditions $(i) - (iv)$ of Theorem 1.1.4 are satisfied, so for any $y \in \mathcal{Y}$,

$$\begin{aligned} f_Y(y) &= f_X(-\sqrt{y}) \left| -\frac{1}{2\sqrt{y}} \right| + f_X(\sqrt{y}) \left| \frac{1}{2\sqrt{y}} \right| \\ &= \frac{1}{\sqrt{2\pi y}} e^{-y/2}. \end{aligned}$$

Thus $f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2} 1_{(0,\infty)}(y)$.

Note that the above example, the distribution of the random variable X can be divided into two intervals that are each strictly monotone. Many distributions require more than two intervals to ensure strict monotonicity for each interval.

Theorem 1.1.6. *(Probability integral transformation). Let X have continuous CDF $F_X(x)$ and define a random variable*

$$Y = F_X(X).$$

Then $Y \sim Unif(0, 1)$; that is, $F_Y(y) = \mathbb{P}(Y \leq y) = y,\ 0 \leq y \leq 1$.

Proof. Here we only prove a special case that $F_X(x)$ is invertible. Since $F_X(x)$ is increasing, so is its inverse $F_X^{-1}(x)$. Therefore we have for any $y \in [0, 1]$,

$$\begin{aligned} \mathbb{P}(Y \leq y) =& \mathbb{P}(F_X(X) \leq y) \\ =& \mathbb{P}(F_X^{-1}(F_X(X)) \leq F_X^{-1}(y)) \\ =& \mathbb{P}(X \leq F_X^{-1}(y)) \\ =& F_X(F_X^{-1}(y)) \\ =& y. \end{aligned}$$

□

One use of the probability integral transformation is when generating random samples. Suppose X has CDF $F_X(x)$. Theorem 2 tells us how to create a random sample of X:

(1) Generate a random sample $Y \sim \text{Unif}(0, 1)$.

(2) Then $X = F_X^{-1}(Y)$ has CDF $F_X(x)$.

For certain distributions (such as normal or gamma) there is a simple procedure to generate random samples directly, but the steps above can apply to any distribution. We will not dive further into random sampling in this class, but suffice to show that theorem 2 is particularly useful in cases such as this one.

1.2 Location and Scale Families

Theorem 1.2.1. *Let $f(x)$ be any PDF and let $\mu \in \mathbb{R}$, $\sigma > 0$. Given:*

$$g(x \mid \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma} \right),$$

If $X \sim f(x)$, then $Z = \sigma X + \mu$ has PDF $g(x \mid \mu, \sigma)$.

Proof. We first write Z as a function of X to get $X = \frac{z-u}{\sigma}$. Substituting this into the expression for the PDF results in:

$$f_Z(z) = \frac{1}{\sigma} f_X\left(\frac{z-\mu}{\sigma}\right)$$

This represents f(x) as part of a **location-scale family**, where μ determines the shift of the function, while σ is the scale parameter.

□

Example 1.2.2 (Normal Distribution)**.** Let

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad X \sim N(0,1).$$

Given our transformation $X = \frac{Z-\mu}{\sigma}$ from **Theorem 3.1**, we find that:

$$f_Z(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = g(x \mid \mu, \sigma),$$

This represents the PDF of $X \sim N(\mu, \sigma^2)$.

Example 1.2.3 (Location Exponential Distribution)**.** Let

$$f_X(x) = e^{-x} 1_{(0,\infty)}(x), \quad X \sim \text{Exp}(1).$$

Now consider the location family of

$$g(x \mid \mu) = f_x(x-u) = e^{-(x-\mu)} 1_{(0,\infty)}(x-\mu).$$

This differs **Example 3.2** as we only change the location so it is in the **location-family** of the original function while maintaining the scale. As a result:

$$\begin{cases} e^{-(x-\mu)}, & x > u \\ 0, & x < u \end{cases}$$

Observation: If $\sigma = 1$ and μ is the parameter, the family is a **location family**. If $\mu = 0$ and σ is the parameter, the family is a **scale family**.

Theorem 1.2.4. *Let Z have pdf $f_z(z)$. If X has a PDF $\frac{1}{\sigma} f_z\left(\frac{x-\mu}{\sigma}\right)$, then:*

$$\mathbb{E}[X] = \sigma\mathbb{E}[Z] + \mu, \qquad \text{Var}(X) = \sigma^2 \text{Var}(Z).$$

The theorem states that the expected value of X is σ times the expected value of Z, plus μ. The variance of X, however, does not depend on μ and is σ^2 times the variance of Z

Proof. First, we let X be a transformed variable $X = \sigma Z + \mu$ where both μ and σ are constants. The expectation of X is:

$$\mathbb{E}[X] = \mathbb{E}[\sigma Z + \mu]$$

Since the expectation is linear, we can separate $\mathbb{E}[\sigma Z + \mu]$ into $\sigma\mathbb{E}[Z] + \mathbb{E}[\mu]$. $E[\mu] = \mu$ as μ is a constant, resulting in:

$$\mathbb{E}[X] = \sigma\mathbb{E}[Z] + \mu,$$

We do the same for variance, where we get $\text{Var}[X] = \text{Var}[\sigma Z + \mu]$. We can move the scale σ outside by squaring it and constant μ doesn't affect variance so it can be taken out. Simplifying:

$$\text{Var}(X) = \sigma^2 \text{Var}[Z].$$

□

1.3 Bivariate Transformations

So far, we have covered how to find distributions of one-dimension, but what about two-dimensional R.V.s?

Let (X, Y) be a bivariate random vector, and (U, V) be a transformation using $U = g_1(X, Y)$ and $V = g_2(X, Y)$. The goal of this section is to calculate the distribution of (U, V) using (X, Y).

1.3.1 Discrete Case

Example 1.3.1 (Sum of Binomial distribution)**.** Let X_1, X_2 be independent with $X_1 \sim Bin(n_1, p), X_2 \sim Bin(n_2, p)$, and suppose $U = X_1 + X_2$. We know U is a discrete distribution with sample space

$$\mathcal{U} = \{0, 1, \dots n_1 + n_2\}.$$

For $u \in \mathcal{U}$, we have

$$\begin{aligned} f_U(u) =& \mathbb{P}(U = u) \\ =& \mathbb{P}(X_1 + X_2 = u) \\ =& \sum_{x_1 + x_2 = u} \mathbb{P}(X_1 = x_1, X_2 = x_2) \end{aligned}$$

To determine the range of x, note $x \in \{0, 1, \dots, n_1\}$ and $y \in \{0, 1, \dots, n_2\}$. Since $y = u - x$, we have $0 \le u - x \le n_2$, which implies $u - n_2 \le x \le u$. Combining these constraints with the original support:

$$\max\{0, u - n_2\} \le x \le \min\{n_1, u\}$$

Using independence:

$$\begin{aligned} f_U(u) &= \sum_{x=\max\{0,u-n_2\}}^{\min\{n_1,u\}} \binom{n_1}{x} p^x (1-p)^{n_1 - x} \binom{n_2}{u-x} p^{u-x} (1-p)^{n_2 - (u-x)} \\ &= p^u (1-p)^{n_1 + n_2 - u} \sum_{x=\max\{0,u-n_2\}}^{\min\{n_1,u\}} \binom{n_1}{x} \binom{n_2}{u-x} \end{aligned}$$

Applying the **VanderMonde Identity**, $\sum_x \binom{n_1}{x}\binom{n_2}{u-x} = \binom{n_1+n_2}{u}$. Thus:

$$f_U(u) = \binom{n_1 + n_2}{u} p^u (1-p)^{n_1 + n_2 - u}, \quad u \in \mathcal{U}$$

Therefore, $U \sim$ Binomial$(n_1 + n_2, p)$.

1.3.2 Continuous Case

Now we consider the case where (X, Y) are continuous random variables. Let $U = g_1(X, Y)$ and $V = g_2(X, Y)$. Assuming we can find the inverse functions $x = h_1(u, v)$ and $y = h_2(u, v)$, the joint pdf of (U, V) is:

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v)) \cdot |J|$$

where the Jacobian J is calculated as:

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}$$

Example 1.3.2. Let X, Y be independent with $X \sim \text{Gamma}(\alpha_1, \beta)$ and $Y \sim \text{Gamma}(\alpha_2, \beta)$. The marginal pdf for X is

$$\text{pdf}_X(x) = \frac{1}{\Gamma(\alpha_1)\beta^{\alpha_1}} x^{\alpha_1 - 1} e^{-x/\beta} \mathbf{1}_{\{x \in (0,\infty)\}}.$$

Similarly, the marginal pdf for Y is

$$\text{pdf}_Y(y) = \frac{1}{\Gamma(\alpha_2)\beta^{\alpha_2}} y^{\alpha_2 - 1} e^{-y/\beta} \mathbf{1}_{\{y \in (0,\infty)\}}.$$

Consider the transformation $U = \frac{X}{X+Y}$ and $V = X + Y$. The sample space for X, Y is $(0, \infty)$, resulting in $\mathcal{U} = (0, 1)$ and $\mathcal{V} = (0, \infty)$.

Finding the inverse transformation: Solving for X and Y in terms of U and V:

$$X = U \cdot (X + Y) = UV = h_1(u, v)$$
$$Y = V - X = V - UV = V(1 - U) = h_2(u, v)$$

Computing the Jacobian: We compute each partial derivative:

$$\frac{\partial x}{\partial u} = \frac{\partial (uv)}{\partial u} = v, \quad \frac{\partial x}{\partial v} = \frac{\partial (uv)}{\partial v} = u$$
$$\frac{\partial y}{\partial u} = \frac{\partial (v(1-u))}{\partial u} = -v, \quad \frac{\partial y}{\partial v} = \frac{\partial (v(1-u))}{\partial v} = 1 - u$$

Therefore, the Jacobian is:

$$J = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v(1-u) - (u)(-v) = v(1-u) + uv = v - uv + uv = v$$

Since $v > 0$ on the domain $\mathcal{V} = (0, \infty)$, we have $|J| = v$.

Finding the joint pdf: Using the transformation formula and independence of X and Y:

$$\begin{aligned} f_{U,V}(u, v) &= f_X(h_1(u, v)) \cdot f_Y(h_2(u, v)) \cdot |J| \\ &= f_X(uv) \cdot f_Y(v(1-u)) \cdot v \\ &= \left[\frac{1}{\Gamma(\alpha_1)\beta^{\alpha_1}} (uv)^{\alpha_1 - 1} e^{-uv/\beta} \right] \left[\frac{1}{\Gamma(\alpha_2)\beta^{\alpha_2}} (v(1-u))^{\alpha_2 - 1} e^{-v(1-u)/\beta} \right] \cdot v \end{aligned}$$

Expanding the terms:

$$f_{U,V}(u, v) = \frac{1}{\Gamma(\alpha_1)\beta^{\alpha_1}} \cdot \frac{1}{\Gamma(\alpha_2)\beta^{\alpha_2}} \cdot u^{\alpha_1 - 1} v^{\alpha_1 - 1} \cdot (1-u)^{\alpha_2 - 1} v^{\alpha_2 - 1} \cdot e^{-uv/\beta} \cdot e^{-v(1-u)/\beta} \cdot v$$

Combining the exponential terms: $e^{-uv/\beta} \cdot e^{-v(1-u)/\beta} = e^{-(uv + v - uv)/\beta} = e^{-v/\beta}$.

Combining the powers of v: $v^{\alpha_1 - 1} \cdot v^{\alpha_2 - 1} \cdot v = v^{\alpha_1 + \alpha_2 - 1}$.

Therefore:

$$f_{U,V}(u, v) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1 + \alpha_2}} u^{\alpha_1 - 1} (1-u)^{\alpha_2 - 1} v^{\alpha_1 + \alpha_2 - 1} e^{-v/\beta}$$

for $u \in (0, 1)$ and $v > 0$.

Observation (Independence of U and V): Notice that the joint pdf can be written as a product of two functions—one depending only on u and one depending only on v:

$$f_{U,V}(u, v) = \underbrace{\left(u^{\alpha_1 - 1} (1-u)^{\alpha_2 - 1} \right)}_{\text{function of } u \text{ only}} \cdot \underbrace{\left(\frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1 + \alpha_2}} v^{\alpha_1 + \alpha_2 - 1} e^{-v/\beta} \right)}_{\text{function of } v \text{ only}}$$

This factorization implies that U and V are **independent**.

Finding the marginal pdf of V:
Method 1: Notice that

$$f_V(v) \propto v^{\alpha_1+\alpha_2-1}e^{-v/\beta},$$

and by comparison to the pdf of Gamma$(\alpha_1+\alpha_2, \beta)$, we know:

$$f_V(v) = \frac{1}{\Gamma(\alpha_1+\alpha_2)\beta^{\alpha_1+\alpha_2}} v^{\alpha_1+\alpha_2-1}e^{-v/\beta}, \quad v > 0$$

Method 2: To find $f_V(v)$, we integrate $f_{U,V}(u,v)$ over u:

$$\begin{aligned} f_V(v) &= \int_0^1 f_{U,V}(u,v)\,du \\ &= \int_0^1 \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} u^{\alpha_1-1}(1-u)^{\alpha_2-1}v^{\alpha_1+\alpha_2-1}e^{-v/\beta}\,du \\ &= \frac{v^{\alpha_1+\alpha_2-1}e^{-v/\beta}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} \int_0^1 u^{\alpha_1-1}(1-u)^{\alpha_2-1}\,du \end{aligned}$$

The integral $\int_0^1 u^{\alpha_1-1}(1-u)^{\alpha_2-1}\,du$ is the **Beta function** $B(\alpha_1,\alpha_2)$, which satisfies:

$$B(\alpha_1,\alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$$

Substituting this result:

$$\begin{aligned} f_V(v) &= \frac{v^{\alpha_1+\alpha_2-1}e^{-v/\beta}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} \cdot \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)} \\ &= \frac{1}{\Gamma(\alpha_1+\alpha_2)\beta^{\alpha_1+\alpha_2}} v^{\alpha_1+\alpha_2-1}e^{-v/\beta}, \quad v > 0 \end{aligned}$$

This is the pdf of a Gamma$(\alpha_1+\alpha_2,\beta)$ distribution. Therefore, $V = X+Y \sim \text{Gamma}(\alpha_1+\alpha_2,\beta)$.

Finding the marginal pdf of U: Since U and V are independent, we can find $f_U(u)$ by dividing the joint pdf by $f_V(v)$:

$$\begin{aligned} f_U(u) &= \frac{f_{U,V}(u,v)}{f_V(v)} \\ &= \frac{\frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} u^{\alpha_1-1}(1-u)^{\alpha_2-1}v^{\alpha_1+\alpha_2-1}e^{-v/\beta}}{\frac{1}{\Gamma(\alpha_1+\alpha_2)\beta^{\alpha_1+\alpha_2}} v^{\alpha_1+\alpha_2-1}e^{-v/\beta}} \end{aligned}$$

Simplifying (the terms involving v, β, and the exponential cancel):

$$f_U(u) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} u^{\alpha_1-1}(1-u)^{\alpha_2-1}, \quad u \in (0,1)$$

This is the pdf of a Beta(α_1,α_2) distribution. Therefore, $U = \frac{X}{X+Y} \sim \text{Beta}(\alpha_1,\alpha_2)$.

1.3.3 Non-Invertible Mappings

What if the map is not invertible? Suppose there exists a partition $A_0, A_1, \ldots, A_k$ of the sample space such that the map on each partition A_i is invertible and $P(A_0) = 0$. This is similar to the one-dimensional case where we partition the domain to handle non-invertibility.

In this case, the joint pdf of (U, V) is the sum of the contributions from each partition:

$$f_{U,V}(u,v) = \sum_{i=1}^{k} f_{X,Y}(h_1^{(i)}(u,v), h_2^{(i)}(u,v)) \cdot |J_i|$$

where $h_1^{(i)}, h_2^{(i)}$ are the inverse functions on partition A_i and J_i is the corresponding Jacobian.

Example 1.3.3 (Distribution of the ratio of Normal random variables)**.** Consider independent standard Normal variables $X, Y \sim N(0,1)$ and the transformation $U = \frac{X}{Y}$, $V = |Y|$. We want to find the joint distribution of (U, V) and the marginal distribution of U.

The sample space for (U, V) is $\mathcal{U} = \mathbb{R}$ and $\mathcal{V} = (0, \infty)$.

We partition the original sample space into:

- $A_1 = \{(x, y) : y > 0\}$
- $A_2 = \{(x, y) : y < 0\}$
- $A_0 = \{(x, y) : y = 0\}$

Note that $P((X, Y) \in A_0) = P(Y = 0) = 0$ since Y is a continuous random variable.

On A_1 (where $y > 0$): Since $V = |Y| = Y$ when $y > 0$, we have:

$$\begin{aligned} Y = V \quad &\Rightarrow \quad y = v = h_2(u,v) \\ X = U \cdot Y = UV \quad &\Rightarrow \quad x = uv = h_1(u,v) \end{aligned}$$

Computing the partial derivatives:

$$\frac{\partial x}{\partial u} = \frac{\partial (uv)}{\partial u} = v, \quad \frac{\partial x}{\partial v} = \frac{\partial (uv)}{\partial v} = u$$

$$\frac{\partial y}{\partial u} = \frac{\partial v}{\partial u} = 0, \quad \frac{\partial y}{\partial v} = \frac{\partial v}{\partial v} = 1$$

The Jacobian is:

$$J_1 = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v \cdot 1 - u \cdot 0 = v$$

On A_2 (where $y < 0$): Since $V = |Y| = -Y$ when $y < 0$, we have $Y = -V$:

$$\begin{aligned} Y = -V \quad &\Rightarrow \quad y = -v = h_2'(u,v) \\ X = U \cdot Y = U(-V) = -UV \quad &\Rightarrow \quad x = -uv = h_1'(u,v) \end{aligned}$$

Computing the partial derivatives:

$$\frac{\partial x}{\partial u} = \frac{\partial (-uv)}{\partial u} = -v, \quad \frac{\partial x}{\partial v} = \frac{\partial (-uv)}{\partial v} = -u$$

$$\frac{\partial y}{\partial u} = \frac{\partial (-v)}{\partial u} = 0, \quad \frac{\partial y}{\partial v} = \frac{\partial (-v)}{\partial v} = -1$$

The Jacobian is:

$$J_2 = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} -v & -u \\ 0 & -1 \end{vmatrix} = (-v)(-1) - (-u)(0) = v$$

Note that on both partitions, $|J| = v$ (since $v > 0$).

Finding the joint pdf: Since X and Y are independent standard Normal, their joint pdf is:

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2} = \frac{1}{2\pi} e^{-(x^2+y^2)/2}$$

The joint pdf of (U, V) is the sum of contributions from both partitions:

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(uv, v) \cdot |J_1| + f_{X,Y}(-uv, -v) \cdot |J_2| \\ &= \frac{1}{2\pi} e^{-\frac{(uv)^2+v^2}{2}} \cdot v + \frac{1}{2\pi} e^{-\frac{(-uv)^2+(-v)^2}{2}} \cdot v \\ &= \frac{1}{2\pi} e^{-\frac{u^2v^2+v^2}{2}} \cdot v + \frac{1}{2\pi} e^{-\frac{u^2v^2+v^2}{2}} \cdot v \end{aligned}$$

Since $(-uv)^2 = u^2v^2$ and $(-v)^2 = v^2$, both terms are identical. Therefore:

$$f_{U,V}(u,v) = 2 \cdot \frac{1}{2\pi} v e^{-\frac{v^2(u^2+1)}{2}} = \frac{1}{\pi} v e^{-\frac{v^2(u^2+1)}{2}}, \quad u \in \mathbb{R},\ v > 0$$

Observation: Note that U and V are **not independent** because the joint pdf cannot be factored into a product of a function of u only and a function of v only. The term $e^{-\frac{v^2(u^2+1)}{2}}$ contains both u and v in the exponent in a way that cannot be separated.

Finding the marginal pdf of U: Since U and V are not independent, we must integrate out V:

$$f_U(u) = \int_0^\infty f_{U,V}(u,v)\, dv = \int_0^\infty \frac{1}{\pi} v e^{-\frac{v^2(u^2+1)}{2}}\, dv$$

We use the substitution $z = v^2$, so $dz = 2v\, dv$, which gives $v\, dv = \frac{1}{2} dz$. When $v = 0$, $z = 0$; when $v \to \infty$, $z \to \infty$. Also, $v = \sqrt{z}$.

Substituting:

$$\begin{aligned} f_U(u) &= \frac{1}{\pi} \int_0^\infty \sqrt{z}\, e^{-z(u^2+1)/2} \cdot \frac{1}{2\sqrt{z}}\, dz \\ &= \frac{1}{2\pi} \int_0^\infty e^{-\frac{(u^2+1)}{2} z}\, dz \\ &= \frac{1}{2\pi} \cdot \left(-\frac{2}{u^2+1} e^{-z(u^2+1)/2} \Big|_{z=0}^{\infty} \right) \\ &= \frac{1}{2\pi} \cdot \frac{2}{u^2+1} \\ &= \frac{1}{\pi(1+u^2)}, \quad u \in \mathbb{R} \end{aligned}$$

Therefore,

$$U \sim \text{Cauchy}(0, 1),$$

where the standard Cauchy distribution has PDF

$$f_U(u) = \frac{1}{\pi(1+u^2)}, \quad u \in \mathbb{R}.$$

Note that the Cauchy distribution does not have an expected value, as the integral used to compute it does not converge. For this example, we will not calculate the marginal distribution of V.

Chapter 2

Properties of a Random Sample

2.1 Random Samples

Definition 2.1.1. Random variables $X_1, X_2, \ldots, X_n$ are called **random samples of size n from the population** $f(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal PDF or PMF of each X_i is the same function $f(x)$. Note that we write $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$.

Definition 2.1.2. Let $X_1, \ldots, X_n$ be a random sample. Let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called a **statistic**. The distribution of this random variable Y is called a **sampling distribution** of Y.

In this lecture, we are particularly interested in studying the sample mean and sample variance.

Definition 2.1.3. The **sample mean** is

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Definition 2.1.4. The **sample variance** is

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

Definition 2.1.5. The **sample standard deviation** is

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

Note that all of the above definitions are functions of the random sample. Each statistic is a random variable.

2.2 Properties of Sample Mean and Sample Variance

Theorem 2.2.1. *Let $x_1, \ldots, x_n$ be any numbers and let*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Then:

(a)

$$\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

(b)

$$(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Proof. We prove both statements.

Proof of (a):

For any real number a, we write

$$\begin{aligned}\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}\left[(x_i - \bar{x}) + (\bar{x} - a)\right]^2 \\ &= \sum_{i=1}^{n}\Big((x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - a) + (\bar{x} - a)^2\Big) \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - a) + \sum_{i=1}^{n}(\bar{x} - a)^2.\end{aligned}$$

Consider the middle term:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - a) &= (\bar{x} - a)\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= (\bar{x} - a)\left(\sum_{i=1}^{n} x_i - n\bar{x}\right) \\ &= 0.\end{aligned}$$

Therefore,

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - a)^2. \tag{$*$}$$

Since it is a quadratic function of a, it is minimized when $a = \bar{x}$. Hence,

$$\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Proof of (b):

Taking $a = 0$ in $(*)$, we obtain

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

Rearranging terms gives

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2. \tag{$**$}$$

Since $(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$, the result follows. □

Lemma 2.2.2. *If X_1 and X_2 are independent random variables, then*

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2).$$

Lemma 2.2.3. *Let $X_1, \ldots, X_n$ be a random sample. Let $g(x)$ be a function such that $\mathbb{E}[g(X_1)]$ and $\mathrm{Var}(g(X_1))$ exist. Then*

$$\mathbb{E}\left(\sum_{i=1}^{n} g(X_i)\right) = n\,\mathbb{E}[g(X_1)],$$

and

$$\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\ \mathrm{Var}(g(X_1)).$$

Proof.

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} g(X_i)\right) &= \sum_{i=1}^{n} \mathbb{E}[g(X_i)] \\
&= \sum_{i=1}^{n} \mathbb{E}[g(X_1)] \quad \text{(identically distributed)} \\
&= n\,\mathbb{E}[g(X_1)].
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) &= \sum_{i=1}^{n} \mathrm{Var}(g(X_i)) \quad \text{(independence)} \\
&= \sum_{i=1}^{n} \mathrm{Var}(g(X_1)) \quad \text{(identically distributed)} \\
&= n\ \mathrm{Var}(g(X_1)).
\end{aligned}$$

□

Theorem 2.2.4. *Let $X_1, \ldots, X_n$ be a random sample with mean μ and variance $\sigma^2 < \infty$. Then:*

(a) $\mathbb{E}(\bar{X}) = \mu$.

(b) $\mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$.

(c) $\mathbb{E}(S^2) = \sigma^2$.

Proof. We prove each part.

Proof of (a): Recall that

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

Taking expectation and using linearity of expectation,

$$\begin{aligned}
\mathbb{E}(\bar{X}) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) \\
&= \frac{1}{n} \cdot n\mu \\
&= \mu.
\end{aligned}$$

Proof of (b): Using the definition of $\bar{X}$,

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right).\end{aligned}$$

By independence,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\mathrm{Var}(X_i) = n\sigma^2.$$

Therefore,

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

Proof of (c):
Recall

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

$$\begin{aligned}\mathbb{E}(S^2) &= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right) \\ &= \mathbb{E}\left[\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\right] \quad \text{by } (**) \\ &= \frac{1}{n-1}\left(\mathbb{E}\sum_{i=1}^{n} X_i^2 - n\,\mathbb{E}(\bar{X}^2)\right).\end{aligned}$$

By the lemma with $g(x) = x^2$,

$$\mathbb{E}\sum_{i=1}^{n} X_i^2 = n\,\mathbb{E}(X^2).$$

$$\mathbb{E}(\bar{X}^2) = \mathrm{Var}(\bar{X}) + (\mathbb{E}\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2, \quad \mathbb{E}(X^2) = \sigma^2 + \mu^2.$$

Therefore,

$$\begin{aligned}\mathbb{E}(S^2) &= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] \\ &= \frac{1}{n-1}(n-1)\sigma^2 \\ &= \sigma^2.\end{aligned}$$

□

Recall that

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

The use of $n-1$ in the definition of S^2 may have seemed unintuitive. Now we see that, with this definition, $ES^2 = \sigma^2$. If S^2 were defined as the usual average of the squared deviations with n rather than $n-1$ in the denominator, then ES^2 would be $\frac{n-1}{n}\sigma^2$.

2.2.1 Moment Generation Functions of sample mean

The **moment generating function (MGF)** of a random variable X is defined as:

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

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_1, \ldots, X_n$ are i.i.d. Then:

$$M_{\bar{X}}(t) = \mathbb{E}\left[e^{t\bar{X}}\right]$$

We can rewrite the exponent with $\bar{X}$:

$$\mathbb{E}\left[e^{t\bar{X}}\right] = \mathbb{E}\left[e^{\frac{t}{n}\sum_{i=1}^{n} X_i}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{\frac{t}{n}X_i}\right]$$

Because the X_i's are independent, we move the summation outside the expectation:

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

X_i is also identically distributed, with every expectation being the same. Simplifying:

$$M_{\bar{X}}(t) = \mathbb{E}\left[e^{\frac{t}{n}X_i}\right]^n = \left(M_X\left(\frac{t}{n}\right)\right)^n.$$

Example 2.2.5. Suppose $X_1, \ldots, X_n$ is a random sample (meaning i.i.d) from $N(\mu, \sigma^2)$. The MGF of a Normal Distribution is:

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

Substituting in for the formula to find $M_{\bar{X}}(t)$, we change t to $\frac{t}{n}$ to get:

$$M_{\bar{X}}(t) = (e^{\mu\frac{t}{n} + \frac{1}{2}\sigma^2(\frac{t}{n})^2})^n$$

Applying the outer n simplifies to:

$$M_{\bar{X}}(t) = e^{\mu t + \frac{\sigma^2 t^2}{2n}},$$

This identifies $\bar{X}$ as following normal distribution with mean μ and variance $\frac{\sigma^2}{n}$:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

2.2.2 Sum of independent random variables

Theorem 2.2.6. *Let X and Y be independent continuous random variables with p.d.f of $f_x(x)$ and $f_y(y)$, respectively. Given $Z = X + Y$, the PDF of Z is:*

$$f_z(z) = \int_{-\infty}^{\infty} f_x(w) f_y(z - w)\, dw.$$

Proof. Let $W = X$ and $Z = X + Y$ so our variables change from (X, Y) to (W, Z). The inverse function is $X = W$ and $Y = Z - W$. First, we solve for the Jacobian Determinant to adjust the PDF:

$$J = \begin{vmatrix} \frac{dx}{dw} & \frac{dx}{dz} \\ \frac{dy}{dw} & \frac{dy}{dz} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1(1) - (-1)(0) = 1$$

Since the Jacobian determinant is 1, the change in variables does not change the PDF. The joint PDF of W and Z is then: $f_{WZ}(w,z) = f_X(w)f_Y(z-w)$. To find $f_Z(z)$, integrate out w to get:

$$f_Z(z) = \int_{-\infty}^{\infty} f_x(w) f_y(z-w)\, dw.$$

□

Example 2.2.7 (Sum of Cauchy Random Variables)**.** Consider $U \sim \text{Cauchy}(0,\sigma)$ and $V \sim \text{Cauchy}(0,\tau)$. Their PDFs are:

$$f_U(u) = \frac{1}{\pi\sigma}\frac{1}{1+(u/\sigma)^2}, \quad f_V(v) = \frac{1}{\pi\tau}\frac{1}{1+(v/\tau)^2}.$$

Let Z = U + V. We want to determine the distribution of Z. From **Theorem 2.1**, we find the PDF as:

$$f_Z(z) = \int_{-\infty}^{\infty} f_U(w) f_V(z-w)\, dw = \int_{-\infty}^{\infty} \left(\frac{1}{\pi\sigma}\frac{1}{1+(\frac{w}{\sigma})^2}\right)\left(\frac{1}{\pi\tau}\frac{1}{1+(\frac{z-w}{\tau})^2}\right) dw$$

The result of the integral is:

$$\frac{1}{\pi(\sigma+\tau)}\frac{1}{1+(\frac{z}{\sigma+\tau})^2}$$

This simplifies to the PDF of Z, which has a Cauchy distribution with location 0 and scale $\sigma+\tau$:

$$Z \sim \text{Cauchy}(0, \sigma+\tau).$$

Now consider $X_1, \ldots, X_n \overset{iid}{\sim} \text{Cauchy}(0,1)$. Then by the above example, $\sum_{i=1}^n X_i \sim \text{Cauchy}(0,n)$. Then by simple calculation (or by the scale family density in the next topic) $\bar{X} \sim \text{Cauchy}(0,1)$. Recall that from last lecture, we learned that

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

if the random sample follows iid from a distribution with variance $\sigma^2 < \infty$. So the variance decreases when n increases. However, the Cauchy sample above has sample mean $\bar{X} \sim \text{Cauchy}(0,1)$ invariant to the sample size!

Observation: For Cauchy distribution, the $\mathbb{E}[X]$ and $\text{Var}(X)$ are both ∞. As a result, the general results

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

does not apply to Cauchy samples.

