Introduction to Statistical Inference

Winter 2025, Math 357 Course Notes

McGill University

Jiajun Zhang

July 11, 2025

Acknowledgments

I would like to extend my deepest thanks and appreciation to the following people,
without whose support this note would not have been possible:

[Abbas Khalili], *Professor, McGill University*
I would like to express my deepest gratitude to Professor Abbas,
the instructor for this course, whose guidance and expertise
were invaluable throughout the development of my notes.

Despite all efforts, there may still be some typos, unclear explanations, etc. If you find potential mistakes, or any suggestions regarding concepts or formats, etc., feel free to reach out to the author at *zhangjohnson729@gmail.com*.

Contents

Chapter 1

From Probability to Statistics

Random Samples

Suppose we seek information about certain characteristic of a group or a collection of elements or units, called population. Due to cost, we may not be able to study every unit of a population, and we restrict to a “random sample” from a population. Then the purpose of statistics is to make “inference” or educated guess about the characteristics of interest.

In general, there are two types of statistical inference, namely parametric inference and non-parametric inference. In a parametric inference, we assume that we already know the functional form of the distribution of the random sample X up to some unknown parameter $\boldsymbol{\theta}$, for example we may know X is normal with parameters $\boldsymbol{\theta} = (\mu, \sigma^2)^T$, or $X \sim Exp(\theta)$ with $\theta > 0$. In this case, we would make inference about θ based on our random sample. In a non-parametric inference, we basically know nothing about the function form of the distribution.

Definition

Definition 1. *Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed random variables with a common distribution F, written as $X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} F$. This collection $X_1, X_2, \cdots, X_n$ is called a random sample from the population of interest.*

And our goal is to use $X_1, \cdots, X_n$ to make inference of F. Here are some examples:

Example 1: (Parametric Inference)

Assume we have $X \sim N(\mu, \sigma^2)$, then we would like to find $\boldsymbol{\theta} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$ such that it fits the sample best.

We already know that

$$F_{\theta}(x) = \mathbb{P}_{\theta}(X \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx$$

And now if we have a random sample $X_1, \cdots X_n$, by WLLN (Weak Law of Large Numbers) and CMT (Continuous Mapping Theorem), we have

$$\hat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu,$$

$$\hat{\sigma}_n^2 = S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2 \xrightarrow{P} \sigma^2.$$

So our inference is successful, and the guess is

$$\boldsymbol{\theta} = \begin{pmatrix} \hat{\mu}_n \\ \\ \hat{\sigma}_n^2 \end{pmatrix} = \begin{pmatrix} \overline{X}_n \\ \\ S_n^2 \end{pmatrix}.$$

Example 2: (Non-parametric Inference)

We may not assume any parametric form for F, suppose all we know is $X \sim F, F(x) = \mathbb{P}(X \le x), \forall x \in \mathbb{R}$. Now in this random sample $X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} F$, we may define our Empirical CDF (ECDF) as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\chi(X_i \le x), \qquad \chi(X_i \le x) = \begin{cases} 1 & X_i \le x \\ \\ 0 & \text{otherwise} \end{cases}$$

Again, according to WLLN, we have $F_n(x) \xrightarrow{P} F(x), \forall x \in \mathbb{R}, n \to \infty$, and we actually have a stronger statement, by Glivenlco-Cantelli Theorem:

$$\sup_{x\in\mathbb{R}} \Big| F_n(x) - F(x) \Big| \xrightarrow{a.s} 0, n \to \infty.$$

Definition

Definition 2. *Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} F$ and let*

$$T : \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}^k$$

be a Borel measurable function. Then $T(X_1, X_2, \ldots, X_n)$ is called a statistic provided that it does not depend on any unknown, i.e T only depends on $X_1, X_2, \cdots, X_n$.

For example, if $X_1, \cdots, X_n \overset{i.i.d}{\sim} F$, then

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i; S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2; F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\chi(X_i \le x)$$

are all statistics. But note that $\overline{X}_n - \mu$ is not a statistic, since it is also a function of μ, but it is still a random variable. We call $T(x_1, x_2, \cdots, x_n)$ an observed value of the statistic $T(X-1, X_2, \cdots, X_n)$.

We would like to discuss some general results:

Theorem

Theorem 1. *Let* $X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} F$ *and* $g : \mathbb{R} \to \mathbb{R}$ *be a Borel measurable function with* $\mathbf{Var}(g(X_i)) < +\infty$, *then*

$$\mathbb{E}\left\{\sum_{i=1}^{n} g(X_i)\right\} = n\mathbb{E}(g(X_i)) \;\; \textit{and} \;\; \mathbf{Var}\left\{\sum_{i=1}^{n} g(X_i)\right\} = n\mathbb{E}\{(g(X_i) - \mathbb{E}(g(X_i)))^2\}$$

Proof. ■

Theorem

Theorem 2. *Let* $X_1, \cdots, X_n \overset{i.i.d}{\sim} F$, *and* $\mu = \mathbb{E}X_i, \sigma^2 = \mathbf{Var}(X_i) < +\infty$, *then:*

(i) $\mathbb{E}(\overline{X}_n) = \mu$;

(ii) $\mathbf{Var}(\overline{X}_n) = \dfrac{\sigma^2}{n}$

(iii) $\mathbb{E}(S_n^2) = \sigma^2$, *here* $S_n^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

(iv) $M_{\overline{X}_n}(t) = \left[M_{X_i}\left(\dfrac{t}{n}\right)\right]^n$

An example is that if we have $X_1, \cdots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, then $\overline{X}_n \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$.

Theorem

Theorem 3. *Let* $X_1, \cdots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, *then*

(i) $\overline{X}_n \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$

(ii) $\overline{X}_n$ *and* S_n^2 *are independent*

(iii) $\dfrac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$.

Proof. (i) is easy, we will prove (ii) and (iii). For (ii), note that we have

$$\begin{aligned} S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2 &= \frac{1}{n-1}\left\{\sum_{i=2}^{n}(X_i-\overline{X}_n)^2+(X_1-\overline{X}_n)^2\right\} \\ &= \frac{1}{n-1}\left\{\sum_{i=2}^{n}(X_i-\overline{X}_n)^2+\left[\sum_{i=2}^{n}(X_i-\overline{X}_n)\right]^2\right\} \end{aligned}$$

So S_n^2 is a function of $(X_2-\overline{X}_n,\cdots,X_n-\overline{X}_n)$, and hence it is equivalent to show that $\overline{X}_n$ and $(X_2-\overline{X}_n,\cdots,X_n-\overline{X}_n)$ are independent. We will include a transformation technique, let

$$\begin{cases} Y_1 &= \overline{X}_n \\ Y_2 &= X_2-\overline{X}_n \\ \quad\vdots \\ Y_n &= X_n-\overline{X}_n \end{cases} \Longrightarrow \begin{cases} X_1 &= Y_1-\sum_{i=1}^{n} Y_i \\ X_2 &= Y_2+Y_1 \\ \quad\vdots \\ X_n &= Y_n+Y_1 \end{cases}$$

