

Dipartimento Matematica

LECTURE NOTES FOR STATISTICS

Professore
Stefano Favaro

Laureando
Claudio Meggio

Academic Year 2018/2019

Contents

Sommario	2
1 Introduction	3
1.1 Inequalities	3
1.2 Common distributions	3
1.3 Linear Algebra	5
1.4 Multivariate analysis	6
1.5 Basic Concepts of Random Samples	8
2 Concentration Measure	11
2.1 Concentration for sum of random variables	12
3 Likelihood Function	17
3.1 Likelihood principles	17
3.2 Condition of Regularity	18
3.3 Properties of the Likelihood Function	19
3.4 Exponential Families	20
3.5 Natural Exponential Families	22
4 Statistics	23
4.1 Estimators	27
4.2 Properties of Estimators	28
5 Exercises	31
Bibliografia	35

Informazioni sul corso

mail: stefano.favaro@unito.it

book: [1]

1 Introduction

1.1 Inequalities

Markow's Inequality:

Let Y be a non negative random variable with finite expected value then

$$\mathbb{P}(Y \geq t) \leq \frac{\mathbb{E}[Y]}{t}$$

Chebyshev's Inequality:

Let X be a random variable with finite second moment and let $\sigma = \sqrt{\text{Var}(x)}$, then for any positive real h

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq h\sigma) \leq \frac{1}{h^2} \quad (1.1.1)$$

Theorem 1.1.1. Schwartz If X, Y are random variables with finite second moment then:

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$

If X is a random variable taking values in a set I with $\mathbb{E}[X] = \mu$ and $f(\cdot)$ is convex on I , then $f(x) \geq f(\mu) + h(x - \mu)$ holds with probability 1 for some choice of h . By integrating both sides of the inequality with respect to the distribution of X we obtain:

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[X]) \quad (\text{Jensen's Inequality})$$

1.2 Common distributions

Gaussian:

A continuous random variable Y is said to have a Gaussian distribution with parameters μ and σ^2 if the density function at t is:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

Y is unimodal and symmetric around the mode $t = \mu$ and we write

$$Y \sim N(\mu, \sigma^2)$$

It's *characteristic function* is

$$\mathbb{E}[e^{tiy}] = \exp \left\{ it\mu - \frac{\sigma^2 t^2}{2} \right\}$$

The derivative of the characteristic function valued in $t = 0$ give us the non centred moments of Y .

If $Y \sim N(\mu\sigma^2)$ and $a, b \in \mathbb{R}$ then $(a + bY) \sim N(a + b\mu, b^2\sigma^2)$. This mean that the entire family of distribution can be generated by linear transformations starting from any member of the family (i.e. is a *location scale family*)

If $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$ and $Y_1 \perp Y_2$ then $Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. This result can be extended to linear combinations of Gaussian random variables.

Uniform distribution:

A continuous random variable can Y with density function:

$$f(t; a, b) = \frac{1}{b-a} \mathbb{1}_{[a, b]}(t)$$

is said to be *uniformly distributed* in $[a, b]$ and we write $Y \sim U(, b)$.

Theorem 1.2.1. Integral transformation theorem

If Z is a continuous random variable with distribution function F then the random variable

$$W := F(Z) \sim U(0, 1)$$

Proof.

$$\begin{aligned} \mathbb{P}(W \leq t) &= \mathbb{P}(F(Z) \leq t) \\ &= \mathbb{P}(Z \leq F^{-1}(t)) \\ &= F(F^{-1}(t)) \\ &= t \end{aligned}$$

Which is the distribution function of uniform in $[0, 1]$

□

Gamma distribution:

The *Gamma function* is:

$$\Gamma(x) := \int_0^{+\infty} t^{x-1} e^{-t} dt$$

Some properties of this function are:

- $\Gamma(x+1) = x\Gamma(x)$
- if x is a positive integer $\Gamma(x) = (x-1)!$
- $\Gamma(1) = 1$
- $\Gamma\left(-\frac{1}{2}\right) = \sqrt{\pi}$

Stirling's Approximation

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

Definition 1.2.1. We say that a continuous random variable X has a *Gamma distribution* with shape parameter w and scale parameter λ ($X \sim \text{Gamma}(w, \lambda)$) if its density function is:

$$f(t; w, \lambda) = \frac{\lambda^w}{\Gamma(w)} t^{w-1} e^{-\lambda t} \mathbb{1}_{[\mathbb{R}^+]}(t)$$

Proposition 1.2.1. If $Y_1 \sim \text{Gamma}(w_1, \lambda)$ and $Y_2 \sim \text{Gamma}(w_2, \lambda)$ and $Y_1 \perp\!\!\!\perp Y_2$ then:

$$Y_1 + Y_2 \sim \text{Gamma}(w_1 + w_2, \lambda)$$

Other distributions:

1. Beta distribution
2. Binomial Distribution
3. Hypergeometric Distribution
4. Negative Binomial distribution

1.3 Linear Algebra

Matrix:

Consider A, B two $n \times n$ squared matrix

Notation:

I_n is the identity matrix of order n

1_n is the $n \times 1$ (column) vector with all elements equal to 1

\bigcirc is a matrix with all element equal to zero

$|A|$ denotes the determinant of A

Definitions / Properties:

Definition 1.3.1. A is called *symmetric matrix* if

$$A = A^T$$

Proposition 1.3.1. for two conformable matrix A, B we have:

$$|AB| = |A||B|$$

Definition 1.3.2. if $|A| \neq 0$, A is called *non singular* or *invertible* and there exist a matrix A^{-1} called *inverse* such that

$$AA^{-1} = A^{-1}A = I_n$$

Definition 1.3.3. A *diagonal matrix* is a matrix with all elements outside the main diagonal equal to zero

Definition 1.3.4. A matrix A is called *invertible* if there exist an invertible matrix P such that $P^{-1}AP$ is a diagonal matrix

Proposition 1.3.2. It holds:

$$(A^T)^{-1} = (A^{-1})^T$$

$$(AB)^{-1} = B^{-1}A^{-1}$$

Definition 1.3.5. A symmetric matrix A is said to be *positive semi-definite* if

$$v^T Av \geq 0, \forall v \in \mathbb{R}^n$$

Definition 1.3.6. A matrix A is called *orthogonal* if:

$$A^{-1} = A^T$$

Definition 1.3.7. Given a matrix A , we call *trace of A* the sum of all the elements on the main diagonal:

$$Tr(A) := \sum_{i=1}^n a_{ii}$$

Proposition 1.3.3. For any matrix A, B we have:

$$Tr(AB) = Tr(BA)$$

Definition 1.3.8. An *idempotent matrix* A is a matrix which, when multiplied by itself, yields itself i.e.:

$$AA = A$$

Proposition 1.3.4. properties of an idempotent matrix A :

1. $I - M$ is also an idempotent matrix
2. A is idempotent if and only if for all positive integers k , $A^k = A$
3. an idempotent matrix is always diagonalizable and its eigenvalues are either 0 or 1
4. the trace of an idempotent matrix is always an integer and it is equal to its rank

Proposition 1.3.5. Two identities:

1. $(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$
2. $(A + bd^T)^{-1} = A^{-1} - \frac{1}{1+d^T A^{-1}b} A^{-1}bd^T A^{-1}$

Theorem 1.3.1. Spectral theorem

Let A be a symmetric $n \times n$ matrix, then there exist an orthogonal matrix Q such that:

$$A = Q\Lambda Q^T$$

where Λ is a diagonal matrix whose diagonal elements are the eigenvalues $\lambda_1, \dots, \lambda_n$ of A .

Corollary 1.3.1.

$$|A| = |\Lambda| = \prod_{i=1}^n \lambda_i$$

1.4 Multivariate analysis

Definition 1.4.1. Take X_1, \dots, X_n random variables defined on the same probability space, we define the *random vector* or the *multivariate random variable* X as:

$$X = (x_1, \dots, x_n)^T$$

Definition 1.4.2. The *mean vector* of X is obtained by forming the vector of the mean values of the components

$$\mathbb{E} = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n])^T$$

Definition 1.4.3. Similarly we can define the *variance matrix* as:

$$\text{Var}[X] := \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \dots & \text{Var}(X_n) \end{bmatrix}$$

Definition 1.4.4. A generic element of the *correlation matrix* is defined as following:

$$\text{Corr}(x_i, x_j) := \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(x_i)\text{Var}(X_j)}}$$

Lemma 1.4.1. Let $A = a_{ij}$ be a $k \times n$ matrix, $b = (b_1 \dots b_n)^T$ a $n \times 1$ vector and $x = (X_1 \dots X_n)$ a random vector with $\mathbb{E}[x] = \mu$, $\text{Var}(x) = V$ define

$$Y := Ax + b$$

then

$$\mathbb{E}[Y] = A\mu + b$$

$$\text{Var}[Y] = AvA^T$$

Lemma 1.4.2. The variance matrix V of the random vector X is positive semi-definite and it is positive definite if there exist no vectors b such that $b^T X$ is a degenerate random variable.

Lemma 1.4.3. If $A = (a_{ij})$ is a $n \times n$ matrix then:

$$\mathbb{E}[X^T A X] = \mu^T A \mu + \text{Tr}(A V)$$

Proof.

$$\begin{aligned} \mathbb{E}[X^T A X] &= \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n x_i a_{ij} x_j\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n a_{ij} \mathbb{E}[x_i x_j] \\ &= \sum_{i=1}^n \sum_{j=1}^n a_{ij} \mu_i \mu_j + v_{ij} \\ &= \mu^T A \mu + \sum_{i=1}^n (A V)_{ii} \\ &= \mu^T A \mu + \text{Tr}(A V) \end{aligned}$$

□

Multivariate Gaussian distribution:

Consider a vector $Z = (Z_1 \dots Z_k)^T$ where Z_1, \dots, Z_k are independent and identically distributed standard Gaussian random variables. Now set

$$Y = A Z + \mu$$

Where A is a non singular $k \times k$ matrix and μ is a $k \times 1$ vector.

It is natural to define Y as a *k-generated distribution of the Gaussian distribution*.

We start from:

$$f_z = \frac{1}{(2\pi)^{k/2}} \exp \left\{ -\frac{1}{2} t^T t \right\}$$

(Z are independent so we simply multiplied them).

Since $Z = A^{-1}(Y - \mu)$, the Jacobian of the transformation is:

$$\left| \frac{dz_i}{dy_i} \right| = |A|^{-1} = |V|^{-1/2}$$

Taking into account that $|V| = |A A^T| = |A|^2$.

Setting $Y = A t + \mu \implies t = A^{-1}(Y - \mu)$ we obtain:

$$t^T t = \{A^{-1}(Y - \mu)\}^T \{A^{-1}(Y - \mu)\} = (Y - \mu)^T V^{-1} (Y - \mu)$$

Therefore the density of Y is:

$$f_Y(y) = \frac{1}{(2\pi)^{k/2} |V|^{1/2}} \exp \left\{ -\frac{1}{2} (y - \mu)^T V^{-1} (y - \mu) \right\}$$

We say that the random variable $Y = (Y_1 \dots Y_n)^T$ with density function f_Y is a multivariate Gaussian random variable with mean μ and variance V . $Y \sim N_k(\mu, V)$.

Now we will explore *marginal and conditional* distribution of Y .