2.2.3 Location and Scale Families of Distributions

Given random variables

$$X_1, \ldots, X_n \overset{iid}{\sim} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right),$$

we want to determine the distribution of the sample mean, $\bar{X}$. Since this represents a location-scale family with location μ and scale σ,

$$X_i = \sigma Z_i + \mu,$$

where random variables $Z_1, \ldots, Z_n \overset{iid}{\sim} f(x)$. (Proven in lecture 5, Theorem 3.1.) Then,

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

This means that since Z_i's are part of the same location-scale family as X_i, we can just find the distribution of $\bar{Z}$ and use that to determine the distribution of $\bar{X}$:

1. Find the distribution of $\bar{Z}$, say with pdf $g(x)$.
2. Then the distribution of $\bar{X} = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right)$.

Example 2.2.8. Given $X_1, \ldots, X_n \overset{iid}{\sim} \text{Cauchy}(\mu, \sigma)$, find the distribution of $\bar{X}$.
Following the steps above,

1. Consider $Z_1, \ldots, Z_n \overset{iid}{\sim} \text{Cauchy}(0, 1)$.
   From example 2.2 in lecture 5, we know that

   $$\sum_{i=1}^{n} Z_i \sim \text{Cauchy}(0, n).$$

   Note that we had previously only shown that this was true when $\mu = 0$, but not general μ. Next, Confirm for yourself that

   $$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \sim \text{Cauchy}(0, 1).$$

2. $\bar{X}$ then follow from the location-scale family of the distribution of $\bar{Z}$ in step 1. So,

   $$\bar{X} \sim \text{Cauchy}(\mu, \sigma).$$

2.3 Sampling from the Normal Distribution

2.3.1 χ^2 distribution

Recall If $X_1, \ldots, X_n \overset{iid}{\sim} f(x)$ with mean μ and finite variance σ^2, then

1. $\mathbb{E}\left[\bar{X}\right] = \mu$.
2. $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$.
3. $\mathbb{E}\left[S^2\right] = \sigma^2$.

This is true for any pdf $f(x)$. If $f(x)$ is Normal, what else can we conclude about $\bar{X}$ and S^2?

Theorem 2.3.1. *If $X_1, \ldots, X_n$ are a random sample from $N\left(\mu, \sigma^2\right)$, then:*

a) $\bar{X}$ and S^2 are independent random variables.

b) $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

c) $\dfrac{(n-1)\,S^2}{\sigma^2}$ *has a chi-squared distribution with* $n-1$ *degrees of freedom.*

Proof. a)

$$\begin{aligned} S^2 &\stackrel{\text{def}}{=} \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \\ &= \frac{1}{n-1}\left[\left(X_1 - \bar{X}\right)^2 + \sum_{i=2}^{n}\left(X_i - \bar{X}\right)^2\right] \end{aligned}$$

Since

$$\begin{aligned} \left(X_i - \bar{X}\right) + \sum_{i=2}^{n}\left(X_i - \bar{X}\right) &= \sum_{i=1}^{n}\left(X_i - \bar{X}\right) \\ &= \sum_{i=1}^{n} X_i - n\bar{X} \\ &= 0, \end{aligned}$$

we can say

$$S^2 = \frac{1}{n-1}\left[\left[-\sum_{i=2}^{n}\left(X_i - \bar{X}\right)\right]^2 + \sum_{i=2}^{n}\left(X_i - \bar{X}\right)^2\right]$$

This shows that S^2 can be written as a function of the vector $\left(X_2 - \bar{X}, \ldots, X_n - \bar{X}\right)$. So, to prove that S^2 is independent of $\bar{X}$, it will suffice to show that $\left(X_2 - \bar{X}, \ldots, X_n - \bar{X}\right)$ is independent of $\bar{X}$.

Consider a new random variable where

$$\begin{aligned} y_1 &= \bar{x} \\ y_2 &= x_2 - \bar{x} \\ &\vdots \\ y_n &= x_n - \bar{x} \end{aligned}$$

It is easy to see that

$$\begin{aligned} x_2 &= y_1 - y_2 \\ &\vdots \\ x_n &= y_1 - y_n. \end{aligned}$$

Then

$$\begin{aligned} x_1 &= ny_1 - \sum_{i=2}^{n} x_i \\ &= ny_i - \sum_{i=2}^{n}(y_1 + y_i) \\ &= y_1 - \sum_{i=2}^{n} y_i. \end{aligned}$$

Thus the inverses would be

$$\begin{aligned} x_1 &= y_1 - \sum_{i=2}^{n} y_i \\ x_2 &= y_1 - y_2 \\ &\vdots \\ x_n &= y_1 - y_n \end{aligned}$$

Taking the partial derivatives gives us

$$\begin{aligned} J &= \begin{vmatrix} 1 & -1 & -1 & -1 & \cdots & -1 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 1 & 0 & \cdots & \cdots & 0 & 1 \end{vmatrix} \\ &= \begin{vmatrix} n & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 1 & 0 & \cdots & \cdots & 0 & 1 \end{vmatrix} \quad \text{(by adding the } i\text{-th row to the first row for } i = 2, \ldots, n) \\ &= n \end{aligned}$$

By the change of variables formula,

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = f_{X_1,\ldots,X_n}\left(y_1 - \sum_{i=2}^{n} y_i, y_1 + y_2, \ldots, y_1 + y_n\right) n$$

The joint pdf of $(X_1, \ldots, X_n)$ is

$$f(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

We can write

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \mu)^2 &= (y_1 - \sum_{i=2}^{n} y_i - \mu)^2 + \sum_{i=2}^{n}(y_1 + y_i - \mu)^2 \\ &= (y_1 - \mu)^2 - 2(y_1 - \mu)\sum_{i=2}^{n} y_i + \left(\sum_{i=2}^{n} y_i\right)^2 \\ &\quad + \sum_{i=2}^{n}\left[(y_1 - \mu)^2 + 2(y_1 - \mu)y_i + y_i^2\right] \\ &= n(y_1 - \mu)^2 + \left(\sum_{i=2}^{n} y_i\right)^2 + \sum_{i=2}^{n} y_i^2 \end{aligned}$$

Substituting this into the joint pdf gives us

$$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) = n\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{n(y_1-\mu)^2}{2\sigma^2}\right)\exp\left(-\frac{1}{2\sigma^2}\left[\left(\sum_{i=2}^n y_i\right)^2 + \sum_{i=2}^n y_i^2\right]\right)$$

This factorization shows that $Y_1 = \bar{X}$ is independent of $(Y_2,\ldots,Y_n) = (X_2-\bar{X},\ldots,X_n-\bar{X})$, and hence $\bar{X}$ is independent of S^2.

Remark 2.3.2. The independence of $\bar{X}$ and S^2 is a special property of the normal distribution. In general, for non-normal populations, the sample mean and sample variance are not independent.

b) From the factorization above, the marginal pdf of $\bar{X}$ is

$$f_{Y_1}(y_1) \propto \exp\left(-\frac{n(y_1-\mu)^2}{2\sigma^2}\right),$$

which is the pdf of a normal distribution with mean μ and variance σ^2/n. Therefore,

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

□

For part (c), we will first need to introduce the chi-squared distribution. The chi-squared distribution with p degrees of freedom has the pdf

$$f(x) = \frac{1}{\Gamma(\frac{p}{2})2^{\frac{p}{2}}} x^{\frac{p}{2}-1}e^{-\frac{x}{2}}\mathbf{1}_{(0,\infty)}(x)$$

Note that this is a special case of the Gamma distribution with shape $\frac{p}{2}$ and scale 2. We will use the notation χ^2_p to denote the chi-squared distribution with p degrees of freedom.

Lemma 2.3.3. *Facts about chi squared random variables*

a) If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$. (See lecture 1, example 2.2 for proof)

b) If $X_1,\ldots,X_n$ are independent and $X_i \sim \chi^2_{p_i}$, then $\sum_{i=1}^n X_i \sim \chi^2_{\sum_{i=1}^n p_i}$

Proof. b) Since $X_i \sim \chi^2_{p_i}$, $X_i \sim \text{Gamma}(\frac{p_i}{2}, 2)$. By example 3.1 in lecture 3, we know

$$\sum_{i=1}^n X_i \sim \text{Gamma}\left(\frac{\sum_{i=1}^n p_i}{2}, 2\right),$$

which is equivalent to $\chi^2_{\sum_{i=1}^n p_i}$. □

This lemma will be used in the proof of part c in the next lecture. We will also use the following lemma in the proof of part (c).

Lemma 2.3.4. *Let $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1,\ldots,n$, and suppose X_i and X_j are independent for every $i \neq j$. Then*

$$\sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i\mu_i,\ \sum_{i=1}^n a_i^2\sigma_i^2\right).$$

Proof. By moment generating function. □

Proof of part (c) by mathematical induction. We use the identity (which will be a homework problem)

$$(n-1)S_n^2 = (n-2)\,S_{n-1}^2 + \frac{n-1}{n}\left(X_n - \bar{X}_{n-1}\right)^2. \tag{2.1}$$

Base case ($n = 2$): By (2.1),

$$S_2^2 = \frac{1}{2}(X_2 - X_1)^2 = \left[\frac{1}{\sqrt{2}}(X_2 - X_1)\right]^2.$$

Then

$$\frac{S_2^2}{\sigma^2} = \left[\frac{1}{\sqrt{2}\,\sigma}(X_2 - X_1)\right]^2.$$

Since $X_1, X_2 \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, by Lemma 2.3.4 we have

$$\frac{1}{\sqrt{2}\,\sigma}(X_2 - X_1) \sim N\left(0,\ \left(\frac{1}{\sqrt{2}\,\sigma}\right)^2\sigma^2 + \left(\frac{1}{\sqrt{2}\,\sigma}\right)^2\sigma^2\right) = N(0,1).$$

Therefore $\dfrac{S_2^2}{\sigma^2} \sim \chi_1^2$, which verifies the claim when $n = 2$.

Inductive step: Assume that $\dfrac{(k-1)\,S_k^2}{\sigma^2} \sim \chi_{k-1}^2$ holds. We want to show that $\dfrac{k\,S_{k+1}^2}{\sigma^2} \sim \chi_k^2$.

By (2.1) with $n = k+1$,

$$\frac{k\,S_{k+1}^2}{\sigma^2} = \frac{(k-1)\,S_k^2}{\sigma^2} + \frac{k}{k+1}\cdot\frac{\left(X_{k+1} - \bar{X}_k\right)^2}{\sigma^2}.$$

By the induction hypothesis, $\dfrac{(k-1)\,S_k^2}{\sigma^2} \sim \chi_{k-1}^2$.

Next, consider

$$\frac{X_{k+1} - \bar{X}_k}{\sigma} = \frac{1}{\sigma}\left(X_{k+1} - \frac{1}{k}\sum_{i=1}^{k} X_i\right).$$

Since $X_1, \ldots, X_{k+1} \overset{iid}{\sim} N(0,1)$, by Lemma 2.3.4,

$$\frac{X_{k+1} - \bar{X}_k}{\sigma} \sim N\left(0,\ \left(\frac{1}{\sigma}\right)^2\sigma^2 + \left(\frac{1}{k\sigma}\right)^2 \cdot k \cdot \sigma^2\right) = N\left(0,\ 1 + \frac{1}{k}\right) = N\left(0,\ \frac{k+1}{k}\right).$$

Therefore, defining

$$Y = \sqrt{\frac{k}{k+1}}\cdot\frac{X_{k+1} - \bar{X}_k}{\sigma} \sim N(0,1),$$

we have $Y^2 \sim \chi_1^2$.

Now, S_k^2 is independent of X_{k+1} (since S_k^2 depends only on $X_1, \ldots, X_k$), and S_k^2 is independent of $\bar{X}_k$ by part (a). It follows that S_k^2 is independent of Y. Therefore,

$$\frac{k\,S_{k+1}^2}{\sigma^2} = \frac{(k-1)\,S_k^2}{\sigma^2} + Y^2 \sim \chi_k^2.$$

□

2.3.2 Student's t Distribution

By the results of the previous section, if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Consider the statistic

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{(\bar{X}_n - \mu)/(\sigma/\sqrt{n})}{S_n/\sigma} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sqrt{n-1}}{\sqrt{(n-1)S_n^2/\sigma^2}} \overset{d}{=} \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}}, \tag{2.2}$$

where $\overset{d}{=}$ means the random variables on both sides have the same distribution.

Definition 2.3.5. A random variable T has *Student's t distribution* with p degrees of freedom, and we write $T \sim t_p$, if it has pdf

$$f_T(t) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \cdot \frac{1}{\sqrt{p\pi}} \cdot \frac{1}{\left(1 + \frac{t^2}{p}\right)^{\frac{p+1}{2}}}.$$

Remark 2.3.6. If $p = 1$, then

$$f_T(t) = \frac{\Gamma(1)}{\Gamma\left(\frac{1}{2}\right)} \cdot \frac{1}{\sqrt{\pi}} \cdot \frac{1}{1+t^2} = \frac{1}{\pi} \cdot \frac{1}{1+t^2},$$

which is the Cauchy$(0,1)$ distribution.

Theorem 2.3.7. *If $U \sim N(0,1)$ and $V \sim \chi^2_p$, and U, V are independent, then*

$$\frac{U}{\sqrt{V/p}} \sim t_p.$$

Proof. Define the transformation $T = \dfrac{U}{\sqrt{V/p}}$ and $W = V$. Then

$$U = T \cdot \sqrt{\frac{W}{p}}, \qquad V = W.$$

The Jacobian is

$$J = \begin{vmatrix} \dfrac{\partial U}{\partial T} & \dfrac{\partial U}{\partial W} \\ \dfrac{\partial V}{\partial T} & \dfrac{\partial V}{\partial W} \end{vmatrix} = \begin{vmatrix} \sqrt{\dfrac{W}{p}} & \dfrac{T}{p} \cdot \dfrac{1}{2\sqrt{W}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{W}{p}}.$$

The joint density of (T, W) is

$$\begin{aligned} f_{T,W}(t,w) &= f_{U,V}\left(t\sqrt{\frac{w}{p}},\ w\right) \cdot \sqrt{\frac{w}{p}} \\ &= f_U\left(t\sqrt{\frac{w}{p}}\right) \cdot f_V(w) \cdot \sqrt{\frac{w}{p}} \\ &= \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}t^2\frac{w}{p}} \cdot \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2}}\, w^{\frac{p}{2}-1}\, e^{-\frac{w}{2}} \cdot \sqrt{\frac{w}{p}} \cdot 1_{(0,\infty)}(w). \end{aligned}$$

To find the marginal density of T, integrate over w:

$$f_T(t) = \int_0^\infty \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2}} \cdot \sqrt{\frac{1}{p}} \cdot w^{\frac{p}{2}-\frac{1}{2}} \cdot e^{-\frac{w}{2}\left(1+\frac{t^2}{p}\right)}\, dw.$$

The integral $I = \displaystyle\int_0^\infty w^{\frac{p}{2}-\frac{1}{2}}\, e^{-\frac{w}{2}\left(1+\frac{t^2}{p}\right)}\, dw$ can be evaluated by recognizing it as related to the Gamma$\left(\frac{p+1}{2},\ \frac{2}{1+t^2/p}\right)$ distribution. Specifically,

$$I = \Gamma\left(\frac{p+1}{2}\right) \cdot \left(\frac{2}{1+t^2/p}\right)^{\frac{p+1}{2}}.$$

Substituting back,

$$\begin{aligned} f_T(t) &= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2}\sqrt{p}} \cdot \Gamma\left(\frac{p+1}{2}\right) \cdot \frac{2^{\frac{p+1}{2}}}{\left(1+\frac{t^2}{p}\right)^{\frac{p+1}{2}}} \\ &= \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \cdot \frac{1}{\sqrt{p\pi}} \cdot \frac{1}{\left(1+\frac{t^2}{p}\right)^{\frac{p+1}{2}}}. \qquad \square \end{aligned}$$

□

By (2.2) and Theorem 2.3.7, we have

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \overset{d}{=} \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} \sim t_{n-1}.$$

Properties of the t distribution (which will be homework problem):

- $\mathbb{E}[T_p] = 0$ if $p > 1$.
- $\mathrm{Var}(T_p) = \dfrac{p}{p-2}$ if $p > 2$.

2.3.3 Snedecor's F Distribution (Variance Ratio Distribution)

Definition 2.3.8. Let $X_1, \ldots, X_n \sim N(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma_Y^2)$ be two independent samples. Consider the ratio

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = \frac{(n-1)S_X^2/\sigma_X^2/(n-1)}{(m-1)S_Y^2/\sigma_Y^2/(m-1)} \overset{d}{=} \frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)}.$$

This has a *Snedecor's F distribution* with $n-1$ and $m-1$ degrees of freedom.

Definition 2.3.9 (Equivalent definition)**.** A random variable has an F distribution with p and q degrees of freedom if its pdf is

$$f_F(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} \cdot \left(\frac{p}{q}\right)^{p/2} \cdot \frac{x^{\frac{p}{2}-1}}{\left(1+\frac{p}{q}x\right)^{\frac{p+q}{2}}} \cdot 1_{(0,\infty)}(x).$$

The equivalence of the above two definitions will be left as a homework.

2.4 Order Statistics

Two statistics we have encountered are the sample mean and sample variance. We now introduce a useful class of statistics: order statistics.

Definition 2.4.1. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (iid) random variables. We define

$$X_{(1)} = \min_{1\le i\le n} X_i, \quad X_{(2)} = \text{second smallest } X_i, \ \ldots\ , \quad X_{(n)} = \max_{1\le i\le n} X_i.$$

Each $X_{(i)}$ is an *order statistic*.

Notice that each $X_{(i)}$ is a random variable.

Definition 2.4.2. Let $X_1, X_2, \ldots, X_n$ be iid. We define the *sample range* to be

$$R = X_{(n)} - X_{(1)}.$$

The sample range is the distance between the largest observation and the smallest observation.

Definition 2.4.3. Let $X_1, X_2, \ldots, X_n$ be iid. We define the *sample median* to be

$$M = \begin{cases} X_{(\frac{n+1}{2})} & \text{if n is odd} \\ \\ \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2} & \text{if n is even.} \end{cases}$$

Example 2.4.4 (Distribution of $X_{(n)}$)**.** Let $X_1, X_2, \ldots, X_n$ be iid with cdf $F(x)$. Then

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_{(n)} \le x) \\ &= P\left(\max_{1\le i\le n} X_i \le x\right) \\ &= P(X_1 \le x,\ X_2 \le x,\ \ldots\ , X_n \le x) \\ &= \big(P(X_1 \le x)\big)^n && (X_1, X_2, \ldots, X_n \text{ are iid}) \\ &= \big(F(x)\big)^n. \end{aligned}$$

If, additionally, $X_1, X_2, \ldots, X_n$ are continuous random variables with pdf $f(x)$, then

$$f_{X_{(n)}}(x) = F'_{X_{(n)}}(x) = n\big(F(x)\big)^{n-1} f(x),$$

where the last equality follows from the chain rule.

Example 2.4.5 (Distribution of $X_{(1)}$)**.** Let $X_1, X_2, \ldots, X_n$ be iid with cdf $F(x)$. Then

$$\begin{aligned} F_{X_{(1)}}(x) &= P(X_{(1)} \le x) \\ &= P\left(\min_{1\le i\le n} X_i \le x\right) \\ &= 1 - P\left(\min_{1\le i\le n} X_i > x\right) \\ &= 1 - P(X_1 > x,\ X_2 > x,\ \ldots\ , X_n > x) \\ &= 1 - \big(P(X_1 > x)\big)^n && (X_1, X_2, \ldots, X_n \text{ are iid}) \\ &= 1 - \big(1 - F(x)\big)^n. \end{aligned}$$

If, additionally, $X_1, X_2, \ldots, X_n$ are continuous random variables with pdf $f(x)$, then

$$f_{X_{(1)}}(x) = F'_{X_{(1)}}(x) = -n\big(1 - F(x)\big)^{n-1}\big(-f(x)\big) = n\big(1 - F(x)\big)^{n-1} f(x).$$

Deriving the distributions of the other order statistics is more complicated.

Theorem 2.4.6. *Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the order statistics of a random sample $X_1, X_2, \ldots, X_n$ from a continuous distribution with cdf $F_X(x)$ and pdf $f_X(x)$. Then the pdf of $X_{(j)}$ is*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x)\big[F_X(x)\big]^{j-1}\big[1-F_X(x)\big]^{n-j}$$

and the cdf of $X_{(j)}$ is

$$F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} \big[F_X(x)\big]^k \big[1-F_X(x)\big]^{n-k}.$$

Proof. Let Y be a random variable that counts how many of $X_1, X_2, \ldots, X_n$ are less than or equal to x. Then, defining a "success" as the event $\{X_i \le x\}$, we see that $Y \sim \text{binomial}\big(n, F_X(x)\big)$. So,

$$\begin{aligned} F_{X_{(j)}}(x) &= P(X_{(j)} \le x) \\ &= \sum_{k=j}^{n} P(\text{exactly } k \text{ of the } X_i \text{ are less than or equal to x}) \\ &= \sum_{k=j}^{n} P(Y=k) \\ &= \sum_{k=j}^{n} \binom{n}{k} \big[F_X(x)\big]^k \big[1-F_X(x)\big]^{n-k}. \end{aligned}$$

Now we find the pdf of $X_{(j)}$:

$$\begin{aligned} f_{X_{(j)}}(x) &= \frac{d}{dx} F_{X_{(j)}}(x) \\ &= \sum_{k=j}^{n} \binom{n}{k} \Big(k\big[F_X(x)\big]^{k-1} f_X(x) \cdot \big[1-F_X(x)\big]^{n-k} + \big[F_X(x)\big]^k \cdot (n-k)\big[1-F_X(x)\big]^{n-k-1}\big(-f_X(x)\big) \Big) \end{aligned}$$

where we have applied the product rule. Splitting the expression into two sums and rearranging factors, we get

$$\sum_{k=j}^{n} \binom{n}{k} k \cdot f_X(x)\big[F_X(x)\big]^{k-1}\big[1-F_X(x)\big]^{n-k} - \sum_{k=j}^{n-1} \binom{n}{k} (n-k) f_X(x)\big[F_X(x)\big]^k\big[1-F_X(x)\big]^{n-k-1}$$

where we have dropped the $k=n$ term of the second sum because the $(n-k)$ factor causes the term to vanish. Now we re-index both sums: for the first sum, we use $s = k-1$, and for the second sum, we use $s = k$. We get

$$\sum_{s=j-1}^{n-1} \binom{n}{s+1} (s+1) f_X(x)\big[F_X(x)\big]^s\big[1-F_X(x)\big]^{n-s-1} - \sum_{s=j}^{n-1} \binom{n}{s} (n-s) f_X(x)\big[F_X(x)\big]^s\big[1-F_X(x)\big]^{n-s-1}.$$

Now, since

$$\binom{n}{s+1}(s+1) = \frac{n!}{(s+1)!(n-s-1)!}(s+1) = \frac{n!}{s!(n-s-1)!} = \frac{n!}{s!(n-s)!}(n-s) = \binom{n}{s}(n-s),$$

our expression equals

$$\sum_{s=j-1}^{n-1} \binom{n}{s+1}(s+1)f_X(x)\big[F_X(x)\big]^s\big[1-F_X(x)\big]^{n-s-1} - \sum_{s=j}^{n-1} \binom{n}{s+1}(s+1)f_X(x)\big[F_X(x)\big]^s\big[1-F_X(x)\big]^{n-s-1}.$$

All terms cancel except for the $s = j-1$ term in the first sum, so we get

$$\begin{aligned} f_{X_{(j)}}(x) &= \binom{n}{j} j \cdot f_X(x)\big[F_X(x)\big]^{j-1}\big[1-F_X(x)\big]^{n-j} \\ &= \frac{n!}{(j-1)!(n-j)!} f_X(x)\big[F_X(x)\big]^{j-1}\big[1-F_X(x)\big]^{n-j}. \end{aligned}$$

□

Exercise 2.4.7. Check that the formulas given in Theorem 2.4.6 match what we proved earlier for the cases $j = 1$ and $j = n$.

Exercise 2.4.8. Does the CDF formula in the above theorem applies to discrete $X_1, \ldots, X_n$? Why?

Remark 2.4.9. A result similar to Theorem 2.4.6 for discrete distributions is given by Theorem 5.4.3 in [1].

Example 2.4.10 (Uniform order statistic pdf)**.** Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{uniform}(0,1)$. Then each X_i has pdf $f_X(x) = 1_{(0,1)}(x)$ and cdf

$$F_X(x) = \begin{cases} 0, & x \le 0 \\ x, & x \in (0,1) \\ 1, & x \ge 1. \end{cases}$$

By the previous theorem,

$$\begin{aligned} f_{X_{(j)}}(x) &= \frac{n!}{(j-1)!(n-j)!} 1_{(0,1)}(x)\big[F_X(x)\big]^{j-1}\big[1-F_X(x)\big]^{n-j} \\ &= \frac{n!}{(j-1)!(n-j)!} 1_{(0,1)}(x) \cdot x^{j-1}(1-x)^{n-j} \end{aligned}$$

where we were able to replace $F_X(x)$ with x because the $1_{(0,1)}(x)$ factor makes $f_{X_{(j)}}(x)$ vanish outside of $(0,1)$. Thus, the j^{th} order statistic has a beta$(j, n-j+1)$ distribution. It follows that $\mathbb{E}X_{(j)} = \frac{j}{n+1}$ and $\text{Var } X_{(j)} = \frac{j(n-j+1)}{(n+1)^2(n+2)}$. We observe that the expectations of the order statistics are evenly spaced: $\mathbb{E}X_{(1)} = \frac{1}{n+1}$, $\mathbb{E}X_{(2)} = \frac{2}{n+1}$, $\ldots$, $\mathbb{E}X_{(n)} = \frac{n}{n+1}$.

Theorem 2.4.11. *Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the order statistics of a random sample $X_1, X_2, \ldots, X_n$ from a continuous distribution with cdf $F_X(x)$ and pdf $f_X(x)$. Then the joint pdf of $X_{(i)}$ and $X_{(j)}$, where $1 \le i < j \le n$, is*

$$f_{X_{(i)},X_{(j)}}(u,v) = \begin{cases} \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u)f_X(v)\big[F_X(u)\big]^{i-1}\big[F_X(v)-F_X(u)\big]^{j-1-i}\big[1-F_X(v)\big]^{n-j}, & u < v \\ 0, & \textit{otherwise.} \end{cases}$$

The proof of the above theorem, not given here, is similar to that of Theorem 2.4.6; it uses a multinomial random variable rather than a binomial one.

Theorem 2.4.12. *Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the order statistics of a random sample $X_1, X_2, \ldots, X_n$ from a continuous distribution with cdf $F_X(x)$ and pdf $f_X(x)$. Then the joint pdf of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is*

$$f_{X_{(1)},\ldots,X_{(n)}}(y_1, ..., y_n) = \begin{cases} n! \cdot f_X(y_1) \cdots f_X(y_n), & y_1 < \cdots < y_n \\ 0, & \textit{otherwise.} \end{cases}$$

Notice, the signs defining the indicator function's preferred domain are $<$, not $\leq$. This is because when X is a *continuous* r.v. and hence

$$P[X_i = X_j] = 0,$$

so we can also say there is a 0 probability for any of the random variables to be equal to each other.

Proof. The following is the proof for the above theorem for the $n = 2$ case.

$$X_{(1)} = \min(X_1, X_2) \qquad \text{and} \qquad X_{(2)} = \max(X_1, X_2),$$

We want the inverse map so that we can write X_1 and X_2 in terms of $X_{(1)}$ and $X_{(1)}$. To achieve this, we can consider the two possible cases for $X_{(1)}$ and $X_{(2)}$. The first, A_1, where $X_1 < X_2$, and the second, A_2, where $X_1 > X_2$.

$$A_1 = \{(x_1, x_2) : x_1 < x_2\}.$$

Then on A_1, the map is

$$A_1 : \begin{cases} y_1 = x_1 \\ y_2 = x_2 \end{cases} \Rightarrow \begin{cases} x_1 = y_1 \\ x_2 = y_2 \end{cases}$$

Compute the Jacobian:

$$J_1 = \det \begin{pmatrix} \frac{\partial x_1}{\partial x_1} & \frac{\partial x_1}{\partial x_2} \\ \frac{\partial x_2}{\partial x_1} & \frac{\partial x_2}{\partial x_2} \end{pmatrix} = \det \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 1, \qquad |J_1| = 1.$$

Also,

$$A_2 = \{(x_1, x_2) : x_1 > x_2\}.$$

Then on A_2, the map is

$$A_2 : \begin{cases} y_1 = x_2 \\ y_2 = x_1 \end{cases} \Rightarrow \begin{cases} x_2 = y_1 \\ x_1 = y_2 \end{cases}$$

Compute the Jacobian:

$$J_2 = \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix} = \det \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = -1, \qquad |J_2| = 1.$$

Note also our third case, $A_0 = \{x_1 = x_2\}$, but recall again that for continuous r.v. the possibility of X_1 and X_2 being the same value is 0, so

$$\begin{aligned} f_{X_{(1)},X_{(2)}}(y_1, y_2) &= (f_{X_1,X_2}(y_1, y_2)|J_1| + f_{X_1,X_2}(y_2, y_1)|J_2|)\mathbf{1}_{\{y_1<y_2\}} \\ &= (f_X(y_1)f_X(y_2) + f_X(y_2)f_X(y_1))\mathbf{1}_{\{y_1<y_2\}} \\ &= 2f_X(y_1)f_X(y_2)\mathbf{1}_{\{y_1<y_2\}} \\ &= n!f_X(y_1)...f_X(y_n)\mathbf{1}_{\{y_1<\ldots<y_n\}} \text{ for n = 2} \end{aligned}$$

Remark. For general n, the split subsets are

$$\{x_1 < x_2 < ... < x_n\}$$
$$\{x_1 < x_2 < ... < x_n < x_{n+1}\}$$
$$.$$
$$.$$
$$.$$

There are totally $n!$ subsets (excluding the zero probability sets). The Jacobian will always be positive or negative 1 on each such subset. Thus the joint pdf will simplify to:

$$n! f_X(x_1)...f_X(x_n), \qquad \text{for any n}$$

Example 2.4.13. Consider the following continuous, independent, and identically distributed random variables,

$$X_1, X_2, X_3 \overset{iid}{\sim} \text{Exp}(1), \qquad f_{X_i}(x) = e^{-x}\mathbf{1}_{\{0<x<\infty\}}$$

Find the distribution of $Y = (X_{(1)}, X_{(2)}, X_{(3)})$,

$$\begin{aligned}\forall x_1 < x_2 < x_3, \quad f_{X_{(1)},X_{(2)},X_{(3)}}(x_1, x_2, x_3) &= 3! e^{-x_1}\mathbf{1}_{\{0<x_1<\infty\}} \\ &\quad \times e^{-x_2}\mathbf{1}_{\{0<x_2<\infty\}} \\ &\quad \times e^{-x_3}\mathbf{1}_{\{0<x_3<\infty\}} \\ &= 6e^{-(x_1+x_2+x_3)}\mathbf{1}_{\{0<x_1<x_2<x_3<\infty\}}\end{aligned}$$

Example 2.4.14. the distribution of midrange and range

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, a)$. Let $X_{(1)} = \min\{X_i\}$ and $X_{(n)} = \max\{X_i\}$. Define

$$\text{Range} \qquad R = X_{(n)} - X_{(1)}, \qquad \text{Midrange} \qquad V = \frac{X_{(1)} + X_{(n)}}{2}.$$

We want to find the joint density of $(X_{(1)}, X_{(n)})$

By the order-statistics result from the previous lecture,

$$\begin{aligned}f_{X_{(1)},X_{(n)}}(x_1, x_n) &= \frac{n(n-1)}{a^2}\left(\frac{x_n - x_1}{a}\right)^{n-2}\mathbf{1}_{\{0<x_1<x_n<a\}} \\ &= \frac{n(n-1)}{a^n}(x_n - x_1)^{n-2}\mathbf{1}_{\{0<x_1<x_n<a\}}.\end{aligned}$$

Change of variables to (R, V): from the definitions,

$$X_{(1)} = V - \frac{R}{2}, \qquad X_{(n)} = V + \frac{R}{2}.$$

The constraints $0 < X_{(1)} < X_{(n)} < a$ become

$$0 < V - \frac{R}{2} < V + \frac{R}{2} < a \quad \Rightarrow \begin{cases} 0 < V - \frac{R}{2}, \\ V - \frac{R}{2} < V + \frac{R}{2}, \\ V + \frac{R}{2} < a \end{cases} \quad \Rightarrow \begin{cases} \frac{R}{2} < V, \\ 0 < \frac{R}{2}, \\ V < a - \frac{R}{2} \end{cases} \quad \Rightarrow \begin{cases} 0 < R < a, \\ \frac{R}{2} < V < a - \frac{R}{2} \end{cases}$$

So, the support set is

$$A = \left\{(r, v) : 0 < r < a, \ \frac{r}{2} < v < a - \frac{r}{2}\right\}.$$

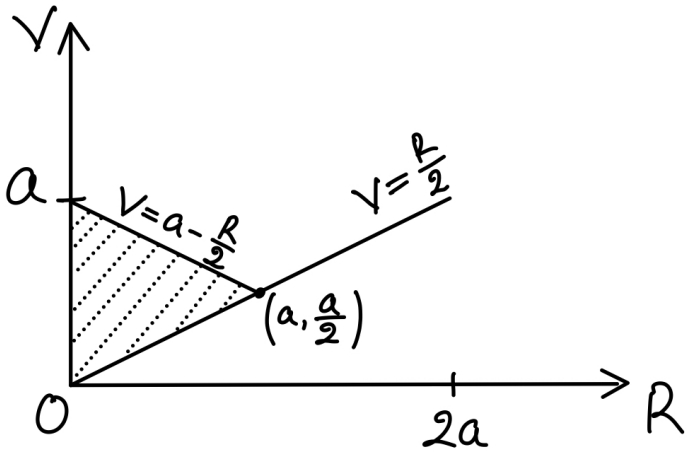


Figure 2.1: Feasible region for (R, V).