The Jacobian is given by

$$|J| = \det\begin{pmatrix} 1 & -1 & -1 & \cdots & -1 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix} = n$$

Thus we have

$$\begin{aligned} f_{Y_1,Y_2,\cdots,Y_n}(y_1,y_2,\cdots,y_n) &= |J|\cdot f_{X_1,X_2,\cdots,X_n}(x_1(y_1,\cdots,y_n),\cdots,x_n(y_1\cdots,y_n)) \\ &= n\cdot\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left[x_i(y_1,\cdots,y_n)-\mu\right]^2\right\} \\ &= K\cdot\exp\left(\frac{-n(y_1-\mu)^2}{2\sigma^2}\right)\cdot\exp\left\{-\frac{1}{2\sigma^2}\left[\left(\sum_{i=2}^{n}y_i\right)^2+\sum_{i=2}^{n}y_i^2\right]\right\}, \end{aligned}$$

for some constant K. Since their joint pdf splits, so $\overline{X}_n$ and S_n^2 are independent.

For (iii), note that

$$\begin{aligned} \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 &= \frac{\sum_{i=1}^{n}(X_i-\overline{X}_n+\overline{X}_n-\mu)^2}{\sigma^2} \\ &= \frac{\sum_{i=1}^{n}(X_i-\overline{X}_n)^2+\sum_{i=1}^{n}(\overline{X}_n-\mu)^2+2\sum_{i=1}^{n}(X_i-\overline{X}_n)(\overline{X}_n-\mu)}{\sigma^2} \\ &= \frac{\sum_{i=1}^{n}(X_i-\overline{X}_n)^2}{\sigma^2}+\frac{n(\overline{X}_n-\mu)^2}{\sigma^2}. \end{aligned}$$

Here we use the fact that if $X\sim N(0,1)$ then $X^2\sim\chi_1^2$. So

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}\sim\chi_n^2;\frac{n(\overline{X}_n-\mu)^2}{\sigma^2}\sim\chi_1^2$$

and thus

$$\frac{\sum_{i=1}^{n}(X_i-\overline{X}_n)^2}{\sigma^2}\sim\chi_{n-1}^2 \Longrightarrow \frac{(n-1)S_n^2}{\sigma^2}\sim\chi_{n-1}^2.$$

■

t Distribution and F distribution

1.2.1 t Distribution

Definition

Definition 3. *Let X be a continuous random variable, we say X forms a t distribution (student distribution), denote as $X \sim t(\nu)$ where ν is a parameter, if its pdf is given by*

$$f_X(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Here are some remarks:

(i) If $\nu = 1$, it is called a Cauchy distribution, in this case $\mathbb{E}X$ does not exist. In general if $\nu > 2$, we have $\mathbb{E}X = 0, \mathbf{Var}(X) = \dfrac{\nu}{\nu-2}$;

(ii) If $Z \sim N(0,1)$, $V = \chi^2_\nu$ and Z, V are independent, then

$$T = \frac{Z}{\sqrt{\frac{V}{\nu}}} \sim t(\nu)$$

Theorem

Theorem 4. *Let $X-1, \cdots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, then*

$$T = \frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}} \sim t(n-1).$$

Proof. By

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim N(0,1) \text{ and } \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

it is easy to see that theorem 4 holds.

■

1.2.2 F distribution

Suppose $U \sim \chi^2_m, V = \chi^2_n$ and U, V are independent, then

$$F = \frac{\chi^2_m/m}{\chi^2_n/n} \sim F(m,n).$$

Theorem

Theorem 5. *Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, \cdots, Y_n \overset{i.i.d}{\to} N(\mu_2, \sigma_2^2)$ be two independent random samples, then*

$$F = \frac{S_m^2/\sigma_1^2}{S_n^2/\sigma_2^2} \sim F(m-1, n-1).$$

Asymptotic Theories

What happens to a "statistic" or its "distribution" when $n \to \infty$? We will first state some theorems that are useful:

Definition

Definition 4. *Let $\{X_n : n \geq 1\}$ be a sequence of random variables, and define $T(X_1, \cdots, X_n) = T_n$, we say T_n convergent to θ in probability, denote as $T_n \overset{P}{\to} \theta$, if*

$$\forall \varepsilon > 0, \lim_{n\to\infty} \mathbb{P}\{|T_n - \theta| > \varepsilon\} = 0.$$

From this definition, we may state the weak law of large numbers (WLLN):

Theorem

Theorem 6. *(Weak Law of Large Numbers)*

Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} F$ with $\mathbb{E}X_i = \mu$, and $\mathbf{Var}(X_i) < +\infty$, then $\overline{X}_n \overset{P}{\to} \mu$ as $n \to \infty$.

Proof. We will show that $\forall \varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{1}{n}(X_1 + \cdots + X_n) - \mu\right| > \varepsilon\right\}.$$

Since $\mathbb{E}\overline{X}_n = \mu, \mathbf{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$, by Chebyshev's inequality, we have

$$\mathbb{P}\{|\overline{X}_n - \mu| > \varepsilon\} \leq \frac{\mathbf{Var}\overline{X}_n}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon}\frac{1}{n} \to 0, \text{as } n \to \infty.$$

■

Definition

Definition 5. *Let $\{X_n : n \geq 1\}$ be a sequence of random variables, and X_i has a distribution F_{X_i}. We say the sequence $\{F_{X_i} : n \geq 1\}$ convergent to F in distribution, denote by $F_{X_n} \xrightarrow{d} F$ or $X_n \xrightarrow{d} X$, if*

$$\lim_{n\to\infty} F_{X_i}(x) = F(x), \forall x$$

The most used application of convergence in distribution is central limit theorem, which states that if we have $X_1, \cdots, X_n \overset{i.i.d}{\sim} F$ and $\mu = \mathbb{E}X_i, \sigma^2 = \mathbf{Var}(X_i) < +\infty$, then

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

as $n \to \infty$.