Proposition 1.4.1. If A is a $k \times k$ positive matrix and b is a $k \times 1$ vector then:

$$\int_{\mathbb{R}^k} \frac{1}{(2\pi)^{k/2}} \exp \left\{ -\frac{1}{2}(y^T A y - 2b^T y) \right\} dy = \frac{\exp\{1/2 b^T A^{-1} b\}}{|A|^{1/2}}$$

Proof. Let $\mu A^{-1} b$ and within the integral expand \exp by adding and subtracting $\frac{1}{2} \mu^T A \mu$ so that

$$\int_{\mathbb{R}^k} \frac{1}{(2\pi)^{k/2}} \exp \left\{ -\frac{1}{2}(y^T A y - 2b^T y) \right\} dy = |A^{-1}|^{1/2} \exp \left\{ \frac{1}{2} \mu^T A \mu \right\} \int_{\mathbb{R}^k} g(y)$$

□

1.5 Basic Concepts of Random Samples

Definition 1.5.1. Let $X_1 \dots X_n$ independent and identically distributed random variables with distribution $\sim f_{X_i}(x_i; \theta)$. We call $X := (X_1 \dots X_n)$ *random sample*

According with the definition of random sample the distribution of X will be:

$$f_X(X; \theta) = \prod_{i=1}^n f_{x_i}(x_i; \theta)$$

Definition 1.5.2. We denote by $x = (x_1 \dots x_n)$ the observed sample

Definition 1.5.3. A *statistical model* is defined as following

$$\{f_X(x; \theta) : \theta \in \Theta\}$$

Where Θ is the *parametric space*

Usually Θ will be a open subset of \mathbb{R}^n .

Definition 1.5.4. Let X_1, \dots, X_n be a random sample of size n from a population and let $T(x_1, \dots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of (X_1, \dots, X_n) . Then the random variable or random vector $T_n = T(X_1, \dots, X_n)$ is called a *statistic*. The probability distribution of a statistic T_n is called the *sampling distribution* of T_n .

Some examples of statistic are:

- Sample mean: $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n x_i$
- Sample variance: $\tilde{S}^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$
- Corrected sample variance: $S^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2$
- Sample moments of order r : $M_{r,n} := \frac{1}{n} \sum_{i=1}^n x_i^r$
- Sample moments of order r : $\bar{M}_{r,n} := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^r$
- Ordered statistic: $X_{(m)}$
- Sample min: $X_{(1)}$
- Sample max: $X_{(n)}$
- Sample median: $Me := \begin{cases} \frac{1}{2}(X_{(n/2)} + X_{(n/2+1)}) & \text{if } n - \text{even} \\ X_{(\frac{n+1}{2})} & \text{if } n - \text{odd} \end{cases}$

Finding the distribution of T_n in general it is complex. We can make it easier by putting constraints. Suppose for example $X \sim N(\mu, \sigma^2) \leftarrow$ fair assumption because there is the Central Limit Theorem.

Theorem 1.5.1. Fisher Cochran

Let Q, Q_1, Q_2 random variables such that $Q = Q_1 + Q_2$ and let $Q \sim \chi_g^2$ and $Q_2 \sim \chi_{g_1}^2$. Then

$$Q_2 \sim \chi_{g_2}^2 \quad \text{where } g_2 = g - g_1,$$

and $Q_1 \perp\!\!\!\perp Q_2$

Proposition 1.5.1. Let $X_i \sim N(\mu, \sigma^2)$ and $X = (X_1 \dots X_n)$. Then

1. $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n x_i \sim N(\mu, \frac{\sigma^2}{n})$
2. $\tilde{S}_n := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 \sim \frac{\sigma^2}{n} \chi_{n-1}^2$

where χ_{n-1}^2 is the Chi-squared distribution with $n - 1$ degrees of freedom

Proof. 1. the first one is easily checked using the linearity of the Gaussian distribution.

2. Consider the random variable $\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$ and proceed as following:

$$\begin{aligned} \frac{n\tilde{S}^2}{\sigma^2} &= \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2 \\ &= \sum_{i=1}^n \left(\frac{x_i - \mu + \bar{x}_n - \bar{x}_n}{\sigma} \right)^2 \\ &= \sum_{i=1}^n \left(\frac{x_i - \bar{x}_n}{\sigma} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{x}_n - \mu}{\sigma} \right)^2 + 2 \sum_{i=1}^n \left(\frac{x_i - \bar{x}_n}{\sigma} \right) \left(\frac{\bar{x}_n - \mu}{\sigma} \right) \end{aligned}$$

Now consider separately the three terms of the sum:

$$\sum_{i=1}^n \left(\frac{x_i - \bar{x}_n}{\sigma} \right) = n\bar{x}_n \sum_{i=1}^n x_i = n\bar{x}_n - n\bar{x}_n = 0$$

$$\implies 2 \sum_{i=1}^n \left(\frac{x_i - \bar{x}_n}{\sigma} \right) \left(\frac{\bar{x}_n - \mu}{\sigma} \right) = 0$$

For the other two terms consider:

$$\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2 \sim \chi_1^2$$

And

$$\sum_{i=1}^n \left(\frac{\bar{x}_n - \mu}{\sigma} \right)^2 = \left(\frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \right)^2 \sim \chi_n^2$$

So, using the theorem 1.5.1

$$\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n \left(\frac{x_i - \bar{x}_n}{\sigma} \right)^2 + \left(\frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \right)^2 \sim \chi_{n-1}^2$$

$$\implies \tilde{S}_n^2 \sim \frac{\sigma^2}{n} \chi_{n-1}^2$$

□

Let $X = (X_1 \dots X_n)$ be a random sample where $X_i \sim N(\mu, \sigma^2)$, consider the statistic:

$$T_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{(\bar{X}_n - \mu)/\sigma/\sqrt{n}}{\sqrt{S_n/\sigma^2}} = \frac{Z}{\sqrt{R/(n-1)}}$$

where $Z \sim N(0, 1)$, $R \sim (\chi_{n-1}^2)$ We can say that $T_n \sim T$ - Student with $(n - 1)$ degrees of freedom only if $Z \perp\!\!\!\perp R$. We can they are independent because of the following:

Theorem 1.5.2. *If $(x_1 \dots x_n)$ random sample with $X_i \sim N(\mu\sigma^2)$ then*

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

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

are independent

The other way around is also true:

Theorem 1.5.3. *If $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n x_i$, $S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X}_n)^2$ are independent then $X = (x_1 \dots x_n)$ is random sample where $X_i \sim N(\mu\sigma^2)$*

2 Concentration Measure

We're now going to investigate some methods to study the tail of a distribution.

Consider a non negative random variable and let $t > 0$. Then by Markow inequality we have:

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

We can try to improve this inequality using a function Φ that is strictly increasing with non negative values. Then we can write

$$\mathbb{P}(X \geq t) = \mathbb{P}(\Phi(X) \geq \Phi(t)) \leq \frac{\mathbb{E}[\Phi(X)]}{\Phi(t)}$$

In particular we can take $\Phi(x) = x^q$, $X \geq 0, q > 0$ so we have

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^q]}{t^q}$$

In specific examples one can choose the value of q that optimize the upper bound.

A related idea is a the basis of **Chernoff's bounding method**: taking $\Phi(X) = e^{sX}$ where s is an arbitrary positive number for any random variable X and $t \in \mathbb{R}$ we have:

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{sX} \geq e^{st}) \leq \frac{\mathbb{E}[e^{sX}]}{e^{st}} \quad (2.0.1)$$

So we can bound the probability using the characteristic function which is usually easier than $\mathbb{E}[X^q]$ to compute.

However it can be proven that the bounding given form $\Phi(X) = x^q$ is always better than the one given by $\Phi(X) = E^{sX}$.

Theorem 2.0.1. *Cauchy Swartz inequality*

Given two random variables with finite second moments then:

$$|\mathbb{E}[XY]|^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$

Theorem 2.0.2. *let $t \geq 0$ then*

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \frac{\text{Var}(X)}{\text{Var}(X) + t^2}$$

Proof. We assume that $\mathbb{E}[X] = 0$ (the proof for the general case is the same).

For all t we can write

$$t = \mathbb{E}[t] = \mathbb{E}[t] - \mathbb{E}[X] = \mathbb{E}[t - X] \leq \mathbb{E}[(t - X)\mathbb{1}_{[X < t]}(X)]$$

Then for $t \geq 0$ from Cauchy Swartz inequality:

$$\begin{aligned} t^2 &\leq \mathbb{E}[(t - X)^2]\mathbb{E}[(\mathbb{1}_{[X < t]}(X))^2] \\ &= \mathbb{E}[(t - X)^2]\mathbb{P}(X < t) \\ &= (\text{Var}(X) + t^2)\mathbb{P}(x < t) \end{aligned}$$

$$\implies \mathbb{P}(X < t) \geq \frac{t^2}{\text{Var}(X) + t^2}$$

$$\implies \mathbb{P}(X \geq t) = 1 - \mathbb{P}(X < t) \leq 1 - \frac{t^2}{\text{Var}(X) + t^2} = \frac{\text{Var}(X)}{\text{Var}(X) + t^2} \quad \square$$

Theorem 2.0.3. *Let f, g be non decreasing real valued functions defined on the real line. If X is a real valued random variable then:*

$$\mathbb{E}[f(x)g(x)] \geq \mathbb{E}[f(x)]\mathbb{E}[g(x)]$$

If f is non increasing and g is non decreasing then:

$$\mathbb{E}[f(x)g(x)] \leq \mathbb{E}[f(x)]\mathbb{E}[g(x)]$$

Proof. Let Y be a random variable with the same distribution of X and $X \perp\!\!\!\perp Y$. Because f, g are non decreasing functions we have $(f(x) - f(y))(g(x) - g(y)) \geq 0$

$$\implies 0 \leq \mathbb{E}[(f(x) - f(y))(g(x) - g(y))] = \mathbb{E}[f(x)g(x) - f(x)g(y) - f(y)g(x) + f(y)g(y)]$$

\implies

$$\begin{aligned} \mathbb{E}[f(x)g(x)] &\geq \mathbb{E}[f(x)g(y)] + \mathbb{E}[f(y)g(x)] - \mathbb{E}[f(x)g(y)] \\ &= \mathbb{E}[f(x)g(y)] \\ &= \mathbb{E}[f(x)]\mathbb{E}[g(y)] \\ &= \mathbb{E}[f(x)]\mathbb{E}[g(x)] \end{aligned}$$

The second part of the theorem can be proved in the same way. \square

The previous theorem can be generalized as following:

Theorem 2.0.4. *Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ be non increasing functions. Let $X_1 \dots X_n$ be independent real valued random variables and define the random variable $X = (X_1 \dots X_n)$ that take values in \mathbb{R}^n then:*

$$\mathbb{E}[f(x)g(x)] \geq \mathbb{E}[f(x)]\mathbb{E}[g(x)]$$

If f is non increasing and g is non decreasing then:

$$\mathbb{E}[f(x)g(x)] \leq \mathbb{E}[f(x)]\mathbb{E}[g(x)]$$

2.1 Concentration for sum of random variables

We want to bound the probability $\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t)$ where $S_n = \sum_{i=1}^n X_i$ and $X_1 \dots X_n$ are independent random variables real valued.