Compute the Jacobian:

$$J = \det \begin{pmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial v} \\ \frac{\partial x_n}{\partial r} & \frac{\partial x_n}{\partial v} \end{pmatrix} = \det \begin{pmatrix} -\frac{1}{2} & 1 \\ \frac{1}{2} & 1 \end{pmatrix} = -1, \qquad |J| = 1.$$

Therefore, for $(r, v) \in A$,

$$\begin{aligned} f_{R,V}(r,v) &= f_{X_{(1)},X_{(n)}}\left(v - \frac{r}{2},\, v + \frac{r}{2}\right)|J| \\ &= \frac{n(n-1)}{a^n}\left(\left(v + \frac{r}{2}\right) - \left(v - \frac{r}{2}\right)\right)^{n-2} \mathbf{1}_{\{(r,v)\in A\}} \\ &= \frac{n(n-1)}{a^n} r^{\,n-2}\, \mathbf{1}_{\{\ 0<r<a,\ \frac{r}{2}<v<a-\frac{r}{2}\ \}}. \end{aligned}$$

Notice that the density $f_{R,V}(r,v)$ does *not* actually factor into a function of r times a function of v because the domain couples r and v. Hence R and V are **not independent**.

Marginal distribution of R

Integrate out v:

$$\begin{aligned} f_R(r) &= \int_{r/2}^{a-r/2} \frac{n(n-1)}{a^n} r^{n-2}\, dv \\ &= \frac{n(n-1)}{a^n} r^{n-2}\left(a - \frac{r}{2} - \frac{r}{2}\right)\mathbf{1}_{\{0<r<a\}} \\ &= \frac{n(n-1)}{a^n} r^{n-2}(a-r)\,\mathbf{1}_{\{0<r<a\}}. \end{aligned}$$

Marginal distribution of V

From the support $\frac{r}{2} < v < a - \frac{r}{2}$, for a fixed v the valid r range depends on whether $v < a/2$ or $v > a/2$:

- If $0 < v < \frac{a}{2}$, then $r < 2v$.
- If $\frac{a}{2} < v < a$, then $r < 2(a - v)$.

Thus,

$$f_V(v) = \int f_{R,V}(r, v)\, dr = \int \frac{n(n-1)}{a^n} r^{n-2}\, dr$$

with the appropriate upper limit.

Case 1: $0 < v < \frac{a}{2}$**.**

$$\begin{aligned} f_V(v) &= \int_0^{2v} \frac{n(n-1)}{a^n} r^{n-2}\, dr \\ &= \frac{n(n-1)}{a^n} \cdot \frac{(2v)^{n-1}}{n-1} \mathbf{1}_{\{0<v<\frac{a}{2}\}} \\ &= \frac{n(2v)^{n-1}}{a^n} \mathbf{1}_{\{0<v<\frac{a}{2}\}}. \end{aligned}$$

Case 2: $\frac{a}{2} < v < a$**.**

$$\begin{aligned} f_V(v) &= \int_0^{2(a-v)} \frac{n(n-1)}{a^n} r^{n-2}\, dr \\ &= \frac{n(2(a-v))^{n-1}}{a^n} \mathbf{1}_{\{\frac{a}{2}<v<a\}} \\ &= \frac{n(2a-2v)^{n-1}}{a^n} \mathbf{1}_{\{\frac{a}{2}<v<a\}}. \end{aligned}$$

Combining, we have,

$$f_V(v) = \frac{n(2v)^{n-1}}{a^n} \mathbf{1}_{\{0<v<\frac{a}{2}\}} \; + \; \frac{n(2a-2v)^{n-1}}{a^n} \mathbf{1}_{\{\frac{a}{2}<v<a\}}.$$

Chapter 3

Point Estimation

3.1 Introduction

Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f(x \mid \theta)$, for example $f(x \mid \theta) = \mathcal{N}(\theta, 1)$. In practice, we observe data, but we want to understand the underlying structure of the data (i.e., the unknown parameter θ).

Definition: A **point estimator** is any function

$$W(X_1, X_2, \ldots, X_n),$$

of a sample, that is, any statistic is a point estimator. However, the estimator $W(X_1, X_2, \ldots, X_n)$ **cannot depend on** the unknown parameter θ.

3.2 Maximum Likilihood estimators

Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f(x \mid \theta)$, where $f(x \mid \theta)$ can be a pmf or a pdf and $\theta \in \Omega$ is an unknown parameter.

Definition 3.2.1 (Likelihood Function)**.** The likelihood function is

$$L(\theta; x_1, \ldots, x_n) = L(\theta \mid X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i \mid \theta).$$

viewed as a function of $\theta \in \Omega$ (the parameter space).

Definition of the MLE
We estimate the true value of θ by choosing the parameter that maximizes the likelihood:

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Omega} L(\theta; x_1, \ldots, x_n).$$

Equivalently,

$$L(\widehat{\theta}_{\text{MLE}}; x_1, \ldots, x_n) = \max_{\theta \in \Omega} L(\theta; x_1, \ldots, x_n).$$

Intuition: pick θ such that the chance / likelihood of observing $(x_1, \ldots, x_n)$ is the largest.

Log-likelihood

Definition 3.2.2 (Log-Likelihood)**.** The log-likelihood function is

$$\ell(\theta \mid X_1, \ldots, X_n) = \log L(\theta \mid X_1, \ldots, X_n) = \sum_{i=1}^{n} \log f(X_i \mid \theta).$$

The log-likelihood helps reduce the burden of differentiation (product becomes sum):

$$\begin{aligned} \ell(\theta; x_1, \ldots, x_n) &:= \log L(\theta; x_1, \ldots, x_n) \\ &= \log \left(\prod_{i=1}^{n} f(x_i \mid \theta) \right) \\ &= \sum_{i=1}^{n} \log \big(f(x_i \mid \theta) \big). \end{aligned}$$

Since $\log(\cdot)$ is increasing, maximizing L is equivalent to maximizing ℓ:

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Omega} \ell(\theta; x_1, \ldots, x_n).$$

Lemma 3.2.3 (Maximum Likelihood Estimator)**.** *The MLE is*

$$\hat{\theta} = \arg\max_{\theta \in \Omega} L(\theta \mid X_1, \ldots, X_n) = \arg\max_{\theta \in \Omega} \ell(\theta \mid X_1, \ldots, X_n).$$

The above holds only when $\ell(\theta \mid X_1, \ldots, X_n)$ has the same domain Ω. Sometimes taking the logarithm will shrink the domain and the analysis will become more complicated (see Example 3.2.8).

Remark 3.2.4. $\hat{\theta}(X_1, \ldots, X_n)$ is a random variable since it depends on $X_1, \ldots, X_n$.

Example 3.2.5 (Example: Location Normal MLE)**.** Consider

$$X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1).$$

Goal: estimate θ

$$L(\theta \mid X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i \mid \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_i - \theta)^2}{2} \right) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (X_i - \theta)^2 \right).$$

Method 1 (Algebraic)
Since the constant does not depend on θ, maximizing $L(\theta \mid X_1, \ldots, X_n)$ is equivalent to minimizing

$$\sum_{i=1}^{n} (X_i - \theta)^2.$$

By the lemma in the lecture note of sample mean and sample variance, the minimizer is

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

Method 2 (General)
The log-likelihood is

$$\ell(\theta) = \ell(\theta \mid X_1, \ldots, X_n) = n \log(\frac{1}{\sqrt{2\pi}}) + (-\frac{1}{2} \sum_{i=1}^{n} (X_i - \theta)^2).$$

First derivative:

$$\begin{aligned}\ell'(\theta) &= -\frac{1}{2}\sum_{i=1}^{n} 2(\theta - X_i) \\ &= -\sum_{i=1}^{n}(\theta - X_i) \\ &= -(n\theta - \sum_{i=1}^{n} X_i) \\ &= n(\frac{1}{n}\sum_{i=1}^{n} X_i - \theta) \\ &= n(\bar{X}_n - \theta)\end{aligned}$$

Setting $\ell'(\theta) = 0$ gives the following critical point:

$$\theta = \bar{X}_n$$

It is not completed since we need to justify the critical point is maximizer. There are two methods: first derivative test and second derivative test.

Method 2a: First Derivative Test

Now analyze the sign of $\ell'(\theta)$.

- If $\theta < \bar{X}_n$, then:

$$\ell'(\theta) > 0.$$

  Thus $\ell(\theta)$ is increasing.

- If $\theta > \bar{X}_n$, then:

$$\ell'(\theta) < 0.$$

  Thus $\ell(\theta)$ is decreasing.

Therefore, $\ell(\theta)$ increases up to $\bar{X}_n$ and decreases afterward.

Hence,

$$\theta = \bar{X}_n$$

is the maximizer.

Method 2b: Second Derivative Test

Recall

$$\ell''(\theta) = -n.$$

Since

$$\ell''(\theta) = -n < 0,$$

the log-likelihood is strictly concave.

Therefore, the critical point

$$\theta = \bar{X}_n$$

is the unique global maximizer.

Thus, the MLE is:

$$\hat{\theta}_n = \bar{X}_n.$$

Example 3.2.6 (Example: Location Normal with Constraint)**.** Suppose

$$X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1),$$

but now the parameter space is

$$\Omega = [0, \infty).$$

From the unconstrained problem, we know the critical point satisfies

$$\hat{\theta}_{\text{unconstrained}} = \bar{X}_n.$$

Recall from the first derivative test:

- If $\theta < \bar{X}_n$, then $\ell'(\theta) > 0$ and $\ell(\theta)$ is increasing.
- If $\theta > \bar{X}_n$, then $\ell'(\theta) < 0$ and $\ell(\theta)$ is decreasing.

Thus $\ell(\theta)$ increases up to $\bar{X}_n$ and decreases afterward.
Case 1: $\bar{X}_n > 0$ $\ell(\theta)$ is increasing on $[0, X_n)$ and decreasing on (X_n, ∞), and thus the maximizer is

$$\theta = \bar{X}_n.$$

Case 2: $\bar{X}_n \leq 0$
$\ell(\theta)$ is decreasing on $[0, \infty)$. The maximum over $\Omega = [0, \infty)$ occurs at the boundary:

$$\theta = 0.$$

Conclusion
Combining both cases, the MLE is

$$\hat{\theta} = \begin{cases} \bar{X}_n & \text{if } \bar{X}_n > 0, \\ 0 & \text{if } \bar{X}_n \leq 0. \end{cases}$$

Equivalently,

$$\hat{\theta} = \max\{\bar{X}_n, 0\}.$$

Exercise 3.2.7. what if $\Omega = (0, \infty)$?

Example 3.2.8 (Example: Bernoulli MLE)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$, where $0 \leq p \leq 1$.
The pmf of a Bernoulli random variable is

$$f(x \mid p) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases} = p^x (1 - p)^{1 - x}.$$

Likelihood Function is

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} f(X_i \mid p) \\ &= \prod_{i=1}^{n} p^{X_i} (1 - p)^{1 - X_i} \\ &= p^{\sum_{i=1}^{n} X_i} (1 - p)^{n - \sum_{i=1}^{n} X_i}. \end{aligned}$$

Log-Likelihood
For $0 < p < 1$, take the logarithm:

$$\begin{aligned}\ell(p) &= \log L(p) \\ &= \sum_{i=1}^{n} X_i \log p + \left(n - \sum_{i=1}^{n} X_i\right) \log(1-p).\end{aligned}$$

Note the domain for the log-likelihood function is smaller than the likelihood function!

Derivative and Critical Point
For $0 < p < 1$, differentiate $\ell(p)$:

$$\ell'(p) = \frac{\sum_{i=1}^{n} X_i}{p} - \frac{n - \sum_{i=1}^{n} X_i}{1-p}.$$

Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\begin{aligned}\ell'(p) &= \frac{n\bar{X}_n}{p} - \frac{n(1-\bar{X}_n)}{1-p} \\ &= n\left(\frac{\bar{X}_n}{p} - \frac{1-\bar{X}_n}{1-p}\right) \\ &= n \cdot \frac{\bar{X}_n(1-p) - p(1-\bar{X}_n)}{p(1-p)} \\ &= n \cdot \frac{\bar{X}_n - p}{p(1-p)}.\end{aligned}$$

Setting $\ell'(p) = 0$ gives the critical point

$$p = \bar{X}_n.$$

First Derivative Test
From

$$\ell'(p) = n \cdot \frac{\bar{X}_n - p}{p(1-p)},$$

we have:

- If $p < \bar{X}_n$, then $\ell'(p) > 0$, so $\ell(p)$ is increasing.
- If $p > \bar{X}_n$, then $\ell'(p) < 0$, so $\ell(p)$ is decreasing.

If $\bar{X}_n \in (0,1)$, then $\bar{X}_n$ is the maximizer of $\ell(p)$ on $(0,1)$.
Maximizer on $[0,1]$ (Boundary Cases)
Case 1: $\bar{X}_n \in (0,1)$.

$$L(\bar{X}_n) = \max_{0<p<1} L(p) = \max_{0\le p\le 1} L(p)$$

since the likelihood function $L(p)$ is defined at $p = 0$ and $p = 1$ and right and left continuous, respectively (actually $L(0) = L(1) = 0$). Hence,

$$L(0) = 0, \qquad L(1) = 0,$$

and $L(p)$ is continuous on $[0,1]$. Therefore, $\bar{X}_n$ is also the maximizer of $L(p)$ on $[0,1]$.

Case 2: $\bar{X}_n = 0$. Then $X_1 = \cdots = X_n = 0$, so

$$L(p) = (1-p)^n,$$

which is decreasing on $[0,1]$. Thus the maximizer is $p = 0 = \bar{X}_n$.

Case 3: $\bar{X}_n = 1$**.** Then $X_1 = \cdots = X_n = 1$, so

$$L(p) = p^n,$$

which is increasing on $[0,1]$. Thus the maximizer is $p = 1 = \bar{X}_n$.

Conclusion

We conclude the MLE is

$$\hat{p}_n = \bar{X}_n.$$

Exercise 3.2.9. Will we have the same issue for continuous distribution?

However, we may not be able to find the explicit solution of MLE.

Example 3.2.10. (Logistic Distribution) Consider a random sample $X_1, \ldots, X_n \overset{iid}{\sim} Logistic(\theta, 1)$. The probability density function (pdf) is given by:

$$f(x;\theta) = \frac{e^{-(x-\theta)}}{(1+e^{-(x-\theta)})^2} = \frac{e^{-x+\theta}}{(1+e^{-x+\theta})^2}. \tag{3.1}$$

The likelihood function is

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \frac{e^{-x_i+\theta}}{(1+e^{-x_i+\theta})^2} = \frac{e^{-\sum_{i=1}^n x_i + n\theta}}{\prod_{i=1}^n (1+e^{-x_i+\theta})^2}.$$

The log-likelihood simplifies to:

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= -\sum_{i=1}^{n} x_i + n\theta - 2\sum_{i=1}^{n} \log\left(1+e^{-x_i+\theta}\right). \end{aligned} \tag{3.2}$$

To find the MLE, we take the derivative with respect to θ:

$$\begin{aligned} l'(\theta) &= n - 2\sum_{i=1}^{n} \frac{e^{-x_i+\theta}}{1+e^{-x_i+\theta}} \\ &= n - 2\sum_{i=1}^{n} \left(1 - \frac{1}{1+e^{-x_i+\theta}}\right) \\ &= -n + 2\sum_{i=1}^{n} \frac{1}{1+e^{-x_i+\theta}}. \end{aligned} \tag{3.3}$$

The second derivative is:

$$l''(\theta) = -2\sum_{i=1}^{n} \frac{e^{-x_i+\theta}}{(1+e^{-x_i+\theta})^2} < 0. \tag{3.4}$$

Since $l''(\theta) < 0$, the first derivative $l'(\theta)$ is strictly decreasing. Thus, $l'(\theta) = 0$ has a unique solution, denoted by θ_1. When $\theta < \theta_1, l'(\theta) > 0$ thus $l(\theta)$ is increasing, and when $\theta > \theta_1, l'(\theta) < 0$ thus $l(\theta)$ is decreasing. Therefore, by the first derivative test the maximizer is θ_1, i.e., $\theta_1 = \hat{\theta}_n$.

In all previous examples, we have dealt with cases where the likelihood function is differentiable. However, there are cases where the likelihood function is non-differentiable or discontinuous. In such cases, we need to investigate the increase and decrease of the likelihood or log-likelihood function to find the maximum likelihood estimator as the following example.

Example 3.2.11. (MLE for the Double Exponential (Laplace) Distribution) Let $(X_1, \ldots, X_n) \overset{iid}{\sim} \text{DE}(\theta, 1))$, with pdf:

$$f(x;\theta) = \frac{1}{2}e^{-|x-\theta|}. \tag{3.5}$$

The likelihood function is:

$$\begin{aligned} L(\theta; x_1, \ldots, x_n) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= \prod_{i=1}^{n} \left(\frac{1}{2}e^{-|x_i-\theta|}\right) \\ &= \left(\prod_{i=1}^{n}\frac{1}{2}\right) \cdot \left(\prod_{i=1}^{n} e^{-|x_i-\theta|}\right) \\ &= \left(\frac{1}{2}\right)^n e^{\sum_{i=1}^{n} -|x_i-\theta|} \\ &= \frac{1}{2^n}e^{-\sum_{i=1}^{n}|x_i-\theta|}. \end{aligned}$$

Maximizing this likelihood is equivalent to minimizing the sum of absolute deviations $S(\theta) = \sum_{i=1}^{n}|x_i-\theta|$.

Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the order statistics. Define $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$. For a given r, when θ falls in the interval $[X_{(r)}, X_{(r+1)})$, the log-likelihood can be written as:

$$\begin{aligned} l(\theta) &= \log\left(\frac{1}{2^n}\right) - \left(\sum_{i=1}^{r}(\theta - X_{(i)}) + \sum_{i=r+1}^{n}(X_{(i)} - \theta)\right) \\ &= \log\left(\frac{1}{2^n}\right) - \left(r\theta - \sum_{i=1}^{r}X_{(i)}\right) - \left(\sum_{i=r+1}^{n}X_{(i)} - (n-r)\theta\right) \\ &= \log\left(\frac{1}{2^n}\right) + \sum_{i=1}^{r}X_{(i)} - \sum_{i=r+1}^{n}X_{(i)} + (n-2r)\theta. \end{aligned} \tag{3.6}$$

This is a linear function in θ on the interval $[X_{(r)}, X_{(r+1)})$. The slope of this line is $n-2r$.

- If $n-2r > 0$ (i.e., $r < n/2$), $l(\theta)$ is increasing on the interval.
- If $n-2r < 0$ (i.e., $r > n/2$), $l(\theta)$ is decreasing on the interval.
- If $n-2r = 0$ (i.e., $r = n/2$), the function is constant on the interval.

For an odd sample size $n = 2m+1$, the slope changes from positive to negative at $X_{(m+1)}$, making this order statistic the unique maximizer. For an even sample size $n = 2m$, the slope is zero on the interval $[X_{(m)}, X_{(m+1)}]$, meaning any point in that interval maximizes the likelihood.

- **If** $n = 2m+1$ **(odd):** The MLE is the middle order statistic, $\hat{\theta}_n = X_{(m+1)}$.
- **If** $n = 2m$ **(even):** The MLE is not unique. Any θ in the interval between the two middle order statistics, $[X_{(m)}, X_{(m+1)}]$, minimizes $S(\theta)$.

Lastly, there are cases where the maximum likelihood estimator is not uniquely determined.

Example 3.2.12. (Non-Uniqueness of MLE) There are cases where the maximum likelihood estimator is not unique. Consider a sample $X_1, \ldots, X_n \overset{iid}{\sim} \text{Uniform}[\theta - 1, \theta + 1]$. The pdf is:

$$f(x;\theta) = \frac{1}{2}\mathbf{1}_{[\theta-1,\theta+1]}(x). \tag{3.7}$$

The likelihood function $L(\theta)$ is the product of the individual pdfs for every x_i:

$$\begin{aligned} L(\theta; x_1, \ldots, x_n) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= \prod_{i=1}^{n} \left(\frac{1}{2}\mathbf{1}_{\{(\theta-1,\theta+1)\}}(x_i) \right) \\ &= \left(\frac{1}{2}\right)^n \prod_{i=1}^{n} \mathbf{1}_{\{(\theta-1,\theta+1)\}}(x_i) \end{aligned}$$

This likelihood is positive if and only if *all* observations x_i lie within the interval $[\theta - 1, \theta + 1]$. This is equivalent to requiring that the largest observation $x_{(n)}$ and the smallest observation $x_{(1)}$ satisfy:

$$\theta - 1 \leq x_{(1)} \quad \text{and} \quad x_{(n)} \leq \theta + 1. \tag{3.8}$$

Rearranging these inequalities gives:

$$x_{(n)} - 1 \leq \theta \leq x_{(1)} + 1. \tag{3.9}$$

Therefore, any θ in the interval $[x_{(n)} - 1, x_{(1)} + 1]$ will make the likelihood equal to $1/2^n$, which is its maximum value. Since $x_{(1)} \leq x_{(n)}$, this interval is valid. The MLE is not unique; it can be any number in this interval, often written as $[X_{(n)} - 1, X_{(1)} + 1]$.

3.2.1 Invariance Property of MLEs

A key property of maximum likelihood estimators is their invariance under transformation.

Theorem 3.2.13 (Invariance Property)**.** *If $\hat{\theta}_n$ is the MLE of θ, then for any function $T(\theta)$, the MLE of $\eta = T(\theta)$ is $T(\hat{\theta}_n)$.*

Example 3.2.14 (Normal Distribution)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} N(0, 1)$, and the MLE $\hat{\theta}_n = \bar{X}_n$. Consequently, by the invariance property:

- Then the MLE for θ^2 is equal to $(\bar{X}_n)^2$

Example 3.2.15 (Bernoulli Distribution)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} Bernoulli(p)$. The MLE for the success probability p is $\hat{p}_n = \bar{X}_n$. The MLE for $\sqrt{p(1-p)}$ is then:

$$\widehat{\sqrt{p(1-p)}} = \sqrt{\hat{p}_n(1-\hat{p}_n)}. \tag{3.10}$$

3.2.2 Two-Parameter Maximum Likelihood Estimator

If θ is multidimensional, we must maximize the function at several different variables in order to find the MLE. For a differentiable likelihood function, we can set the first partial derivatives equal to 0, but verifying through a second derivative may lead to our work becoming needlessly tedious and muddled. Instead, let us try the method of successive maximizations.

Example 3.2.16 (Example 2.1 (Normal MLEs, μ and σ^2 unknown))**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ where $\mu \in \mathbb{R}, \sigma^2 > 0$. Find the MLE of $\theta = (\mu, \sigma^2)^T$
The likelihood function is:

$$\begin{aligned} L(\mu, \sigma^2 | x_1, \ldots, x_n) &= \prod_{i=1}^{n} pdf(x_i | \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \end{aligned}$$

We want to maximize this function with respect to both μ and σ^2. To do so, we first find the log likelihood function:

$$\begin{aligned} l(\mu, \sigma^2 | x_1, \ldots, x_n) &= \log L(\mu, \sigma^2 | x_1, \ldots, x_n) \\ &= n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \\ &= -\frac{1}{2} n \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned}$$

Now, we want to maximize the function with respect to one variable at a time. Recognize that in order to maximize $l(\mu, \sigma^2)$ with respect to μ, we only need to minimize $\sum_{i=1}^{n}(x_i - \mu)^2$ with respect to μ. Based on the result from a previous lecture, we recall it is minimized at $\hat{\mu} = \bar{x}_n$, the sample mean. **In general of successive maximization, the maximizer may depends on the other parameter σ. The fact that this maximizer does not depends on the other parameter σ significantly simplifies the calculations.**

After maximizing our two-parameter function with respect to one parameter, we can use the result to reduce the function to one parameter and find the remaining maximization:

$$\max_{(\mu, \sigma^2)} l(\mu, \sigma^2) = \max_{\sigma^2} l(\hat{\mu}, \sigma^2)$$

$$l(\hat{\mu}, \sigma^2) = -\frac{1}{2} n \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

Let $\xi = \sigma^2$:

$$l(\hat{\mu}, \xi) = -\frac{1}{2} n \log(2\pi\xi) - \frac{1}{2\xi}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

Now, we take the partial derivative of the log likelihood function $l(\hat{\mu}, \xi)$ with respect to ξ,

$$\begin{aligned} \frac{\partial l(\hat{\mu}, \xi)}{\partial \xi} &= -\frac{1}{2} n \cdot \frac{1}{2\pi\xi} \cdot 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \left(-\frac{1}{\xi^2}\right) \\ &= -\frac{1}{2} n \cdot \frac{1}{\xi} + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \left(-\frac{1}{\xi^2}\right) \\ &= \frac{1}{2} \cdot \frac{-n\xi + \sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{\xi^2} \end{aligned}$$

First Derivative Test Set $\frac{\partial(\hat{\mu}^u,\xi)}{\partial\xi} = 0$ and solve for ξ to find the critical point.

$$\begin{aligned}
\frac{\partial l(\hat{\mu},\xi)}{\partial\xi} &= 0 \\
\frac{1}{2}\cdot\frac{-n\xi+\sum_{i=1}^n(x_i-\bar{x}_n)^2}{\xi^2} &= 0 \\
-n\xi+\sum_{i=1}^n(x_i-\bar{x}_n)^2 &= 0 \\
\xi = \frac{1}{n}\sum_{i=1}^n(x_i-\bar{x}_n)^2 & \qquad \text{is the critical point}
\end{aligned}$$

- When $\xi < \frac{1}{n}\sum_{i=1}^n(x_i-\bar{x}_n)^2$,

$$\frac{\partial l(\hat{\mu},\xi)}{\partial\xi} > 0.$$

So, $l(\hat{\mu},\xi)$ is increasing.

- When $\xi > \frac{1}{n}\sum_{i=1}^n(x_i-\bar{x}_n)^2$,

$$\frac{\partial l(\hat{\mu},\xi)}{\partial\xi} < 0$$

.

So, $l(\hat{\mu},\xi)$ is decreasing.

Thus, by the first derivative test we have verified that $\xi = \frac{1}{n}\sum_{i=1}^n(x_i-\bar{x}_n)^2$ is the maximizer. We see that $l(\hat{\mu},\xi)$ increases to ξ and decreases after reaching it, making it the highest, or maximum, point.

Therefore, we have found the MLE of $\theta = (\mu,\sigma^2)^T$ is $(\hat{\mu},\hat{\sigma}^2) = (\bar{x}_n, \frac{1}{n}\sum_{i=1}^n(x_i-\bar{x}_n)^2)$.
Recall the sample variance,

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^n(x_i-\bar{x}_n)^2 \\
\mathbb{E}(S^2) &= \sigma^2
\end{aligned}$$

Note the MLE and the sample variance are different. The coefficient of the sample variance $\frac{1}{n-1}$ is to make the expectation to be σ^2.

3.3 Method of Moments

The method of moments (MME) is an alternative way to construct estimators by matching theoretical moments to sample moments. Instead of maximizing a likelihood, we equate functions of the parameters to corresponding empirical quantities.

Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} f(x \mid \theta)$. Suppose the distribution has theoretical moments $m_r = \mathbb{E}[X^r]$ that can be expressed in terms of θ.

First step

Write the parameter in terms of the moments. This means we first find algebraic relationships between θ and the moments m_r, then solve those equations for θ.

Example: $X_1,\ldots,X_n \overset{\text{iid}}{\sim} N(\mu,\sigma^2)$. Here $\theta = (\mu,\sigma^2)^T$.

The first two theoretical moments of a normal random variable are

$$m_1 = \mu, \qquad m_2 = \sigma^2 + \mu^2.$$

The first moment is the mean, and the second moment equals variance plus squared mean: $m_2 = \mathrm{Var}(X) + (\mathbb{E}X)^2$.

We now solve these equations for μ and σ^2:

$$\mu = m_1, \qquad \sigma^2 = m_2 - m_1^2.$$

So, once we know the first two moments, we can recover the parameters.

Second step

Replace moments with sample moments. We approximate each theoretical moment m_r by the empirical average of X_i^r:

$$\hat{m}_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r.$$

This uses the idea that, by the law of large numbers, sample moments converge to their population counterparts.

(Recall that for a general function g, $\mathbb{E}[g(X)] \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i)$.) Thus, we plug $\hat{m}_r$ into the relationships obtained in the first step.

Applying this to the normal example, the MME estimators are

$$\hat{\mu} = \hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n,$$

which matches the sample mean, and

$$\hat{\sigma}^2 = \hat{m}_2 - \hat{m}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Here $\hat{\sigma}^2$ is the sample second moment about zero minus the squared sample mean, mimicking the identity $m_2 = \sigma^2 + \mu^2$.