Then we introduce a very useful theorem:

Theorem

Theorem 7. *(Slutsky's Theorem)*

Assume that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} a \in \mathbb{R}$ as $n \to \infty$, then we have:

(i) $X_n + Y_n \xrightarrow{d} X + A$;

(ii) $X_n \cdot Y_n \xrightarrow{d} aX$;

(iii) $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X_n}{a}, a \neq 0,$

as $n \to \infty$

Theorem

Theorem 8. *(Continuous Mapping Theorem)*

Suppose $X_n \xrightarrow{P} X$ and g is a continuous function on the set C such that $\mathbb{P}(X \in C) = 1$, then

$$g(X_n) \xrightarrow{P} g(X), \text{ as } n \to \infty.$$

Corollary

Corollary 1. *Let $X - 1, \cdots, X_n \overset{i.i.d}{\to} F$, $\mu = \mathbb{E}X_i, \sigma^2 = \mathbf{Var}(X_i) < +\infty$, then*

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \xrightarrow{d} N(0,1), \text{ as } n \to \infty.$$

Proof. First, we have

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} = \frac{\sqrt{n}(\overline{X}_n - \mu)/\sigma}{S_n/\sigma}$$

where the numerator $\xrightarrow{d} N(0,1)$ by central limit theorem, and in the denominator, since $S_n^2 \xrightarrow{P} \sigma^2$, then by continuous mapping theorem $S_n \xrightarrow{P} \sigma$, so the denominator $\xrightarrow{P} \sigma$, finally using Slutsky's theorem we finished the proof. ■

Theorem

Theorem 9. *(First Order Delta Method)*

Let $\sqrt{n}(X_n - \mu) \xrightarrow{d} V$ *as* $n \to \infty$, *and let* g *be a real valued function such that* g' *exists at* $x = \mu$, $g'(\mu) \neq 0$, *then*

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} g'(\mu) \cdot V$$

as $n \to \infty$.

In particular, if $V \sim N(0, \sigma^2)$, then

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, (g'(\mu))^2\sigma^2).$$

Theorem

Theorem 10. *(Second Order Delta Method)*

Let $\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ *as* $n \to \infty$, $g'(\mu) = 0$, $g''(\mu) \neq 0$. *Then*

$$n\{g(x_n) - g(\mu)\} \xrightarrow{d} \frac{\sigma^2 g''(\mu)}{2}\chi_1^2$$

Chapter 2

Theory of Point Estimation

Unbiased Estimator

In what follows we mainly focus on parametric inference, $X \sim F_\theta$ given by

$$\mathscr{F} := \{F_\theta : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\} \;\; \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{pmatrix}$$

where Θ is the parameter space and $X_1, \cdots, X_n \overset{i.i.d}{\sim} F$ to be the random sample.

Definition

Definition 6. *We define $\hat{\theta}(X_1, \cdots, X_n) \equiv \hat{\theta}_n$ is an estimator of a parameter θ if it is a statistic, that is it does not depend on any unknown parameter.*

Thus an estimator $\hat{\theta}$ is also a random variable.

Remark: Here our key assumption is that $n >> d$, that is, the sample size is really large compared to the known data. We define $\hat{\theta}(x_1, \cdots, x_n)$ (post experimental data) to be an estimate of θ, and $\hat{\theta}(X_1, \cdots, X_n)$ to be an estimator of θ.

Example: 1 Let X be a random selected chip from a large sample and denote

$$X = \begin{cases} 1 & \text{if the chip is operational} \\ 0 & \text{otherwise} \end{cases}$$

We may assume $X_1, \cdots, X_n \overset{i.i.d}{\sim} Bernoulli(\theta)$ and θ is the probability that the randomly selected chip is operational. In this case we have the following possible estimators of θ, as long as it is a function of $X_1, \cdots, X_n$:

$$\overline{X}_n; \frac{X_1+X_2}{2}; X_5$$

But $X_1 + \theta$ is not an estimator! Since it is also a function of θ.

We may come up with many possible estimators for a parameter θ, but how do we choose the good estimators? In general, we would like to choose estimators that work well on average. For example, consider the two estimations below:

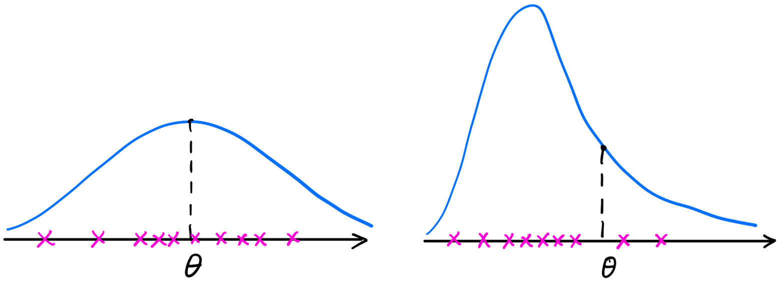

Figure 2.1: In this figure, the blue curve represents the distribution of $\hat{\theta}_n$, and θ is the target parameter, the pink cross is all possible values of $\hat{\theta}_n$.

We would prefer the estimator on the left. And thus our first criterion based on which we could choose an estimator is called "unbiased-ness".

Definition

Definition 7. *An estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \cdots, X_n)$ is an "unbiased" estimator of θ if $\forall \theta \in \Theta$, $\mathbb{E}_\theta\{\hat{\theta}_n\} = \theta$.*

We also define the biased estimator to be

$$Bias(\hat{\theta}_n) = \mathbb{E}_\theta\{\hat{\theta}_n\} - \theta.$$

Recall from the previous example where we have $X_1, \cdots, X_n \overset{i.i.d}{\sim} Bernoulli(\theta)$, and the estimators we introduced before, i.e $\overline{X}_n; \frac{X_1+X_2}{2}, X_5$ are all unbiased indeed!

Example 2: Suppose we have $X_1, \cdots, X_n \overset{i.i.d}{\sim} F_\theta$ where $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \mu = \mathbb{E}X_i, \sigma^2 = \textbf{Var}(X_i)$, then

- $\hat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of μ.

- $\hat{\sigma}_n^2 = S_n^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ is an unbiased estimator of σ^2. To see this, we have

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

Thus

$$\begin{aligned}
\mathbb{E}_\theta\{\hat{\sigma}_n^2\} &= \frac{1}{n-1}\mathbb{E}_\theta\left\{\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}\mathbb{E}_\theta(X_i^2) - n\mathbb{E}_\theta(\overline{X}_n^2)\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}(\sigma^2+\mu^2) - n\left[\frac{\sigma^2}{n}+\mu^2\right]\right\} \\
&= \frac{n-1}{n-1}\sigma^2 = \sigma^2.
\end{aligned}$$

Remark: $\dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ is also a good estimator for the variance, but this estimator is biased!

Example 3: Suppose we have $X_1, \cdots, X_n \overset{i.i.d}{\sim} Uniform(0,\theta)$, $\theta > 0$ is out target parameter, then the estimator $\hat{\theta}_{n,1} = 2X_3, \hat{\theta}_{n,2} = 2\overline{X}_n$ are all unbiased, but $\hat{\theta}_{n,3} = X_{(n)} := \max_{1\le i\le n}(X_1, \cdots, X_n)$ is biased, since $\mathbb{E}\{\hat{\theta}_{n,3}\} = \frac{n}{n+1}\theta$. A way to correct the bias of $\hat{\theta}_{n,3}$ would be $\hat{\theta}_{n,4} = \frac{n+1}{n}X_{(n)}$.

As we see from the previous examples, for a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ we could come up with unbiased estimators. Another criterion that we could bring into our assessment of an estimator is its variance $\mathbf{Var}_\theta\{\hat{\theta}_n\}$, if exists.