An application of the Chebyshev's inequality give us:

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq \frac{\text{Var}(S_n)}{t^2} = \frac{\sum_{i=1}^n \text{Var}(X_i)}{t^2}$$

Applying the Chebyshev's inequality to $\frac{1}{n} \sum_{i=1}^n x_i$ we get

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{n} \left(\sum_{i=1}^n x_i - \mathbb{E}[X_i]\right)\right| \geq \epsilon\right) &= \mathbb{P}\left(|S_n - \mathbb{E}[S_n]| \geq \epsilon n\right) \\ &\leq \frac{\sum_{i=1}^n \text{Var}(X_i)}{\epsilon^2 n^2} \end{aligned}$$

If we define $\sigma^2 := \frac{1}{n} \sum_{i=1}^n \text{Var}(X_i)$ then:

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n x_i - \mathbb{E}[X_i]\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2} \quad (2.1.1)$$

To understand why the equation 2.1.1 is unsatisfying recall what appens with the *Central Limit Theorem*:

$$\mathbb{P}\left(\sqrt{\frac{n}{\sigma^2}}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X_i]\right) \geq y\right) \xrightarrow{n \rightarrow \infty} 1 - \Phi(y) \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-y^2/2}}{y}$$

(where Φ is the CDF of the standard Gaussian distribution)

so

$$\mathbb{P}\left(\sqrt{\frac{n}{\sigma^2}}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X_i]\right) \geq \epsilon\right) \lesssim \exp\left\{\frac{-n\epsilon^2}{2\sigma}\right\}$$

So for $\mathbb{P}\left(\sqrt{\frac{n}{\sigma^2}}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X_i]\right) \geq \epsilon\right)$ we have:

$$\exp\left\{\frac{-n\epsilon^2}{2\sigma}\right\} \leftarrow \text{from Central Limit Theorem}$$

$$\frac{\sigma^2}{n\epsilon^2} \leftarrow \text{from Chebyshev's inequality}$$

From here we can see that the Chebyshev's inequality doesn't work well for the sum of n random variables when n is large. Meanwhile the Chebyshev's inequality works better than the Central Limit Theorem for small n .

Another instrument previously introduced that can be helpful for bounding tail probabilities of sum of independent random variables is the **Chernoff bounding** 2.0.1:

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-st} \mathbb{E}[\exp\{s \sum_{i=1}^n (x_i - \mathbb{E}[X_i])\}] = e^{-st} \prod_{i=1}^n \mathbb{E}[\exp\{s(x_i - \mathbb{E}[X_i])\}] \quad (2.1.2)$$

(remember that s is an arbitrary positive number)

Now the problem of finding bound on the tail probability reduces to the problem of finding (upper) bounds for the moments generating function of $X_i - \mathbb{E}[X_i]$.

As we saw Chebyshev's inequality 1.1.1 does not work well for sums of random variables. In this section we will see a partial solution given by *Hoeffding's Inequality*, then a more complete solution given by *Bernstein Inequality*.

Lemma 2.1.1. *Let X be a with $\mathbb{E}[X] = 0$ (actually it can be generalized for a random variable with any expected value), $a \leq X \leq b$ (X bounded random variable). Then*

$$\mathbb{E}[e^{sx}] \leq \exp\left\{\frac{s^2(b-a)^2}{8}\right\} \quad \text{for } s > 0$$

Proof. By the convexity of the exp function we have

$$e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa} \quad \text{with } a \leq x \leq b$$

Using $\mathbb{E}[X] = 0$ and defining $p := \frac{-a}{b-a}$ we obtain

$$\begin{aligned}
\mathbb{E}[e^{sx}] &\leq \mathbb{E}\left[\frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}\right] \\
&\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\
&= \frac{b-a+a}{b-a}e^{sa} + pe^{sb} \\
&= (1-p)e^{sa} + pe^{sb} \\
&= (1-p)e^{sa} + pe^{s(b-a+a)} \\
&= (1-p)e^{sa} + pe^{s(b-a)}e^{sa} \\
&= (1-p + pe^{s(b-a)})e^{sa} \\
&= (1-p)e^{sa} + pe^{s(b-a)}e^{sa\frac{b-a}{b-a}} \\
&= (1-p)e^{sa} + pe^{s(b-a)}e^{-ps(b-a)}
\end{aligned}$$

Then defining

$$\mu = s(s-a)$$

$$\Phi(\mu) = -p\mu + \ln(1-p + pe^\mu)$$

so we have that the last equality $(1-p)e^{sa} + pe^{s(b-a)}e^{-ps(b-a)} = e^{\Phi(\mu)}$

It is possible to show

$$\Phi'(X) = -p + \frac{p}{p + (1-p)e^{-\mu}}$$

therefore $\Phi(\mu) = \Phi'(0) = 0$, moreover

$$\Phi(\mu) = \frac{p(1-p)e^{-\mu}}{(p + (1-p)p^{-\mu})^4} \leq \frac{1}{4}$$

by Taylor's theorem we have:

$$\Phi(x) = \Phi(0) + \mu\Phi'(0) + \frac{\mu^2}{2}\Phi''(\sigma) \leq \frac{\mu^2}{8} = \frac{s^2(b-a)^2}{8}$$

with $\sigma \in [0, \mu]$. □

We're now ready for the **Hoeffding's Inequality**

Theorem 2.1.1. *Let $(x_1 \dots x_n)$ be independent random variable such that $x_i \in [a_i, b_i]$ then for any $t > 0$*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \leq -t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

Proof. Using the *Chernoff's bounding* for sums of random variables 2.1.2 and the precedent lemma 2.1.1 we obtain

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-st} \prod_{i=1}^n e^{\frac{s^2(b_i-a_i)^2}{8}} = e^{-st} e^{\frac{s^2}{8} \sum_{i=1}^n (b_i-a_i)^2} = e^{-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}}$$

where we chose $s = \frac{4t}{\sum_{i=1}^n (b_i-a_i)^2}$ □

This inequality has the same form as the one based on the central limit theorem except that the average variance σ^2 is replaced by the upper bound $\frac{1}{4} \sum_{i=1}^n (b_i - a_i)^2$. Next we will see *Bernstein*

Inequality an inequality that take into account also the variance.

Lemma 2.1.2. Assume that $\mathbb{E}[X_i] = 0$ then if for all X_i , $|X_i| \leq c$ (X_i are bounded):

$$\mathbb{E}[e^{sx_i}] \leq \exp \left\{ s^2 \sigma_i^2 \frac{e^{sc} - 1 - sc}{sc} \right\}$$

where $\sigma_i^2 := \mathbb{E}[X_i^2]$

Proof. define $F_i = \sum_{r=2}^{\infty} s^{r-2} \frac{\mathbb{E}[x_i^r]}{r! \sigma_i^2}$.

Since (for Taylor) $e^{sx} = 1 + sx + \sum_{r=2}^{\infty} s^r \frac{x^r}{r!}$ then taking into account $\mathbb{E}[X_i] = 0$

$$\mathbb{E}[e^{sX_i}] = 1 + s\mathbb{E}[X_i] \sum_{r=2}^{\infty} s^r \frac{\mathbb{E}[x_i^r]}{r!} = 1 + s^2 \sigma_i^2 F_i \leq e^{s^2 \sigma_i^2 F_i}$$

Because we supposed $|X_i| \leq c$ for each index r we have

$$\mathbb{E}[X_i^r] = \mathbb{E}[X_i^{r-2} X_i^2] \leq \mathbb{E}[c^{r-2} X_i^2] = c^{r-2} \sigma_i^2$$

Thus

$$\begin{aligned} F_i &\leq \sum_{r=2}^{\infty} \frac{s^{r-2} c^{r-2} \sigma_i^2}{r! \sigma_i^2} \\ &= \frac{1}{(sc)^2} \sum_{r=2}^{\infty} \frac{(sc)^r}{r!} \\ &= \frac{e^{sc} - 1 - sc}{(sc)^2} \end{aligned}$$

where in the last step we recognized the summation as the exponential wrote in Taylor series missing the first two terms \square

Theorem 2.1.2. Bernstein Inequality

Let $(x_1 \dots x_n)$ be independent real valued random variables with $\mathbb{E}[X_i] = 0$ and $|X_i| \leq c$. Set $\sigma^2 = \frac{1}{n} \sum_{i=1}^n \text{Var}[X_i]$ (note that $\text{Var}[X_i] = \mathbb{E}[X_i^2]$ because $\mathbb{E}[X_i] = 0$). Then for $t > 0$

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp \left\{ -\frac{n\sigma^2}{c^2} h\left(\frac{ct}{n\sigma^2}\right) \right\}$$

where $h(\mu) = (1 + \mu) \ln(1 + \mu) - \mu$ for $\mu \geq 0$.

Proof. Using the Chernoff's bounding for sums of random variables 2.1.2 we obtain and the precedent lemma 2.1.2 we obtain

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp \left\{ \frac{n\sigma^2(e^{sc} - 1 - sc)}{c^2} - st \right\}$$

and the bound is minimized by $s = \frac{1}{c} \ln \left(1 + \frac{tc}{n\sigma^2} \right)$ \square

Corollary 2.1.1. Referring to the Bernstein Inequality there is a lower bound for h :

$$h(\mu) \geq \frac{\sigma^2}{2 + 2\mu_\epsilon}$$

so for $\epsilon > 0$ the Bernstein Inequality becomes:

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left\{-\frac{n\epsilon}{2\sigma^2 + \frac{2}{3}c\epsilon}\right\}$$

This result is extremely useful in hypothesis testing ($\mathbb{P}(T_n > t) = \alpha$) because usually to do the test we have to invert the CDF of T_n . With this result we can instead use the second term of the *Bernstein Inequality* as α and then we can isolate the ϵ to find the small t . Sadly this work only if T_n is a sum of independent random variables which however is the most common situation.

We consider now the problem of deriving inequalities for the Variance of functions of independent random variables.

Lemma 2.1.3. *Let \mathcal{X} be some set and let $g : \mathcal{X}^n \rightarrow \mathbb{R}$ be a measurable function. Define $Z := g(X_1 \dots X_n)$ where $(x_1 \dots x_n)$ are independent random variables in \mathcal{X} and $\mathbb{E}_i Z$ the expected value of Z with respect to X_i that is $\mathbb{E}_i Z = \mathbb{E}[Z | X_1 \dots X_{i-1}, X_{i+1} \dots X_n]$. Then*

$$\text{Var}(Z) \leq \sum_{i=1}^n \mathbb{E}[(Z - \mathbb{E}_i Z)^2]$$

Directly from this lemma follows

Theorem 2.1.3. Efron-Stein Inequality *Let $X'_1 \dots X'_n$ be from an independent copy of $X_1 \dots X_n$ and define $Z'_i = g(X_1 \dots X_{i-1}, X'_i, X_{i+1} \dots X_n)$ then*

$$\text{Var}(Z) \leq \sum_{i=1}^n \mathbb{E}[(Z - Z'_i)^2]$$

when $g(X_1 \dots X_n) = \sum_{i=1}^n X_i$ the inequality becomes an equality.

3 Likelihood Function

The *likelihood function* is a function that contains all the statistical information required to make inference.