3.3.1 Non-Unique MMEs

This section shows that the method of moments need not give a unique estimator; different moment equations can lead to different solutions. We illustrate this with the Poisson distribution.

Example: $X_1, \ldots, X_n \sim \mathrm{Poisson}(\lambda)$, with $\lambda > 0$. The Poisson distribution has the property that its mean and variance are equal to λ.

From the first moment,

$$m_1 = \lambda, \qquad \hat{\lambda}_1 = \hat{m}_1 = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

So matching the mean gives the natural estimator: the sample average.

When using the second moment, note that

$$\mathrm{Var}(X) = \lambda, \qquad m_2 = \lambda + \lambda^2.$$

The second moment is the sum of the variance and squared mean, both equal to λ.

Thus we obtain the quadratic equation

$$\lambda^2 + \lambda - m_2 = 0,$$

so solving for λ gives

$$\lambda = \frac{-1 \pm \sqrt{1 + 4m_2}}{2}.$$

These are two algebraic solutions for λ in terms of the second moment.

Replacing m_2 by the sample second moment $\hat{m}_2$ yields two MMEs:

$$\hat{\lambda}_{2,1} = \frac{-1 + \sqrt{1 + 4\hat{m}_2}}{2}, \quad \hat{\lambda}_{2,2} = \frac{-1 - \sqrt{1 + 4\hat{m}_2}}{2}.$$

Both satisfy the moment equation, but one of them may violate the parameter constraints.

Remember that we have the constraint $\lambda > 0$. Therefore $\hat{\lambda}_{2,2}$ is usually discarded because it is negative, while $\hat{\lambda}_{2,1}$ is kept as a candidate estimator.

Another way, using the variance identity, is to note that

$$\lambda = \mathrm{Var}(X) = m_2 - m_1^2 \quad \Rightarrow \quad \hat{\lambda}_3 = \hat{m}_2 - \hat{m}_1^2.$$

Here we use both the first and second moments simultaneously to construct a different estimator.

So we have multiple method-of-moments estimators for λ. This example emphasizes that MME is a recipe rather than a unique rule.

3.4 Midterm review

Change of Variables

Problem 3.4.1. Let $X_1, X_2 \overset{\text{iid}}{\sim} N(0, \sigma^2)$, or equivalently, $(X_1, X_2) \sim N(0, \sigma^2 I_2)$. This means that the vector (X_1, X_2) follows a two-dimensional Normal distribution with mean vector $0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance matrix σ^2 times identity matrix $I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

(a) Find the joint distribution of two new random variables, Y_1 and Y_2, where $Y_1 = X_1^2 + X_2^2$ and $Y_2 = \dfrac{X_1}{\sqrt{Y_1}}$.

(b) Prove that Y_1 and Y_2 are independent and interpret that result geometrically.

(a) First, we must understand what Y_1 and Y_2 mean geometrically. If we draw a plot with X_1 and X_2 as the axes, then Y_1 is the squared distance from any random point to the origin. In other words, it is the radius squared. Y_2 is then $\cos(\theta)$, where θ is the angle made by the radius.

From this, we also have that the range of Y_1 is $Y_1 > 0$, and using the definition of cosine, $-1 < Y_2 < 1$. This will be the domain of the joint density function.

To derive the distribution, we start by finding the inverse function. X_1 can be written in terms of Y_1 and Y_2 easily:

$$X_1 = \sqrt{Y_1} Y_2$$

However, as X_2 is squared, we will need to determine if it is positive or negative. Therefore, we will consider three regions:

$$A_0 = \{(X_1, X_2) : X_2 = 0\}$$
$$A_1 = \{(X_1, X_2) : X_2 > 0\}$$
$$A_2 = \{(X_1, X_2) : X_2 < 0\}$$

Since X_2 is continuous, $P((X_1, X_2) \in A_0) = P(X_2 = 0) = 0$. On A_1, X_2 is positive, so we have:

$$A_1 : \begin{cases} x_1 = \sqrt{y_1} y_2 \\ x_2 = \sqrt{y_1 - y_1 y_2^2} = \sqrt{y_1}\sqrt{1 - y_2^2} \end{cases}$$

Now we find the determinant of the Jacobian for case A_1:

$$\begin{aligned} J &= \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} \\ &= \begin{vmatrix} \frac{y_2}{2\sqrt{y_1}} & \sqrt{y_1} \\ \frac{\sqrt{1 - y_2^2}}{2\sqrt{y_1}} & -\frac{\sqrt{y_1} y_2}{\sqrt{1 - y_2^2}} \end{vmatrix} \\ &= -\frac{y_2^2}{2\sqrt{1 - y_2^2}} - \frac{\sqrt{1 - y_2^2}}{2} \\ &= -\frac{y_2^2}{2\sqrt{1 - y_2^2}} - \frac{1 - y_2^2}{2\sqrt{1 - y_2^2}} \\ &= -\frac{1}{2\sqrt{1 - y_2^2}} \end{aligned}$$

Then we find the inverse for A_2, where X_2 is negative:

$$A_2 : \begin{cases} x_1 = \sqrt{y_1} y_2 \\ x_2 = -\sqrt{y_1 - y_1 y_2^2} = -\sqrt{y_1}\sqrt{1 - y_2^2} \end{cases}$$

Notice that this is the same as A_1, save for the negative sign. Therefore, we can do the same calculations as last time for the Jacobian to get:

$$J = \frac{1}{2\sqrt{1 - y_2^2}}$$

We'll be using the absolute value of the determinant, so these two are essentially the same. Now we can plug these results into the change of variables formula. Since we have two domains (A_1 and A_2), the joint density of Y_1 and Y_2 is a summation over the two regions, like so:

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) = {} & f_{X_1,X_2}(\sqrt{y_1} y_2, \sqrt{y_1}\sqrt{1 - y_2^2}) \frac{1}{2\sqrt{1 - y_2^2}} \\ & + f_{X_1,X_2}(\sqrt{y_1} y_2, -\sqrt{y_1}\sqrt{1 - y_2^2}) \frac{1}{2\sqrt{1 - y_2^2}} \end{aligned}$$

Since X_1 and X_2 are independent of each other and follow a Normal distribution with mean 0 and variance σ^2, we have:

$$\begin{aligned} f_{X_1,X_2}(x_1,x_2) &= f_{X_1}(x_1)f_{X_2}(x_2) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x_1^2}{2\sigma^2}}\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x_2^2}{2\sigma^2}}\right) \end{aligned}$$

Additionally, notice that since X_2 follows a Normal distribution with mean 0, it is symmetric about the y-axis. That is, $f_{X_2}(x_2) = f_{X_2}(-x_2)$. Using this, we have:

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= f_{X_1}(\sqrt{y_1}y_2)f_{X_2}(\sqrt{y_1}\sqrt{1-y_2^2})\frac{1}{2\sqrt{1-y_2^2}} \\ &\quad + f_{X_1}(\sqrt{y_1}y_2)f_{X_2}(-\sqrt{y_1}\sqrt{1-y_2^2})\frac{1}{2\sqrt{1-y_2^2}} \\ &= 2\cdot f_{X_1}(\sqrt{y_1}y_2)f_{X_2}(\sqrt{y_1}\sqrt{1-y_2^2})\frac{1}{2\sqrt{1-y_2^2}} \\ &= 2\cdot\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{y_1y_2^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{y_1(1-y_2^2)}{2\sigma^2}}\frac{1}{2\sqrt{1-y_2^2}} \\ &= \frac{1}{2\pi\sigma^2}e^{-\frac{y_1}{2\sigma^2}}\frac{1}{\sqrt{1-y_2^2}}1_{(y_1>0,-1<y_2<1)}(y_1,y_2) \end{aligned}$$

which completes part (a).

(b) Notice that the joint density found in (a) can be factorized into two terms, where the first depends only on Y_1 and the second depends only on Y_2. Therefore, Y_1 and Y_2 are independent.

Recall the geometric meaning of Y_1 and Y_2. Independence then means that the length of the radius and the cosine of its radius do not affect each other.

Order Statistics

Problem 3.4.2. Let U_i, $i = 1, 2, \ldots$ be independent random variables that follow Unif$(0, 1)$. Let X have distribution $P(X = x) = \frac{c}{x!}$, $x = 1, 2, \ldots$ where $c = \frac{1}{e-1}$. Find the distribution of $Z = min\{U_1, U_2, \ldots, U_X\}$.

Note that for this example, the sample size is not fixed, and instead depends on X. To get the distribution of Z, we must condition it on X so the sample size becomes fixed.

As each U_i is continuous, Z is continuous as well. Also, since the range of each U_i is $(0, 1)$, we have that the range of Z is also $(0, 1)$.

If we find $P(Z > z)$, then $1 - P(Z > z)$ will be the CDF, and the PDF may be derived from that. To get this probability, we will condition Z on X. So, for $z \in (0, 1)$:

$$P(Z > z) = \sum_{x=1}^{\infty} P(Z > z \mid X = x)P(X = x)$$

If we know that $X = x$, then $Z = min\{U_1, U_2, \ldots, U_x\}$, and we want $Z > z$. If the minimum of U_i, $i = 1, 2, \ldots, x$ is greater than z, that is equivalent to $U_1 > z, U_2 > z, \ldots, U_x > z$, so we have:

$$P(Z > z) = \sum_{x=1}^{\infty} P(U_1 > z, U_2 > z, \ldots, U_x > z) P(X = x)$$

And since all U_i are independent and identically distributed:

$$P(Z > z) = \sum_{x=1}^{\infty} (P(U_1 > z))^x P(X = x)$$

Notice that $P(U_1 > z) = 1 - P(U_1 \leq z)$. $P(U_1 \leq z)$ is the CDF of U_1, so $P(U_1 \leq z) = z$ since $z \in (0, 1)$. Plugging this and the PMF for X in, we get:

$$\begin{aligned} P(Z > z) &= \sum_{x=1}^{\infty} (1-z)^x \frac{1}{(e-1)x!} \\ &= \frac{1}{e-1} \sum_{x=1}^{\infty} \frac{(1-z)^x}{x!} \end{aligned}$$

Finally, we need to do the summation over x. Recall that the Taylor Expansion of e^y is $e^y = \sum_{x=0}^{\infty} \frac{y^x}{x!}$. If we adjust the bounds of the summation and let $y = 1 - z$, we have:

$$\begin{aligned} P(Z > z) &= \frac{1}{e-1} \left(\sum_{x=0}^{\infty} \frac{(1-z)^x}{x!} - 1 \right) \\ &= \frac{1}{e-1} (e^{1-z} - 1) \end{aligned}$$

Now we may find the CDF from $1 - P(Z > z)$, and from there get the distribution of Z.

Method of Moments

Problem 3.4.3. Suppose $X_1, \ldots, X_n$ are iid with density

$$f(x \mid \theta) = \frac{1}{\theta} \mathbf{1}_{\{0 < x < \theta\}}, \qquad \theta > 0.$$

Find the method-of-moments estimator of θ.

This is the $\text{Unif}(0, \theta)$ model. Its first population moment is

$$\mathbb{E}[X_1] = \frac{\theta}{2}.$$

The method of moments equates this to the sample first moment $\bar{X}_n$, so

$$\bar{X}_n = \frac{\theta}{2}.$$

Solving for θ yields the estimator

$$\widehat{\theta}_{\text{MM}} = 2\bar{X}_n.$$

A useful takeaway from this example is that it helps to know the first few moments of common distributions, since method-of-moments estimators are often obtained immediately from those formulas.

Maximum Likelihood Estimation

Problem 3.4.4. Suppose there is one observation $X \sim N(0, \sigma^2)$ with $\sigma > 0$. Find the maximum likelihood estimator of σ.

The likelihood is

$$L(\sigma \mid x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right), \qquad \sigma > 0.$$

The log-likelihood is

$$\ell(\sigma \mid x) = -\log(\sqrt{2\pi}\sigma) - \frac{x^2}{2\sigma^2}.$$

Differentiate with respect to σ:

$$\ell'(\sigma \mid x) = -\frac{1}{\sigma} + \frac{x^2}{\sigma^3} = \frac{x^2 - \sigma^2}{\sigma^3}.$$

The critical point satisfies

$$\ell'(\sigma \mid x) = 0 \quad \Longleftrightarrow \quad x^2 = \sigma^2 \quad \Longleftrightarrow \quad \sigma = |x|\,,$$

since $\sigma > 0$.

To check that this critical point is a maximizer, note that

$$\ell'(\sigma \mid x) > 0 \text{ when } 0 < \sigma < |x|\,, \qquad \ell'(\sigma \mid x) < 0 \text{ when } \sigma > |x|\,.$$

Thus the log-likelihood increases up to $\sigma = |x|$ and decreases afterward. By the first derivative test,

$$\widehat{\sigma}_{\text{MLE}} = |X|\,.$$

3.5 Exponential family

Previously, we discussed how to estimate parameters by introducing 'Maximum Likelihood Estimators' and 'Method of Moment Estimators'. Now, we will discuss a particular family of distributions that will allow us to easily derive the MLE. This distribution allows a shortcut from our typical procedure for finding MLE (discussed in a previous lecture).

An exponential family is an important class of probability distributions with a specific form, which is equipped with its mathematical convenience by its algebraic properties, such as the calculation of expectations and covariances through differentiation. More importantly, we will show that this class of distributions naturally determines complete and sufficient statistics in the later sections. In this section, even though we only consider continuous variable cases for the proof, it can include discrete variable cases.

The probability density function of an natural exponential family is given by

$$\mathrm{f}(x;\eta) = \begin{cases} \exp\{\eta T(x) - A(\eta) + S(x)\}, & x \in \mathcal{X}, \\ 0 & \text{otherwise} \end{cases} = \exp\{\eta T(x) - A(\eta) + S(x)\} I_{\mathcal{X}}(x)$$

where $\eta \in \Omega$ and the following properties are satisfied:

(i) The support $\mathcal{X} = \{x : \mathrm{pdf}(x;\eta) > 0\}$ does not change with respect to the parameter η.

(ii) The parameter space Ω is an open interval on the real line.

(iii) $\mathrm{Var}_\eta(T(X_1)) > 0$, i.e., $T(x)$ is not a constant function.

We call η the natural parameter, and $A(\eta)$ the log-partition function. Later, We will see that $T(x)$ is a sufficient statistic for η. **When we have the R.V. following from a family distribution, we use the subscript to emphasize the expectation is taken when the R.V. is from a pdf with a particular parameter. For example, $\mathbb{E}_\eta$ and $\mathbf{Var}_\eta$ in (iii) above means $X_1 \sim \mathbf{pdf}(x;\eta)$.**

In many cases, the parameter η used in the expression of the exponential family is given by a one-to-one transformation g as $\eta = g(\theta), \quad \theta \in \Theta$, thus, the probability density function parameterized with the natural parameter θ is given by

$$\bar{\mathrm{f}}(x;\theta) = \exp\left\{g(\theta)T(x) - A(g(\theta)) + S(x)\right\}, \quad x \in \mathcal{X}, \quad \theta \in \Omega.$$

(i) The support $\mathcal{X} = \{x : \bar{\mathrm{f}}(x;\theta) > 0\}$ does not change with respect to the parameter η.
(ii) The parameter space Θ is an open interval on the real line.
(iii) $\mathrm{Var}_\theta(T(X_1)) > 0$, i.e., $T(x)$ is not a constant function.

Remark 3.5.1. *Although we only wrote out the probability density function above, implying that we only care about the continuous RV, the same will apply for discrete distribution i.e. probability mass function (pmf). Our support will change to reflect discrete points, but everything else will remain the same.*

Many common families introduced in the previous section are exponential families. These include the continuous families—normal, gamma, and beta, and the discrete families—binomial, Poisson, and negative binomial.

Example 3.5.2 (Bernoulli)**.** To show that Bernoulli distribution is part of the Exponential Family, we must first write out the distribution in its normal form, and check that properties 1 - 3 are satisfied.

$$pmf(x;p) = p^x(1-p)^{1-x} = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}.$$

We will now try to write our pmf in the particular form given in part 2. To start we will use the fact that $a = e^{log(a)}$. Our new probability mass function is as follows.

$$pmf(x;p) = e^{xlog(p)}e^{(1-x)log(1-p)} = e^{xlog(p)+(1-x)log(1-p)}$$

From here, we compare to the Exponential Family pdf. If we had stuck with the Natural Exponential Family, there would be no way for us to write the pmf as a function of p times the function of x, so instead we will rewrite it in the form of the Exponential Family pdf.

$$\begin{aligned} pmf(x;p) &= e^{xlog(p)+(1-x)log(1-p)} && (3.11)\\ &= e^{xlog(p)+log(1-p)-xlog(1-p)} && (3.12)\\ &= e^{xlog(\frac{p}{1-p})+log(1-p)} && (3.13) \end{aligned}$$

So

$$\begin{aligned} T(x) &= x \\ A(g(\theta)) &= -\log(1-p) \\ S(x) &= 0 \end{aligned}$$

$\mathcal{X} = \{0, 1\}$, this satisfies property 1)
p ∈ (0,1), this satisfies property 2). *It is important to note that if p ∈ [0,1], then it would not be an exponential family.*

If X ∼ pmf(x;p), then we have

$$Var_p(T(X)) = Var_p(X) = p(1-p) > 0$$

which satisfies property 3).
Thus we conclude Bernoulli(p) with p ∈ (0,1) belongs to exponential family.

Example 3.5.3. Example 3.5.2 (cont)

For natural exponential family formulation, we apply the transformation

$$\eta = \log \frac{p}{1-p}.$$

Then,

$$e^{\eta} = \frac{p}{1-p}$$

$$e^{\eta}(1-p) = p$$

$$e^{\eta} = p(1+e^{\eta})$$

$$p = \frac{e^{\eta}}{1+e^{\eta}}$$

Then the probability density function with η becomes

$$\begin{aligned} pmf(x;\eta) &= e^{x\eta + log(1-\frac{e^{\eta}}{1+e^{\eta}})} && (3.14)\\ &= e^{\eta x + log(\frac{1}{1+e^{\eta}})} && (3.15)\\ &= e^{\eta x - log(1+e^{\eta})} && (3.16) \end{aligned}$$

That is,

$\text{pmf}(x;\eta) = \exp\left\{\eta x - \log\left(1+e^{\eta}\right)\right\}, \quad x = 0, 1, \quad -\infty < \eta < \infty$ (excercise: check the space of η by yourself)

So

$$\begin{aligned} T(x) &= x \\ A(\eta) &= \log(1+e^{\eta}) \\ S(x) &= 0 \end{aligned}$$

$\mathcal{X} = \{0, 1\}$, this satisfies property 1)
$\eta \in \mathbb{R}$, this satisfies property 2)

$$Var_{\eta}(T(X)) = Var_{\eta}(X) = p(1-p)|_{p=\frac{e^{\eta}}{1+e^{\eta}}} = \frac{e^{\eta}}{1+e^{\eta}} \frac{1}{1+e^{\eta}} > 0$$

this satisfies property 3)

Example 3.5.4. Note unif($[0, \theta]$) is not exponential family since the support depends on the parameter.

Lemma 3.5.5. *Suppose X has pdf in the natural exponential family*

$$pdf(x;\eta) = \exp(\eta T(x) - A(\eta) + S(x)) 1_{\mathcal{X}}(x).$$

Then

$$E_{\eta}[T(X)] = A'(\eta),$$

and

$$Var_{\eta}(T(X)) = A''(\eta).$$

Proof.

$$M_{T(X)}(s) = E[e^{sT(X)}]$$

$$= \int_{-\infty}^{\infty} e^{sT(x)} pdf(x;\eta)\, dx$$

$$= \int_{-\infty}^{\infty} e^{sT(x)} e^{\eta T(x) - A(\eta) + S(x)} 1_{\mathcal{X}}(x)\, dx$$

$$= \int_{-\infty}^{\infty} e^{(s+\eta)T(x) - A(\eta) + S(x)} 1_{\mathcal{X}}(x)\, dx$$

$$= \int_{-\infty}^{\infty} e^{(s+\eta)T(x) - A(s+\eta) + S(x)} 1_x(x) e^{A(s+\eta) - A(s)}\, dx$$

$$= e^{A(s+\eta)-A(\eta)} \int_{-\infty}^{\infty} pdf(x; s+\eta) dx$$

$$= e^{A(\eta+s)-A(\eta)}.$$

Method 1: Using the moment generating function. Taking derivatives,

$$E_\eta[T(X)] = \frac{d}{ds} mgf_{T(X)}(s)\Big|_{s=0} = e^{A(s+\eta)-A(\eta)} A'(s+\eta)|_{s=0} = A'(\eta).$$

Method 2: Using the cumulant generating function.

$$cgf_{T(X)}(s) = \log M_{T(X)}(s)$$

$$= A(\eta + s) - A(\eta)$$

Thus

$$E_\eta[T(X)] = \frac{d}{ds} cgf_{T(X)}(s)\Big|_{s=0} = A'(s+\eta)|_{s=0} = A'(\eta)$$

and

$$Var_\eta(T(X)) = \frac{d^2}{ds} cgf_{T(X)}(s)\Big|_{s=0} = A''(s+\eta)|_{s=0} = A''(\eta).$$

□

Example 3.5.6. Bernoulli Distribution (using example 3.1 set up)

$$pmf(x;p) = exp(xlog\frac{p}{1-p}) + log(1-p)); x = 0, 1$$

$$\eta = log\frac{p}{1-p} <=> p = \frac{e^x}{1+e^x}$$

$$pmf(x;\eta) = exp(x\eta - log(1+e^\eta)); x = 0, 1$$

$$\text{then: } T(X) = x, A(\eta) = log(1+e^\eta), S(x) = 0$$

By Lemma,

$$X \sim pmf(x;\eta)$$

$$E_\eta T(X) = E_\eta X = A'(\eta) = \frac{1}{1+e^\eta} e^\eta = \frac{e^\eta}{1+e^\eta}$$

$$X \sim pmf(x;p)$$

Then,

$$E_p X = E_\eta T(X) = \frac{e^\eta}{1+e^\eta} = p$$

Also by Lemma,

$$\mathrm{Var}_\eta T(X) = \mathrm{Var}_\eta X = A''(\eta) = \frac{d}{d\eta}(1 - \frac{1}{1+e^\eta}) = -\frac{d}{d\eta}((1+e^\eta)^{-1}) = -(-1)(1+e^\eta)^{-2}e^\eta = \frac{e^\eta}{(1+e^\eta)^2}$$

$$X \sim pmf(x;p)$$

Then,

$$\mathrm{Var}_p(X) = \frac{e^\eta}{1+e^\eta}\frac{1}{1+e^\eta} = p(\frac{1}{1+e^\eta}) = p(1 - \frac{e^\eta}{1+e^\eta}) = p(1-p)$$

3.5.1 Maximum Likelihood Estimation for exponential family

Recall the standard procedure for finding the maximum likelihood estimator.

1. The likelihood function is given by

$$L(\theta; X_1, \ldots, X_n) = \prod_{i=1}^n f(X_i;\theta).$$

2. The log-likelihood is

$$\ell(\theta; X_1, \ldots, X_n) = \sum_{i=1}^n \log f(X_i;\theta).$$

3. Differentiate and solve the likelihood equation

$$\frac{d}{d\theta}\,\ell(\theta; X_1, \ldots, X_n) = 0.$$

4. Use the second derivative test (or other arguments) to verify that the critical point is a maximizer.

For distributions in the exponential family, these steps can often be simplified and reduced to solving a single equation.

Theorem 3.5.7. *Let $X_1, \ldots, X_n$ be i.i.d. with pdf*

$$f(x;\eta) = \exp\{\eta T(x) - A(\eta) + S(x)\}\,\mathbf{1}_{\mathcal{X}}(x).$$

Then the likelihood equation reduces to

$$A'(\eta) = \mathbb{E}_\eta[T(X)] = \frac{1}{n}\sum_{i=1}^n T(X_i).$$

If a solution $\eta = h(X_1, \ldots, X_n)$ exists, then it is the maximum likelihood estimator of η.

Thus, instead of carrying out the usual maximum likelihood procedure step by step, it suffices to solve

$$A'(\eta) = \frac{1}{n}\sum_{i=1}^{n} T(X_i).$$

This allows us to bypass the usual steps of computing and differentiating the likelihood, and if a solution exists, the theorem tells us that it is the MLE of η.

Proof. Let $X_1, ...X_n$ be i.i.d with pdf

$$f(x;\eta) = \exp\{\eta T(x) - A(\eta) + S(x)\}\,\mathbf{1}_{\mathcal{X}}(x).$$

Then the likely hood and log-likelihood functions are as follows

$$L(\eta, X_1, ..., X_n) = \prod_{i=1}^{n} \exp(\eta T(x) - A(\eta) + S(x))\mathbf{1}_{\mathcal{X}}(x)$$

$$L(\eta, X_1, ..., X_n) = \exp(\sum_{i=1}^{n} \eta T(x_i) - nA(\eta) + \sum_{i=1}^{n} S(x_i)) \prod_{i=1}^{n} \mathbf{1}_{\mathcal{X}}(x_i)$$

$$\ell(\eta; X_1, \ldots, X_n) = \sum_{i=1}^{n} \eta T(x_i) - nA(\eta) + \sum_{i=1}^{n} S(x_i)$$

$$\frac{\partial}{\partial \eta}\ell(\eta; X_1, \ldots, X_n) = \sum_{i=1}^{n} \eta T(x_i) - nA'(\eta)$$

So, to find the critical point we set $\frac{\partial}{\partial \eta}\ell(\eta; X_1, \ldots, X_n) = 0$

$$\sum_{i=1}^{n} \eta T(x_i) - nA'(\eta) = 0$$

or equivalently

$$\sum_{i=1}^{n} \eta T(x_i) = nA'(\eta) <=> \frac{1}{n}\sum_{i=1}^{n} \eta T(x_i) = A'(\eta)$$

As you can see, this matches Theorem 2.3. So, a critical point is defined by the equation above. The next step is to prove that this critical point is the maximum. We will show this with a second derivative test.

$$\frac{\partial^2}{\partial \eta^2}\ell(\eta; X_1, \ldots, X_n) = \frac{\partial}{\partial \eta}(\sum_{i=1}^{n} \eta T(x_i) - nA'(\eta)) = -nA''(\eta)$$

By the assumption that this distribution is of the exponential family, we can say that

$$A''(\eta) = Var_{\eta}(T(X)) > 0$$

So $-nA(\eta) = -nVar_{\eta}(T(X)) < 0$ Therefore, by the second derivative test, the critical point is the MLE $\square$

Example 3.5.8. Let $X_1, ...X_n$ be i.i.d with pmf

$$pmf(x;p) = p^x(1-p)^{1-x}\mathbf{1}_{\mathcal{X}}(x), \mathcal{X} = 0, 1$$

In the general exponential family

$$pmf(x;p) = \exp(x\log(\frac{p}{1-p}) + \log(1-p))\mathbf{1}_{\mathcal{X}}(x)$$

To get it in natural exponential family form $\eta = g(p)$ where $g(p) = log(\frac{p}{1-p})$. This means that $p = \frac{e^\eta}{1+e^\eta}$ By substitution we get

$$f(x;\eta) = \exp(\eta x - \log(1+e^\eta))\mathbf{1}_{\mathcal{X}}(x)$$

And

$$T(x) = x, A(\eta) = \log(1+e^\eta), S(x) = 0$$

Next we apply the theorem

$$A'(\eta) = \frac{1}{n}\sum_{i=1}^{n} T(x_i) <=> \frac{e^\eta}{1+e^\eta} = \frac{1}{n}\sum_{i=1}^{n} T(x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n$$

Solving for η we get

$$\eta = \log(\frac{\bar{x}_n}{1-\bar{x}_n}).$$

So the MLE is

$$\hat{\eta} = \log(\frac{\bar{X}_n}{1-\bar{X}_n}).$$

Denote $E_\theta(h(X))$ when $X \sim \bar{f}(x|\theta)$ the general exponential family. It is a slight abuse of notation, since $E_\eta(h(X))$ when $X \sim f(x|\eta)$ the natural exponential family.

Lemma 3.5.9. *Let $X_1, ...X_n$ be i.i.d with $\bar{f}(x|\theta)$ the general exponential family. Then $E_\theta(T(X)) = A'(g(\theta))$*

Proof. By change of variables since $\eta = g(\theta)$

$$E_\theta(T(X)) = E_\eta(T(X)) = A'(\eta) = A'(g(\theta)),$$

where the first step follows since $X \sim \bar{f}(x|\theta)$ is exactly the same as $X \sim f(x|\eta)$. □

Theorem 3.5.10. *Let $X_1, \ldots, X_n$ be i.i.d. with pdf*

$$\bar{f}(x;\eta) = \exp\{g(\theta)T(x) - A(\eta) + S(x)\}\,\mathbf{1}_{\mathcal{X}}(x).$$

Then the likelihood equation reduces to

$$A'(g(\eta)) = \mathbb{E}_\theta[T(X)] = \frac{1}{n}\sum_{i=1}^{n} T(X_i).$$

If a solution $\theta = h(X_1, \ldots, X_n)$ exists, then it is the maximum likelihood estimator of η.

Proof. Recall $\eta = g(\theta)$, $\theta = g^{-1}(\eta)$. By the invariance principle

$$\hat{\theta}_{MLE} = g^{-1}(\hat{\eta}_{MLE})$$

$$A'(\eta_{MLE}) = \frac{1}{n}\sum_{i=1}^{n} T(x_i) <=> A'(g(\theta_{MLE})) = \frac{1}{n}\sum_{i=1}^{n} T(x_i).$$

So $\hat{\theta}_{MLE}$ is the solution to the following equation

$$E_\theta(T(X)) = A'(g(\theta)) = \frac{1}{n}\sum_{i=1}^{n} T(x_i)$$

□

Example 3.5.11. By the above theorem, the MLE for p is the solution of $E_p(T(X)) = \frac{1}{n}\sum_{i=1}^{n} T(x_i)$. Since $T(x) = x$, the equation becomes: $E_p(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$. We also have $E_p(X) = p$. Thus $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$.

Example 3.5.12. This is an example for the exponential distribution, not to be confused for the exponential family. Let $X_1, ...X_n$ be i.i.d with $pdf(x \mid \theta) = \frac{1}{\theta} e^{\frac{-x}{\theta}} \mathbf{1}_{(0,\infty)}(x)$ This is obviously of the exponential family, with general exponential form of

$$\exp(\log(\frac{1}{\theta}) - \frac{x}{\theta})\mathbf{1}_{(0,\infty)}(x)$$

Let $g(\theta) = -\frac{1}{\theta} = \eta$, so the natural exponential form is

$$\exp(\log(-\eta) + \eta x)\mathbf{1}_{(0,\infty)}(x)$$

$T(x) = x, A(\eta) = -\log(-\eta), S(x) = 0$, support $= (0, \infty)$ The three properties of the exponential family are satisfied by the exponential distribution so the above distribution is of the exponential family.

1. The support (0,∞) does not depend on θ
2. The parameter space $\Omega = (0, \infty)$ is an open interval on the real line
3. $Var_\eta(T(X_1)) > 0$, or equivalently T(x) = x is not a constant function.