In the class of unbiased estimator, we could choose the one that has the smallest variance.

Another criterion is MSE (Mean Square Error), namely

$$MSE_\theta(\hat{\theta}_n) = \mathbb{E}_\theta[(\hat{\theta}_n^2 - \theta)^2].$$

Note that

$$MSE_\theta(\hat{\theta}_n) = \mathbb{E}_\theta[(\hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n) + \mathbb{E}_\theta(\hat{\theta}_n) - \theta)^2] = \mathbf{Var}_\theta(\hat{\theta}_n) + [Bias(\hat{\theta}_n)]^2.$$

Definition

Definition 8. *An estimator $\hat{\theta}_n$ of θ is called consistent if $\hat{\theta} \xrightarrow{P} \theta$ as $n \to \infty$.*

Uniformly Minimum Variance Unbiased Estimator (UMVUE)

Now we would like to focus on the class of unbiased estimators of a real valued parameter $\tau(\theta)$, $\tau : \Theta \to \mathbb{R}$. We would like to introduce the method of Uniformly Minimum Variance Unbiased Estimator (UMVUE):

Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$ be a random vector with a joint pdf (pmf) given by $p_\theta(x_1, \cdots, x_n) = p_\theta(\boldsymbol{x})$, and the vectoer $\boldsymbol{x}$ given by $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$.

Definition

Definition 9. *An estimator $T(\boldsymbol{X})$ of a real valued parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ where $\tau(\theta) : \Theta \to \mathbb{R}$ is said to be a UMVUE of $\tau(\theta)$ if:*

① $\mathbb{E}_\theta\{T(\boldsymbol{X})\} = \tau(\theta), \forall \theta \in \Theta$;

② *For any other unbiased estimator of $\tau(\theta)$, call it $T^*(\boldsymbol{X})$, where $\mathbb{E}_\theta\{T^*(\boldsymbol{X})\} = \tau(\theta), \forall \theta \in \Theta$, and*

$$\mathbf{Var}_\theta\{T(\boldsymbol{X})\} \leq \mathbf{Var}_\theta\{T^*(\boldsymbol{X})\}, \forall \theta \in \Theta.$$

How do we actually find a UMVUE? There are two ways, one is by Cramér-Rao Lower Bound, one is by Rao-Blackwell & Lehmann-Scheffé theorem. We first introduce the Cramér-Rao Lower Bound (CRLB). Assume $d = 1$, $\Theta \subseteq \mathbb{R}$, $\boldsymbol{X} \sim p_\theta$ (joint pdf / pmf).

(i) Assume that the family $\{p_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ has a common support, i.e

$$\mathscr{S} := \left\{ \boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} : p_\theta(\boldsymbol{x}) > 0 \right\}$$

and does not depend on θ. (In this case $Uniform(0, \theta)$ is excluded);

(ii) Furthermore, assume for $\boldsymbol{x} \in \mathscr{S}$, $\theta \in \Theta \subseteq \mathbb{R}$, $\dfrac{d}{d\theta} \log p_\theta(\boldsymbol{x})$ exists;

(iii) Also, assume for any statistics $h(\boldsymbol{X})$ with $\mathbb{E}\{|h(\boldsymbol{X})|\} < +\infty, \forall\theta \in \Theta$ we have

$$\frac{d}{d\theta}\int_{\mathscr{S}} h(\boldsymbol{x})p_\theta(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathscr{S}} h(\boldsymbol{x})\frac{d}{d\theta}p_\theta(\boldsymbol{x})d\boldsymbol{x}$$

whenever the *R.H.S* is finite.

Theorem

Theorem 11. *(Cramér-Rao Lower Bound)*

Assume (i),(ii),(iii) hold, let $T(\boldsymbol{X})$ be such that

$$\mathbf{Var}_\theta(T(\boldsymbol{X})) < +\infty, \mathbb{E}_\theta\{T(\boldsymbol{X})\} = \tau(\theta), \forall\theta \in \Theta,$$

then if $0 < \mathbb{E}_\theta\left\{\left[\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right]^2\right\} < +\infty, \forall\theta \in \Theta$*, then*

$$\mathbf{Var}_\theta(T(\boldsymbol{X})) \geq \frac{[\tau'(\theta)]^2}{\mathbb{E}_\theta\left\{\left[\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right]^2\right\}}.$$

Here, we denote $I(\theta) = \mathbb{E}_\theta\left\{\left[\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right]^2\right\}$, it is called the *Fisher Information*.

Proof. W.L.O.G, we will only prove the continuous case. Note $\tau(\theta) = \mathbb{E}_\theta\{T(\boldsymbol{X})\}, \forall\theta \in \Theta$, so

$$\begin{aligned}\frac{d}{d\theta}\tau(\theta) &= \frac{d}{d\theta}\{\mathbb{E}_\theta[T(\boldsymbol{X})]\}\\ &= \frac{d}{d\theta}\left\{\int_{\mathscr{S}} T(\boldsymbol{x})p_\theta(\boldsymbol{x})d\boldsymbol{x}\right\}\\ &= \int_{\mathscr{S}} T(\boldsymbol{x})\cdot\frac{d}{d\theta}p_\theta(\boldsymbol{x})d\boldsymbol{x}\\ &= \int_{\mathscr{S}} T(\boldsymbol{x})\frac{d}{d\theta}[\log p_\theta(\boldsymbol{x})]p_\theta(\boldsymbol{x})d\boldsymbol{x}.\end{aligned}$$

Therefore,

$$\tau'(\theta) = \mathbb{E}_\theta\left\{T(\boldsymbol{X})\frac{d}{d\theta}[\log p_\theta(\boldsymbol{x})]\right\}, \forall\theta \in \Theta.$$

On the other hand, by (iii), if $h(\boldsymbol{x}) \equiv 1$, then

$$0 = \int_{\mathscr{S}} \frac{d}{d\theta}p_\theta(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathscr{S}}\left[\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right]p_\theta(\boldsymbol{x})d\boldsymbol{x}, \forall\theta \in \Theta.$$

Thus we have

$$\mathbb{E}_\theta\left\{\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right\} = 0, \forall\theta \in \Theta.$$

The equation above is also known as the Bartlett Identity.