Definition 3.0.1. Consider a random sample $(X_1 \dots X_n)$ from $X \sim f_X(X; \theta)$, then the distribution of $(X_1 \dots X_n)$ will be:

$$f_{\underline{X}}(\underline{x}; \theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta)$$

when we see $f_{\underline{X}}(\underline{x}; \theta)$ as a function of θ for fixed \underline{x} we call it **likelihood function**

$$\mathcal{L}(\theta, \underline{x}) = \prod_{i=1}^n f_{X_i}(x_i; \theta)$$

An important function related to the likelihood function is the **log likelihood function**

Definition 3.0.2. The \ln of the likelihood function is said *log likelihood function*

$$V_n(\theta) = \ln \mathcal{L}(\theta, \underline{x}) = \ln \left(\prod_{i=1}^n f_{X_i}(x_i; \theta) \right) = \sum_{i=1}^n \ln(f_{X_i}(x_i; \theta))$$

Definition 3.0.3. Score function:

$$V'_n = \frac{d}{d\theta} V_n(\theta) = \frac{\mathcal{L}'(\theta, \underline{x})}{\mathcal{L}(\theta, \underline{x})}$$

Note that if we fix θ then $\mathcal{L}(\theta, \underline{x})$ is (related to) the probability that the particular value we fixed for θ has generated \underline{x} .

Suppose we fix two value $\theta, \theta_2 \in \Theta$ and $\mathcal{L}(\theta_1, \underline{x}) > \mathcal{L}(\theta_2, \underline{x})$ then we say \underline{x} is "*more likely*" generated under θ_1 .

Note that the same meaning is also applicable to the log likelihood function.

It is because of this that we usually search for the maximum of the likelihood function. Often to find it we just derive, but sometimes \mathcal{L} is not regular enough so we have to "*regularize*" it.

3.1 Likekihood principles

The statistical inference based on the likelihood function is a consequence of two principles:

1. **Week likelihood principle:** for a fixed parametric model $X \sim F_X(x, \theta)$ if two observed samples \underline{x} and \underline{y} are such that

$$\mathcal{L}(\theta, \underline{x}) \propto \mathcal{L}(\theta, \underline{y})$$

then the two likelihood functions are equivalent i.e. must produce the same inference result on θ .

2. **Strong likelihood principle:** let \underline{x} be an observed sample under the model $X \sim F_X(x, \theta)$ with likelihood function $\mathcal{L}(\theta, \underline{x})$ and let \underline{y} be an observed sample under the model $X \sim F_Y(y, \theta)$ with likelihood function $\mathcal{L}(\theta, \underline{y})$.

If $\mathcal{L}(\theta, \underline{x}) \propto \mathcal{L}(\theta, \underline{y})$ the the two samples provides with the same inference.

The fundamental difference between *Probability* and *Statistic* is that in the first one the goal is to find the chance of a random variable to take a particular value, statistic instead given the results of a experiment, try to find the distribution where it came from.

Example 3.1.1. Take (X_1, X_2, X_3) from one of the following distribution:

- (a) $X \sim \text{Ber}(\theta_1), \theta_1 = \frac{1}{2}$
- (b) $X \sim \text{Ber}(\theta_2), \theta_2 = \frac{1}{3}$
- (c) $X \sim \text{Ber}(\theta_3), \theta_3 = \frac{1}{4}$

$X \in [0, 1] \quad \Theta = [0, 1]$.

We can imagine (x_1, x_2, x_3) as the results of a experiment where we had to flip a coin 3 times. Now we want to know the parameter θ of the coin we flipped tree times and we have tree possibilities: $\theta_1 = \frac{1}{2}, \theta_2 = \frac{1}{3}, \theta_3 = \frac{1}{4}$.

So $(x_1, x_2, x_3) \in \{0, 1\}^3$

x_1, x_2, x_3	$\theta_1 = \frac{1}{2}$	$\theta_2 = \frac{1}{3}$	$\theta_3 = \frac{1}{4}$
0,0,0	$\frac{1}{8}$	$\frac{8}{27}$	$\frac{27}{64}$
0,0,1	$\frac{1}{8}$	$\frac{4}{27}$	$\frac{9}{64}$
0,1,0	$\frac{1}{8}$	$\frac{4}{27}$	$\frac{9}{64}$
1,0,0	$\frac{1}{8}$	$\frac{4}{27}$	$\frac{9}{64}$
0,1,1	$\frac{1}{8}$	$\frac{2}{27}$	$\frac{3}{64}$
1,0,1	$\frac{1}{8}$	$\frac{2}{27}$	$\frac{3}{64}$
1,1,0	$\frac{1}{8}$	$\frac{2}{27}$	$\frac{3}{64}$
1,1,1	$\frac{1}{8}$	$\frac{1}{27}$	$\frac{1}{64}$

Once we know the result of the throw we will "guess" the value of θ choosing the one that give us more probability for the given result.

3.2 Condition of Regularity

In our investigations on θ we will assume some condition of regularity for our model.

Given $X \sim F_x(x, \theta)$

1. we assume that $\theta \in \Theta$ where Θ is a open real set
2. for any $\theta \in \Theta$ there exist the derivative of $\mathcal{L}(\theta; z)$ with respect to θ at least up to the third order
3. for any $\theta_0 \in \Theta$ there exist tree functions g, h, H that are integrable in a neighborhood of θ_0 and

- $\left| \frac{d}{d\theta} f_X(x, \theta) \right| \leq g(x)$
- $\left| \frac{d^2}{d^2\theta} f_X(x, \theta) \right| \leq h(x)$
- $\left| \frac{d^3}{d^3\theta} \ln(f_X(x, \theta)) \right| \leq H(x)$

4. for any $\theta \in \Theta$

$$0 < \mathbb{E}[(\ln(\mathcal{L}(\theta, \underline{X})))^2] < \infty$$

(With \underline{X} we're tanking it as random variable).

In addition there is the condition of identifiability.

5. We say that a statistical model is identifiable if for every θ_1, θ_2 there is al least one event E such that:

$$\mathbb{P}(X \in E | \theta_1) \neq \mathbb{P}(X \in E | \theta_2)$$

(We will always take 5. as granted)

3.3 Properties of the Likelihood Function

Proposition 3.3.1. *Some properties of the score function 3.0.3 are:*

1. $\mathbb{E}[V'_n(\theta)] = 0$
2. $\text{Var}(V'_n(\theta)) = \mathbb{E}[(V'_n(\theta))^2] = -\mathbb{E}[V''_n(\theta)]$

Proof. 1.

$$\begin{aligned}
 \mathbb{E}[V'_n(\theta)] &= \int_{\mathbb{R}^n} V'_n(\theta) f_{\underline{X}}(\underline{x}, \theta) d\underline{x} \\
 &= \int_{\mathbb{R}^n} \frac{f'_{\underline{X}}(\underline{x}, \theta)}{f_{\underline{X}}(\underline{x}, \theta)} f_{\underline{X}}(\underline{x}, \theta) d\underline{x} \\
 &= \int_{\mathbb{R}^n} \frac{d}{d\theta} f_{\underline{X}}(\underline{x}, \theta) d\underline{x} \\
 &= \frac{d}{d\theta} \int_{\mathbb{R}^n} f_{\underline{X}}(\underline{x}, \theta) d\underline{x} \\
 &= \frac{d}{d\theta} 1 \\
 &= 0
 \end{aligned}$$

Where in the in the fourth equal we used Leibniz and for the fifth recall that $f_{\underline{X}}(\underline{x}, \theta)$ is the PDF of \underline{X}

2. Start by showing that $V''_n(\theta) = \frac{f''_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{x}; \theta)} - \left(\frac{f'_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{x}; \theta)} \right)^2$

$$\begin{aligned}
 V''_n(\theta) &= \frac{d^2}{d\theta^2} \mathcal{L}(\theta, \underline{x}) \\
 &= \frac{d}{d\theta} \frac{f'_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{x}; \theta)} \\
 &= \frac{f''_{\underline{X}}(\underline{x}; \theta) f_{\underline{X}}(\underline{x}; \theta) - f'_{\underline{X}}(\underline{x}; \theta) f'_{\underline{X}}(\underline{x}; \theta)}{|f_{\underline{X}}(\underline{x}; \theta)|^2} \\
 &= \frac{f''_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{x}; \theta)} - \left(\frac{f'_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{x}; \theta)} \right)^2 \\
 &= \frac{f''_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{x}; \theta)} - (V'_n(\theta))^2
 \end{aligned}$$

So now

$$\begin{aligned}
 \mathbb{E}[V''_n(\theta)] &= \int_{\mathbb{R}^n} \frac{f''_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{x}; \theta)} f_{\underline{X}}(\underline{x}; \theta) d\underline{x} - \mathbb{E}[V'_n(\theta)^2] \\
 &= \int_{\mathbb{R}^n} \frac{d^2}{d\sigma^2} f_{\underline{X}}(\underline{x}; \theta) d\underline{x} - \mathbb{E}[V'_n(\theta)^2] \\
 &= \frac{d^2}{d\sigma^2} \int_{\mathbb{R}^n} f_{\underline{X}}(\underline{x}; \theta) d\underline{x} - \mathbb{E}[V'_n(\theta)^2] \\
 &= -\mathbb{E}[V'_n(\theta)^2]
 \end{aligned}$$

□

Definition 3.3.1. we define the **Fisher Information** as:

$$\mathcal{I}_n(\theta) = -\mathbb{E}[V''_n(\theta)]$$

Note that this is the definition of Fisher information just for a particular case, there exist a more general one.

The Fisher information has a central role in statistic because it can be shown that for *unbiased estimators* $\tilde{\theta}$ it holds: $Var(\tilde{\theta}) \geq \frac{1}{\mathcal{I}_n(\theta)}$. So if we can find an estimator with $Var(\tilde{\theta}) = \frac{1}{\mathcal{I}_n(\theta)}$ we are sure that it is the one with the lowest variance.