So we can do the following.

$$E_\theta(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Since $E_\theta(X) = \theta$, $\hat{\theta}_{MLE} = \overline{X}_n$

Chapter 4

Sufficient Statistics and comparisons of estimators

4.1 Sufficient statistics

When estimating parameters using a random sample $X_1, \ldots, X_n$, if the estimator can be fully and precisely determined by using only some function of the sample, it may be advantageous because we only need the value of the function.

For example, to store $\bar{X}_n$ and S_n^2 needs much less space than the whole sample $(X_1, \ldots, X_n)$.

This concept of data reduction leads to the notion of sufficiency. Suppose X_1,...,X_n are iid with probability distribution function $f(x|\theta)$. The central idea is to find some statistic such that information about θ is preserved.

Definition 4.1.1 (Sufficiency Principle)**.** If $T(x_1, ..., x_n)$ is a sufficient statistic for θ, then any inference about θ should depend on $(X_1, ..., X_n)$ only through the value of $T(X_1, ..., X_n)$. That is, if $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$ are two sample points such that $T(x_1, ..., x_n) = T(y_1, ..., y_n)$, then the inference about θ should be the same, whether $(X_1, ...X_n) = (x_1, ..., x_n)$ or $(X_1, ..., X_n) = (y_1, ..., y_n)$

Definition 4.1.2. For a probability density function $f(x; \theta)$ with parameter $\theta \in \Omega$ from the sample $X_1, \ldots, X_n$, the statistic $Y = u(X_1, \ldots, X_n)$ is sufficient for θ if

$$P_{\theta_1}((X_1, \ldots, X_n)^T \in A \mid Y = y) = P_{\theta_2}((X_1, \ldots, X_n)^T \in A \mid Y = y) \quad \forall A, y, \text{ and } \forall \theta_1, \theta_2 \in \Omega.$$

Equivalently, if the conditional distribution of $(X_1, \ldots, X_n)^T$ given the statistic $Y = u(X_1, \ldots, X_n) = y$ does not depend on $\theta \in \Omega$, we call Y a sufficient statistic for $\theta \in \Omega$.

In simpler terms, this means that θ does not depend on $(X_1, ...X_n)$, if Y is known. This is discussed in more detail in the next lecture.

Provide some intuition.

Example 4.1.3. Consider a random sample $X_1, ..., X_n$ from a Bernoulli distribution with parameter θ, $0 < \theta < 1$. Then,

$Y = \sum_{i=1}^n X_i$ follows Bin(n, θ) and

$$P_\theta(X_1 = x_1, ..., X_n = x_n | Y = y) = \frac{P(X_1 = x_1, ..., X_n = x_n, Y = y)}{P(Y = y)}$$

$$= \frac{P(X_1 = x_1, ..., X_n = x_n, Y = y)}{P(Y = y)} * 1(y = \sum_{i=1}^n x_i)$$

$$= \frac{\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} * 1_{y=\sum_{i=1}^{n} x_i}}{\binom{n}{y}\theta^{y}(1-\theta)^{n-y}}$$

$$= \frac{\theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} * 1(y = \sum_{i=1}^{n} x_i)}{\binom{n}{y}\theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}}$$

$$= \frac{1(y = \sum_{i=1}^{n} x_i)}{\binom{n}{y}}$$

Since this does not depend on θ, Y is a sufficient statistic for θ by definition.

Given the likelihood function, how can we find a sufficient statistic?

Theorem 4.1.4. *Suppose X_1,...,X_n are iid with probability distribution function $f(x|\theta)$ for $\theta \in \Re$. Then, $Y = u(X_1, ..., X_n)$ is a sufficient statistic for $\theta \in \Re$ if and only if there exists two functions k_1, k_2 satisfying*

$$L(\theta|x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta) = k_1(u(x), \theta) * k_2(x)$$

$x = (x_1, ..., x_n), \forall \theta \in \Re$.

Proof. We will only prove the case when $X_1, .., X_n$ are discrete R.V..
($\Leftarrow$):

$$P_\theta(X_1 = x_1, ..., X_n = x_n|Y = y)$$

$$= \frac{P_\theta(X_1 = x_1, ..., X_n = x_n, Y = y)}{P(Y = y)}$$

$$= \frac{P_\theta(X_1 = x_1, ..., X_n = x_n, Y = y) * 1(y = u(x_1, ..., x_n))}{P(Y = y)}$$

$$= \frac{P_\theta(X_1 = x_1, .., X_n = x_n) * 1_{y=u(x_1,...,x_n)}}{P(Y = y)}$$

$$= \frac{k_1(u(x_1, ..., x_n), \theta)k_2(x_1, ..., x_n) * 1_{y=u(x_1,...,x_n)}}{P(u(X_1, .., X_n) = y)}$$

$$Z = \{(z_1, ..., z_n) : u(z_1, ..., z_n) = y)\}$$

$$= \frac{k_1(u(x_1, ..., x_n), \theta)k_2(x_1, ..., x_n) * 1_{y=x_1,...,x_n}}{\sum_{(z_1,...,z_n)\in Z} P(X_1 = z_1, ..., X_n = z_n)}$$

$$= \frac{k_1(u(x_1, ..., x_n), \theta)k_2(x_1, ..., x_n) * 1_{y=u(x_1,...,x_n)}}{\sum_{(z_1,...,z_n)\in Z} k_1(u(z_1, ...z_n), \theta)k_2(z_1, ..., z_n)}$$

$$= \frac{k_2(x_1, ..., x_n) * 1_{y=u(x_1,...,x_n)}}{\sum_{(z_1,...,z_n)\in Z} k_2(z_1, ..., z_n)}$$

Therefore, $Y = u(X_1, ..., X_n)$ is a sufficient statistic for $\theta \in \Omega$
($\Rightarrow$):
Assume $Y = u(X_1, ..., X_n)$ is sufficient for $\theta \in \Re$. Therefore,

$$k_2(x_1, ..., x_n) = P_\theta(X_1 = x_1, ..., X_n = x_n|Y = u(x_1, ..., x_n))$$

is independent of θ. Additionally, we let

$$k_1(y, \theta) := P_\theta(Y = y).$$

Then,

$$P_\theta(X_1 = x_1, ..., X_n = x_n)$$

$$= P_\theta(Y = u(x_1, ..., x_n)) * P(X_1 = x_1, ..., X_n = x_n | Y = u(x_1, ..., x_n))$$

$$= k_1(u(x_1, ..., x_n), \theta) * k_2(x_1, ..., x_n)$$

,

□

Remark 4.1.5. The above theorem also applies that when the data are not iid. As long as the joint distribution factorizes, the conclusion still holds.

Example 4.1.6. Consider a random sample $X_1, ..., X_n$ from a Poisson distribution with parameter $\theta > 0$.

$$\prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} (\frac{e^{-\theta} * \theta^{x_i}}{x_i!}) = \frac{e^{-n\theta} * \theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n}(x_i!)}$$

$$= e^{-n\theta} * \theta^{\sum_{i=1}^{n} x_i} * \frac{1}{\prod_{i=1}^{n}(x_i!)}$$

Define

$$k_1(y, \theta) = e^{-n\theta}\theta^y$$

and

$$k_2(x_1, \ldots, x_n) = \frac{1}{\prod_{i=1}^{n}(x_i!)}$$

Thus

$$\prod_{i=1}^{n} f(x_i; \theta) = k_1\left(\sum_{i=1}^{n} x_i, \theta\right) \prod_{i=1}^{n} \frac{1}{x_i!}.$$

Therefore, by the Factorization Theorem, $Y = \sum_{i=1}^{n} X_i$ is a sufficient statistic for $\theta > 0$.

When the support depends on the parameter, one has to be careful by using indicator function.

Example 4.1.7 (Example 6.28 (Uniform Sufficient Statistic))**.** Consider a random sample $X_1, ..., X_n$ from Unif$(1, 2, ..., \theta)$ with $pmf(x|\theta) = \frac{1}{\theta}$, $x \in \{1, 2, \ldots, \theta\} := Z_\theta$.

$$\prod_{i=1}^{n} pmf(x_i|\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} 1_{Z_\theta(x_i)}$$

$$\prod_{i=1}^{n} 1_{z_\theta(X_i)} = (\prod_{i=1}^{n} 1_N(X_i)) 1_{max(x_i \leq \theta)}$$

where N is the natural number.

$$\prod_{i=1}^{n} pmf(x_i|\theta) = \frac{1}{\theta^n} (\prod_{i=1}^{n} 1_N(x_i)) * (1_{(0,\theta)}(max(X_i)))$$

If we let

$$k_1(y,\theta) = \frac{1}{\theta^n} 1_{(0,\theta)}(y)$$

and

$$u(x_1, ..., x_n) = maxX_i = X_{(n)}$$

and

$$k_2(x_1, \ldots, x_n) = \prod_{i=1}^{n} 1_N(x_i)$$

then

$$\frac{1}{\theta^n}(\prod_{i=1}^{n} 1_N(x_i)) * (1_{(0,\theta)}(max(X_i))) = k_1(X_{(n)}, \theta) * k_2(x_1, ..., x_n)$$

Thus, $Y = X_{(n)}$ is a sufficient statistic for θ.

Remark 4.1.8. (1) $Y = u(X_1, \ldots, X_n)$ can be a vector; θ can also be a vector. Typically, if θ has k components, we need k sufficient statistics. For instance, if

$$X_1, \ldots, X_n \sim \text{Gamma}(\alpha, \beta),$$

then $\theta = (\alpha, \beta)$ and we need a vector of sufficient statistics.

(2) In the last lecture, we assumed the data are i.i.d.:

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

In fact, the i.i.d. assumption is not needed. The only thing required is that the joint pdf can be factorized into the two terms k_1 and k_2. As long as we can write down the joint pdf, the conclusion still holds, even when the data are not i.i.d. That is, $Y = u(X_1, \ldots, X_n)$ is sufficient for θ if and only if

$$f(x_1, \ldots, x_n \mid \theta) = k_1\big(u(x_1, \ldots, x_n), \theta\big)\, k_2(x_1, \ldots, x_n).$$

Example. Consider observing the order statistics $X_{(1)}, \ldots, X_{(n)}$ instead of the original i.i.d. data. The order statistics are not i.i.d., but we still have a formula for their joint density:

$$f_{X_{(1)},\ldots,X_{(n)}}(y_1, \ldots, y_n \mid \theta) = \cdots$$

Since we can write down this joint pdf, we can apply the factorization theorem directly to find sufficient statistics, even though the data are dependent.

4.1.1 k-parameter Natural Exponential Family

We can also define the multivariate exponential family. The probability density function is given by

$$\text{pdf}(x;\eta) = \exp\left(\sum_{j=1}^{k} \eta_j T_j(x) - A(\eta) + S(x)\right), \quad x \in \mathcal{X}, \quad \eta = (\eta_1, \ldots, \eta_k)^T \in \Omega$$

where the parameter η satisfies

(i) The support of the distribution $\mathcal{X} = \{x : \text{pdf}(x;\eta) > 0\}$ does not change with respect to η.

(ii) The sample space Ω contains an open k-dimensional region in $\mathbb{R}^k$ defined by the Cartesian product of intervals $(a_1, b_1) \times \cdots \times (a_k, b_k)$.

(iii) For any real constants $c_1, \ldots, c_k$, the linear combination $c^T T(x) = c_1 T_1(x) + \cdots + c_k T_k(x)$ is not a constant. That is, for any nonzero vector $c \in \mathbb{R}^k$, the variance is positive:

$$\mathrm{Var}(c^T T(X)) > 0 \quad \forall\, c \in \mathbb{R}^k,\, c \neq 0.$$

This condition is equivalent to that $\mathrm{cov}(T(X), T(X))$ is positive definite.

Key idea. Previously we defined the exponential family with a single natural parameter η (that is, $k = 1$), so the exponent contained just $\eta T(x)$. Now we generalize to k parameters. This is important because many common distributions, such as $N(\mu, \sigma^2)$, have more than one parameter.

Equivalently, in terms of the original parameters $\theta_1, \ldots, \theta_k$,

$$\bar{f}(x \mid \theta_1, \ldots, \theta_k) = \exp\left(\sum_{j=1}^{k} g_j(\theta_1, \ldots, \theta_k) T_j(x) - A\big(g_1(\theta_1, \ldots, \theta_k), \ldots, g_k(\theta_1, \ldots, \theta_k)\big) + S(x) \right) \mathbf{1}_X(x).$$

The change between this two parametrizations are

$$\eta_j = g_j(\theta_1, \ldots, \theta_k), \qquad j = 1, \ldots, k.$$

Example 4.1.9 (Example: $N(\mu, \sigma^2)$)**.** We want to show that $N(\mu, \sigma^2)$ belongs to a 2-parameter exponential family. Start with

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right).$$

Step 1: Combine into a single exponential. Using $a = e^{\log a}$, rewrite the leading constant:

$$f(x \mid \mu, \sigma^2) = \exp\big(-\log(\sqrt{2\pi}) - \log\sigma\big) \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right),$$

so

$$f(x \mid \mu, \sigma^2) = \exp\left(-\log(\sqrt{2\pi}) - \log\sigma - \frac{(x-\mu)^2}{2\sigma^2} \right).$$

Step 2: Expand and rearrange.

$$\begin{aligned} f(x \mid \mu, \sigma^2) &= \exp\left(-\log(\sqrt{2\pi}) - \log\sigma - \frac{x^2 - 2\mu x + \mu^2}{2\sigma^2} \right) \\ &= \exp\left(-\log(\sqrt{2\pi}) - \log\sigma - \frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2} x - \frac{\mu^2}{2\sigma^2} \right) \\ &= \exp\left(\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \log\sigma - \frac{\mu^2}{2\sigma^2} - \log(\sqrt{2\pi}) \right). \end{aligned}$$

Reading off the exponential family components:

$$\underbrace{\frac{\mu}{\sigma^2}}_{g_1(\mu,\sigma^2)} \underbrace{x}_{T_1(x)} + \underbrace{\left(-\frac{1}{2\sigma^2}\right)}_{g_2(\mu,\sigma^2)} \underbrace{x^2}_{T_2(x)} - \underbrace{\left(\log\sigma + \frac{\mu^2}{2\sigma^2}\right)}_{A(g_1,g_2)} + \underbrace{\left(-\log\sqrt{2\pi}\right)}_{S(x)}.$$

Therefore,

$$N(\mu, \sigma^2)$$

belongs to a 2-parameter exponential family. □

4.1.2 Sufficient Statistics for Exponential Families

Theorem 4.1.10. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples from a distribution belonging to an exponential family*

$$f(x \mid \theta) = \exp\left(\sum_{j=1}^{k} g_j(\theta)T_j(x) - A\big(g_1(\theta), \ldots, g_k(\theta)\big) + S(x)\right)\mathbf{1}_{\mathcal{X}}(x).$$

Then the sufficient statistic for $\theta \in \Theta$ is

$$\left(\sum_{i=1}^{n} T_1(X_i), \sum_{i=1}^{n} T_2(X_i), \ldots, \sum_{i=1}^{n} T_k(X_i)\right).$$

Proof.

$$\prod_{i=1}^{n} f(x_i \mid \theta) = \prod_{i=1}^{n}\left[\exp\left(\sum_{j=1}^{k} g_j(\theta)T_j(x_i) - A\big(g_1(\theta), \ldots, g_k(\theta)\big) + S(x_i)\right)\mathbf{1}_{\mathcal{X}}(x_i)\right].$$

So,

$$\begin{aligned}\prod_{i=1}^{n} f(x_i \mid \theta) &= \exp\left(\sum_{j=1}^{k} g_j(\theta)\sum_{i=1}^{n} T_j(x_i) - nA\big(g_1(\theta), \ldots, g_k(\theta)\big) + \sum_{i=1}^{n} S(x_i)\right)\prod_{i=1}^{n}\mathbf{1}_{\mathcal{X}}(x_i) \\ &= \underbrace{\exp\left(\sum_{j=1}^{k} g_j(\theta)\sum_{i=1}^{n} T_j(x_i) - nA\big(g_1(\theta), \ldots, g_k(\theta)\big)\right)}_{k_1(Y_1, \ldots, Y_k, \theta)} \underbrace{\exp\left(\sum_{i=1}^{n} S(x_i)\right)\prod_{i=1}^{n}\mathbf{1}_{\mathcal{X}}(x_i)}_{k_2(x_1, \ldots, x_n)},\end{aligned}$$

where

$$Y_j = \sum_{i=1}^{n} T_j(X_i), \qquad j = 1, \ldots, k.$$

By the Factorization Theorem, the proof is complete. □

Example 4.1.11 (Example: $N(\mu, \sigma^2)$ continued.)**.** We found

$$T_1(x) = x, \qquad T_2(x) = x^2.$$

So by the theorem, the sufficient statistic for $\theta = (\mu, \sigma^2)$ is

$$\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right).$$

Question. Will $(\overline{X}_n, S_n^2)$ be sufficient for (μ, σ^2) in the previous example? More generally, is the sufficient statistic unique?

4.1.3 Non-Uniqueness of Sufficient Statistics

Theorem 4.1.12. *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x \mid \theta)$, $\theta \in \Theta$. Then for any one-to-one function h,*

$$W = h(Y)$$

is also a sufficient statistic for $\theta \in \Theta$.

Proof. For all A and for all y,

$$P_\theta\big((X_1,\ldots,X_n)\in A \mid Y=y\big)$$

does not depend on θ since Y is sufficient. Now For all A and for all w

$$\begin{aligned} P_\theta\big((X_1,\ldots,X_n)\in A \mid W=w\big) &= P_\theta\big((X_1,\ldots,X_n)\in A \mid h(Y)=w\big) \\ &= P_\theta\big((X_1,\ldots,X_n)\in A \mid Y=h^{-1}(w)\big), \end{aligned}$$

which does not depend on θ.

Hence W is sufficient for $\theta\in\Theta$. □

Example 4.1.13 (Example: $N(\mu,\sigma^2)$ continued.)**.** We know that

$$(Y_1,Y_2) := \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right).$$

is sufficient for (μ,σ^2). It holds that

$$\bar{X}_n = \frac{1}{n}Y_1$$

and

$$S_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\bar{X}_n^2\right) = \frac{1}{n-1}\left(Y_2 - \frac{1}{n}Y_1^2\right).$$

It is easy to see that the two-dimensional map from (Y_1,Y_2) to $(\bar{X}_n,S_n^2)$ are one-to-one. So by the above theorem, $(\bar{X}_n,S_n^2)$ is sufficient for (μ,σ^2).

4.2 Methods of Evaluating Estimators (Section 7.3)

We will continue the topic explored briefly in the last lecture: methods of evaluating estimators. Our motivation is to compare different types of estimators. Previously, we introduced MLE and MME. However, there are more types of estimators out there.

Key Questions. Since we have different options for a single parameter we seek to estimate, which estimator should we choose? In many cases, different methods of finding estimators will yield different results. How do we decide which method is better?

Definition 4.2.1. (Bias). The *bias* of a point estimator W of a parameter θ is

$$\text{Bias}_\theta W = E_\theta W - \theta.$$

An estimator with $\text{Bias}_\theta W = 0$ for all θ is called *unbiased*.

The bias of an estimator is a function of θ if we let θ changes.

Example. Let $X_1,\ldots,X_n$ be from a distribution with mean μ and variance $\sigma^2<\infty$. Then

$$E[\overline{X}_n]=\mu, \qquad E[S_n^2]=\sigma^2.$$

So $\overline{X}_n$ is an unbiased estimator for μ, and S_n^2 is an unbiased estimator for σ^2. □

Mean Squared Error

$$\begin{aligned}\mathrm{MSE}(W,\theta) &:= E_\theta\big[(W-\theta)^2\big] = E_\theta\big[(W - E_\theta W + E_\theta W - \theta)^2\big] \\ &= E_\theta\big[(W - E_\theta W)^2\big] + (E_\theta W - \theta)^2\end{aligned}$$

because the cross-term vanishes. Therefore,

$$\mathrm{MSE}(W,\theta) = \mathrm{Var}_\theta(W) + \big(\mathrm{Bias}_\theta(W)\big)^2.$$

This gives the following interpretation:

$$\mathrm{MSE} = \mathrm{Variance} + \mathrm{Bias}^2.$$

Definition 4.2.2. (MSE). The *mean squared error* (MSE) of an estimator W of a parameter θ is the function of θ defined by

$$E_\theta(W-\theta)^2 = Var_\theta W + (E_\theta W - \theta)^2 = Var_\theta W + (Bias_\theta W)^2.$$

The MSE of an estimator is a function of θ if we let θ changes.

MSE has 2 main advantages over other distance measures:

1. It has a nice interpretation - if the MSE is small, then our estimator has a small variance. This means our estimator is close to its mean. Moreover, the bias is also small. Hence, the mean is close to the true parameter ($E_\theta W \approx \theta$).
2. It is easy to calculate (easy to differentiate w.r.t. θ) since it is squared.

Remark 4.2.3. You may consider some other possible distance. For instance, we may choose to calculate the absolute value instead:

$$E_\theta|W-\theta|.$$

However, it will not be easy to differentiate (we lose the 2nd advantage of MSE). Thus, although there are many other possible candidates, MSE is most popular in practice. In this course, we will focus on MSE.

4.2.1 Calculating the MSE

Example 4.2.4 (Normal MSE)**.** Let $X_1,\ldots,X_n$ be i.i.d. $N(\mu,\sigma^2)$. The sample mean $\bar{X}_n$ and sample variance S_n^2 satisfy

$$E_{(\mu,\sigma^2)}\bar{X}_n = \mu, \qquad E_{(\mu,\sigma^2)}S_n^2 = \sigma^2,$$

which was established previously. That is, $\bar{X}_n$ and S_n^2 are unbiased estimators of μ and σ^2, respectively.

Since $\bar{X}_n$ is unbiased,

$$E_{(\mu,\sigma^2)}(\bar{X}_n-\mu)^2 = Var(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Check previous lecture notes for this calculation. Therefore,

$$MSE(\bar{X}_n) = \frac{\sigma^2}{n}.$$

even if the normality assumption is dropped.

Only under normality (normal distribution),

$$Var(S_n^2) \overset{(exercise)}{=} \frac{2\sigma^4}{n-1}.$$

Therefore,

$$MSE(S_n^2) = \frac{2\sigma^4}{n-1}.$$

Although unbiased estimators are often reasonable from the standpoint of MSE, controlling bias alone does not guarantee that the MSE is small. In some cases, a small increase in bias can produce a larger decrease in variance and hence a smaller MSE (a concept known as the bias-variance tradeoff).

Example 4.2.5 (Continuation of Example 2.1)**.** For $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, the maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n}S_n^2.$$

Its expectation is

$$E\hat{\sigma}^2 = \frac{n-1}{n}E(S_n^2) = \frac{n-1}{n}\sigma^2,$$

where $\hat{\sigma}^2$ is a biased estimator of σ^2 that on average always underestimate σ^2.

Its variance is

$$\begin{aligned} Var(\hat{\sigma}^2) &= Var\left(\frac{n-1}{n}S_n^2\right) = \left(\frac{n-1}{n}\right)^2 Var(S_n^2) \\ &= \left(\frac{n-1}{n}\right)^2 \cdot \frac{2\sigma^4}{n-1} = \frac{2(n-1)}{n^2}\sigma^4. \end{aligned}$$

Hence,

$$\begin{aligned} MSE(\hat{\sigma}^2) &= E[(\hat{\sigma}^2 - \sigma^2)^2] \\ &= Var(\hat{\sigma}^2) + [(\text{Bias}(\hat{\sigma}^2)]^2 \\ &= \frac{2(n-1)}{n^2}\sigma^4 + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 \\ &= \frac{2(n-1)}{n^2}\sigma^4 + \frac{1}{n^2}\sigma^4 \\ &= \frac{2n-1}{n^2}\sigma^4. \end{aligned}$$

and

$$\begin{aligned} MSE(S_n^2) &= Var(S_n^2) + [\text{Bias}(S_n^2)]^2 \\ &= Var(S_n^2) \\ &= \frac{2\sigma^4}{n-1}. \end{aligned}$$

Since

$$\frac{2n-1}{n^2} < \frac{2}{n-1},$$

we obtain

$$MSE(\hat{\sigma}^2) < MSE(S_n^2) \quad \text{for all } \sigma^2,$$

showing the biased estimator $\hat{\sigma}^2$ has smaller MSE than the unbiased estimator S_n^2. Thus, MSE is improved by trading off variance for bias.

This comparison does not imply that S_n^2 should be abandoned as an estimator of σ^2. The argument only shows that, on average and under squared-error loss, $\hat{\sigma}^2$ is closer to σ^2 than S_n^2. Since $\hat{\sigma}^2$ systematically underestimates σ^2, one may still prefer S_n^2 in some contexts.

Thus, the goal is to gather more information about the estimators in hopes of choosing a good estimator (since no absolute answer can be obtained).

Example 4.2.6. In example 2.2, we have a function uniformly smaller than the other function. Here, MLE is clearly a better method than sample variance. Note that this is only true on a particular domain, where one MSE is smaller than the other. We provide an example where such a comparison is not straightforward. Let $X_1, \ldots, X_n$ be i.i.d $N(\theta, 1)$ where $\theta \in R$. We know the typical estimator is

$$\hat{\theta}_1 := \bar{X}_n.$$

Let's define our second estimator

$$\hat{\theta}_2 := 0.$$

We predict our unknown parameter to be 0 (although in principal, this estimator is naive). Now, by the previous example

$$MSE(\hat{\theta}_1) = \frac{1}{n}.$$

Let's calculate MSE of the second estimator. By definition,

$$MSE(\hat{\theta}_2) = \mathbb{E}(\hat{\theta}_2 - \theta)^2 = \mathbb{E}(0 - \theta)^2 = \theta^2.$$

If $|\theta| < \frac{1}{\sqrt{n}}$,

$$MSE(\hat{\theta}_2) < MSE(\hat{\theta}_1).$$

If $|\theta| > \frac{1}{\sqrt{n}}$,

$$MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2).$$

The MSE of the function parameter θ is only smaller on some domain and depends on the value of the true parameter, which is unknown. In other words, it is typically hard for us to determine the best estimator in terms of MSE, because it may vary by the domain.

4.2.2 Best Unbiased Estimators

One way to approach this problem is to constrain the space of all possible estimators. Thus, we can constrain our candidates to be unbiased estimators.

Let us consider only unbiased estimators. What is the "best" unbiased estimator? Recall that if W is unbiased of θ,

$$MSE_\theta(W) = Var_\theta(W).$$

Among all possible unbiased estimators, we want to find the estimator with the smallest MSE. Therefore, we essentially want to find the estimator with the smallest variance. We want this to hold for any θ.

Definition 4.2.7. An estimator W^* is a best unbiased estimator of $\tau(\theta)$ if it satisfies:

1. $E_\theta W^* = \tau(\theta)$
2. $Var_\theta(W^*) \leq Var_\theta(W)$ for any other W unbiased estimator of $\tau(\theta)$. This holds for all θ.

W^* is also called a uniform minimum variance unbiased estimator (UMVUE) of $\tau(\theta)$.

T/F Question. $\tau(\theta)$ is a UMVUE.
False. $\tau(\theta)$ is not an estimator. Recall that an estimator is a function of $(X_1, \ldots, X_n)$ that does not depend on θ.

4.3 Cramer-Rao Lower Bound

In this lecture, we introduce the concept of the Uniformly Minimum Variance Unbiased Estimator (UMVUE), define the Fisher information number, and state and prove the Cramér–Rao inequality. This provides a fundamental lower bound on the variance of any unbiased estimator, which we use to identify UMVUEs. We close with an example for the Poisson distribution.

Before we prove the Cramér–Rao Inequality, we will introduce some concepts

4.3.1 Fisher Information

Definition 4.3.1. Consider $(X_1, \ldots, X_n) \sim \bar{f}(x_1, \ldots, x_n|\theta)$, where $\bar{f}$ denotes the joint density (or mass) function. Define:

$$I_n(\theta) := \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log \bar{f}(X_1, \ldots, X_n|\theta) \right)^2 \right]$$

This is called the Fisher information number.

The quantity

$$S(\theta|x_1, \ldots, x_n) := \frac{\partial}{\partial \theta} \log \bar{f}(x_1, \ldots, x_n \mid \theta)$$

is called the *score function*.

Proposition 4.3.2. *Let* $(X_1, \ldots, X_n) \sim \bar{f}(x_1, \ldots, x_n|\theta)$.