So now we have

$$\tau'(\theta) = \mathbf{Cov}\left(T(\boldsymbol{X}), \frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right), \forall \theta \in \Theta$$

and hence by Cauchy-Schwarz inequality, we have

$$\begin{aligned}[\tau'(\theta)]^2 &= \mathbf{Cov}^2\left(T(\boldsymbol{X}), \frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right) \\ &\leq \mathbf{Var}_\theta(T(\boldsymbol{X})) \cdot \mathbf{Var}_\theta\left(\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right).\end{aligned}$$

Now using Bartlett's identity, we have

$$[\tau'(\theta)]^2 \leq \mathbf{Var}_\theta(T(\boldsymbol{X})) \cdot \mathbb{E}_\theta\left\{\left[\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right]^2\right\}.$$

■

Note that if we have $X_1, \cdots, X_n \overset{i.i.d}{\sim} f(x,\theta)$, then by definition

$$p_\theta(\boldsymbol{x}) = \prod_{i=1}^n f(\theta, x_i)$$

hence in this case

$$\mathbb{E}\left\{\left[\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right]^2\right\} = n\mathbb{E}\left\{\left[\frac{d}{d\theta}\log f(\theta, x_i)\right]^2\right\}.$$

and now

$$\mathbf{Var}_\theta(T(\boldsymbol{X})) \geq \frac{[\tau'(\theta)]^2}{nI_i(\theta)}$$

sometimes it is easier to compute the Fisher information for one sample, we may do so and multiply by n in the end, with the assumption that we are working with random sample (i.i.d).

Corollary

Corollary 2. *(Second Bartlett's Identity) We hvae*

$$\mathbb{E}\left\{\left[\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right]^2\right\} = -\mathbb{E}\left\{\frac{d^2}{d\theta^2}\log p_\theta(\boldsymbol{x})\right\}.$$

Proof. Using Chain rule, we have

$$\frac{d^2}{d\theta^2}\log p_\theta(\boldsymbol{x}) = \frac{p''_\theta(\boldsymbol{x})}{p_\theta(\boldsymbol{x})} - \left[\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right]^2$$

and taking expectation on both sides, we have

$$\mathbb{E}\left\{\frac{d^2}{d\theta^2}\log p_\theta(\boldsymbol{x})\right\} = \mathbb{E}\left\{\frac{p_\theta''(\boldsymbol{x})}{p_\theta(\boldsymbol{x})}\right\} - \mathbb{E}\left\{\left[\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right]^2\right\}$$

where the first term on the right hand side of the equation equals 0, because

$$\frac{d}{d\theta}\mathbb{E}\left\{\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right\} = \int_{\mathscr{S}}\frac{d}{d\theta}\left\{\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right\} p_\theta(\boldsymbol{x})dx.$$

■

Example: Suppose we have a random sample $X_1,\cdots,X_n \overset{i.i.d}{\sim} Bernoulli(\theta)$, its probability mass function is given by $f(x,\theta) = \theta^x(1-\theta)^{1-x}, x \in \{0,1\}$ and $\theta \in (0,1)$. Find the CRLB of all unbiased estimator of $\tau(\theta) = \theta$.

Solution: We first find the Fisher information. We investigate the function

$$\log f(\theta,x_i) = x_i \log\theta + (1-x_i)\log(1-\theta)$$

and

$$\frac{d}{d\theta}\log f(\theta,x_i) = \frac{x_i}{\theta} - \frac{1-x_i}{1-\theta}, \quad \frac{d^2}{d\theta^2}\log f(\theta,x_i) = -\frac{x_i}{\theta^2} + \frac{1-x_i}{(1-\theta)^2}$$

by Bartlett's identity, we have

$$I_i(\theta) = -\mathbb{E}\left\{-\frac{X_i}{\theta^2} + \frac{1-X_i}{(1-\theta)^2}\right\} = \frac{1}{\theta(1-\theta)}.$$

Now $\tau(\theta) = \theta$ so $[\tau'(\theta)]^2 = 1$ hence the CRLB is given by

$$\mathbf{Var}_\theta(T(\boldsymbol{X})) \geq \frac{1}{nI_i(\theta)} = \frac{\theta(1-\theta)}{n}.$$

An unbiased estimator of $\tau(\theta)$ might be the UMVUE, but its variance could be larger than the CRLB. That is, CRLB is not a sharp lower bound.

Theorem

Theorem 12. *Suppose that $\boldsymbol{X} = (X_1,\cdots,X_n)$ and we have a joint pdf $p_\theta(\boldsymbol{x})$, and $T(\boldsymbol{X})$ be unbiased for $\tau(\theta)$, then $T(\boldsymbol{X})$ attains the CRLB if and only if*

$$a(\theta)\cdot\{T(\boldsymbol{X}) - \tau(\theta)\} = \frac{d}{d\theta}\log p_\theta(\boldsymbol{x})$$

for some function $a(\theta)$ and for all $\theta \in \Theta$.

Example: We say $f_\theta(x)$ belongs to exponential family *(note that you may find me use $f_\theta(x), f(x,\theta), f(x|\theta)$, they are all the same notation, meaning θ is our parameter, but the density function is a function of x)*, if its density function takes the form

$$f_\theta(x) = h(x)\cdot c(\theta)\cdot \exp\{\omega(\theta)\cdot T(x)\}$$

for some non-negative function $h(x)$ of x and $c(\theta)$ of θ, and $T(x)$ is a function of x, also its support does not depend on θ. Then show that if a random sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} f_\theta(x)$ where $f_\theta(x)$ belongs to exponential family, then

$$T(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n T(x_i)$$

is the UMVUE of $\tau(\theta) = -c'(\theta)/(c(\theta)\omega'(\theta))$.

Solution: Fist of all, the joint pdf of the sample takes the form

$$p_\theta(\boldsymbol{x}) = \left(\prod_{i=1}^n h(x_i)\right) c(\theta)^n \cdot \exp\left\{\omega(\theta) \cdot \sum_{i=1}^n T(x_i)\right\}$$

so we have

$$\frac{d}{d\theta}\log p_\theta(\boldsymbol{x}) = n \cdot \frac{c'(\theta)}{c(\theta)} + \omega'(\theta) \cdot \sum_{i=1}^n T(x_i) = \omega'(\theta) \cdot \left\{\sum_{i=1}^n T(x_i) - \frac{-nc'(\theta)}{c(\theta)\omega'(\theta)}\right\}.$$

By the previous theorem, it suffies to check $T(\boldsymbol{X})$ is unbiased. Using Bartlett's identity, we know that $\mathbb{E}\left\{\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right\} = 0$, which means

$$\mathbb{E}\left\{\sum_{i=1}^n T(\boldsymbol{X}_i)\right\} = n\tau(\theta)$$

hence $T(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^n T(X_i)$ is unbiased, and by the previous theorem, it is the UMVUE.

Example: Let a random sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} Poisson(\theta)$ with density $f(x,\theta) = e^{-\theta}\frac{\theta^x}{x!}$, $x \in \mathbb{N}_0$.

(i) find the UMVUE of θ.

(ii) Find the CRLB of all estimates of θ.