P

Proposition 3.3.2. Consider $(x_1 \dots x_n)$ from $X \sim F_X(x; \theta)$ regular then

$$\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$$

Example 3.3.1. Consider $(x_1 \dots x_n)$ from $X \sim Ber(\theta)$

$$\mathcal{L}_n(\theta, \underline{x}) = \prod_{i=1}^n f_{X_i}(x_i; \theta) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$$

$$V_n(\theta) = \ln(\mathcal{L}_n(\theta, \underline{x})) = \ln(\theta) \sum_{i=1}^n x_i + (\ln(1 - \theta)) \left(n - \sum_{i=1}^n x_i \right)$$

$$V'_n(\theta) = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1 - \theta}$$

$$V''_n(\theta) = \frac{-\sum_{i=1}^n x_i}{\theta^2} - \frac{-\sum_{i=1}^n x_i}{(1 - \theta)^2}$$

$$\begin{aligned} \mathbb{E}[V_n(\theta)] &= \frac{1}{\theta} \sum_{i=1}^n \mathbb{E}[x_i] - \frac{1}{1 - \theta} \left(n - \sum_{i=1}^n \mathbb{E}[x_i] \right) \\ &= \frac{n\theta}{\theta} - \frac{1}{1 - \theta} (n - n\theta) \\ &= n - n = 0 \end{aligned}$$

$$\begin{aligned} [V''_n(\theta)] &= -\frac{1}{\theta^2} \sum_{i=1}^n \mathbb{E}[x_i] - \frac{1}{(1 - \theta)^2} \left(n - \sum_{i=1}^n \mathbb{E}[x_i] \right) \\ &= -\frac{n\theta}{\theta^2} - \frac{n - n\theta}{(1 - \theta)^2} \\ &= -\frac{n}{\theta} - \frac{n(1 - \theta)}{(1 - \theta)^2} \\ &= -\frac{n\theta}{(1 - \theta)\theta} \end{aligned}$$

$$\begin{aligned} \mathcal{I}_n(\theta) &= -\mathbb{E}[V''_n(\theta)] \\ &= \frac{n}{(1 - \theta)\theta} \end{aligned}$$

3.4 Exponential Families

Definition 3.4.1. We say that the distribution of a random variable is an element of an **Exponential Family** $X \sim EF(\theta)$ if its PDF can be written as follow:

$$f_X(x; \theta) = \exp\{Q(\theta)A(x) + C(x) - k(\theta)\}$$

The generalization to random sample becomes:

Definition 3.4.2. We say that the distribution of a random sample $(x_1 \dots x_n)$ from $X \sim EF(\theta)$ is an

element of an **Exponential Family** $X \sim EF(\theta)$ if its PDF can be written as follow:

$$f_{\underline{X}}(\underline{x}; \theta) = \exp\{Q(\theta) \sum_{i=1}^n A(x_i) + \sum_{i=1}^n C(x_i) - nK(\theta)\}$$

Example 3.4.1. $X \sim Ber(\theta)$, $X \in \{0, 1\}$, $\Theta = (0, 1)$

$$\begin{aligned} P_X(x) &= \theta^x (1\theta)^{1-x} \mathbb{1}_{\{0,1\}}(x) \\ &= \exp\{x \ln(\theta) + (1-x) \ln(1-\theta)\} \\ &= \exp\{x \ln(\theta) + \ln(1-\theta) - x \ln(1-\theta)\} \\ &= \exp\left\{x \ln\left(\frac{\theta}{1-\theta}\right) + \ln(1-\theta)\right\} \end{aligned}$$

so we get

$$Q(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$$

$$A(x) = x$$

$$C(x) = 0$$

$$K(\theta) = -\ln(1-\theta)$$

Note that for $K(\theta)$ we had to put a $-$ because in the definition we have $-K$.

Proposition 3.4.1. Let $X \sim EF(\theta)$ then

1. $\mathbb{E}[A(X)] = \frac{K'(\theta)}{Q'(\theta)}$
2. $Var(A(X)) = \frac{K(\theta)}{(Q'(\theta))^2} - \frac{Q''(\theta)}{(Q'(\theta))^2} \frac{K'(\theta)}{Q'(\theta)}$

Note that this proposition gives us only the expectation and variance for $A(X)$, but it is not a problem because usually $A(X) = X$.

Proof. 1. because the exponential family is regular we can use Leibniz so

$$\begin{aligned} 0 &= \frac{d}{d\theta} 1 \\ &= \frac{d}{d\theta} \int f_X(x; \theta) dx \\ &= \int \frac{d}{d\theta} f_X(x; \theta) dx \\ &= \int (A(x)Q'(\theta) - K'(\theta)) f_X(x; \theta) dx \\ &= Q'(\theta) \int A(x) f_X(x; \theta) dx - K'(\theta) \int f_X(x; \theta) dx \\ &= Q'(\theta) \mathbb{E}[A(X)] - K'(\theta) \\ &\implies \mathbb{E}[A(X)] = \frac{K'(\theta)}{Q'(\theta)} \end{aligned}$$

2. because the exponential family is regular we can use Leibniz so

$$\begin{aligned}
0 &= \frac{d^2}{d\theta^2} \int f_X(x; \theta) \\
&= \int \frac{d^2}{d\theta^2} f_X(x; \theta) \\
&= \int (A(x)Q''(\theta) - K''(\theta))f_X(x; \theta) + (A(x)Q'(\theta) - K'(\theta))^2 f_X(x; \theta) dx \\
&= Q''(\theta)\mathbb{E}[A(X)] - K''(\theta) + (Q'(\theta))^2 \int \left(A(x) - \frac{K'(\theta)}{Q'(\theta)} \right)^2 f_X(x; \theta) dx \\
&= Q''(\theta) \frac{K'(\theta)}{Q'(\theta)} - K''(\theta) + (Q'(\theta))^2 \int (A(x) - \mathbb{E}[A(X)])^2 f_X(x; \theta) dx \\
&= Q''(\theta) \frac{K'(\theta)}{Q'(\theta)} - K''(\theta) + (Q'(\theta))^2 \text{Var}(A(X)) \\
\implies \text{Var}(A(X)) &= \frac{K(\theta)}{(Q'(\theta))^2} - \frac{Q''(\theta)}{(Q'(\theta))^2} \frac{K'(\theta)}{Q'(\theta)} = \frac{K(\theta)}{(Q'(\theta))^2} - \frac{Q''(\theta)}{(Q'(\theta))^2} \mathbb{E}[A(X)]
\end{aligned}$$

□

Observation 1. If $Q(\theta) = \theta$ we get

$$\begin{aligned}
\mathbb{E}[A(X)] &= K'(\theta) \\
\text{Var}(A(X)) &= K''(\theta)
\end{aligned}$$

3.5 Natural Exponential Families

Definition 3.5.1. We say that the distribution of a random sample $(x_1 \dots x_n)$ from $X \sim NEF(\theta)$ is an element of a **Natural Exponential Family** $X \sim NEF(\theta)$ if its PDF can be written as follow:

$$f_X(x; \nu) = \exp\{\nu x + C(x) - K(\nu)\}$$

4 Statistics

The notation of statistic was introduced by fisher (1920).

The importance of sufficiency is that ot can be found in any statistical decision (point estimation, testing, confidential bound)

Definition 4.0.1. Let $X = (X_1 \dots X_n)$ be a random sample from a parametric model $X \sim f_X(x, \theta)$ for some $\theta \in \Theta$ unknown.

We say that $T_n = T(X)$ **sufficient for the parameter** θ if the conditional distribution of X given T_n does not depend of θ i.e.

- $f_{\underline{X}|T_n=t}(X; t, \theta)$ conditional distribution of \underline{X} given T_n
- $h_{\underline{X}, T_n}(z, t, \theta)$ joint distribution of \underline{X} and T_n
- $g_{T_n}(t, \theta)$ marginal distribution of T_n

then T_n is **sufficient for the parameter** θ only if $f_{\underline{X}|T_n=t}(X; t, \theta)$ does not depend of θ .
Note that

$$f_{\underline{X}|T_n=t}(X; t, \theta) = \frac{h_{\underline{X}, T_n}(z, t, \theta)}{g_{T_n}(t, \theta)}$$

Example 4.0.1. $(X_1 \dots X_n) \in \{0, 1\}^n$ from a $Ber(\theta)$, $\theta \in (0, 1)$.
Define $T_n = \sum_{i=1}^n X_i$ then we want to verify if T_n is sufficient.

- $f_{\underline{X}}(\underline{x}; \theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$
- $g_{T_n}(t, \theta) = \binom{n}{t} \sigma^t (1 - \sigma)^{n-t} \mathbb{1}_{0,1 \dots n}(t)$
- $h_{\underline{X}, T_n}(z, t, \theta) \mathbb{P}(X = \underline{x}, T_n = t) = \sigma^t (1 - \sigma)^{n-t}$

so

$$f_{\underline{X}|T_n=t} = \frac{\sigma^t (1 - \sigma)^{n-t}}{\binom{n}{t} \sigma^t (1 - \sigma)^{n-t}} = \frac{1}{\binom{n}{t}}$$

So T_n is a sufficient statistic for θ .

This is a really special case because all the X_i are already in function of T_n .

Observation 2. If T_n is sufficient for θ then all the statistical information of θ is contained in the random sample relocated in T_n . In the example above we just need $\sum_{i=1}^n X_i$.

Observation 3. The notation of sufficiency derive from the probability structure of the parametric family $X \sim f_X(x; \theta)$ We can talk about sufficiency for a parameter θ only after we have specified $X \sim f_X(x; \theta)$

The definition of sufficiency based on conditional probability is not of practical use because we need this two distributions $\begin{cases} g_{T_n}(\cdot) \\ h_{\underline{X}, T_n}(\cdot, \cdot) \end{cases}$ that can be difficult to find. To avoid that there is a corollary of the *Fisher Factorization Theorem*:

Corollary 4.0.1. Let $\underline{X} = (X_1 \dots X_n)$ from $X \sim f_X(x, \theta)$. Then a statistic T_n is sufficient for θ if and only if there exist two non negative functions $g(\cdot), h(\cdot)$ such that $\mathcal{L}_n(\theta, \underline{x}) = g(T(\underline{X}); \theta) h(\underline{X})$

Observation 4. • g is a function of the observed sample via T_n

- h is a function of the observed sample and does not depend on θ

Example 4.0.2. Recall The example 4.0.1 $(X_1 \dots X_n) \in \{0, 1\}^n$ from a $Ber(\theta)$, $\theta \in (0, 1)$. Define $T_n = \sum_{i=1}^n X_i$ then we want to verify if T_n is sufficient.

- $f_{\underline{X}}(\underline{x}; \theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$
- $h(\underline{X}) = 1$
- $g(T_n(\underline{X}), \theta) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$

Example 4.0.3. $(X_1 \dots X_n) \in \{0, 1\}^n$ from a $N(\theta, 1)$. We want to verify that $T_n = \sum_{i=1}^n X_i$ is a sufficient statistic

$$\begin{aligned} \mathcal{L}_n(\theta, \underline{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}(x_i - \theta)^2 \right\} \\ &= (2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2 \right\} \\ &= (2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n x_i^2 - \frac{n\theta^2}{2} + \theta \sum_{i=1}^n x_i \right\} \\ &= (2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n x_i^2 \right\} \exp \left\{ -\frac{n\theta^2}{2} + \theta \sum_{i=1}^n x_i \right\} \end{aligned}$$

so

- $h(\underline{x}) = \exp \left\{ -\frac{1}{2} \sum_{i=1}^n x_i^2 \right\}$
- $g(\sum_{i=1}^n x_i, \theta) = \exp \left\{ -\frac{n\theta^2}{2} + \theta \sum_{i=1}^n x_i \right\}$

Theorem 4.0.1. Fisher Theorem

If $f_{\underline{X}}(\underline{x}; \theta)$ is the joint density function or the joint probability mass function of \underline{X} and $q(t; \theta)$ is the density function or the probability mass function of $T_n(\underline{X})$, then $T_n(\underline{X})$ is sufficient for θ if for every point in the sample space, the ratio

$$\frac{f_{\underline{X}}(\underline{x}; \theta)}{q(t; \theta)}$$

is a constant function of θ

Proof. MISSING □

We can see now the proof of the corollary 4.0.1

Corollary 4.0.2. Savage Let $f_{\underline{X}}(\underline{x}; \theta)$ be the joint PDF or PMF of a random sample $\underline{X} = (X_1 \dots X_n)$. A statistic T_n is sufficient for θ if and only if there exist two non negative functions $g(t, \theta), h(\underline{x})$ such that for all \underline{x} in the sample space and for all $\theta \in \Theta$

$$f_{\underline{X}}(\underline{x}; \theta) = g(T(\underline{x}); \theta) h(\underline{x})$$

Proof. We are going to prove that in the discrete settings.