Assume we can switch order of differentiation and integration. Then:

(a)

$$\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \log \bar{f}(X_1, \ldots, X_n|\theta) \right] = 0$$

(b)

$$\begin{aligned} I_n(\theta) &= \mathrm{Var} \left(\frac{\partial}{\partial \theta} \log \bar{f}(X_1, \ldots, X_n|\theta) \right) \\ &= \mathbb{E}_\theta \left[-\frac{\partial^2}{\partial \theta^2} \log \bar{f}(X_1, \ldots, X_n|\theta) \right] \end{aligned}$$

Proof. (a) Using the identity $\frac{\partial}{\partial \theta} \log \bar{f} = \frac{\frac{\partial}{\partial \theta} \bar{f}}{\bar{f}}$, we write

$$\begin{aligned} \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \log \bar{f}(X_1, \ldots, X_n \mid \theta) \right] &= \int \frac{\partial}{\partial \theta} \log \bar{f}(x_1, \ldots, x_n \mid \theta)\, \bar{f}(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n \\ &\overset{\text{chain rule}}{=} \int \frac{\frac{\partial}{\partial \theta} \bar{f}(x_1, \ldots, x_n \mid \theta)}{\bar{f}(x_1, \ldots, x_n \mid \theta)} \bar{f}(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n \\ &= \int \frac{\partial}{\partial \theta} \bar{f}(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n \\ &= \frac{\partial}{\partial \theta} \int \bar{f}(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n \\ &= \frac{\partial}{\partial \theta}(1) \\ &= 0 \end{aligned}$$

(b) Since the score function has mean zero by part (a), its variance equals its second moment, which is $I_n(\theta)$ by definition. It remains to show that $I_n(\theta) = \mathbb{E}_\theta \left[-\frac{\partial^2}{\partial \theta^2} \log \bar{f} \right]$. Recall the quotient rule:

$$\frac{d}{d\theta} \left(\frac{u(\theta)}{v(\theta)} \right) = \frac{u'(\theta)v(\theta) - u(\theta)v'(\theta)}{v^2(\theta)}.$$

Differentiating the score function with respect to θ:

$$\frac{\partial^2}{\partial\theta^2}\log\bar{f} = \frac{\frac{\partial^2}{\partial\theta^2}\bar{f}\cdot\bar{f} - \left(\frac{\partial}{\partial\theta}\bar{f}\right)^2}{\bar{f}^2}.$$

Therefore,

$$\mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log\bar{f}(X_1,\ldots,X_n \mid \theta)\right]$$

$$= -\int \frac{\partial^2}{\partial\theta^2}\log\bar{f}(x_1,\ldots,x_n \mid \theta)\,\bar{f}(x_1,\ldots,x_n \mid \theta)\,dx_1\cdots dx_n$$

$$= -\int \left[\frac{\frac{\partial^2}{\partial\theta^2}\bar{f}(x_1,\ldots,x_n \mid \theta)}{\bar{f}(x_1,\ldots,x_n \mid \theta)} - \frac{\left(\frac{\partial}{\partial\theta}\bar{f}(x_1,\ldots,x_n \mid \theta)\right)^2}{\left(\bar{f}(x_1,\ldots,x_n \mid \theta)\right)^2}\right] \times \bar{f}(x_1,\ldots,x_n \mid \theta)\,dx_1\cdots dx_n$$

$$= -\left[\int \frac{\partial^2}{\partial\theta^2}\bar{f}(x_1,\ldots,x_n \mid \theta)\,dx_1\cdots dx_n - \int \frac{\left(\frac{\partial}{\partial\theta}\bar{f}(x_1,\ldots,x_n \mid \theta)\right)^2}{\bar{f}(x_1,\ldots,x_n \mid \theta)}\,dx_1\cdots dx_n\right]$$

$$= -\left[\int \frac{\partial^2}{\partial\theta^2}\bar{f}(x_1,\ldots,x_n \mid \theta)\,dx_1\cdots dx_n - \int \left(\frac{\frac{\partial}{\partial\theta}\bar{f}(x_1,\ldots,x_n \mid \theta)}{\bar{f}(x_1,\ldots,x_n \mid \theta)}\right)^2 \bar{f}(x_1,\ldots,x_n \mid \theta)\,dx_1\cdots dx_n\right]$$

$$= -\left[\frac{\partial^2}{\partial\theta^2}\int \bar{f}(x_1,\ldots,x_n \mid \theta)\,dx_1\cdots dx_n - \int \left(\frac{\partial}{\partial\theta}\log\bar{f}(x_1,\ldots,x_n \mid \theta)\right)^2 \bar{f}(x_1,\ldots,x_n \mid \theta)\,dx_1\cdots dx_n\right]$$

$$= -\left[0 - \int \left(\frac{\partial}{\partial\theta}\log\bar{f}(x_1,\ldots,x_n \mid \theta)\right)^2 \bar{f}(x_1,\ldots,x_n \mid \theta)\,dx_1\cdots dx_n\right]$$

$$= \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log\bar{f}(X_1,\ldots,X_n \mid \theta)\right)^2\right]$$

$$= I_n(\theta).$$

□

Remark 4.3.3. If $X_1,\ldots,X_n \overset{\text{iid}}{\sim} f(x|\theta)$, then:

$$\bar{f}(X_1,\ldots,X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta) = L(\theta \mid x_1,\ldots,x_n)$$

and the log likelihood is

$$\log \bar{f}(X_1, \ldots, X_n|\theta) = \sum_{i=1}^{n} \log f(X_i|\theta) = \ell(\theta|X_1, \ldots, X_n)$$

Then

$$\begin{aligned} I_n(\theta) &= \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2} \log \bar{f}(X_1, \ldots, X_n \mid \theta)\right] \\ &= \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2} \ell(\theta \mid X_1, \ldots, X_n)\right] \\ &= \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2} \sum_{i=1}^{n} \log f(X_i \mid \theta)\right] \\ &= \sum_{i=1}^{n} \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2} \log f(X_i \mid \theta)\right] \\ &= nI_1(\theta). \end{aligned}$$

4.3.2 Cramér–Rao Inequality / Lower Bound

Theorem 4.3.4. ***(a)*** *Let $W = W(X_1, \ldots, X_n)$ be any estimator satisfying*

$$\frac{d}{d\theta}\,\mathbb{E}_\theta[W] = \int \frac{\partial}{\partial\theta}\big[W(x_1, \ldots, x_n)\,\bar{f}(x_1, \ldots, x_n \mid \theta)\big]\,dx_1 \cdots dx_n,$$

and $\mathrm{Var}_\theta(W) < \infty$.

Then

$$\mathrm{Var}_\theta(W) \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[W]\right)^2}{I_n(\theta)}.$$

(b) *If W is unbiased for $\tau(\theta)$, i.e.,*

$$\mathbb{E}_\theta[W] = \tau(\theta),$$

then

$$\mathrm{Var}_\theta(W) \geq \frac{(\tau'(\theta))^2}{I_n(\theta)}.$$

Proof. **(a)**

Recall the Cauchy-Schwarz inequality. For any two random variables X and Y,

$$(\,Cov(X, Y))^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y)$$

which rearranges to $\mathrm{Var}(X) \geq \dfrac{(\mathrm{Cov}(X, Y))^2}{\mathrm{Var}(Y)}$.

Choose

$$X = W \qquad Y = \frac{\partial}{\partial\theta} \log \bar{f}(X_1, \ldots, X_n|\theta)$$

Then

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

Since $\mathrm{Cov}(W,Y) = \mathbb{E}[WY] - \mathbb{E}[W]\mathbb{E}[Y]$ but $\mathbb{E}[Y] = 0$, then:

$$\begin{aligned}
\mathrm{Cov}(W,Y) &= \mathbb{E}[WY] - \mathbb{E}[W]\mathbb{E}[Y] \\
&= \mathbb{E}[WY] \qquad (\mathbb{E}[Y] = 0 \text{ from previous proposition}) \\
&= \mathbb{E}\left[W \cdot \frac{\partial}{\partial\theta} \log \bar{f}(X_1, \ldots, X_n \mid \theta)\right] \\
&= \int w(x_1, \ldots, x_n) \frac{\partial}{\partial\theta} \log \bar{f}(x_1, \ldots, x_n \mid \theta)\, \bar{f}(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n \\
&= \int w(x_1, \ldots, x_n) \frac{\frac{\partial}{\partial\theta}\bar{f}(x_1, \ldots, x_n \mid \theta)}{\bar{f}(x_1, \ldots, x_n \mid \theta)} \bar{f}(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n \\
&= \int w(x_1, \ldots, x_n) \frac{\partial}{\partial\theta} \bar{f}(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n \\
&= \frac{\partial}{\partial\theta} \int w(x_1, \ldots, x_n)\, \bar{f}(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n \\
&= \frac{\partial}{\partial\theta} \mathbb{E}_\theta[W].
\end{aligned}$$

Therefore, by Cauchy–Schwarz,

$$\mathrm{Var}_\theta(W) \geq \frac{(\mathrm{Cov}_\theta(W,Y))^2}{\mathrm{Var}_\theta(Y)} = \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W\right)^2}{I_n(\theta)}.$$

(b)

When W is unbiased for $\tau(\theta)$, we have $\frac{d}{d\theta}\mathbb{E}_\theta W = \tau'(\theta)$. Substituting into part (a) gives

$$\mathrm{Var}_\theta(W) \geq \frac{(\tau'(\theta))^2}{I_n(\theta)}.$$

□

Remark 4.3.5. If we are estimating the parameter θ itself, then $\tau(\theta) = \theta$ and if W is unbiased for θ, then

$$\mathrm{Var}_\theta(W) \geq \frac{1}{I_n(\theta)}.$$

Example 4.3.6 (UMVUE for the Poisson Distribution)**.** We have $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$.

Goal is to find the UMVUE of λ; that is, $\tau(\lambda) = \lambda$.

Step 1: Compute $I_1(\lambda)$. The Poisson PMF is $f(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$.

$$\begin{aligned} l_1(\lambda \mid x_1) &= \log f(x_1 \mid \lambda) \\ &= \log\left(\frac{e^{-\lambda}\lambda^{x_1}}{X_1!}\right) \\ &= -\lambda + x_1 \log\lambda - \log(x_1!) \end{aligned}$$

Then

$$\frac{\partial}{\partial\lambda} l_1(\lambda \mid x_1) = -1 + \frac{x_1}{\lambda}$$

and

$$\frac{\partial^2}{\partial\lambda^2} l_1(\lambda \mid x_1) = -\frac{x_1}{\lambda^2}.$$

Using Proposition 3.2,

$$\begin{aligned} I_1(\lambda) &= \mathbb{E}_\lambda\left[-\frac{\partial^2}{\partial\lambda^2} l_1(\lambda \mid X_1)\right] \\ &= \mathbb{E}_\lambda\left[-\left(-\frac{X_1}{\lambda^2}\right)\right] \\ &= \mathbb{E}_\lambda\left[\frac{X_1}{\lambda^2}\right] \\ &= \frac{1}{\lambda^2}\,\mathbb{E}_\lambda[X_1] \\ &= \frac{1}{\lambda^2}\,\lambda \\ &= \frac{1}{\lambda} \end{aligned}$$

Step 2: Apply the Cramér–Rao lower bound. By the remark above, $I_n(\lambda) = n \cdot I_1(\lambda) = \frac{n}{\lambda}$. Since $\tau'(\lambda) = 1$, for any unbiased estimator W,

$$\mathrm{Var}_\lambda(W) \geq \frac{1}{n/\lambda} = \frac{\lambda}{n}.$$

Step 3: Identify the UMVUE. We suspect $\bar{X}_n$ is the best unbiased estimator. The sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies $\mathbb{E}_\lambda[\bar{X}_n] = \lambda$ and

$$\mathrm{Var}_\lambda(\bar{X}_n) = \frac{\mathrm{Var}_\lambda(X_1)}{n} = \frac{\lambda}{n}.$$

Since $\bar{X}_n$ attains the Cramér–Rao lower bound, it is the UMVUE of λ.

4.4 Lehmann-Scheffé Theorem

In this lecture, we continue to expand on our understanding of the Uniformly Minimum Variance Unbiased Estimator (UMVUE). We are introduced to the Rao-Blackwell Theorem, complete statistics, and the Lehmann-Scheffé Theorem which, in tandem, allow us to construct improved estimators and ultimately the UMVUE.

Cramer-Rao Lower bound presents a scenario in which we seemingly guess an estimator such as the sample mean and then verify that its variance achieves the lower bound. Ideally, we would like a more systematic procedure; this systematic approach is covered in today's lecture.

4.4.1 Recall: Conditional Expectation

For any two random variables X and Y, recall that $E[X \mid Y]$ is a function of Y and is itself a random variable. Here is the reason why it is a function of y: denote $\phi(y) := E[X \mid Y = y]$, then $E[X \mid Y] = \phi(Y)$ and thus is a function of Y. $\mathrm{Var}[X \mid Y]$ is a function of Y by the same logic and it is always non-negative.

Properties of conditional expectation:

$$E[X] = E\big[E[X \mid Y]\big], \qquad \text{(Law of Total Expectation)}$$

$$\mathrm{Var}(X) = \mathrm{Var}\big(E[X \mid Y]\big) + E\big[\mathrm{Var}(X \mid Y)\big]. \qquad \text{(Law of Total Variance)}$$

4.4.2 The Rao–Blackwell Theorem

Theorem 4.4.1 (Rao–Blackwell)**.** *Let W be any unbiased estimator of $\tau(\theta)$, and let T be any random variable. Define $\phi(T) = E[W \mid T]$. Then:*

(a) *$E_\theta[\phi(T)] = \tau(\theta)$ for all θ.*

$\mathrm{Var}_\theta(\phi(T)) \le \mathrm{Var}_\theta(W)$ *for all θ.*

(b) *If T is a sufficient statistic for θ, then $\phi(T)$ is also an estimator and does not depend on θ.*

In other words, the Rao–Blackwell Theorem implies that if we start with one random estimator W, we pick any random variable, and we apply the theorem's conditioning on it, then we get a new random variable for which the mean is preserved and the variance is smaller. If we continue this operation, the conclusion should ultimately converge, but we aren't able to discern how or when this occurs without additional information.

Proof of (a). By (Law of Total Expectation),

$$E_\theta[\phi(T)] = E_\theta\big[E[W \mid T]\big] = E_\theta[W] = \tau(\theta) \ \ \forall\, \theta.$$

By (Law of Total Variance),

$$\mathrm{Var}_\theta(W) = \mathrm{Var}_\theta(\phi(T)) + E_\theta\big[\mathrm{Var}(W \mid T)\big].$$

We know $E_\theta\big(\mathrm{Var}(W \mid T)\big)$ is always a non-negative term so we can construct this inequality.

$$\mathrm{Var}_\theta(\phi(T)) + E_\theta\big[\mathrm{Var}(W \mid T)\big] \ge \mathrm{Var}_\theta(\phi(T)) \tag{4.1}$$

$$\mathrm{Var}_\theta(W) \ge \mathrm{Var}_\theta(\phi(T)) \ \ \forall\, \theta \tag{4.2}$$

□

Since our goal is to get the best estimator, it should not be dependent on the parameter θ by definition. Even if the conditional expectation is employed, there remains a potential to depend on θ and hence, not even be an estimator. Therefore, we must pick our T strategically so we avoid dependency on θ.

Proof of (b). Since T is a sufficient statistic, it follows from definition that $(X_1, \ldots, X_n) \mid T = t$ does not depend on θ. Hence, $W \mid T = t$ does not depend on θ, so

$$\phi(t) = E_\theta[W \mid T = t]$$

is also not dependent on θ and $\phi(T)$ is therefore an estimator. □

4.4.3 Complete Statistics

Definition 4.4.2 (Completeness)**.** The family of distributions $\{f(t \mid \theta)\}_{\theta \in \Omega}$ is called *complete* if for any function g,

$$E_\theta[g(T)] = \int g(t) \cdot f(t \mid \theta)\, dx = 0, \quad \forall\, \theta \in \Omega \tag{4.3}$$

where $T \sim f(t \mid \theta)$, implying that $P_\theta(g(T) = 0) = 1 \ \ \forall\, \theta$. Equivalently, T is called a complete statistic.

Remark 4.4.3. It is clear that $g(T) = 0$ is one solution for system of equations (4.3), the definition of completeness states that $g(T) = 0$ is the only solution.

Example 4.4.4 (Completeness in Binomial Distributions)**.** Let $T \sim \text{Bin}(n, p)$, $p \in (0, 1)$. Suppose $E_p[g(T)] = 0$ for all $p \in (0, 1)$:

$$\begin{aligned} E_p[g(T)] &= \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t} \\ &= \sum_{t=0}^{n} g(t) \binom{n}{t} \frac{p^t}{(1-p)^t} (1-p)^n \\ &= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t \end{aligned}$$

Dividing by $(1-p)^n > 0$ and substituting $z = \frac{p}{1-p} \in (0, \infty)$:

$$\sum_{t=0}^{n} g(t) \binom{n}{t} z^t = 0 \quad \forall\, z \in (0, \infty).$$

Since this polynomial vanishes on $(0, \infty)$, every coefficient is zero: $g(t)\binom{n}{t} = 0$ for $t = 0, \ldots, n$. Since $\binom{n}{t} > 0$, we get $g(t) = 0$ for all $t = 0, \ldots, n$. Thus $g(T) = 0$. Therefore T is complete.

Example 4.4.5 (Completeness in Uniform Distributions)**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$, $\theta \in (0, \infty)$. The order statistics $T = X_{(n)}$ is sufficient for θ (Exercise), with density (by order statistics theorem)

$$f(t \mid \theta) = n t^{n-1} \theta^{-n}\, \mathbf{1}_{[0,\theta]}(t).$$

Suppose $E_\theta[g(T)] = 0$ for all $\theta \in (0, \infty)$. That is

$$0 = \int_0^\theta g(t)\, t^{n-1} n \theta^{-n}\, dt = n\theta^{-n} \int_0^\theta g(t)\, t^{n-1}\, dt \quad \forall\, \theta \in (0, \infty).$$

Since $n\theta^{-n} > 0$,

$$h(\theta) := \int_0^\theta g(t)\, t^{n-1}\, dt = 0 \quad \forall\, \theta \in (0, \infty).$$

Differentiating: $h'(\theta) = g(\theta) \cdot \theta^{n-1} = 0, \forall\, \theta \in (0, \infty)$.. Since $\theta^{n-1} > 0$, we have $g(\theta) = 0$ for all $\theta \in (0, \infty)$. Thus $g(T) = g(X_{(n)}) = 0$. Therefore $T = X_{(n)}$ is complete.

Theorem 4.4.6 (Completeness in the Exponential Family). *Let $X_1, \ldots, X_n$ be iid from an exponential family*

$$f(x \mid \boldsymbol{\theta}) = \exp\left[\sum_{j=1}^{k} \eta_j(\boldsymbol{\theta})\, T_j(x) - B(\boldsymbol{\theta}) + S(x)\right].$$

Then

$$T(X_1, \ldots, X_n) = \left(\sum_{i=1}^{n} T_1(X_i),\ \ldots,\ \sum_{i=1}^{n} T_k(X_i)\right)$$

is complete and sufficient, provided the natural parameter space $\{\boldsymbol{\eta}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{R}\}$ has non-empty interior in $\mathbb{R}^k$.

4.4.4 The Lehmann–Scheffé Theorem

Theorem 4.4.7 (Lehmann–Scheffé). *Let T be a complete sufficient statistic for θ.*

(a) *If $\phi(T)$ satisfies $E_\theta[\phi(T)] = \tau(\theta)$ for all θ, then $\phi(T)$ is the UMVUE of $\tau(\theta)$.*

(b) *If W is any unbiased estimator of $\tau(\theta)$, then $E[W \mid T]$ is the UMVUE.*

Proof of (a). Let W be any unbiased estimator of $\tau(\theta)$. Define

$$g(T) := \phi(T) - E[W \mid T].$$

Both $\phi(T)$ and $E[W \mid T]$ are unbiased of $\tau(\theta)$ by hypothesis and Theorem 4.4.1(a), so $E_\theta[g(T)] = 0$ for all θ. Since T is complete, $g(T) = 0$ with probability 1 for any θ, i.e., $\phi(T) = E[W \mid T]$ with probability 1 for any θ. Then by Theorem 4.4.1(a),

$$\mathrm{Var}_\theta(\phi(T)) = \mathrm{Var}_\theta(E[W \mid T]) \le \mathrm{Var}_\theta(W) \quad \forall\, \theta. \qquad \square$$

Proof of (b). By Theorem 4.4.1(b), $E[W \mid T]$ is a valid estimator and is a function of T. By Theorem 4.4.1(a), $E_\theta[E[W \mid T]] = \tau(\theta)$. Since T is complete, part (a) applies. $\square$

Theorem 4.4.8 (Uniqueness of UMVUE). *If W_0 is a UMVUE of $\tau(\theta)$, then W_0 is unique.*

4.4.5 Indicator Function

Recall the indicator function,

$$\mathbf{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

Some properties of the indicator function include:

1.

$$\begin{aligned} \mathbf{1}_{A_1}(x) \cdot \mathbf{1}_{A_2}(x) &= \begin{cases} 1 \cdot 1 & x \in A_1 \ \&\ x \in A_2 \\ 1 \cdot 0 & x \in A_1 \ \&\ x \notin A_2 \\ 0 \cdot 1 & x \notin A_1 \ \&\ x \in A_2 \\ 0 \cdot 0 & x \notin A_1 \ \&\ x \notin A_2 \end{cases} \\ &= \begin{cases} 1 & x \in A_1 \cap A_2 \\ 0 & x \notin A_1 \cap A_2 \end{cases} \\ &= \mathbf{1}_{A_1 \cap A_2}(x) \end{aligned}$$

2. If $A = \{(x_1, ..., x_n) : x_i = k\}$ then,

$$\begin{aligned}\mathbf{1}_A(x_1, ..., x_n) &= \begin{cases} 1 & (x_1, ..., x_n) \in A \\ 0 & (x_1, ..., x_n) \notin A \end{cases} \\ &= \begin{cases} 1 & x_i = k \\ 0 & x_i \neq k \end{cases} \\ &= \mathbf{1}_k(x_i)\end{aligned}$$

for any $i \in \{1, 2, ..., n\}$ and $k \in \mathbb{R}$
The above can also be extended to apply to restrictions on multiple x_is.

3. If X is a random variable then so is $\mathbf{1}_A(X)$, and $\mathbf{1}_A(X)$ follows a Bernoulli distribution.

$$\mathbf{1}_A(X) = \begin{cases} 1 & \mathbb{P}(X \in A) \\ 0 & \mathbb{P}(X \notin A) \end{cases} \sim Bernoulli\Big(\mathbb{P}(X \in A)\Big).$$

Since $\mathbf{1}_A(X)$ follows a Bernoulli distribution with $p = \mathbb{P}(X \in A)$ then,

$$\mathbb{E}[\mathbf{1}_A(X)] = \mathbb{P}(X \in A)$$

Example 4.4.9 (Example 2.1)**.** Let $X_1, ..., X_n \overset{iid}{\sim} f(x|\theta)$. Suppose we want to estimate $\tau(\theta) = \mathbb{P}_\theta(X > a)$. Then $\mathbf{1}_{(a,\infty)}(X_1)$ is an unbiased estimator of $\tau(\theta)$ as

$$\begin{aligned}\mathbb{E}_\theta[\mathbf{1}_{(a,\infty)}(X_1)] &= \mathbb{P}_\theta\big(X_1 \in (a, \infty)\big) \\ &= \mathbb{P}_\theta(X > a) \\ &= \tau(\theta)\end{aligned}$$

as each X_i is iid (follows the same distribution).
For the same reason, X_1 can be replaced by any X_i for $i \in \{1, ..., n\}$.

4.4.6 Examples

Example 4.4.10 (Example 1 (Uniform Distribution).)**.** Let $X_1, \ldots, X_n$i.i.d.U$(0, \theta)$, $\theta \in (0, \infty)$.

Goal. Find the UMVUE of $\tau(\theta) = \theta$.

Solution (via Theorem 4.4.7). Per the previous lecture, the complete sufficient statistic is $T = X_{(n)} = \max(X_1, \ldots, X_n)$, with density

$$f_{(n)}(x) = \frac{n\, x^{n-1}}{\theta^n}, \quad 0 < x < \theta.$$

Computing the expectation:

$$\big[X_{(n)}\big] = \int_0^\theta x \cdot \frac{n\, x^{n-1}}{\theta^n}\, dx = \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \frac{n\,\theta}{n+1}.$$

Hence $\big[\frac{n+1}{n} X_{(n)}\big] = \theta$ for all $\theta > 0$. By Theorem 4.4.7(a):

$$\phi(T) = \frac{n+1}{n}\, X_{(n)} \quad \text{is the UMVUE of } \theta.$$

Example 4.4.11 (Example 2 (Binomial Distribution).)**.** Let $X_1, \ldots, X_n$i.i.d.Bin(k, θ), $\theta \in (0, 1)$.

Goal. Find the UMVUE of $\tau(\theta) = P(X_1 = 1) = \binom{k}{1}\theta(1-\theta)^{k-1} = k\theta(1-\theta)^{k-1}$.

Step 1. The pmf of each X_i is

$$f(x \mid \theta) = \binom{k}{x}\theta^x(1-\theta)^{k-x}, \quad x = 0, 1, \ldots, k.$$

Rewriting in exponential-family form:

$$\begin{aligned} f(x \mid \theta) &= \binom{k}{x}(1-\theta)^k \left(\frac{\theta}{1-\theta}\right)^x \\ &= \exp\left\{x \log\left(\frac{\theta}{1-\theta}\right) + k\log(1-\theta) + \log\binom{k}{x}\right\}. \end{aligned}$$

The natural parameter is $\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$. As θ ranges over $(0, 1)$, $\eta(\theta)$ ranges over all of $\mathbb{R}$, so the natural parameter space $\{\eta(\theta) | \theta \in (0, 1)\}$ has non-empty interior in $\mathbb{R}$. By the completeness theorem for exponential families, $T = \sum_{i=1}^n X_i$ is complete and sufficient for θ.

Step 2 (via Theorems 4.4.7 (b)). Take $T_1 = \mathbf{1}_{\{X_1=1\}}$ as an initial unbiased estimator:

$$[T_1] = P(X_1 = 1) = k\theta(1-\theta)^{k-1} = \tau(\theta).$$

By Theorem 4.4.1(b), $\phi(T) = [T_1 \mid T]$ is a valid estimator. Since T is also complete, Theorem 4.4.7(b) guarantees it is the UMVUE. We compute:

$$\phi(t) = P(X_1 = 1 \mid T = t) = \frac{P(X_1 = 1)\, P(\sum_{i=2}^n X_i = t-1)}{P(T = t)}.$$

Using $\sum_{i=1}^n X_i \sim \text{Bin}(kn, \theta)$ and $\sum_{i=2}^n X_i \sim \text{Bin}(k(n-1), \theta)$:

$$\phi(t) = \frac{\binom{k}{1}\theta(1-\theta)^{k-1} \cdot \binom{k(n-1)}{t-1}\theta^{t-1}(1-\theta)^{k(n-1)-(t-1)}}{\binom{kn}{t}\theta^t(1-\theta)^{kn-t}} = \frac{k\binom{k(n-1)}{t-1}}{\binom{kn}{t}}.$$

Therefore:

$$\phi(T) = \frac{k\binom{k(n-1)}{T-1}}{\binom{kn}{T}} = \frac{k\binom{k(n-1)}{\sum_{i=1}^n X_i - 1}}{\binom{kn}{\sum_{i=1}^n X_i}} \quad \text{is the UMVUE of } \tau(\theta) = k\theta(1-\theta)^{k-1}.$$

Example 4.4.12 (Example 3 (Bernoulli UMVUE).)**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, $\theta \in (0, 1)$.

Goal: Find the UMVUE of θ.

Step 1: Complete Sufficient Statistic. $\sum_{i=1}^n X_i$ is complete and sufficient for $\theta \in (0, 1)$ (from previous lecture notes; same exponential family argument as the Binomial with $k = 1$).

Step 2: Compute:

$$\left(\sum_{i=1}^n X_i\right) = n\,(X_1) = n\theta.$$

Dividing both sides by n:

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \theta \qquad \text{for all } \theta \in (0,1).$$

Since $(1/n)\sum X_i = \bar{X}$ is a function of the complete sufficient statistic $\sum X_i$, by Theorem:

$$\boxed{\widehat{\theta}_{\text{UMVUE}} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.}$$

So, It is UMVUE.

Remark 4.4.13. In the homework #1 you will use theorem 2 (b) to get the UMVUE.

Example 4.4.14 (Example 4 (Normal UMVUE, two parameters))**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are both unknown.

Goal: Find the UMVUE for $(\mu,\, \sigma^2)$.

Step 1: Write the pdf in Exponential Family Form.

$$\begin{aligned} f(x, \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{\mu}{\sigma^2}\cdot x - \frac{1}{2\sigma^2}\cdot x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right). \end{aligned}$$

Matching to the exponential family template we identify:

- $T_1(x) = x$ with natural parameter $g_1(\mu, \sigma^2) = \mu/\sigma^2$,
- $T_2(x) = x^2$ with natural parameter $g_2(\mu, \sigma^2) = -1/(2\sigma^2)$.

Step 2: Check the Non-Empty Interior Condition. The natural parameter set is:

$$\begin{aligned} \left\{(g_1(\mu,\sigma^2),\, g_2(\mu,\sigma^2)) : \mu \in \mathbb{R},\ \sigma^2 > 0\right\} &= \left\{\left(\frac{\mu}{\sigma^2},\, -\frac{1}{2\sigma^2}\right) : \mu \in \mathbb{R},\ \sigma^2 > 0\right\} \\ &= (-\infty, \infty) \times (-\infty, 0) \ \subset\ \mathbb{R}^2. \end{aligned}$$

This is the lower half-plane in $\mathbb{R}^2$ (first coordinate unrestricted; second coordinate strictly negative), which **contains a non-empty interior**. The condition is satisfied.

Step 3: Conclude the Complete Sufficient Statistic. By Theorem, the complete sufficient statistic is:

$$Y = \left(\sum_{i=1}^{n} T_1(X_i),\ \sum_{i=1}^{n} T_2(X_i)\right) = \left(\sum_{i=1}^{n} X_i,\ \sum_{i=1}^{n} X_i^2\right).$$

Y is **complete and sufficient** for (μ, σ^2).

Step 4: Use Known Unbiased Estimators. Recall the following facts:

- Sample Mean $[\bar{X}_n] = \mu$, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\cdot Y_1$ is a function of Y.
- Sample Variance $[S_n^2] = \sigma^2$ and

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right) = \frac{1}{n-1}\left(Y_2 - n\left(\frac{1}{n}\cdot Y_1\right)^2\right)$$

(Check the previous lecture note). So, $(\bar{X}_n,\, S_n^2)$ is a function of Y.

Since both $\bar{X}_n$ and S_n^2 are unbiased functions of the complete sufficient statistic Y, by Theorem 2 (b): $(\bar{X}_n,\, S_n^2)$ is the UMVUE of $(\mu,\, \sigma^2)$.

Example 4.4.15 (Curved normal distribution)**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \mu^2)$, where $\mu > 0$. We want to find a complete and sufficient statistic for μ. So the pdf:

$$f(x \mid \mu) = \frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{(x-\mu)^2}{2\mu^2}\right)$$

Then,

$$f(x \mid \mu) = \exp\left(\frac{1}{\mu}\, x - \frac{1}{2\mu^2}\, x^2 - \frac{1}{2}\log(2\pi\mu^2) - \frac{1}{2}\right)$$

This has $k = 2$ natural parameters even though there is only 1 free parameter μ.

Identify the natural parameters and sufficient statistics:

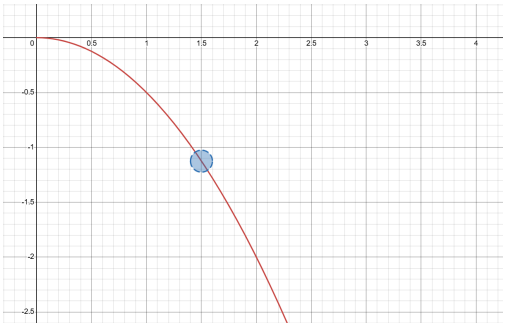
$$g_1(\mu) = \frac{1}{\mu}, \quad T_1(x) = x; \qquad g_2(\mu) = -\frac{1}{2\mu^2}, \quad T_2(x) = x^2$$

Check that the natural parameter space has an non-empty interior:

Let $y_1 = \frac{1}{\mu}$ and $y_2 = -\frac{1}{2\mu^2}$. Then

$$\mu = \frac{1}{y_1}, \qquad y_2 = \frac{1}{2\left(\frac{1}{y_1}\right)^2} = -\frac{y_1^2}{2}$$

The natural parameter space $\{(y_1, y_2) \mid \mu > 0\}$ lies on the curve $y_2 = -\frac{y_1^2}{2}$, which does not have an interior in $\mathbb{R}^2$ since for any point on the curve and any open ball around that point, we are able to find points in the ball that are not on the curve. Therefore, we cannot apply the Exponential Family theorem to conclude completeness. **Always check this condition before applying theorems.**



Recall that the two-parameter family $\mathcal{N}(\mu, \sigma^2)$. Define

$$Y = \left(\sum_{i=1}^n X_i,\, \sum_{i=1}^n X_i^2\right),$$

which is complete and sufficient for (μ, σ^2) if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. One may wonder whether Y is complete for one-parameter problem $\mathcal{N}(\mu, \mu^2)$ in this example. It turns out that Y is not complete. The reason is as below.