Solution:

(i) We rewrite the density of Poisson distribution as

$$f(x,\theta) = \frac{e^{-\theta}}{x!}e^{x\log\theta}$$

let $h(x) = 1/x!, c(\theta) = e^{-\theta}, \omega(\theta) = \log(\theta), T(x) = x$, and its support $x \in \mathbb{N}_0$ does not depend on θ, so it belongs to exponential family, then we use the result from the previous example, now $\tau(\theta) = \theta$, so the UMVUE is just the sample mean, given by

$$\textbf{UMVUE}(\theta) = \frac{1}{n}\sum_{i=1}^n X_i.$$

(ii) By direct computation, we have

$$\frac{d}{d\theta}\log f(x,\theta) = -1+\frac{x}{\theta}, \frac{d^2}{d\theta^2}\log f(x,\theta) = -\frac{x}{\theta^2}$$

so using Bartlett's identity, the Fisher information of one sample is given by

$$I(\theta) = -\mathbb{E}\left\{\frac{d^2}{d\theta^2}\log f(X,\theta)\right\} = \mathbb{E}\left\{-\frac{X}{\theta^2}\right\} = \frac{1}{\theta}$$

since $X \sim Poisson(\theta)$, hence the CRLB can be computed by

$$\textbf{CRLB} = \frac{[\tau'(\theta)]^2}{nI(\theta)} = \frac{\theta}{n}.$$

Sufficiency

Often times, our raw data can be very large and long, and it would be inefficient to work with them directly. We may think of some kind of data reduction to compress our large data into a short and manageable way without losing any information. A statistic $T(\boldsymbol{X})$ is a form of data reduction or summary, thus we are interested in those statistics that do not discard important information about an unknown parameter θ. That's why we introduced the sufficiency principle.

Definition

Definition 10. *Suppose the random sample $X_1,\cdots,X_n$ has a joint pdf (pmf) $p_\theta(\boldsymbol{x})$, $\theta\in\Theta$, a statistic $T(\boldsymbol{X})$ is sufficient of θ, if the conditional distribution of $X_1,\cdots,X_n$ given $T(\boldsymbol{X})=t$ for any $\theta\in\Theta$, $t\in S_T$ such that $f_{\boldsymbol{X}|T(\boldsymbol{X})=t}(x_1,\cdots,x_n)$ does not depend on θ.*

In general, the distribution $p_\theta(\boldsymbol{x}\in\mathscr{X}|T(\boldsymbol{X})=t)$ does not depend on θ.

Example: Let $X_1,\cdots,X_n \overset{i.i.d}{\sim} Bernoulli(\theta)$, we claim that $T(\boldsymbol{X})=\sum_{i=1}^{n} X_i$ is a sufficient statistic of θ, this is because

$$\begin{aligned} f_\theta(x_1,\cdots,x_n|T(\boldsymbol{X})=t) &= \frac{p_\theta(X_1=x_1,\cdots,X_n=x_n,\sum X_i=t)}{p_\theta(\sum X_i=t)} \\ &= \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{n-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \begin{cases} \frac{1}{\binom{n}{t}} & \sum_{i=1}^n X_i = t \\ 0 & \text{elsewhere} \end{cases} \end{aligned}$$

we see that the conditional distribution does not depend on θ.

Note that a sufficient statistic can also be a vector. For example, later we will see if $X_1, \cdots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, a sufficient statistic of (μ, σ^2) is

$$\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right).$$

Corollary

Corollary 3. *Let $T(\boldsymbol{X})$ be a sufficient statistic of θ, g be a one-to-one transformation, then $g(T(\boldsymbol{X}))$ is a sufficient statistic of θ.*

So in the previous Bernoulli example, $T(\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i$ is also a sufficient statistic of θ.

Theorem

Theorem 13. *(Neyman-Fisher Factorization Theorem) Let a random sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} f$ with a joint pdf (pmf) $p_\theta(\boldsymbol{x}) = p_\theta(x_1, \cdots, x_n)$, a statistic $T(\boldsymbol{X})$ is sufficient of θ if and only if*

$$p_\theta(x_1, \cdots, x_n) = g(T(\boldsymbol{x}), \theta) \cdot h(x_1, \cdots, x_n)$$

for all $\theta \in \Theta$, and function g depends on $T(\boldsymbol{x}), \theta$ while $h(\boldsymbol{x})$ does not depend on θ.

The proof of this theorem is omitted.

Example: Find a sufficient statistic of the random sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} Bernoulli(\theta)$.

Solution: We derive

$$\begin{aligned} p_\theta(x_1, \cdots, x_n) &= \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}, x_i \in \{0, 1\} \\ &= \theta^{\sum X_i} \cdot (1-\theta)^{n - \sum x_i} \cdot \prod_{i=1}^{n} \mathbf{1}\{X_i\} \end{aligned}$$

we let the first term to be $g(T(\boldsymbol{x}), \theta)$ and the second term to be $h(\boldsymbol{x})$, and we apply Neyman-Fisher theorem, $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ is a sufficient statistic of θ.

Note: A sufficient statistic doesn't mean a good (unbiased) estimator!

Example: Find a sufficient statistic of the random sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} Uniform(0, \theta)$, $\theta > 0$.

Solution: The pdf of a uniform distribution is given by

$$f(x) = \begin{cases} \frac{1}{\theta} & x \in (0, \theta) \\ 0 & \text{otherwise} \end{cases}$$

then we derive

$$\begin{aligned} p_\theta(\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{\theta} \cdot \chi\{0 < x_i < \theta\} \\ &= \left(\frac{1}{\theta}\right)^n \cdot \mathbf{1}\{0 < X_{(n)} < \theta\} \cdot \mathbf{1}\{0 < X_{(1)}\} \end{aligned}$$

where $X_{(n)} = \max\{X_1, \cdots, X_n\}, X_{(1)} = \min\{X_1, \cdots, X_n\}$, and let the first two terms to be $g(T(\boldsymbol{X}), \theta)$ and the last term to be $h(\boldsymbol{x})$, then by Neyman-Fisher theorem, $T(\boldsymbol{X}) = X_{(n)}$ is a sufficient statistic of θ.

Corollary

Corollary 4. *If $X_1, \cdots, X_n \overset{i.i.d}{\sim} f(\theta, x)$ where f belongs to exponential family, then a sufficient statistic of θ is $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$.*

Proof. Verify it yourself. ■

Example: Find a sufficient statistic of the random sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$.

Solution: Note that the joint pdf can be derived as

$$\begin{aligned} p_\theta(\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \cdot \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[x_i^2 - 2x_i\mu + \mu^2\right]\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \cdot \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2\right]\right) \end{aligned}$$

here, the sufficient statistic takes a vector form, since we have 2 parameters so it makes sense that we have two terms in our sufficient statistic. We let

$$T(\boldsymbol{X}) = \left(\sum_{I=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right)$$

to be the sufficient statistic of $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Note that we cannot separate them, we can't say $\sum X_i$ is a sufficient statistic of μ because in this case Neyman-Fisher won't work.

In fact there are many ways to describe a sufficient statistic, for example recall the Bernoulli example, we can make

$$T(\boldsymbol{X}) = \left(\sum_{i=1}^{m} X_i, \sum_{j=m+1}^{n} X_i\right), 1 \leq m \leq n-1$$

as our sufficient statistic. So although sufficient statistic is a form of data reduction, but it matters how far we can make our reduction. We would want to make as much reductions as possible, and hence we shall introduce minimum sufficient statistic.