\Rightarrow Suppose that $T(\underline{X})$ is sufficient for θ .

- $g(t, \theta) = \mathbb{P}(T(\underline{X}) = t)$
- $h(\underline{x}) = \mathbb{P}(\underline{X} = \underline{x} | T(\underline{X}) = T(\underline{x}))$

Because $T(\underline{X})$ is sufficient for θ the conditional probability defining $h(\underline{x})$ does not depend on θ . Hence the choice of $g(t, \theta)$ and $h(\underline{x})$ is legitimate and for this choice we have

$$\begin{aligned}\mathbb{P}(\underline{X} = \underline{x}) &= \mathbb{P}(\underline{X} = \underline{x} \wedge T(\underline{X}) = T(\underline{x})) \\ &= \mathbb{P}(T(\underline{X}) = T(\underline{x}))\mathbb{P}(\underline{X} = \underline{x} | T(\underline{X}) = T(\underline{x})) \\ &= g(t, \theta)h(\underline{x})\end{aligned}$$

So we have the factorization and in particular we can see that

$$\mathbb{P}(T(\underline{X}) = T(\underline{x})) = g(t, \theta)$$

$\implies g(T(\underline{x}), \theta)$ is the PMF of $T(s)$

\Leftarrow We assume that the factorization holds.

Let $q(t, \theta)$ be the PMF of $T(\underline{X})$ we study the ratio

$$\frac{f_{\underline{X}}(\underline{x}; \theta)}{q(T(\underline{x}); \theta)}$$

in particular define

$$A_{T(\underline{x})} = \{\underline{y} | T(\underline{y}) = T(\underline{x})\}$$

Then

$$\begin{aligned}\frac{f_{\underline{X}}(\underline{x}; \theta)}{q(T(\underline{x}); \theta)} &= \frac{g(T(\underline{x}); \theta)h(\underline{x})}{q(T(\underline{x}); \theta)} \\ &= \frac{g(T(\underline{x}); \theta)h(\underline{x})}{\sum_{\underline{y} \in A_{T(\underline{x})}} g(T(\underline{x}); \theta)h(\underline{y})} \\ &= \frac{g(T(\underline{x}); \theta)h(\underline{x})}{g(T(\underline{x}); \theta) \sum_{\underline{y} \in A_{T(\underline{x})}} h(\underline{y})} \\ &= \frac{h(\underline{x})}{\sum_{\underline{y} \in A_{T(\underline{x})}} h(\underline{y})}\end{aligned}$$

This is constant with respect to θ .

Then by the Fisher Theorem $T(\underline{X})$ is sufficient for θ .

□

Example 4.0.4. In with similar condition of the example 4.0.3 we want to find if $T_n = \frac{1}{n} \sum_{i=1}^n X_i$ is a sufficient statistic for μ . $(X_1 \dots X_n) \in \{0, 1\}^n$ from a $N(\mu, \sigma^2)$, σ^2 known.

$$\begin{aligned}f_{\underline{X}}(\underline{x}; \mu\sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x}_n - \bar{x}_n - \mu)^2 \right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu)^2 \right) \right\}\end{aligned}$$

we already have the distribution of $T(\underline{x}) = \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ which is $\bar{X}_n \sim N(\mu, \sigma^2/n)$. so we try o

apply Fisher theorem to the ratio:

$$\frac{(2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu)^2 \right) \right\}}{(2\pi\sigma^2/n)^{-1/2} \exp \left\{ -\frac{n}{2\sigma^2} (\bar{x}_n - \mu)^2 \right\}}$$

So by Fisher Theorem $T_n = \frac{1}{n} \sum_{i=1}^n X_i$ is sufficient for μ

Observation 5. Until now we found only one sufficient statistic for a fixed parametric model. However we can define many sufficient statistic.

For example the statistic given by the identity $T(\underline{x}) = \underline{x}$ is always a sufficient statistic indeed we can factorize the distribution $f_X(x, \theta)$ with

- $h(x) = 1$
- $g(T(x); \theta) = f_X(x, \theta)$

Another way to produce more sufficient statistics given one is thru a one to one function. Suppose $T(\underline{x})$ is a sufficient statistic for θ , and define $T * (\underline{x}) = r(T(\underline{x}))$ where r is a one to one function with inverse r^{-1} .

By Savage's Theorem there exist g, h such that

$$\mathcal{L}_n(\theta, \underline{x}) = g(T(\underline{x}), \theta)h(\underline{x}) = g(r^{-1}T * (\underline{x}), \theta)h(\underline{x})$$

So defining $g * (t\theta) = g(r^{-1}(t), \theta)$ we have that

$$\mathcal{L}_n(\theta, \underline{x}) = g * (T * (\underline{x}), \theta)h(\underline{x})$$

\implies by Savage Theorem we have that $T * (\underline{x})$ is a sufficient statistic.

We saw that in principle we can define many sufficient statistics so it is natural to define a tool that allows us to decide when a sufficient statistic is better than another.

Recall that the purpose of statistic is to achieve data reduction without loss of information.

Therefore a statistic that achieve the most data reduction while still retaining all of the information about θ might be preferable.

Observation 6. We saw in example 4.0.3 that if $(X_1 \dots X_n) \in \{0, 1\}^n$ from a $N(\theta, 1)$, $T_n = \sum_{i=1}^n x_i$ is a sufficient statistic. Instead of $\sum_{i=1}^n X_i$ we can use $T'(\underline{x}) \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2 \right)$. Clearly $T(X)$ is a greater data reduction than $T'(\underline{x})$ since we do not need to know the sample variance if we want to know $T(\underline{x})$. Moreover we can write $T(\underline{x})$ as a function of $T'(\underline{x})$ by defining the function $r(a, b) = a$, then we can write

$$T(\underline{x}) = \bar{x}_n = r(\bar{x}_n, S_n^2) = r(T'(\underline{x}))$$

Since $T(\underline{x})$ and $T'(\underline{x})$ are both sufficient they contains the same information about μ . In other terms the additional information given by the sample variance is null.

Definition 4.0.2. A sufficient statistic $T(\underline{x})$ is called minimal if for any other sufficient statistic $T'(\underline{x})$, $T(\underline{x})$ is a function of $T'(\underline{x})$.

To say that $T(\underline{x})$ is a function of $T'(\underline{x})$ simply means that if $T'(x) = T'(y)$ then $T(x) = T(y)$. In other terms if $\{B_t\}$ where $B_t := \{t' : T'(t) = T'(t')\}$ is the partition set induced by T' and $\{A_t\}$ where $A_t := \{t' : T(t) = T(t')\}$ is the partition set induced by T then for every t , $B_t \subseteq A_t$. \implies the partition of the sample space induced by a minimal statistic is the partition with the smallest cardinality.

Theorem 4.0.2. Lehmann and Sheffe let $f_{\underline{X}}(\underline{x}; \theta)$ be the joint density function or joint probability mass function of a random sample $\underline{X} = (X_1 \dots X_n)$. Suppose there exist a function T such that for any two sample points $\underline{x}, \underline{y}$ the ratio

$$\frac{f_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{y}; \theta)}$$

is constant as a function of θ if and only if $T(\underline{x}) = T(\underline{y})$

Then T is a minimal sufficient statistic for θ

Proof. To simplify the proof we assume $f_{\underline{X}}(\underline{x}; \theta) > 0 \forall \underline{x}, \forall \theta$.

First we show that $T(\underline{x})$ is sufficient.

Define \mathcal{T} as the image of the sample space under the function $T(\underline{x})$.

$$\mathcal{T} := \{t : t = T(\underline{x}) \text{ for some } \underline{x} \text{ in the sample space}\}$$

Define $\{A_t\}$ where $A_t := \{\underline{x} : T(\underline{x}) = t\}$, the partition set induced by T . For each A_t choose and fix some elements $\underline{x}_t \in A_t$. For any point in the space $\underline{x}_{T(\underline{x})}$ is the fixed element that is in the same set A_t , as \underline{x} . Since \underline{x} and $\underline{x}_{T(\underline{x})}$ are in the same set A_t then $T(\underline{x}) = T(\underline{x}_{T(\underline{x})})$ so by the assumptions the ratio

$$\frac{f_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{x}_{T(\underline{x})}; \theta)}$$

Does not depend on θ . Thus we can define $h(\underline{x}) := \frac{f_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{x}_{T(\underline{x})}; \theta)}$.

Then define the function $g(\underline{x}, \theta) f_{\underline{X}}(\underline{x}; \theta)$, so we have:

$$f_{\underline{X}}(\underline{x}; \theta) = \frac{f_{\underline{X}}(\underline{x}_{T(\underline{x})}; \theta) f_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{x}_{T(\underline{x})}; \theta)} = g(T(\underline{x}), \theta) h(\theta)$$

and by Savage Theorem $T(\underline{x})$ is sufficient for θ .

Now we show that $T(\underline{x})$ is minimal sufficient.

Let $T'(\underline{x})$ be another sufficient statistic. By Savage Theorem we know that exist h', g' such that

$$f_{\underline{X}}(\underline{x}; \theta) = g'(T'(\underline{x}), \theta) h'(\theta)$$

Let $\underline{x}, \underline{y}$ be two sample points such that $T'(\underline{x}) = T'(\underline{y})$ then we can study the ratio:

$$\frac{f_{\underline{X}}(\underline{x}; \theta)}{f_{\underline{X}}(\underline{y}; \theta)} = \frac{g'(T'(\underline{x}); \theta) h'(\theta)}{g'(T'(\underline{y}); \theta) h'(\theta)} = \frac{h'(\underline{x})}{h'(\underline{y})}$$

Since the ratio does not depend on θ , by the assumption (the other implication of the IIF) implies $T(\underline{x}) = T(\underline{y})$. So we can say that $T(\underline{x})$ is a function of $T'(\underline{x})$ therefore $T(\underline{x})$ is minimal. \square

4.1 Estimators

Definition 4.1.1. Suppose there is a fixed parameter θ that needs to be estimated. Then an **estimator** is a function that maps the sample space to a set of sample estimates. An estimator of θ is usually denoted by the symbol $\bar{\theta}$.

Now we are going to introduce some definition of "good" estimators.

Definition 4.1.2. $T_n(\underline{X})$ is said to be **unbiased** for θ if $\mathbb{E}[T_n(\underline{X})] = \theta$

Definition 4.1.3. Bias

$$Bias_{\theta}(T_n(\underline{X})) = \mathbb{E}[T_n(\underline{X}) - \theta]$$

Observation 7. In the definition we used the expected value because it is one of the linearity of the operator.

When we ask an estimator to be unbiased basically we are requiring it to be centred around θ . Another parameter that give us information about the goodness of an estimator is the variance. We can interpret the variance as a measure of the dispersion around the expected value, so before check the variance we must be sure that the expected value overlap with θ . The less is the variance the best is the estimator.

Observation 8. Variance is a good parameter to watch only if the estimator is unbiased

To avoid this problem we can introduce the *Mean Squared Error*

Definition 4.1.4. Mean Squared Error(MSE)

$$\mathbb{E}[(T_n(\underline{X}) - \theta)^2]$$

The importance of this quantity comes from the *Chebyshev's Inequality* 1.1.1

$$\mathbb{P}(|T_n(\underline{x}) - \theta| < k) > 1 - \frac{\mathbb{E}[(T_n(\underline{X}) - \theta)^2]}{k^2}$$

Indeed we notice that the smaller the MSE the greater is $\mathbb{P}(|T_n(\underline{x}) - \theta| < k)$.