Compute the following expectations under $\mathcal{N}(\mu, \mu^2)$:

$$\mathbb{E}_\mu\left(\frac{Y_2}{n}\right) = \mathbb{E}_\mu[X_1^2] = \mathrm{Var}_\mu(X_1) + \left(\mathbb{E}_\mu X_1\right)^2 = \mu^2 + \mu^2 = 2\mu^2$$

Then,

$$\mathbb{E}_\mu\left(\left(\frac{Y_1}{n}\right)^2\right) = \mathbb{E}_\mu[\bar{X}_n^2] = \mathrm{Var}_\mu(\bar{X}_n) + \left(\mathbb{E}_\mu \bar{X}_n\right)^2 = \frac{\mu^2}{n} + \mu^2 = \frac{n+1}{n}\,\mu^2$$

Both expressions are proportional to μ^2. So, Define

$$g(Y_1, Y_2) = \frac{Y_2}{2n} - \frac{n}{n+1}\left(\frac{Y_1}{n}\right)^2$$

Then,

$$\mathbb{E}_\mu[g(Y_1, Y_2)] = \frac{2\mu^2}{2} - \frac{n}{n+1} \cdot \frac{n+1}{n}\,\mu^2 = \mu^2 - \mu^2 = 0, \quad \forall\, \mu > 0.$$

Since $g(Y_1, Y_2) \not\equiv 0$ but $\mathbb{E}_\mu[g(Y_1, Y_2)] = 0$ for all $\mu > 0$, Y is **not complete**. □

Example 4.4.16 (Bernoulli: UMVUE of $\theta(1-\theta)$)**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Bernoulli(θ), $\theta \in (0,1)$. Find the UMVUE of $\tau(\theta) = \theta(1-\theta)$.

From the example in the last lecture, we know that $T = \sum_{i=1}^n X_i$ is complete and sufficient for $\theta \in (0,1)$, and as the summation of n distributions following Bernoulli(θ), $T \sim$ Binomial(n, θ). So, we need to find ϕ(T) such that $\mathbb{E}_\theta[\phi(T)] = \theta(1-\theta)$. Calculate

$$\begin{aligned}
\theta(1-\theta) = \mathbb{E}_\theta[\phi(T)] &= \sum_{y=0}^n \phi(y)\binom{n}{y}\theta^y(1-\theta)^{n-y} \\
&= (1-\theta)^n \sum_{y=0}^n \phi(y)\binom{n}{y}\left(\frac{\theta}{1-\theta}\right)^y, \quad \forall \theta \in (0,1)
\end{aligned}$$

Multiply both sides by $(1-\theta)^{-n}$ and set $\lambda = \frac{\theta}{1-\theta} \in (0, \infty)$, or equivalently, $\theta = \frac{\lambda}{1+\lambda}$:

$$\begin{aligned}
\sum_{y=0}^n \phi(y)\binom{n}{y}\lambda^y &= \frac{\lambda}{1+\lambda}\left(\frac{1}{1-\frac{\lambda}{1+\lambda}}\right)^{n-1} \\
&= \frac{\lambda}{1+\lambda}(1+\lambda)^{n-1} \\
&= \lambda(1+\lambda)^{n-2}.
\end{aligned}$$

Expand the right-hand side using the binomial theorem:

$$\lambda(1+\lambda)^{n-2} = \lambda \sum_{z=0}^{n-2} \binom{n-2}{z} \lambda^z = \sum_{y=1}^{n-1} \binom{n-2}{y-1} \lambda^y$$

where y = z + 1.

Therefore,

$$\sum_{y=0}^{n} \phi(y) \binom{n}{y} \lambda^y = \sum_{y=1}^{n-1} \binom{n-2}{y-1} \lambda^y$$

Now, in order for the polynomials to be equivalent, we know that

- $y = 0$: $\phi(0)\binom{n}{0} = 0 \implies \phi(0) = 0$.
- $y = n$: $\phi(n)\binom{n}{n} = 0 \implies \phi(n) = 0$.
- $1 \le y \le n-1$: $\phi(y)\binom{n}{y} = \binom{n-2}{y-1}$.

For $1 \le y \le n-1$:

$$\phi(y) = \frac{\binom{n-2}{y-1}}{\binom{n}{y}} = \frac{\frac{(n-2)!}{(y-1)!\,(n-1-y)!}}{\frac{n!}{y!\,(n-y)!}} = \frac{y(n-y)}{n(n-1)}$$

Note that for $y = 0$,

$$\frac{y(n-y)}{n(n-1)} = \frac{0(n)}{n(n-1)} = 0$$

And for $y = n$,

$$\frac{y(n-y)}{n(n-1)} = \frac{n(0)}{n(n-1)} = 0$$

Hence $\phi(y) = \dfrac{y(n-y)}{n(n-1)}$ for all $y \in \{0, 1, \ldots, n\}$.

By Theorem 1(a), (Lehmann–Scheffé), the UMVUE is:

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

Remark 4.4.17. In the homework #5 you will use Lehmann–Scheffé (b) to get the UMVUE.

Example 4.4.18 (Poisson: UMVUW of $e^{-2\theta}$)**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$, $\theta > 0$. We know, $Y = \sum_{i=1}^{n} X_i$ is complete and sufficient for $\theta > 0$, and as the summation of n distributions each following Poisson(θ), $Y \sim \text{Poisson}(n\theta)$. Find the UMVUE of $\tau(\theta) = e^{-2\theta}$.

Set $\mathbb{E}_\theta[\phi(Y)] = \tau(\theta) = e^{-2\theta}$ for all $\theta > 0$. $Y \sim \text{Poisson}(n\theta)$. Consider

$$\mathbb{E}_\theta[\phi(Y)] = \sum_{y=0}^{\infty} \phi(y) \frac{e^{-n\theta}(n\theta)^y}{y!} = e^{-2\theta}$$

Then,

$$\sum_{y=0}^{\infty} \phi(y) \frac{(n\theta)^y}{y!} = e^{-2\theta+n\theta} = e^{n\theta-2\theta}$$

Let $\lambda = n\theta \in (0, \infty)$:

$$\sum_{y=0}^{\infty} \phi(y) \frac{\lambda^y}{y!} = e^{\lambda - \frac{2\lambda}{n}} = e^{\lambda\left(1-\frac{2}{n}\right)}, \quad \forall, \lambda > 0$$

Applying Taylor Expansion: Using $e^x = \sum_{z=0}^{\infty} \frac{x^z}{z!}$:

$$e^{\lambda\left(1-\frac{2}{n}\right)} = \sum_{z=0}^{\infty} \frac{\left[\lambda\left(1-\frac{2}{n}\right)\right]^z}{z!}$$

Now expand the right-hand side as a power series:

$$e^{\lambda(1-\frac{2}{n})} = \sum_{y=0}^{\infty} \frac{\left[\lambda(1-\frac{2}{n})\right]^y}{y!} = \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \left(1-\frac{2}{n}\right)^y.$$

Comparing coefficients gives

$$\phi(y) = \left(1-\frac{2}{n}\right)^y, \quad y = 0, 1, \ldots.$$

Therefore,

$$\phi(Y) = \left(1-\frac{2}{n}\right)^Y = \left(1-\frac{2}{n}\right)^{\sum_{i=1}^n X_i}$$

is the UMVUE of

$$\tau(\theta) = e^{-2\theta}.$$

Chapter 5

Hypothesis Testing

5.1 Introduction to Hypothesis Testing

In hypothesis testing, we are interested in making decisions about an unknown parameter θ based on observed data $X_1, \ldots, X_n$. These observations are assumed to arise from a probability distribution that depends on θ. Thus, different values of θ give different probabilistic explanations of the same observed data.

We consider two competing hypotheses:

$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0.$$

Here H_0 is called the null hypothesis, and H_1 is called the alternative hypothesis. The goal is to decide, based on the observed data, whether there is sufficient evidence to reject H_0 in favor of H_1.

A test is defined by specifying a rejection region R. If

$$(x_1, \ldots, x_n) \in R,$$

then we reject H_0. Otherwise, we do not reject H_0.

Geometrically, the sample space can be viewed as being divided into two parts: the rejection region, where the data is considered inconsistent with H_0, and the non-rejection region, where the data is not considered sufficiently unusual under H_0.

Typically, we construct a test using a statistic

$$W = W(X_1, \ldots, X_n),$$

and define a rejection region (or critical region) of the form

$$R = \{(x_1, \ldots, x_n) : W(x_1, \ldots, x_n) \leq c\}.$$

The statistic W summarizes the data into a quantity useful for deciding between H_0 and H_1, and c is a constant determining the boundary of the rejection region.

The main question is therefore how to choose a statistic W and a constant c so that the resulting test distinguishes well between the null and alternative hypotheses.

5.2 Likelihood Ratio Test

5.2.1 Likelihood Function

Let $X_1, \ldots, X_n$ be independent and identically distributed with pdf or pmf $f(x|\theta)$. The likelihood function is defined by

$$L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

The likelihood function measures how plausible a given parameter value θ is after observing the data. In a probability function, the parameter is fixed and the data varies. In a likelihood function, the data is fixed and θ varies. The likelihood function measures how plausible a given value of θ is, given the observed data. In particular, larger values of the likelihood indicate that the corresponding parameter value provides a better explanation of the observed data.

Thus, larger values of $L(\theta \mid x_1, \ldots, x_n)$ indicate that the corresponding parameter value gives a better explanation of the observed data. This is the same idea used in maximum likelihood estimation.

5.2.2 Likelihood Ratio Test

Definition 5.2.1. The likelihood ratio statistic for testing

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$$

is defined by

$$\lambda(x_1, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta \mid x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} L(\theta \mid x_1, \ldots, x_n)}.$$

The numerator is the largest likelihood that can be achieved while restricting θ to lie in the null parameter space Θ_0. In other words, it is the best possible fit to the data under the null hypothesis.

The denominator is the largest likelihood that can be achieved over the entire parameter space Θ. This represents the best possible fit to the data without imposing the null hypothesis restriction.

Since $\Theta_0 \subseteq \Theta$, the denominator is always at least as large as the numerator. Hence,

$$0 \leq \lambda(x_1, \ldots, x_n) \leq 1.$$

If $\lambda(x_1, \ldots, x_n)$ is close to 1, then the best fit under H_0 is nearly as good as the best fit over all possible parameter values. In this case, the data does not provide strong evidence against H_0.

If $\lambda(x_1, \ldots, x_n)$ is small, then even the best parameter value under H_0 gives a much worse fit than the best parameter value in the full space. This suggests that the data is more consistent with parameter values outside Θ_0.

Thus, the likelihood ratio test rejects H_0 when

$$\lambda(x_1, \ldots, x_n) \leq c, \quad c \in (0, 1).$$

Thus, the rejection region (or critical region) is

$$R = \{(x_1, \ldots, x_n) : \lambda(x_1, \ldots, x_n) \leq c\}, \quad 0 < c < 1.$$

5.2.3 Interpretation of the Rejection Rule

Suppose

$$\lambda(x_1, \ldots, x_n) \leq c.$$

Then, by the definition of λ,

$$\frac{\sup_{\theta \in \Theta_0} L(\theta \mid x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} L(\theta \mid x_1, \ldots, x_n)} \leq c.$$

Equivalently,

$$\sup_{\theta \in \Theta_0} L(\theta \mid x_1, \ldots, x_n) \leq c \cdot \sup_{\theta \in \Theta} L(\theta \mid x_1, \ldots, x_n).$$

This says that the maximum likelihood under the null hypothesis is at most a fraction c of the maximum likelihood over the entire parameter space. Since $c \in (0, 1)$, this means that imposing the null hypothesis causes a substantial loss in likelihood.

Therefore, if the likelihood ratio is small, the observed data is better explained by a value of θ outside the null hypothesis. In this case, we reject H_0.

5.2.4 Computation of the Likelihood Ratio

To compute the likelihood ratio, we generally perform two maximum likelihood estimation procedures.

First, we compute the maximum likelihood estimate under the null hypothesis:

$$\hat{\theta}_{\text{MLE},0} = \arg\max_{\theta \in \Theta_0} L(\theta \mid x_1, \ldots, x_n).$$

This is a constrained MLE because θ is required to lie in Θ_0.

Second, we compute the maximum likelihood estimate over the full parameter space:

$$\hat{\theta}_{\text{MLE},1} = \arg\max_{\theta \in \Theta} L(\theta \mid x_1, \ldots, x_n).$$

This is the unrestricted MLE.

Then the likelihood ratio can be written as

$$\lambda(x_1, \ldots, x_n) = \frac{L(\hat{\theta}_{\text{MLE},0} \mid x_1, \ldots, x_n)}{L(\hat{\theta}_{\text{MLE},1} \mid x_1, \ldots, x_n)}.$$

Thus, the likelihood ratio statistic is essentially obtained by carrying out maximum likelihood estimation under two parameter spaces and plugging the resulting maximizers into the likelihood function.

In practice, one often does not need to carry out two completely separate MLE calculations. Usually, we first compute the MLE over the full parameter space by using the standard MLE procedure, such as writing down the log-likelihood, differentiating, setting the derivative equal to zero, and solving. Then the constrained MLE under Θ_0 can often be obtained by modifying the unrestricted solution to satisfy the constraint.

This is why likelihood ratio tests rely heavily on maximum likelihood estimation. Even when the topic is hypothesis testing, we still need to know how to compute MLEs from earlier material.

A particularly simple case occurs when the null hypothesis consists of only one parameter value:

$$H_0 : \theta = \theta_0.$$

Then the null parameter space is

$$\Theta_0 = \{\theta_0\}.$$

In this case, the numerator of the likelihood ratio does not require a true optimization procedure. Since there is only one possible value under H_0, the restricted maximizer is simply

$$\hat{\theta}_{\text{MLE},0} = \theta_0.$$

Therefore, the numerator is just the likelihood function evaluated at θ_0:

$$\sup_{\theta \in \Theta_0} L(\theta \mid x_1, \ldots, x_n) = L(\theta_0 \mid x_1, \ldots, x_n).$$

The denominator still requires maximizing over the full parameter space. In common examples, such as the normal model, the unrestricted MLE may already be known from previous calculations.

Example 5.2.2. Let

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1).$$

We test

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0.$$

Here the variance is known and equal to 1, while the mean θ is unknown. The null hypothesis says that the mean is exactly θ_0, while the alternative says that the mean differs from θ_0.

The density of a single observation is

$$f(x_i|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i-\theta)^2}{2}\right).$$

Since the observations are independent, the likelihood is the product of the individual densities:

$$L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i-\theta)^2}{2}\right).$$

Combining the constants and exponential terms,

$$L(\theta \mid x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right).$$

Under H_0, the parameter space contains only the point θ_0. Therefore,

$$\hat{\theta}_{\text{MLE},0} = \theta_0.$$

Over the full parameter space $\Theta = \mathbb{R}$, we maximize the likelihood. From previous lectures, the maximizer is the sample mean:

$$\hat{\theta}_{\text{MLE},1} = \bar{x}_n.$$

Therefore,

$$\lambda(x_1, \ldots, x_n) = \frac{L(\theta_0 \mid x_1, \ldots, x_n)}{L(\bar{x}_n \mid x_1, \ldots, x_n)}.$$

Substituting the likelihood expressions gives

$$\lambda(x_1, \ldots, x_n) = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta_0)^2\right)}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\right)}.$$

The constants cancel, so

$$\lambda(x_1, \ldots, x_n) = \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{n}(x_i-\theta_0)^2 - \sum_{i=1}^{n}(x_i-\bar{x}_n)^2\right)\right].$$

To simplify this expression, use the identity

$$\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 = \sum_{i=1}^{n}(x_i-a)^2 - n(\bar{x}_n-a)^2, \quad \forall a.$$

Taking $a = \theta_0$, we get

$$\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 = \sum_{i=1}^{n}(x_i-\theta_0)^2 - n(\bar{x}_n-\theta_0)^2.$$

Substituting this into the likelihood ratio,

$$\lambda(x_1, \ldots, x_n) = \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{n}(x_i-\theta_0)^2 - \left[\sum_{i=1}^{n}(x_i-\theta_0)^2 - n(\bar{x}_n-\theta_0)^2\right]\right)\right].$$

The two sums cancel, giving

$$\lambda(x_1, \ldots, x_n) = \exp\left(-\frac{n}{2}(\bar{x}_n-\theta_0)^2\right).$$

Step 2: Rejection Region for the Normal Example
The likelihood ratio test rejects H_0 when

$$\lambda(x_1, \ldots, x_n) \leq c, \quad c \in (0, 1).$$

Using the expression

$$\lambda(x_1, \ldots, x_n) = \exp\left(-\frac{n}{2}(\bar{x}_n - \theta_0)^2\right),$$

the rejection rule becomes

$$\exp\left(-\frac{n}{2}(\bar{x}_n - \theta_0)^2\right) \leq c.$$

Taking logarithms,

$$-\frac{n}{2}(\bar{x}_n - \theta_0)^2 \leq \log c.$$

Since $0 < c < 1$, we have $\log c < 0$. Rearranging,

$$(\bar{x}_n - \theta_0)^2 \geq -\frac{2\log c}{n}.$$

Taking square roots gives

$$|\bar{x}_n - \theta_0| \geq \frac{\sqrt{-2\log c}}{\sqrt{n}}.$$

Define

$$c_1 = \frac{\sqrt{-2\log c}}{\sqrt{n}}.$$

Then the rejection region can be written as

$$R = \{(x_1, \ldots, x_n) : |\bar{x}_n - \theta_0| \geq c_1\},$$

where $c_1 > 0$.

Thus, the likelihood ratio test rejects the null hypothesis when the sample mean is sufficiently far from the hypothesized value θ_0. This agrees with the usual intuition for testing a normal mean: if the sample mean is close to θ_0, the data is consistent with H_0; if the sample mean is far from θ_0, the data gives evidence against H_0.

The exact choice of the constant c, or equivalently c_1, determines the size of the rejection region. The derivation here shows the form of the rejection region. The method for choosing the constant based on the desired significance level is discussed next.

The likelihood ratio test gives a general method for constructing hypothesis tests. The main idea is to compare the best possible likelihood under the null hypothesis with the best possible likelihood over the full parameter space.

Computationally, this means comparing a constrained MLE with an unrestricted MLE. When the null hypothesis is simple, such as $H_0 : \theta = \theta_0$, the constrained MLE is immediate. In the normal example, the unrestricted MLE is the sample mean, so the likelihood ratio test reduces to checking whether $\bar{x}_n$ is sufficiently far from θ_0.

5.3 Method of evaluating tests

5.3.1 Recall: Hypothesis Testing and the Likelihood Ratio Test

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x \mid \theta)$. We want to test:

$$H_0 : \theta \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \theta \in \Theta \setminus \Theta_0$$

The *likelihood ratio test statistic* is defined as:

$$\lambda(X_1, \ldots, X_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta \mid X_1, \ldots, X_n)}{\sup_{\theta \in \Theta} L(\theta \mid X_1, \ldots, X_n)} = \frac{L(\hat{\theta}_0 \mid X_1, \ldots, X_n)}{L(\hat{\theta} \mid X_1, \ldots, X_n)}$$

where $\hat{\theta}_0$ is the MLE restricted to Θ_0, and $\hat{\theta}$ is the unrestricted MLE over Θ. The rejection region (or critical region) takes the form:

$$R = \{(x_1, \ldots, x_n) \mid \lambda(x_1, \ldots, x_n) \leq c\}$$

The central question of this lecture is: **how do we select the constant c?**

5.3.2 Type I and type II errors

Definition 2.1. There are two main types of errors to consider in hypothesis testing. A *type I error* is when the null hypothesis is rejected even though it is true. A *type II error* is when the null hypothesis is accepted even though it is false.

Truth	**Accept H_0**	**Accept H_1**
H_0	Correct	Type I error
H_1	Type II error	Correct

We want both type I and type II errors to be as small as possible. However, typically, we restrict consideration to tests that control Type I errors at a specified level:

$$\text{Type I error} \leq \alpha \quad (\alpha = 0.1, 0.05, 0.01)$$

If we want to control Type II errors, then switch H_0 and H_1.

Definition 2.2. The *power* function is:

$$\beta(\theta) := \mathbb{P}_\theta\big((X_1, \ldots, X_n) \in \mathcal{R}\big) \subset [0, 1]$$

Power is the probability that a test rejects null hypothesis. In other words:

$$\beta(\theta) = \begin{cases} \text{Type I error} & \text{if } \theta \in \Theta_0 \\ 1 - \mathbb{P}_\theta\big((X_1, \ldots, X_n) \in \mathcal{R}^c\big) & \text{if } \theta \in \Theta \setminus \Theta_0 \end{cases}$$

$$= \begin{cases} \text{Type I error} & \text{if } \theta \in \Theta_0 \\ 1 - \text{Type II error} & \text{if } \theta \in \Theta \setminus \Theta_0 \end{cases}$$

Below is a simple example of an ideal power function where $\Theta_0 = [0, 1]$ and $\Theta = [0, 2]$. On Θ_0, the power is low, meaning the probability of rejecting the null hypothesis (Type I error) is small. On $\Theta \setminus \Theta_0$, the power rapidly increases toward 1, meaning Type II is small.

Definition 2.3. For a random variable X, denote z_α to be the constant such that

$$P(X > z_\alpha) = \alpha.$$

Then, z_α is the $(1 - \alpha)$-*quantile*.

Below is the probability density function of X. The area under the curve to the right of z_α is α.
Similarly, we can also define a value $z_{\frac{\alpha}{2}}$ for the upper tail such that

$$P\big(X > z_{\frac{\alpha}{2}}\big) = \frac{\alpha}{2}.$$

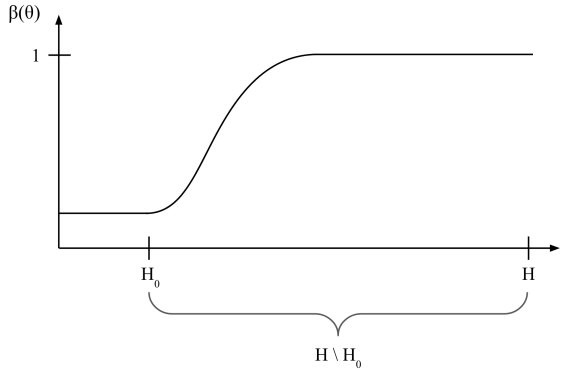


Figure 5.1: Ideal power function

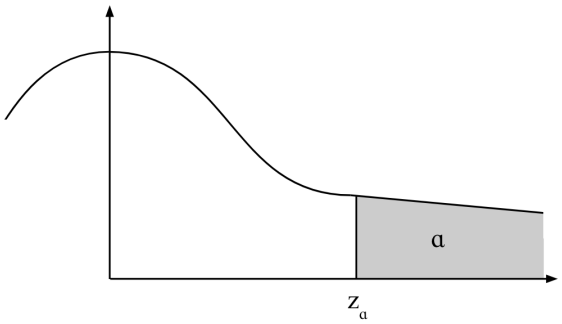


Figure 5.2: z_α (upper quantile)

Remark 5.3.1 (Procedure for LRT Hypothesis testing)**.** Recall: 3 steps for solving LRT problem

1. calculate $\lambda(x)$ the LRT Statistic
2. identify the rejection region
3. set $\alpha =$ type I error

Definition 8.3.5

For any $\alpha \in (0,1)$, a test with power function $\beta(\theta)$ is a size α test if

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha \;=\; \sup_{\theta \in \Theta_0} P_\theta\big((X_1, \ldots, X_n) \in R\big),$$

where R denotes the rejection region and Θ_0 is the parameter set under H_0.

Example 5.3.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$.

The goal is to derive the rejection region of the likelihood ratio test for testing the mean of a normal distribution. Consider the hypotheses:

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0$$

The likelihood ratio test statistic is given by:

$$\text{LRT:} \quad \lambda(x_1, \ldots, x_n) = \exp\Big[-\frac{n}{2}\left(\bar{X}_n - \theta_0\right)^2\Big]$$

The rejection region is given by:

$$\begin{aligned} R &= \{(x_1, \ldots, x_n) \mid \lambda(x_1, \ldots, x_n) \leq c\} \\ &= \left\{(x_1, \ldots, x_n) \mid \left|\bar{X}_n - \theta_0\right| \geq \sqrt{\frac{-2\log c}{n}}\right\} \\ &= \left\{(x_1, \ldots, x_n) \mid \left|\bar{X}_n - \theta_0\right| \geq c'\right\} \end{aligned}$$

We want to choose c' such that the probability of rejection (Type I error) equals α.

$$\begin{aligned} \sup_{\theta \in \Theta_0} \mathbb{P}_\theta((X_1, \ldots, X_n) \in R) &= \alpha \\ \mathbb{P}_{\theta_0}((X_1, \ldots, X_n) \in R) &= \alpha \\ \mathbb{P}_{\theta_0}\left(\left|\bar{X}_n - \theta_0\right| \geq c'\right) &= \alpha \end{aligned}$$

If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta_0, 1)$, then

$$\bar{X}_n \sim N\left(\theta_0, \frac{1}{n}\right)$$

$$\Rightarrow \ \bar{X}_n - \theta_0 \sim N\left(0, \frac{1}{n}\right)$$

$$\Rightarrow \ \sqrt{n}\,(\bar{X}_n - \theta_0) \sim N(0, 1)$$

Now, we have a standard normal distribution. So,

$$\mathbb{P}_{\theta_0}\left(\sqrt{n}\,\left|\bar{X}_n - \theta_0\right| \geq \sqrt{n}\,c'\right) = \alpha.$$

Let $Z \sim N(0, 1)$. Then,

$$\mathbb{P}_{\theta_0}\left(|Z| \geq \sqrt{n}\,c'\right) = \alpha.$$

Since nothing inside the probability depends on θ_0, we can drop the subscript and write

$$\mathbb{P}\left(|Z| \geq \sqrt{n}\,c'\right) = \alpha.$$

$$\begin{aligned} &\Rightarrow 2\mathbb{P}\left(Z \geq \sqrt{n}\,c'\right) = \alpha \\ &\Rightarrow \mathbb{P}\left(Z \geq \sqrt{n}\,c'\right) = \frac{\alpha}{2} \\ &\Rightarrow \sqrt{n}\,c' = z_{\frac{\alpha}{2}} \end{aligned}$$

Now we have a quantile of $N(0, 1)$.

$$c' = \frac{z_{\alpha/2}}{\sqrt{n}}.$$

Thus,

$$R = \left\{ (x_1, \ldots, x_n) \mid \left| \bar{X}_n - \theta_0 \right| \geq \frac{z_{\alpha/2}}{\sqrt{n}} \right\}.$$

We reject H_0 if $|\bar{X}_n - \theta_0| \geq \frac{z_{\alpha/2}}{\sqrt{n}}$, and fail to reject H_0 otherwise.

Remark 5.3.3. The Type I error (size) of the test is

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta((X_1, \ldots, X_n) \in R) = \alpha.$$

Thus, we choose the constant c' such that the test has significance level α.

Example 5.3.4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x \mid \theta)$ where

$$f(x \mid \theta) = e^{-(x-\theta)} \cdot \mathbf{1}_{(\theta, \infty)}(x)$$

is the *location exponential distribution*. Note that if $Y \sim \text{Exp}(1)$, then

$$f_Y(y) = e^{-y} \cdot \mathbf{1}_{\{y \geq 0\}},$$

and setting $X = Y + \theta$ gives $X \sim f(x \mid \theta)$. Consider the hypotheses:

$$H_0 : \theta \leq \theta_0 \qquad H_1 : \theta > \theta_0$$

The likelihood function is:

$$\begin{aligned} L(\theta \mid x_1, \ldots, x_n) &= \prod_{i=1}^{n} f(x_i \mid \theta) \\ &= \prod_{i=1}^{n} \left[e^{-(x_i - \theta)} \cdot \mathbf{1}_{(\theta, \infty)}(x_i) \right] \\ &= \exp\left\{ -\sum_{i=1}^{n} (x_i - \theta) \right\} \cdot \mathbf{1}_{(\theta, \infty)}(x_{(1)}) \\ &= \begin{cases} \exp\left(n\theta - \sum_{i=1}^{n} x_i \right) & \text{if } \theta < x_{(1)} \\ 0 & \text{if } \theta \geq x_{(1)} \end{cases} \end{aligned}$$

where $x_{(1)} = \min_i x_i$. Note that $L(\theta \mid x_1, \ldots, x_n)$ is increasing in θ for $\theta < x_{(1)}$ and drops to 0 at $\theta = x_{(1)}$. Since $\Theta = \mathbb{R}$, the unrestricted supremum is:

$$\sup_{\theta \in \mathbb{R}} L(\theta \mid x_1, \ldots, x_n) = \exp\left\{ n x_{(1)} - \sum_{i=1}^{n} x_i \right\}$$

For the restricted supremum over $\Theta_0 = (-\infty, \theta_0]$:

$$\sup_{\theta \leq \theta_0} L(\theta \mid x_1, \ldots, x_n) = \begin{cases} \exp\left\{ n x_{(1)} - \sum_{i=1}^{n} x_i \right\} & \text{if } \theta_0 \geq x_{(1)} \\ \exp\left\{ n\theta_0 - \sum_{i=1}^{n} x_i \right\} & \text{if } \theta_0 < x_{(1)} \end{cases}$$

$$= \exp\left\{ n(\theta_0 \wedge x_{(1)}) - \sum_{i=1}^{n} x_i \right\}$$

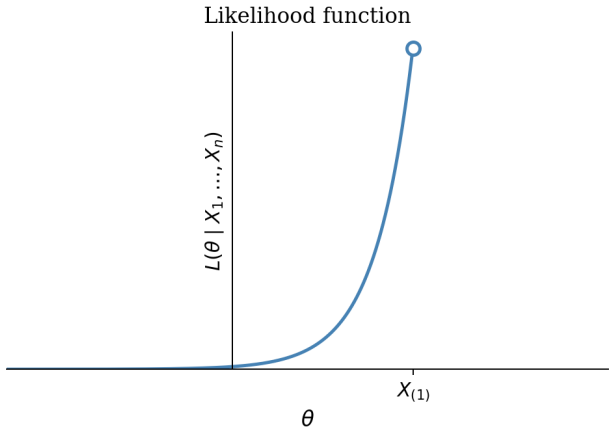


Figure 5.3: Likelihood function $L(\theta \mid x_1, \ldots, x_n)$, increasing in θ for $\theta < x_{(1)}$ and dropping to 0 at $\theta = x_{(1)}$ (open circle).

where $a \wedge b = \min\{a, b\}$. Therefore, the likelihood ratio statistic is:

$$\lambda(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } \theta_0 \geq x_{(1)} \\ \exp\{n(\theta_0 - x_{(1)})\} & \text{if } \theta_0 < x_{(1)} \end{cases}$$

Since $c \in (0, 1)$, $\lambda \leq c$ is only binding in the second branch, the rejection region is:

$$\begin{aligned} R &= \{(x_1, \ldots, x_n) \mid \lambda(x_1, \ldots, x_n) \leq c\}, \quad c \in (0, 1) \\ &= \{(x_1, \ldots, x_n) \mid \theta_0 < x_{(1)} \text{ and } \exp\{n(\theta_0 - x_{(1)})\} \leq c\} \end{aligned}$$

$\exp\{n(\theta_0 - x_{(1)})\} \leq c$ is equivalent to:

$$\begin{aligned} n(\theta_0 - x_{(1)}) &\leq \log c \\ x_{(1)} &\geq \theta_0 - \frac{1}{n} \log c \\ x_{(1)} &\geq \theta_0 + c' \end{aligned}$$

where $c' = -\frac{1}{n} \log c > 0$ (since $c \in (0, 1)$ implies $\log c < 0$). Note that the condition $x_{(1)} \geq \theta_0 + c'$ with $c' > 0$ makes $x_{(1)} > \theta_0$ redundant. Hence:

$$R = \{(x_1, \ldots, x_n) \mid x_{(1)} \geq \theta_0 + c'\}, \quad c' > 0$$

The Type I error is:

$$\begin{aligned} &= \sup_{\theta \in \Theta_0} \mathbb{P}_\theta((X_1, \ldots, X_n) \in R) \\ &= \sup_{\theta \leq \theta_0} \mathbb{P}_\theta\big(X_{(1)} \geq \theta_0 + c'\big) \end{aligned}$$

Denote $g(\theta) := P_\theta\big(X_1 \geq \theta_0 + c'\big)$. Is $g(\theta)$ increasing?

$$g(\theta) = P_\theta\big(X_1 \geq \theta_0 + c', \ldots, X_n \geq \theta_0 + c'\big).$$

Because the X_i are i.i.d.,

$$g(\theta) = \big(P_\theta(X_1 \geq \theta_0 + c')\big)^n.$$

Write $X_1 = Y + \theta$ (so that the distribution of Y does not depend on θ). Then

$$P_\theta(X_1 \geq \theta_0 + c') = P\big(Y + \theta \geq \theta_0 + c'\big) = P\big(Y \geq \theta_0 - \theta + c'\big).$$

$$g(\theta) = \big(P\big(Y \geq \theta_0 - \theta + c'\big)\big)^n.$$

As θ increases, $\theta_0 - \theta + c'$ decreases, so the event $\{Y \geq \theta_0 - \theta + c'\}$ becomes easier to satisfy and the probability increases. Hence $P_\theta(X_1 \geq \theta_0 + c')$ (and therefore $g(\theta)$) is increasing in θ.