Definition

Definition 11. *A statistic $T(\boldsymbol{X})$ is a minimal sufficient statistic of θ if and only if:*

(i) $T(\boldsymbol{X})$ is sufficient;

(ii) For any other sufficient statistic $T^*(\boldsymbol{X})$ of θ, $T(\boldsymbol{X})$ is a function of $T^*(\boldsymbol{X})$, i.e. $T(\boldsymbol{X}) = \varphi(T^*(\boldsymbol{X}))$ where φ is some measurable function.

The following theorem can verify minimal sufficient statistic easily:

Theorem

Theorem 14. *(Lehmann-Scheffe) For a random sample $X_1, \cdots, X_n$ with joint density $p_\theta(\boldsymbol{x})$, suppose a statistic $T(\boldsymbol{X})$ is such that $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathscr{X} \subseteq \mathbb{R}^n$, $T(\boldsymbol{x}) = T(\boldsymbol{y})$ if and only if $\dfrac{p_\theta(\boldsymbol{x})}{p_\theta(\boldsymbol{y})}$ does not depend on θ, then $T(\boldsymbol{X})$ is a minimum sufficient statistic of θ.*

The proof of this theorem is omitted.

Example: Find a minimal sufficient statistic of the random sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} Uniform(0, \theta)$.

Solution: Previously, we had shown that a sufficient statistic is given by $T(\boldsymbol{X}) = X_{(n)}$, and the joint distribution is given by

$$p_\theta(\boldsymbol{x}) = \left(\frac{1}{\theta}\right)^n \cdot \mathbf{1}\{X_{(n)} < \theta\} \cdot \mathbf{1}\{X_{(1)} > 0\}.$$

Now $\forall \boldsymbol{x}, \boldsymbol{y}$ in our sample space, we compute the ratio

$$\frac{p_\theta(\boldsymbol{x})}{p_\theta(\boldsymbol{y})} = \frac{\mathbf{1}\{X_{(n)} < \theta\} \cdot \mathbf{1}\{X_{(1)} > 0\}}{\mathbf{1}\{Y_{(n)} < \theta\} \cdot \mathbf{1}\{Y_{(1)} > 0\}}$$

which we see will not depend on θ if and only if $X_{(n)} = Y_{(n)}$, which means $T(\boldsymbol{X}) = T(\boldsymbol{Y})$ and $T(\boldsymbol{X})$ is the sufficient statistic, hence $T(\boldsymbol{X}) = X_{(n)}$ is a minimal sufficient statistic.

Exercise: Try to use similar method to verify that

$$T(\boldsymbol{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$$

is a minimal sufficient statistic of the random sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$.

Corollary

Corollary 5. *If $T(\boldsymbol{X})$ is a minimal sufficient statistic of θ, then for any one-to-one function g, $g(T(\boldsymbol{X}))$ is also a minimal sufficient statistic of θ.*

Proof. This proof is trivial. ■

Definition

Definition 12. *Let X be a random variable with pdf (pmf) belonging to a parametric family $\mathscr{F} := \{f_\theta : \theta \in \Theta\}$, this family is said to be complete, if for any measurable function g with $\mathbb{E}[g(\boldsymbol{X})]$ exists, we have $\mathbb{E}[g(\boldsymbol{X})] = 0 \implies \mathbb{P}\{g(X) = 0\} = 1$ almost surely. A statistic $T(\boldsymbol{X})$ is complete, if the family of its distributions is complete.*

Example: Show that the statistic $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ in a Bernoulli random sample is complete.

Solution: Set $\mathbb{E}[g(T(\boldsymbol{X}))] = 0$, by definition we have

$$\begin{aligned}
&\sum_{t=0}^{n} g(t)\binom{n}{t}\theta^t(1-\theta)^{n-t} = 0, \forall \theta \in (0,1)\\
\implies &\sum_{t=0}^{n} g(t)\binom{n}{t}\left(\frac{\theta}{1-\theta}\right)^t = 0, \forall \theta \in (0,1)\\
\implies &g(t)\binom{n}{t} = 0\\
\implies &g(t) = 0, \text{ almost surely.}
\end{aligned}$$

and hence $T(\boldsymbol{X})$ is a complete statistic.

Example: Show that the statistic $T(\boldsymbol{X}) = X^2$ where $X \sim N(0,\theta)$, $\theta > 0$ is complete.

Solution: We first claim that $\frac{X^2}{\theta} \sim \chi^2_{(1)}$, it can be shown using the transformation of functions of random variables introduced in **MATH 356**. Then let $\mathbb{E}[g(T(\boldsymbol{X}))] = 0$ where we have $g(x) = x^2$, and by definition it implies that

$$\begin{aligned}
&\int_0^\infty g(t) f_{T(\boldsymbol{X})}(t,\theta)dt = 0\\
\implies &\int_0^\infty \frac{1}{\sqrt{2\pi\theta}} g(t) t^{-1/2} e^{-\frac{t}{2\theta}} dt = 0,
\end{aligned}$$

which further implies $g(t) \equiv 0$.

Corollary

Corollary 6. *Let a random sample* $X_1, \cdots, X_n \overset{i.i.d}{\sim} f(\theta, x)$ *where* f *belongs to exponential family, then* $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ *is a complete statistic of* θ*.*

Two More Methods for UMVUE: Rao-Blackwell and Lehmann-Scheffe

In this section we propose two more methods to compute the UMVUE.

Theorem

Theorem 15. *(Rao-Blackwell) Let* $U(\boldsymbol{X})$ *be an unbiased estimator of* $\tau(\theta)$ *if exists, let* $T(\boldsymbol{X})$ *be a sufficient statistic of the parameter family, set*

$$\delta(t) = \mathbb{E}\{U(\boldsymbol{X})|T(\boldsymbol{X}) = t\}$$

then we claim that:

(i) $\delta(T(\boldsymbol{X}))$ *is a statistic, also an unbiased estimator of* $\tau(\theta)$*;*

(ii) $\mathbf{Var}\{\delta(T(\boldsymbol{X})) \leq \mathbf{Var}(U(\boldsymbol{X}))\}$.