Proposition 4.1.1.

$$\mathbb{E}[(T_n(\underline{X}) - \theta)^2] = \text{Var}(T_n(\underline{X})) + \text{Bias}_\theta(T_n(\underline{X}))$$

Proof.

$$\begin{aligned} \mathbb{E}[(T_n(\underline{X}) - \theta)^2] &= \mathbb{E}[(T_n(\underline{X}) - \mathbb{E}[T_n(\underline{X})] + \mathbb{E}[T_n(\underline{X})] - \theta)^2] \\ &= \mathbb{E}[(T_n(\underline{X}) - \mathbb{E}[T_n(\underline{X})])^2] + \mathbb{E}[(\mathbb{E}[T_n(\underline{X})] - \theta)^2] + 2\mathbb{E}[(T_n(\underline{X}) - \mathbb{E}[T_n(\underline{X})])(\mathbb{E}[T_n(\underline{X})] - \theta)] \\ &= \mathbb{E}[(T_n(\underline{X}) - \mathbb{E}[T_n(\underline{X})])^2] + \mathbb{E}[(\mathbb{E}[T_n(\underline{X})] - \theta)^2] \\ &= \text{Var}(T_n(\underline{X})) + \text{Bias}_\theta(T_n(\underline{X})) \end{aligned}$$

□

Observation 9. If $\mathbb{E}[T_n(\underline{X})] = \theta$ then $MSE(T_n(\underline{X})) = \text{Var}(T_n(\underline{X}))$.

Definition 4.1.5. Let $X_1..X_n$ from $X \sim f_X(x, \theta)$, T'_n and T''_n estimators for θ . We say T'_n is **more efficient** than T''_n if

$$MSE(T'_n) < MSE(T''_n)$$

Usually we choose the estimator with the lower MSE even if it is not unbiased.

4.2 Properties of Estimators

The problem of the MSE is that we can not be sure that there exist T_n such that $MSE(T_n)$ is the lowest possible.

A solution for this comes from

Theorem 4.2.1. Cramer-Rao Bound

Let $X = (X_1..X_n)$ be a random sample from a parametric model $X \sim f_X(x, \theta)$.

Then, under condition of regularity, for any estimator T_n of θ

$$\text{Var}(T_n) \geq \frac{[1 + b'(T_n)]^2}{\mathcal{I}_n(\theta)}$$

Where

$b(T_n)$ is the bias of T_n

$\mathcal{I}_n(\theta)$ is the Fisher Information

Proof. consider the estimator T_n .

- $\mathbb{E}[T_n] = \theta + b(\theta)$
- $\frac{d}{d\theta}\mathbb{E}[T_n] = 1 + b'(\theta)$
- $\mathbb{E}[V'_n(\theta)] = 0 \leftarrow$ because we suppose our model regular

$$\implies \text{Cov}(T_n V'_n(\theta)) = \mathbb{E}[T_n V'_n(\theta)] - \mathbb{E}[T_n]\mathbb{E}[V'_n] = \mathbb{E}[T_n V'_n(\theta)]$$

$$\begin{aligned} \mathbb{E}[T_n V'_n(\theta)] &= \int_{\mathbb{R}^n} T_n V'_n(\theta) f_{\underline{X}}(\underline{x}; \theta) d\underline{x} \\ &= \int_{\mathbb{R}^n} T_n \frac{f'_{\underline{X}}(\underline{x}, \theta)}{f_{\underline{X}}(\underline{x}; \theta)} f_{\underline{X}}(\underline{x}; \theta) d\underline{x} \\ &= \int_{\mathbb{R}^n} T_n \frac{d}{d\theta} f_{\underline{X}}(\underline{x}; \theta) d\underline{x} \\ &= \frac{d}{d\theta} \int_{\mathbb{R}^n} T_n f_{\underline{X}}(\underline{x}; \theta) d\underline{x} \\ &= \frac{d}{d\theta} \mathbb{E}[T_n] \\ &= 1 + b'(\theta) \end{aligned}$$

So

$$\text{Cov}(T_n, V'_n(\theta)) = 1 + b'(\theta)$$

We know that in general for X, Y random variables such that $\mathbb{E}[X] = \mu, \mathbb{E}[Y] = \nu, \mathbb{E}[X^2] < \infty, \mathbb{E}[Y^2] < \infty$ it holds

$$\begin{aligned} (\text{Cov}(X, Y))^2 &= \left(\mathbb{E}[(X - \mu)(Y - \nu)] \right)^2 \\ &\leq \mathbb{E}[(X - \mu)^2] \mathbb{E}[(Y - \nu)^2] \\ &= \text{Var}(X) \text{Var}(Y) \end{aligned}$$

So replacing X with T_n and Y with $V'_n(\theta)$ we obtain

$$\begin{aligned} \text{Var}(T_n) &\geq \frac{(\text{Cov}(T_n, V'_n(\theta)))^2}{\text{Var}(V'_n(\theta))} \\ &= \frac{(\text{Cov}(T_n, V'_n(\theta)))^2}{\mathbb{E}[V'_n(\theta)]^2 - \mathbb{E}[V'_n(\theta)]^2} \\ &= \frac{(\text{Cov}(T_n, V'_n(\theta)))^2}{\mathbb{E}[V'_n(\theta)]^2} \\ &= \frac{(\text{Cov}(T_n, V'_n(\theta)))^2}{\mathcal{I}_n(\theta)} \end{aligned}$$

□

Corollary 4.2.1. *Under condition of regularity*

$$MSE \geq \frac{(1 + b'(\theta))^2}{\mathcal{I}_n(\theta)} + b^2(\theta)$$

Proof. Directly from Cramer-Rao Bound 4.2.1 remembering that $MSE(T_n) = \text{Var}(T_n) + b^2(\theta)$ □

Corollary 4.2.2. *Let $\underline{X} = (X_1 \dots X_n)$ be a random sample from a regular model $X \sim f_X(x; \theta)$ If there exist a unbiased estimator for θ whose variance is equal to the Cramer-Rao bound, Then T_n is unique*

Proof. Take T_{1n}, T_{2n} be unbiased estimators for θ such that

$$Var(T_{1n}) = Var(T_{2n}) = \frac{1}{\mathcal{I}_n(\theta)}$$

Define $T_n := \frac{T_{1n} + T_{2n}}{2}$
 $\mathbb{E}[T_n] = \frac{\mathbb{E}[T_{1n}] + \mathbb{E}[T_{2n}]}{2} = \frac{2}{2}\theta = \theta$
 $\implies T_n$ is also unbiased for θ
 $\implies Var(T_n) \geq \frac{1}{\mathcal{I}_n(\theta)}$

$$\begin{aligned} Var(T_n) &= Var\left(\frac{T_{1n} + T_{2n}}{2}\right) \\ &= \frac{1}{4} \left[Var(T_{1n}) + Var(T_{2n}) + 2Cov(T_{1n}, T_{2n}) \right] \\ &= \frac{1}{4} \left[Var(T_{1n}) + Var(T_{2n}) + 2Cov(T_{1n}, T_{2n}) \right] \frac{[Var(T_{1n})Var(T_{2n})]^{1/2}}{[Var(T_{1n})Var(T_{2n})]^{1/2}} \\ &= \frac{(1 + Corr(T_{1n}, T_{2n}))}{2} \frac{1}{\mathcal{I}_n(\theta)} \end{aligned}$$

MEMO: $|Corr(X, Y)| \leq 1$

Because $Var(T_n) \leq \frac{1}{\mathcal{I}_n(\theta)}$ then we must have $Corr(T_{1n}, T_{2n}) \geq 1 \implies Corr(T_{1n}, T_{2n}) = 1$.

$\implies T_{2n} = a + bT_{1n}$.

Hence we must have $\theta = \mathbb{E}[T_{1n}] = \mathbb{E}[a + bT_{1n}] = \mathbb{E}[a] + b\mathbb{E}[T_{1n}] = a + b\theta \implies a = 0, b = 1$

$\implies T_{2n} = T_{1n}$ □

Definition 4.2.1. Consider a regular model $X \sim f_x(x\theta)$. We say that an unbiased estimator T_n whose variance is

$$Var(T_n) = \frac{1}{\mathcal{I}_n(\theta)}$$

is **efficient** moreover we define the **efficiency** of an estimator as:

$$Eff(T_n) = \frac{1}{Var(T_n)\mathcal{I}_n(\theta)} \in [0, 1]$$

Theorem 4.2.2. Rao-Blackwell

Let $\underline{X} = (X_1 \dots X_n)$ a random sample from a parametric model $X \sim f_x(x\theta)$.

5 Exercises

Exercise 1. Let X_1 and X_2 two random variables independent and uniformly distributed on the interval $[0, 1]$. Find the distribution of:

1. $Y = X_1 + X_2$

2. $W = \frac{X_1}{X_2}$

3. $Z = X_1 X_2$

Solution:

1. Let's start with the sum of two generic random variables:

we know that $f_{Y|X_1}(y|x_1) = f_{X_2}(y - x_1)$ and the joint distribution on two random variables is:

$$f_{X_1, Y}(x_1, y) = f_{Y|X_1}(y|x_1)f_{X_1}(x_1)$$

so in our case:

$$f_{X_1, Y}(x_1, y) = f_{X_2}(y - x_1)f_{X_1}(x_1)$$

and now we can calculate the PDF of Y simply by integrating the PDF of the joint distribution:

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X_1, Y}(x_1, y) dx_1 \\ &= \int_{-\infty}^{\infty} f_{X_2}(y - x_1) f_{X_1}(x_1) dx_1 \end{aligned}$$

(For a more detailed analysis see the convolution product).