In particular, $\theta_0 - \theta + c'$ is smaller for larger θ, which explains the monotonicity.

so

$$\alpha = g(\theta_0) = P_{\theta_0}\big(X_{(1)} \geq \theta_0 + c'\big)$$

$$= P(Y \geq c')^n$$

$$= \big(e^{-c'}\big)^n \qquad \text{since } Y \text{ is exponential distribution}$$

$$\alpha = e^{-nc'}$$

$$\ln \alpha = -nc'$$

$$c' = \frac{\ln \alpha}{-n} > 0$$

positive because $0 < \alpha \leq 1$

$$\text{so } R = \{(x_1, \ldots, x_n) \mid X_{(1)} \geq \theta_0 - \tfrac{1}{n} \ln \alpha\}.$$

□

Example 5.3.5. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with both parameters unknown.

Consider the hypotheses

$$H_0 : \ \mu \leq \mu_0, \qquad H_1 : \ \mu > \mu_0,$$

with σ^2 unspecified. Equivalently,

$$\Theta_0 = \{(\mu, \sigma^2) : \mu \leq \mu_0, \ \sigma^2 > 0\}.$$

1. Likelihood ratio test (LRT)

The likelihood (for observed $x_1, \ldots, x_n$) is (check this)

$$L(\mu, \sigma^2 \mid x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\Big).$$

To form the likelihood ratio we will need the suprema of L over the relevant parameter sets. The maximum likelihood estimators (unconstrained MLE) are

$$\hat{\mu} = \overline{x}_n, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)^2.$$

Since

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2,$$

we have

$$\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{n}{2}.$$

Thus

$$L(\hat{\mu}, \hat{\sigma}^2 \mid x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\hat{\sigma}}\right)^n \exp(-\frac{n}{2}).$$

Under $H_0 : \mu \leq \mu_0$, we compute

$$\sup_{(\mu,\sigma^2)\in H_0} L(\mu, \sigma^2 \mid x_1, \ldots, x_n) = \sup_{\sigma^2}\sup_{\mu:\mu\leq\mu_0} L(\mu, \sigma^2 \mid x_1, \ldots, x_n)$$

Optimize with respect to μ:

$$\arg\max_{\mu\leq\mu_0} L(\mu, \sigma^2 \mid x_1, \ldots, x_n) = \arg\min_{\mu\leq\mu_0}\sum_{i=1}^{n}(x_i - \mu)^2 = \arg\min_{\mu\leq\mu_0}\left(\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu)^2\right)$$

$$= \begin{cases}\mu_0 & \mu_0 \leq \overline{x}_n \\ \overline{x}_n & \mu_0 \geq \overline{x}_n\end{cases} = \mu_0 \wedge \overline{x}_n$$

Now optimize over σ^2:

$$\sup_{(\mu,\sigma^2)\in H_0} L(\mu, \sigma^2 \mid x_1, \ldots, x_n) = \sup_{\sigma^2} L(\mu_0 \wedge \overline{x}_n, \sigma^2 \mid x_1, \ldots, x_n)$$

(Exercise: Take log, differentiate, find critical point, 1st derivative or 2nd derivative)

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - (\mu_0 \wedge \overline{x}_n))^2$$

Since

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - (\mu_0 \wedge \overline{x}_n))^2$$

we have

$$\frac{1}{2\hat{\sigma}_0^2}\sum_{i=1}^{n}(x_i - \mu_0 \wedge \overline{x}_n)^2 = \frac{n}{2}.$$

Thus

$$L(\mu_0 \wedge \overline{x}_n, \hat{\sigma}_0^2 \mid x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_0}\right)^n \exp(-\frac{n}{2}).$$

Thus,

$$\lambda(x_1, \ldots, x_n) = \frac{L(\mu_0 \wedge \overline{x}_n, \hat{\sigma}_0^2 \mid x_1, \ldots, x_n)}{L(\hat{\mu}, \hat{\sigma}^2 \mid x_1, \ldots, x_n)}$$

Thus the exponential terms cancel:

$$= \left(\frac{\hat{\sigma}}{\hat{\sigma}_0}\right)^n = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} = \begin{cases}1 & \text{if } \mu_0 \geq \bar{x}_n \\ \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} & \text{if } \mu_0 < \bar{x}_n\end{cases}.$$

Thus, for some $c \in (0,1)$, the rejection region is

$$\begin{aligned}
R =& \{(x_1,\ldots,x_n)|\lambda(x_1,\ldots,x_n) \le c\} \\
=& \{(x_1,\ldots,x_n)|\lambda(x_1,\ldots,x_n) \le c, \mu_0 < \bar{x}_n\} \cup \{(x_1,\ldots,x_n)|\lambda(x_1,\ldots,x_n) \le c, \mu_0 \ge \bar{x}_n\} \\
=& \{(x_1,\ldots,x_n)|\lambda(x_1,\ldots,x_n) \le c, \mu_0 < \bar{x}_n\} \cup \emptyset \\
=& \{(x_1,\ldots,x_n)|\lambda(x_1,\ldots,x_n) \le c, \mu_0 < \bar{x}_n\}.
\end{aligned}$$

Under $\mu_0 < \bar{x}_n$, substituting the expressions for $\hat{\sigma}^2$ and $\hat{\sigma}_0^2$,

$$\lambda(x_1,\ldots,x_n) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} \left(\frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)^2}{\frac{1}{n}\sum_{i=1}^n (x_i - \mu_0)^2}\right)^{n/2} = \left(\frac{\sum_{i=1}^n (x_i - \overline{x}_n)^2}{\sum_{i=1}^n (x_i - \mu_0)^2}\right)^{n/2}.$$

Now use

$$\sum_{i=1}^n (x_i - \mu_0)^2 = \sum_{i=1}^n (x_i - \overline{x}_n)^2 + n(\overline{x}_n - \mu_0)^2.$$

Therefore,

$$\begin{aligned}
\lambda(x_1,\ldots,x_n) &= \left(\frac{\sum_{i=1}^n (x_i - \overline{x}_n)^2}{\sum_{i=1}^n (x_i - \overline{x}_n)^2 + n(\overline{x}_n - \mu_0)^2}\right)^{n/2} \\
&= \left(\frac{1}{1 + \frac{n(\overline{x}_n - \mu_0)^2}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}}\right)^{n/2}.
\end{aligned}$$

Define:

$$g(y) = \left(\frac{1}{1 + \frac{y^2}{n-1}}\right)^{n/2}$$

Then:

$$\lambda(x_1,\ldots,x_n) = g\left(\left|\frac{\overline{x}_n - \mu_0}{\sqrt{s_n^2/n}}\right|\right)$$

Note that

$$g(y) = \left(1 + \frac{y^2}{n-1}\right)^{-n/2}.$$

For $y > 0$, the quantity

$$1 + \frac{y^2}{n-1}$$

is increasing in y. Since the exponent $-n/2$ is negative, $g(y)$ is decreasing.

Therefore,

$$\lambda \le c \iff g\left(\left|\frac{\overline{x}_n - \mu_0}{\sqrt{s_n^2/n}}\right|\right) \le c.$$

Since g is decreasing,

$$\left|\frac{\overline{x}_n - \mu_0}{\sqrt{s_n^2/n}}\right| \ge g^{-1}(c) = c_1, \qquad c_1 > 0.$$

Since we are in the case $\mu_0 < \overline{x}_n$, we have

$$\frac{\overline{x}_n - \mu_0}{\sqrt{s_n^2/n}} \ge c_1.$$

Thus the rejection region is

$$R = \left\{(x_1, \ldots, x_n) : \frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} \geq c_1, \mu_0 < \bar{x}_n\right\} = \left\{(x_1, \ldots, x_n) : \frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} \geq c_1\right\}.$$

Final rejection region:

$$R = \left\{(x_1, \ldots, x_n) : \frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} \geq c_1\right\}$$

for some $c_1 > 0$.

Type I Error and the Test Statistic

For a size-α test, we need to know the Type I error. The Type I error is:

$$\sup_{(\mu,\sigma^2):\, \mu \leq \mu_0} P_{\mu,\sigma^2}\left(\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \geq c_1\right) = \alpha,$$

where $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is the sample variance.

Define $g(\mu, \sigma^2) := P_{\mu,\sigma^2}\left(\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \geq c_1\right)$.

Since $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, the following standard distributional rules apply:

$$Z := \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \qquad V := \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1).$$

We rewrite the standardized statistic:

$$\begin{aligned}
\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} &= \frac{\bar{X}_n - \mu + (\mu - \mu_0)}{S_n/\sqrt{n}} \\
&= \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)\frac{\sigma}{S_n} \\
&= \left(Z + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) \cdot \sqrt{\frac{(n-1)\sigma^2}{(n-1)S_n^2}} \\
&= \left(Z + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) \cdot \sqrt{\frac{n-1}{V}}.
\end{aligned}$$

So, the Type I error becomes:

$$\sup_{(\mu,\sigma^2):\, \mu \leq \mu_0} P_{\mu,\sigma^2}\left(\left(Z + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)\sqrt{\frac{n-1}{V}} \geq c_1\right).$$

Then

$$\begin{aligned}
g(\mu, \sigma^2) :=& P\left(\left(Z + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)\sqrt{\frac{n-1}{V}} \geq c_1\right) \\
=& P\left(Z - c_1\sqrt{\frac{V}{n-1}} \geq -\frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right).
\end{aligned}$$

As a function of μ, g is increasing when $\mu \leq \mu_0$, so the supremum is achieved at $\mu = \mu_0$.

Thus

$$\begin{aligned}\sup_{(\mu,\sigma^2):\,\mu\le\mu_0} P_{\mu,\sigma^2}\left(\frac{\bar{X}_n-\mu_0}{S_n/\sqrt{n}}\ge c_1\right) &= \sup_{(\mu,\sigma^2):\,\mu\le\mu_0} g(\mu,\sigma^2)\\ &= \sup_{\sigma^2}\sup_{\mu:\,\mu\le\mu_0} g(\mu,\sigma^2)\\ &= \sup_{\sigma^2} g(\mu_0,\sigma^2)\\ &= \sup_{\sigma^2} P_{\mu_0,\sigma^2}\left(\frac{\bar{X}_n-\mu_0}{S_n/\sqrt{n}}\ge c_1\right).\end{aligned}$$

Since $W := \dfrac{\bar{X}_n-\mu_0}{S_n/\sqrt{n}} \sim t(n-1)$ whenever $X_1,\ldots,X_n \overset{\text{iid}}{\sim} N(\mu_0,\sigma^2)$ (for any $\sigma^2>0$), the probability does not depend on σ^2. So:

$$\sup_{(\mu,\sigma^2):\,\mu\le\mu_0} P_{\mu,\sigma^2}\left(\frac{\bar{X}_n-\mu_0}{S_n/\sqrt{n}}\ge c_1\right) = \sup_{\sigma^2} P_{\mu_0,\sigma^2}(W\ge c_1) = \sup_{\sigma^2} P(W\ge c_1) = P(W\ge c_1) = \alpha,$$

which means $c_1 = t_\alpha(n-1)$, the upper α-quantile of the t-distribution with $n-1$ degrees of freedom. The rejection region is:

$$R = \left\{(x_1,\ldots,x_n)\;\middle|\;\frac{\bar{x}_n-\mu_0}{s_n/\sqrt{n}}\ge t_\alpha(n-1)\right\}.$$

5.4 Randomized test

Sometimes, we need a special treatment to achieve $P(\text{Reject } H_0 \mid H_0)=\alpha$. Let's see the example below.

Example 5.4.1. Let $X_1,\ldots,X_n \sim \text{Pois}(\theta)$. We use the sample mean $\bar{X}$ of a sample of size $n=100$ to test the null hypothesis $H_0:\theta=0.1$ against the alternative hypothesis $H_1:\theta<0.1$ at the significance level $\alpha=0.05$.We want to find the rejection region such that

$$P(\text{Reject } H_0 \mid H_0)=\alpha$$

Since the observed sample mean is expected to be smaller under the alternative hypothesis, the rejection region takes the form $\bar{X}\le c$ for some constant c. From the properties of the Poisson distribution, we know that

$$100\bar{X} = X_1+X_2+\cdots+X_{100} \sim \text{Poisson}(100\theta),$$

thus, under the null hypothesis, we have

$$100\bar{X} = X_1+X_2+\cdots+X_{100} \sim \text{Poisson}(10).$$

we can calculate the rejection probability as

$$P(\text{Reject } H_0) = P_{0.1}(X_1+\cdots+X_{100}\le k) = P(\text{Poisson}(10)\le k),$$

for some k. Using the cumulative probabilities for the Poisson distribution (using R), we summarize the rejection probabilities for different values of k as follows.

k value	0	1	2	3	4	5	$\cdots$
$P_{0.1}(X_1+\cdots+X_{100}\le k)$	0.000	0.000	0.002	0.010	0.029	0.067	$\cdots$

Given the values of k, it is impossible to have a rejection region where $P(\text{Reject } H_0|H_0)$ exactly equals 0.05. However, we can create a rejection region where the rejection probability is slightly less than 0.05, and

then use a randomized method to adjust it to exactly 0.05. This type of test is called a **randomized test**. Here is how we construct the randomized test.

$$\begin{aligned} X_1 + \cdots + X_{100} \leq 4 \quad &\Rightarrow \quad \text{Reject } H_0, \\ X_1 + \cdots + X_{100} = 5 \quad &\Rightarrow \quad \text{Reject } H_0 \text{ with probability } \gamma, \\ X_1 + \cdots + X_{100} \geq 6 \quad &\Rightarrow \quad \text{Not reject } H_0. \end{aligned}$$

The rejection probability for such a randomized test can be calculated as

$$P_{0.1}(X_1 + \cdots + X_{100} \leq 4) + \gamma P(X_1 + \cdots + X_{100} = 5) = 0.029 + \gamma(0.067 - 0.029)$$

To achieve a significance level of $\alpha = 0.05$, we can solve for γ as follows:

$$0.029 + \gamma(0.067 - 0.029) = 0.05,$$

solve for γ

$$\Rightarrow \gamma = \frac{0.05 - 0.029}{0.067 - 0.029} = \frac{21}{38}.$$

Exercise: can you construct a different randomized test above?

Summary: In general, when the distribution $f(x;\theta), \theta \in \Omega$ is given, the null hypothesis H_0 and alternative hypothesis H_1 can be represented as:

$$H_0 : \theta \in \Omega_0 \quad \text{vs.} \quad H_1 : \theta \in \Omega_1 \quad (\Omega_0 \cap \Omega_1 = \emptyset, \Omega_0 \cup \Omega_1 = \Omega).$$

Based on the observations $X_1, \ldots, X_n$, we judge the evidence in favor of or against the null hypothesis H_0. In this type of test, the maximum allowable probability of wrongly rejecting H_0 (i.e., committing a Type I error) is given by the significance level α, and the rejection region is chosen such that

$$P_\theta((X_1, \ldots, X_n) \in C_\alpha) \leq \alpha \quad \text{for all} \quad \theta \in \Omega_0.$$

The objective is to minimize the probability of making a Type II error, which is failing to reject H_0 when it is false. To maximize the rejection region within the significance level, we define a test with

$$\max_{\theta \in \Omega_0} P_\theta((X_1, \ldots, X_n) \in C_\alpha) = \alpha$$

Finally, we define the rejection region C_α and the test as follows.

$$\begin{aligned} (X_1, \ldots, X_n) \in C_\alpha \quad &\Rightarrow \quad \text{Reject } H_0, \\ (X_1, \ldots, X_n) \notin C_\alpha \quad &\Rightarrow \quad \text{Fail to reject } H_0. \end{aligned}$$

In cases where no appropriate non-randomized rejection region can satisfy the conditions (as shown in the Poisson example), we can consider a suitable **randomized test**. The randomized test can be represented by the function $\phi(x_1, \ldots, x_n)$, which indicates the probability of rejecting the null hypothesis H_0 as follows.

$$\begin{aligned} \phi(x_1, \ldots, x_n) = 1 \quad &\Rightarrow \quad H_0 \text{ is rejected}, \\ 0 < \phi(x_1, \ldots, x_n) < 1 \quad &\Rightarrow \quad H_0 \text{ is rejected with probability } \phi(x_1, \ldots, x_n), \\ \phi(x_1, \ldots, x_n) = 0 \quad &\Rightarrow \quad H_0 \text{ is accepted}. \end{aligned}$$

The rejection probability of such a randomized test is given by $E_\theta \phi(X_1, \ldots, X_n)$, which depends on the parameter θ. This function is known as the **test function** $\gamma_\theta(\phi)$, and is expressed as

$$\gamma_\theta(\phi) = E_\theta \phi(X_1, \ldots, X_n), \quad \theta \in \Omega.$$

The maximum value of the rejection probability for all $\theta \in \Omega_0$ under the null hypothesis is

$$\max_{\theta \in \Omega_0} E_\theta \phi(X_1, \ldots, X_n),$$

and this is referred to as the **size** of the test $\phi(X_1, \ldots, X_n)$. The test that satisfies the significance level α is the one where the size does not exceed α. In other words

$$\max_{\theta \in \Omega_0} E_\theta \phi(X_1, \ldots, X_n) \leq \alpha.$$

By determining $\phi(X_1, \ldots, X_n)$ appropriately, we can ensure that the test satisfies the significance level α. The test minimizes the probability of rejecting the null hypothesis H_0 incorrectly (Type I), while maximizing the rejection probability (power) where possible, ensuring that the size **equals** the specified significance level α:

$$\max_{\theta \in \Omega_0} E_\theta \phi(X_1, \ldots, X_n) = \alpha.$$

5.5 Final review

Geometric Distribution: Exponential Family & MLE

Setup. Let X have PMF

$$\text{pmf}(x;\, p) = (1-p)^{x-1} p.$$

(a) Expressing as a nautral exponential family. Using the identities that $a = \exp(\log a)$ and $\log a^b = b \log a$

$$\begin{aligned} \text{pmf}(x;\, p) &= e^{(x-1)\log(1-p)} \cdot e^{\log p} \\ &= \exp\big(x \log(1-p) + \log \tfrac{p}{1-p}\big) \\ &= \exp\big(x \log(1-p) - \log(1-p) + \log p\big) \end{aligned}$$

Let $\eta(p) = \log(1-p)$, therefore $1 - e^\eta = p$, and we obtain the natural exponential family:

$$\text{pmf}(x; \eta) = \exp\big(x\eta - \eta + \log(1 - e^\eta)\big).$$

(b) MLE of p. For an a random variable in the exponential family

$$\mathbb{E}_p(X) = \frac{1}{n} \sum_{i=1}^{n} T(x_i)$$

From the prior part, we have that $T(x) = x$. Since $X \sim \text{Geom}(p)$ gives $\mathbb{E}_p[X] = 1/p$:

$$\begin{aligned} \frac{1}{p} &= \frac{1}{n} \sum_{i=1}^{n} T(x_i) \\ \hat{p} &= \frac{1}{\bar{X}} \end{aligned}$$

Beta$(1,\theta)$: Sufficient Statistic

Setup. Let $X_1,\dots,X_n \sim \text{Beta}(1,\theta)$ for $\theta>0$. Find a sufficient statistic for θ.

Solution.

The likelihood is

$$\begin{aligned} L(\theta;\, x_1,\dots,x_n) &= \prod_{i=1}^{n} \text{pdf}(x_i;\,\theta) \\ &= \prod_{i=1}^{n} \left[\theta\,(1-x_i)^{\theta-1}\,\mathbf{1}_{\{x_i\in(0,1)\}}\right] \\ &= \theta^n \left(\prod_{i=1}^{n}(1-x_i)\right)^{\theta-1} \cdot \prod_{i=1}^{n}\mathbf{1}_{\{x_i\in(0,1)\}} \end{aligned}$$

Define

$$u(x_1,\dots,x_n)=\prod_{i=1}^{n}(1-x_i), \qquad k_1(y,\theta)=\theta\,y^{\theta-1}, \qquad k_2(x)=\prod_{i=1}^{n}\mathbf{1}_{\{x_i\in(0,1)\}}$$

Then

$$L(\theta;\, x_1,\dots,x_n) = k_1(u(x_1,\dots,x_n),\,\theta)\cdot k_2(x)$$

By the **Factorization Theorem**,

$$u(x_1,\dots,x_n)=\prod_{i=1}^{n}(1-x_i) \quad \text{is sufficient for } \theta>0.$$

Fisher Information and Cramér–Rao Lower Bound

Setup. Let $X_1,\dots,X_n \overset{\text{iid}}{\sim} f(x\mid\theta)=\theta x^{\theta-1},\ 0<x<1,\ \theta>0$.

(a) Fisher information. By additivity, $I_n(\theta)=nI_1(\theta)$. The single-observation log-likelihood is

$$\ell_1(\theta)=\log(\theta\,x^{\theta-1})=\log\theta+(\theta-1)\log x.$$

Thus

$$\frac{\partial\ell_1}{\partial\theta}=\frac{1}{\theta}+\log x, \qquad \frac{\partial^2\ell_1}{\partial\theta^2}=-\frac{1}{\theta^2}.$$

Therefore

$$I_1(\theta)=-\mathbb{E}\left[-\frac{1}{\theta^2}\right]=\mathbb{E}\left[\frac{1}{\theta^2}\right]=\frac{1}{\theta^2}, \qquad I_n(\theta)=\frac{n}{\theta^2}.$$

(b) Cramér–Rao lower bound for $\tau(\theta)=\theta^4$.

$$\text{Var}(\hat\tau) \;\geq\; \frac{[\tau'(\theta)]^2}{I_n(\theta)}=\frac{(4\theta^3)^2}{\dfrac{n}{\theta^2}}=\frac{16\,\theta^8}{n}$$

UMVUE for p^4 (Bernoulli Distribution)

Setup. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Ber}(p)$, $p \in (0,1)$. It is known that $\bar{X}_n$ is the UMVUE for p, and therefore so is $\sum X_n$.

(1) An unbiased estimator for p^4. By independence:

$$\begin{aligned} \mathbb{E}[X_i] &= p, \\ \mathbb{E}[X_1 X_2] &= \mathbb{E}[X_1]\,\mathbb{E}[X_2] = p^2, \\ \mathbb{E}\left[\prod_{i=1}^{4} X_i\right] &= p^4. \end{aligned}$$

Hence $U = \prod_{i=1}^{4} X_i$ is an unbiased estimator for p^4.

(2) UMVUE for p^4 via Rao–Blackwell.

From lecture, $T = \sum_{i=1}^{n} X_i$ is complete and sufficient for p. By the Rao–Blackwell / Lehmann–Scheffé theorem, the UMVUE is

$$\eta(T) = \mathbb{E}\left[X_1 X_2 X_3 X_4 \;\middle|\; \sum_{i=1}^{n} X_i \right]$$

We compute $\phi(t) = \mathbb{E}[X_1 X_2 X_3 X_4 \mid \sum_{i=1}^{n} X_i = t]$. Since $X_1 X_2 X_3 X_4 = \mathbf{1}_{\{X_1=1, X_2=1, X_3=1, X_4=1\}}$,

$$\phi(t) = \mathbb{P}\left(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1 \;\middle|\; \sum_{i=1}^{n} X_i = t \right).$$

If $0 \le t \le 3$, then $\phi(t) = 0$.
If $t \ge 4$, then

$$\begin{aligned} \phi(t) &= \frac{\mathbb{P}(X_1 = 1,\, X_2 = 1,\, X_3 = 1,\, X_4 = 1,\, \sum_{i=5}^{n} X_i = t - 4)}{\mathbb{P}(\sum_{i=1}^{n} X_i = t)} \\ &= \frac{p^4 \cdot \binom{n-4}{t-4} p^{t-4} (1-p)^{(n-4)-(t-4)}}{\binom{n}{t} p^t (1-p)^{n-t}} \\ &= \frac{\binom{n-4}{t-4}}{\binom{n}{t}} = \frac{(n-4)!}{(t-4)!\,(n-t)!} \cdot \frac{t!\,(n-t)!}{n!} = \frac{t(t-1)(t-2)(t-3)}{n(n-1)(n-2)(n-3)} \end{aligned}$$

Note that the above formula also includes the case $0 \le t \le 3$.
Therefore the UMVUE of p^4 is

$$\boxed{\hat{\eta}(T) = \phi(T) = \frac{T(T-1)(T-2)(T-3)}{n(n-1)(n-2)(n-3)}}$$

Pareto Distribution: Likelihood Ratio Test (Exercise 8.5 of Casella and Berger 2nd edition)

Setup. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Pareto}(\theta, \nu)$ with PDF

$$f(x \mid \theta, \nu) = \frac{\theta \nu^\theta}{x^{\theta+1}} \mathbf{1}_{(\nu, \infty)}(x), \qquad \theta > 0, \nu > 0.$$

Test $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$.

(1) Likelihood Ratio Test (LRT).

The joint likelihood is

$$L(\theta,\nu \mid x_1,\dots,x_n) = \prod_{i=1}^{n} \frac{\theta\nu^{\theta}}{x_i^{\theta+1}}\,\mathbf{1}_{(\nu,\infty)}(x_i) = (\theta\nu^{\theta})^n \left(\prod_{i=1}^{n}\frac{1}{x_i}\right)^{\theta+1} \mathbf{1}_{\{x_{(1)}\geq\nu\}}$$
$$= \frac{\theta^n\nu^{n\theta}}{(\prod_{i=1}^{n} x_i)^{\theta+1}}\,\mathbf{1}_{\{x_{(1)}\geq\nu\}},$$

where $x_{(1)} = \min_i x_i$. With $\eta = (\theta,\nu)$. Define the null and full parameter spaces

$$\Omega_0 = \{(\theta,\nu) \mid \theta = 1,\ \nu > 0\}, \qquad \Omega = \{(\theta,\nu) \mid \theta > 0,\ \nu > 0\}.$$

Recall

$$\lambda = \frac{\sup_{(\theta,\nu)\in\Omega_0} L(\theta,\nu \mid x_1,\dots,x_n)}{\sup_{(\theta,\nu)\in\Omega} L(\theta,\nu \mid x_1,\dots,x_n)},$$

where $\Theta_0 = \{1\}\times\mathbb{R}^+$ and $\Theta = \mathbb{R}^+\times\mathbb{R}^+$.

Since L is increasing in ν (for fixed $\theta > 0$, $\nu^{n\theta}$ is increasing) and the indicator requires $\nu \leq x_{(1)}$, the supremum over ν is attained at $\hat{\nu} = x_{(1)}$ in both numerator and denominator.

The numerator (restricted to Ω_0, i.e. $\theta = 1$) is

$$\sup_{\eta\in\Omega_0} L(\theta,\nu \mid x_1,\dots,x_n) = \sup_{\nu>0} L(1,\nu \mid x_1,\dots,x_n) = L(1,x_{(1)} \mid x_1,\dots,x_n).$$

The denominator of the LRT statistic is

$$\sup_{\eta\in\Omega} L(\theta,\nu \mid x_1,\dots,x_n) = \sup_{\theta>0}\ \sup_{\nu>0} L(\theta,\nu \mid x_1,\dots,x_n) = \sup_{\theta>0} L(\theta,x_{(1)} \mid x_1,\dots,x_n).$$

To simplify the maximization, take the log:

$$\ell(\theta,x_{(1)} \mid x_1,\dots,x_n) = n\log\theta + n\theta\log x_{(1)} - (\theta+1)\sum_{i=1}^{n}\log x_i.$$

Thus

$$\frac{\partial\ell(\theta,x_{(1)} \mid x_1,\dots,x_n)}{\partial\theta} = \frac{n}{\theta} + n\log x_{(1)} - \sum_{i=1}^{n}\log x_i.$$

Let

$$u(x_1,\dots,x_n) = \sum_{i=1}^{n}\log x_i - n\log x_{(1)},$$

It is clear that $u(x_1,\dots,x_n) > 0$ (with probability 1). Thus

$$\ell(\theta,x_{(1)}) = n\log\theta - u(x_1,\dots,x_n).$$

Set

$$\frac{\partial\ell(\theta,x_{(1)} \mid x_1,\dots,x_n)}{\partial\theta} = 0$$

and we obtain

$$\hat{\theta} = \frac{n}{u(x_1,\dots,x_n)} > 0.$$

When $0 < \theta < \hat{\theta}$, $\frac{\partial \ell(\theta, x_{(1)}|x_1,\ldots,x_n)}{\partial \theta} > 0$. When $\theta > \hat{\theta}$, $\frac{\partial \ell(\theta, x_{(1)}|x_1,\ldots,x_n)}{\partial \theta} < 0$. By 1st derivative test, $\hat{\theta}$ is the maximizer.

Assembling λ. Set $T = u(x_1, \ldots, x_n)$. we have $\prod_{i=1}^n x_i = x_{(1)}^n e^T$. Computing each term in the ratio directly:

$$L(1,\, x_{(1)}) = \frac{x_{(1)}^n}{(\prod x_i)^2} = \frac{x_{(1)}^n}{x_{(1)}^{2n} e^{2T}} = x_{(1)}^{-n}\, e^{-2T},$$

$$L\big(\tfrac{n}{T},\, x_{(1)}\big) = \frac{(n/T)^n\, x_{(1)}^{n^2/T}}{(\prod x_i)^{n/T+1}} = \frac{(n/T)^n\, x_{(1)}^{n^2/T}}{x_{(1)}^{n^2/T+n}\, e^{n+T}} = \Big(\frac{n}{T}\Big)^n\, x_{(1)}^{-n}\, e^{-(n+T)}.$$

Therefore

$$\lambda = \frac{L(1,\, x_{(1)})}{L(n/T,\, x_{(1)})} = \frac{x_{(1)}^{-n}\, e^{-2T}}{(n/T)^n\, x_{(1)}^{-n}\, e^{-(n+T)}}$$

$$= \left(\frac{T}{n}\right)^n e^{-2T+(n+T)} = \left(\frac{T}{n}\right)^n e^{n-T}.$$

Bibliography

[1] G. Casella, and R. Berger, *Statistical inference*, Chapman and Hall/CRC, 2nd edition, 2024.