Furthermore, if $T(\boldsymbol{X})$ *is complete, then the resulting* $\delta(T(\boldsymbol{X}))$ *is the UMVUE of the parameter family.*

Proof. Firstly since $T(\boldsymbol{X})$ is sufficient, so it is easy to verify $\mathbb{E}\{U(\boldsymbol{X}|\boldsymbol{T}(\boldsymbol{X}))\}$ does not depend on θ and is hence a statistic, then

$$\mathbb{E}\{\delta(\boldsymbol{X})\} = \mathbb{E}\{\mathbb{E}\{U(\boldsymbol{X}|T(\boldsymbol{X}))\}\} = \mathbb{E}\{U(\boldsymbol{X})\} = \tau(\theta)$$

which means $\delta(\boldsymbol{X})$ is unbiased. Finally

$$\begin{aligned}\mathbf{Var}\{U(\boldsymbol{X})\} &= \mathbf{Var}\{\mathbb{E}\{U(\boldsymbol{X})|T(\boldsymbol{X})\}\} + \mathbb{E}\{\mathbf{Var}\{U(\boldsymbol{X})|T(\boldsymbol{X})\}\} \\ &= \mathbf{Var}\{\delta(T(\boldsymbol{X}))\} + \mathbb{E}\{\mathbf{Var}(U(\boldsymbol{X}|T(\boldsymbol{X})))\} \\ &\geq \mathbf{Var}\{\delta(T(\boldsymbol{X}))\}.\end{aligned}$$

■

Rao-Blackwell theorem says that, by conditioning ant unbiased estimator on a sufficient statistic will result in a uniform improvement in terms of variance, and if we condition on a complete sufficient statistic, we would result in UMVUE.

Example: Consider a random sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} \mathit{Bernoulli}(p)$, $0 < p < 1$. For $n \geq 4$, find the UMVUE of $\tau(p) = p^4$.

Solution: Firstly, since $\mathbb{E}X_i = p$, so $T^*(\boldsymbol{X}) = X_i$ is simply an unbiased estimator of p, and by independence, we can easily see that $X_1X_2X_3X_4$ is an unbiased estimator for p^4. Now since the joint pdf of the random sample is given by

$$f(\boldsymbol{x}, p) = p^{n\bar{x}_n}(1-p)^{n-n\bar{x}_n} = \left(\frac{p}{1-p}\right)^{n\bar{x}_n} \cdot (1-p)^n$$

so by Neyman-Fisher theorem, $T(\boldsymbol{X})$ is a sufficient statistics, and furthermore it is complete (because it belongs to exponential family and we have established some properties before). So by Rao-Blackwell theorem, the UMVUE for p^4 is given by

$$\begin{aligned}
\delta(\boldsymbol{X}) &= \mathbb{E}\left[X_1X_2X_3X_4 \middle| \sum_{i=1}^n X_i = t\right] \\
&= 1 \times \mathbb{P}\left(X_1X_2X_3X_4 = 1 \middle| \sum_{i=1}^n X_i = t\right) + 0 \times \mathbb{P}\left(X_1X_2X_3X_4 = 0 \middle| \sum_{i=1}^n X_i = t\right) \\
&= \frac{\mathbb{P}\left(X_1X_2X_3X_4 = 1, \sum_{i=1}^n X_i = t\right)}{\mathbb{P}\left(\sum_{i=1}^n X_i = t\right)} \\
&= \frac{p^4 \cdot \binom{n-4}{t-4} p^{t-4}(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} \\
&= \frac{\binom{n-4}{t-4}}{\binom{n}{t}}
\end{aligned}$$

where $t = \sum_{i=1}^n X_i$. This is the UMVUE of $\tau(\theta) = p^4$.

Remark: *It is completely doable if I make $\tau(\theta) = p^{2025}$!*

Example: (Hard) Consider a random sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} \mathit{Bernoulli}(p)$, then define

$$h(p) = \mathbb{P}\left(\sum_{i=1}^{n-1} X_i > X_n\right),$$

find the UMVUE of $h(p)$.

Solution: First we define the estimator as the indicator function

$$T(\boldsymbol{X}) = \mathbf{1}\left\{\sum_{I=1}^{n} X_i > X_{n+1}\right\} = \begin{cases} 1 & X_1 + \cdots + X_n > X_{n+1} \\ 0 & \text{otherwise} \end{cases}$$

It is easy to show $T(\boldsymbol{X})$ is unbiased, since

$$\mathbb{E}[T(\boldsymbol{X})] = 1 \times \mathbb{P}(T(\boldsymbol{X}) = 1) + 0 \times \mathbb{P}(T(\boldsymbol{X}) = 0)) = h(p).$$

Then from the previous example we already showed that $\sum_{i=1}^{n} X_i$ is a complete sufficient statistics of p, then by Rao-Blackwell theorem, the UMVUE s given by

$$\begin{aligned}\delta(\boldsymbol{X}) &= \mathbb{E}\left[h(p)\middle|\sum_{i=1}^{n} X_i = t\right] \\ &= \mathbb{P}\left(\sum_{i=1}^{n} X_i > X_{n+1}\middle|\sum_{i=1}^{n} X_i = t\right) \\ &= \frac{\mathbb{P}\left(\sum_{i=1}^{n} X_i > X_{n+1}, \sum_{i=1}^{n} X_i = t\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = t\right)} \\ &= \frac{\mathbb{P}\left(\sum_{i=1}^{n} X_i > X_{n+1}, \sum_{i=1}^{n} X_i = t\right)}{\binom{n}{t} p^t (1-p)^{n-t}}\end{aligned}$$

We will now investigate the numerator, denote the numerator to be N

If $t = 0$, then $N = 0$;

If $t = 1$, then

$$N = \binom{n}{1} p(1-p)^{n-1} \times (1-p)$$

If $t \geq 2$, then

$$N = \binom{n}{t} p^t (1-p)^{n-t} \times (1-p) + \binom{n}{t} p^t (1-p)^{n-t} \times p$$

So in all, we have

$$\delta(\boldsymbol{X}) = \begin{cases} 0 & t = 0 \\ \dfrac{\binom{n}{1} p(1-p)^{n-1} \times (1-p)}{\binom{n}{t} p^t (1-p)^{n-t}} = \dfrac{\binom{n}{1}}{\binom{n}{t}} & t = 1 \\ \dfrac{\binom{n}{t} p^t (1-p)^{n-t} \times (1-p) + \binom{n}{t} p^t (1-p)^{n-t} \times p}{\binom{n}{t} p^t (1-p)^{n-t}} = 1 & t \geq 2 \end{cases}$$

where $t = \sum_{i=1}^{n} X_i$, and this is the UMVUE.

Another way to find UMVUE is proposed as follows:

Theorem

Theorem 16. *(Lehmann-Scheffe) Let $U(\boldsymbol{X})$ be an unbiased estimator of the parameter family, if $U(\boldsymbol{X}) = g(T(\boldsymbol{X}))$ for some measurable function g and a complete sufficient statistic $T(\boldsymbol{X})$, then $U(\boldsymbol{X})$ is the UMVUE of the parameter family.*

That is, as long as we could write the unbiased estimator as a function of complete sufficient statistic, then the unbiased estimator is the UMVUE.

Example: Again let's consider the Bernoulli random sample, we have shown that $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ is a complete and sufficient statistic, also we have

$$\mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = p$$

which is unbiased. Note that inside the expected value is a function of the complete and sufficient statistic, so by using Lehmann-Scheffe directly, the UMVUE of p is just $U(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$.