Now we can proceed replacing the generic PDF with the one of a uniform distribution on the interval $[0, 1]$ is $\mathbb{1}_{[0,1]}(t)$, we have:

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} \mathbb{1}_{[0,1]}(y - x_1) \mathbb{1}_{[0,1]}(x_1) dx_1 \\ &= \int_0^1 \mathbb{1}_{[0,1]}(y - x_1) dx_1 \end{aligned}$$

and by separating the integral in various cases we can solve it obtaining:

$$f_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ y & \text{if } y \in [0, 1] \\ 2 - y & \text{if } y \in (1, 2] \\ 0 & \text{if } y > 2 \end{cases}$$

2. for the distribution of W we will use the CDF function:

first of all it is easy to prove that for $w \leq 0$, $F_W(w) = 0$, so in the next passage we can assume

$$w > 0$$

$$\begin{aligned}
F_W(w) &= \mathbb{P}[W < w] \\
&= \mathbb{P}\left[\frac{X_2}{X_1} < w\right] \\
&= \mathbb{P}\left[\frac{X_2}{X_1} < w, X_1 > 0\right] + \mathbb{P}\left[\frac{X_2}{X_1} < w, X_1 < 0\right] \\
&= \mathbb{P}[X_2 < X_1 w, X_1 > 0] + 0 \\
&= \int_0^\infty f_{x_1}(x_1) \int_{-\infty}^{wx_1} f_{x_2}(x_2) dx_2 dx_1 \\
&= \int_0^\infty \mathbb{1}_{[0,1]}(x_1) \int_{-\infty}^{wx_1} \mathbb{1}_{[0,1]}(x_2) dx_2 dx_1 \\
&= \int_0^1 \begin{cases} 1 & \text{if } wx_1 > 1 \\ wx_1 & \text{if } wx_1 < 1 \end{cases} dx_1 \\
&= \begin{cases} \int_0^1 wx_1 dx & \text{if } w \leq 1 \\ \int_0^{\frac{1}{w}} wx_1 dx + \int_{\frac{1}{w}}^1 1 dx_1 & \text{if } w > 1 \end{cases} \\
&= \begin{cases} \frac{w}{2} & \text{if } 0 \leq w \leq 1 \\ 1 - \frac{1}{2w} & \text{if } w > 1 \end{cases}
\end{aligned}$$

3. for the distribution of Z we will adopt a more straight forward approach and we will use the CDF function:

$$\begin{aligned}
F_Z(z) &= \mathbb{P}[Z < z] \\
&= \mathbb{P}[X_1 X_2 < z] \\
&= \mathbb{P}[X_1 X_2 < z, X_1 > 0] + \mathbb{P}[X_1 X_2 < z, X_1 < 0] \\
&= \mathbb{P}[X_2 < \frac{z}{X_1}, X_1 > 0] + 0 \\
&= \int_0^\infty f_{X_1}(x_1) \int_{-\infty}^{z/x_1} f_{X_2}(x_2) dx_2 dx_1 \\
&= \int_0^\infty \mathbb{1}_{[0,1]}(x_1) \int_{-\infty}^{z/x_1} \mathbb{1}_{[0,1]}(x_2) dx_2 dx_1 \\
&= \int_0^\infty \mathbb{1}_{[0,1]}(x_1) \begin{cases} 1 & \text{if } \frac{z}{x_1} > 1 \\ \frac{z}{x_1} & \text{if } \frac{z}{x_1} < 1 \end{cases} dx_1 \\
&= \int_0^1 \begin{cases} 1 & \text{if } \frac{z}{x_1} > 1 \\ \frac{z}{x_1} & \text{if } \frac{z}{x_1} < 1 \end{cases} dx_1 \\
&= \begin{cases} \int_0^z 1 dx_1 + \int_z^1 \frac{z}{x_1} dx_1 & \text{if } z < 1 \\ \int_0^1 1 dx_1 & \text{if } z \geq 1 \end{cases} \\
&= \begin{cases} z - z \ln(z) & \text{if } 0 < z < 1 \\ 1 & \text{if } z \geq 1 \end{cases} .
\end{aligned}$$

Exercise 2. Consider X_1, X_2, X_3 three random variables independent and identically distributed with distribution $\sim \text{Exp}(\frac{1}{2})$.

Find the distribution of:

$$1. U = \frac{X_2}{X_1}$$

$$2. W = \sum_{i=1}^3 X_i$$

(in this exercise we will consider the PDF of the exponential $f_X = \lambda e^{-\lambda x}$).

Solution:

1. For the distribution of U we will use the CDF function:

first of all it is easy to prove that for $u \leq 0$, $F_U(u) = 0$, so in the next passage we can assume $u > 0$

$$\begin{aligned}
 F(u) &= \mathbb{P}[U < u] \\
 &= \mathbb{P}\left[\frac{X_2}{X_1} < u\right] \\
 &= \mathbb{P}\left[\frac{X_2}{X_1} < u, X_1 > 0\right] + \mathbb{P}\left[\frac{X_2}{X_1} < u, X_1 < 0\right] \\
 &= \mathbb{P}[X_2 < X_1 u, X_1 > 0] + 0 \\
 &= \int_0^\infty f_{x_1}(x_1) \int_{-\infty}^{ux_1} f_{x_2}(x_2) dx_2 dx_1 \\
 &= \int_0^\infty \lambda e^{-\lambda x_1} \int_0^{ux_1} \lambda e^{-\lambda x_2} dx_2 dx_1 \\
 &= \int_0^\infty \lambda e^{-\lambda x_1} (1 - e^{-\lambda u x_1}) x_1 \\
 &= 1 - \frac{1}{1+u}
 \end{aligned}$$

2. Here we will use the moment-generating function to demonstrate that $W \sim \text{Gamma}(\alpha = 3, \beta = \frac{1}{2})$.

The the moment-generating function of the exponential with parameter 2 is: $M_X(t) = \frac{1}{\frac{1}{2}-t}$.

$$\begin{aligned}
 M_W(t) &= \mathbb{E}[e^{wt}] \\
 &= \mathbb{E}[e^{\sum_{i=1}^3 X_i t}] \\
 &= \mathbb{E}\left[\prod_{i=1}^3 e^{X_i t}\right] \\
 &= \prod_{i=1}^3 \mathbb{E}[e^{X_i t}] \\
 &= \prod_{i=1}^3 \frac{1}{\frac{1}{2}-t} \\
 &= \left(\frac{1}{\frac{1}{2}-t}\right)^3 \\
 &= \left(\frac{\frac{1}{2}-t}{\frac{1}{2}}\right)^{-3} \\
 &= \left(1 - \frac{t}{\frac{1}{2}}\right)^{-3}
 \end{aligned}$$

which is the MGF of a $\text{Gamma}(3, \frac{1}{2})$

Exercise 3. Let (X, Y) be a bivariate random variablesuch that $X \sim U(-1, 1)$ and $Y|X \sim U(x, x+1)$. Find he distribution of $Z = -\ln(Y - X)$

Solution:

$$\begin{aligned}
\mathbb{P}(Z < z) &= \mathbb{P}(-\log(X - Y) < t) \\
&= \mathbb{P}(\log(X - Y) \geq z) \\
&= \mathbb{P}(Y \geq X + e^{-z}) \\
&= \int_{-1}^1 \int_{x+e^{-z}}^{x+1} \frac{1}{2} dx dy \\
&= 1 - e^{-z}
\end{aligned}$$

Which is the distribution function of neg exp

Exercise 4. Let A, B be two independent and identically distributed random variables with distribution $\sim U(0, h)$. Compute the probability that the equation $Z^2 - 2AZ + B = 0$ doesn't admit real solutions.

Solution: We are asked to compute $\mathbb{P}(A^2 - B < 0)$.

Exercise 5. Let $X \sim \text{Gamma}(r, 1), Y \sim \text{Gamma}(s, 1)$ independent random variables. Find the distribution of

1. $W := X + Y$
2. $Z := \frac{X}{W}$
3. (Z, w)

Solution: Did in class.

Exercise 6. Let X_1, X_2, X_3 random variables with distribution:

- $X_1 \sim \text{Gamma}(\alpha_1, 1)$
- $X_2 \sim \text{Gamma}(\alpha_2, 1)$
- $X_3 \sim \text{Gamma}(\alpha_3, 1)$

Define:

- $Z = \frac{X_1}{X_1 + X_2 + X_3}$
- $W = \frac{X_2}{X_1 + X_2 + X_3}$

Find the distribution of (Z, W)

For this exercise we will use the theorem on page 165 of [1]

Solution: To have a $(3, 3)$ parametrization we will add another random variable $S := X_1 + X_2 + X_3$.

Consider the parametrization

- $X_1 = ZS$
- $X_2 = WS$
- $X_3 = S - S(Z + W)$

The Jacobian matrix is defined as:

$$J := \begin{bmatrix} \frac{\partial \phi_1}{\partial x_1} & \cdots & \frac{\partial \phi_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi_N}{\partial x_1} & \cdots & \frac{\partial \phi_N}{\partial x_n} \end{bmatrix}$$

in our situation then:

$$|J| = \left| \begin{bmatrix} S & 0 & Z \\ 0 & S & W \\ -S & -S & 1 - Z - W \end{bmatrix} \right| = S^2$$

So by the previous theorem we have:

$$f_{Z,W,S}(z, w, s) = f_{X_1, X_2, X_3}(zs, ws, s - s(z + w))|J|$$

Remembering that X_1, X_2, X_3 are independent then $f_{X_1, X_2, X_3}(zs, ws, s - s(z + w)) = f_{X_1}(zs)f_{X_2}(ws)f_{X_3}(s - s(z + w))$ so

$$\begin{aligned} f_{Z,W,S}(z, w, s) &= f_{X_1}(zs)f_{X_2}(ws)f_{X_3}(s - s(z + w))s^2 \\ &= \frac{(zs)^{\alpha_1-1}(ws)^{\alpha_2-1}(s - s(z + w))^{\alpha_3-1}e^{-s}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}s^2\mathbb{1}_{[\min(zs, ws, s-s(z+w)), \infty)}(0) \\ &= \frac{z^{\alpha_1-1}w^{\alpha_2-1}(1 - (z + w))^{\alpha_3-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}s^{\alpha_1+\alpha_2+\alpha_3-1}e^{-s} \end{aligned}$$

Notice that the second member is the kernel of a Gamma distribution with parameters $\alpha_1 + \alpha_2 + \alpha_3, 1$.

We know that $s \geq 0$ so $\min(zs, ws, s - s(z + w))$ has the same sign of $\min(z, w, 1 - (z + w))$.

To get the distribution of Z, W we have to integrate $f_{Z,W,S}(z, w, s)$ with respect to s :

$$\begin{aligned} f_{Z,W}(z, w) &= \int_0^\infty f_{Z,W,S}(z, w, s)\mathbb{1}_{[\min(z, w, 1-(z+w)), \infty)}(0)ds \\ &= \mathbb{1}_{[\min(z, w, 1-(z+w)), \infty)}(0) \int_0^\infty \frac{z^{\alpha_1-1}w^{\alpha_2-1}(1 - (z + w))^{\alpha_3-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}s^{\alpha_1+\alpha_2+\alpha_3-1}e^{-s}ds \\ &= \mathbb{1}_{[\min(z, w, 1-(z+w)), \infty)}(0) \frac{z^{\alpha_1-1}w^{\alpha_2-1}(1 - (z + w))^{\alpha_3-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \int_0^\infty s^{\alpha_1+\alpha_2+\alpha_3-1}e^{-s}ds \\ &= \frac{z^{\alpha_1-1}w^{\alpha_2-1}(1 - (z + w))^{\alpha_3-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}\Gamma(\alpha_1 + \alpha_2 + \alpha_3)\mathbb{1}_{[\min(z, w, 1-(z+w)), \infty)}(0) \end{aligned}$$

Exercise 7. Show that the moment generating function of a random variable $X \sim NEF(\nu)$ is

$$\mathbb{E}[e^{sx}] = e^{K(s+\nu)-K\nu}$$

Solution:

$$\begin{aligned} e[e^{sx}] &= \int e^{sx}e^{x\nu+C(x)-K(\nu)}dx = \int e^{x(s+\nu)+C(x)+K(s+\nu)-K(s+\nu)-K(\nu)}dx \\ &= e^{K(s+\nu)-K(\nu)} \int e^{x(s+\nu)+C(x)-K(s+\nu)}dx \\ &= e^{K(s+\nu)-K(\nu)}1 \\ &= e^{K(s+\nu)-K(\nu)} \end{aligned}$$

Bibliography

- [1] Casella and Berger. *Statistical Inerence*. Duxbury press, second edition edition, 2008.