Alexandre St-Aubin

Honours Statistics Notes

FEBRUARY 29, 2024

Taught by Masoud Asgharian

Contents

1 Review of Probability Concepts 3

1.1 Multiple Random Variables 8

1.2 Limit Theorems 10

2 Sample Moments and Their Distributions 11

2.1 Random Sampling 11

2.2 Sample Characteristics and Their Distributions 12

3 Theory of point estimation 14

3.1 Consistency and Bias 14

3.2 Invariance 16

3.3 Sufficient Statistics 16

3.4 Maximum Likelihood 18

3.5 Completeness 19

3.6 UMVUE 19

3.7 Lower Bound for the Variance of an Unbiased Estimate 20

3.8 Review of Estimator Properties 20

1 *Review of Probability Concepts*

Definition 1.1 (Probability Space). The triple (Ω, S, P) is called a **Probability Space**, with (Ω, S) denoting the **Sample Space**, and

(i) Ω is the set of all possible outcomes of the experiment. Any element in Ω is called a **Sample Point**.

(ii) S is a σ-field (same as a $\sigma-$ algebra) of subsets of Ω. Any set $A \in S$ is an **Event**.

(iii) P is a **Probability Measure** if the following hold:

(a) $P(A) \geq 0 \; \forall A \in S$.

(b) $P(\Omega) = 1$.

(c) If $\{A_i\}_I \subset S$ are disjoint, then $P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$

Definition 1.2 (Random Variable). Let (Ω, S) be a *sample space*, then a finite, single-valued function that maps Ω onto $\mathbb{R}$ is called a **Random Variable** (RV) if the inverse images under X of all *Borel* sets in $\mathbb{R}$ are events, that is, if

A RV is called *Discrete* if it takes on countably many values.

$$X^{-1}(B) = \{\omega \mid X(\omega) \in B\} \in S \quad \forall B \in \mathfrak{B}$$

Definition 1.3 (Discrete Probability Measure). Let $S \in \Omega$ be countable, $p_s \in [0, \infty)$, $s \in S$ a collection of numbers such that $\sum_{s \in S} p_s = 1.$, then $P : P(\Omega) \to [0,1]$ such that

$$P(A) = \sum_{s \in S \cap A} p_s = \sum_{s \in S} p_s \chi_A(s)$$

is a probability measure on $(\Omega, P(\Omega))$. It is called a **discrete probability measure** with support S. The function $f_X : \Omega \to \mathbb{R}$ defined by

A discrete probability measure is often called a discrete distribution.

$$f_X(\omega) = \begin{cases} 0, & \text{if } \omega \notin S \\ p_s, & \text{otherwise.} \end{cases}$$

is called the PMF (density of P with respect to the counting measure).

Definition 1.4 (Probability Distribution Function). A probability distribution function F is any function $F : \mathbb{R} \to \mathbb{R}$ such that

A probability measure is called continuous if its distribution function is continuous. In general, F has at most countably many discontinuities

(i) F is non-decreasing.

(ii) F is right-continuous.

(iii) $\lim_{n \to -\infty} F(n) = 0$ and $\lim_{n \to \infty} F(n) = 1$.

Definition 1.5 (Probability Density Function). A function $F_X : \mathbb{R} \to \mathbb{R}$ is called a $\mathcal{R}$ **Probability Density Function** $((\mathcal{R})\ PDF)$ if:

(i) $F_X(x) \geq 0 \; \forall x \in \mathbb{R}$.

(ii) F_X is $\mathcal{R}$-integrable on $\mathbb{R}$ and $\int_{-\infty}^{\infty} f(t)\,dt = 1$.

Definition 1.6 (Distribution of a RV). Let $(\Omega, \mathcal{A}, P)$ a *probability space* and $(\mathcal{X}, \mathcal{B})$ a *measurable space.* Also, let $X : \Omega \to \mathcal{X}$ be $\mathcal{A} - \mathcal{B}$ measurable. Then,

$$P^X : \mathcal{B} \to \mathbb{R}$$
$$P^X(B) = P(X^{-1}(B)) = P(\{\omega \mid X(\omega) \in B\})$$

is a probability measure on $(\mathcal{X}, \mathcal{B})$. It is called the **image measure** of P or **distribution** of X.

A real-valued random variable is continuous if its distribution P^X has a density
For a real-valued RV X : if X is discrete, then P^X is uniquely determined by its PMF f_X. If X is continuous, P^X has density f_X.
$P^X(\mathcal{X}) = P(\{\omega \mid X(\omega) \in \mathcal{X}\}) = P(\Omega) = 1$

Remark 1.7. If X is a discrete real-valued random variable and $g : \mathbb{R} \to \mathbb{R}$ measurable, then $Y = g(x)$ is also discrete and its PMF can be computed as follows if g is 1-1:

$$f_Y(y) = P(Y = y) = P(g(X) = y) = \sum_{x:g(x)=y} f_X(x) = P(X = g^{-1}(y))$$

Definition 1.8 (Expected Value of Discrete RV). Let X be a discrete real-valued RV with PMF f_X, the **Expected Value** of X is given by

$$E[X] = \sum_{x \in \{X(\omega):\omega \in \Omega\}} x f_X(x),$$

provided

$$\sum_{x \in \{X(\omega):\omega \in \Omega\}} |x| f_X(x) < \infty,$$

otherwise it is not defined.

Definition 1.9 (Expected Value of RV with Density). Let X be a RV with $\mathcal{R}$ density f_X (PDF). Then the **Expected Value** of X is given as

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

provided

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx < \infty,$$

otherwise it is not defined.

Lemma 1.10 (Expectation of Symmetric Distributions). Suppose that X has $\mathcal{R}$ density f_X such that

(i) $\int_{-\infty}^{\infty} |x| f_X(x)\,dx < \infty$,

(ii) $\exists a \in \mathbb{R}$ such that $\forall x \in \mathbb{R}$, $f_X(x + a) = f_X(a - x)$,

then

$$E[X] = a$$

Theorem 1.11 (Expectations of Functions of RVs). Let X be a RV and $g : \mathbb{R} \to \mathbb{R}$ borel-measurable, then if $Y = g(X)$,

(i) If X is *Discrete*, and $E[Y]$ exists, then

$$E[Y] = E[g(x)] = \sum g(x) f_X(x)$$

(ii) If X has $\mathcal{R}$ *Density* f_X and g is strictly monotone and continuously differentiable on $(a, b) = \{x : f_X(x) > 0\}$, and $E[g(X)]$ exists, then

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

Lemma 1.12 (Properties of Expectations).

(i) $E[g_1(X) + ... + g_n(X)] = E[g_1(X)] + ... + E[g_n(X)]$, provided each $E[g_i(X)]$ exists.

(ii) $E[aX] = aE[X],\ a \in \mathbb{R}$.

(iii) $E[a] = a,\ a \in \mathbb{R}$.

(iv) if X, Y are independent, then $E[XY] = E[X]E[Y]$, and in particular, for any measurable functions g, f, $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$

(v) $E[X] = E[E[X|Y]]$.

The expected value of the sum of random variables is equal to the sum of their individual expected values, regardless of whether they are independent.

Theorem 1.13 (Properties of Variance). Let X be a RV with $E[|X|] < \infty$, then[1]

(i) $\text{Var}(X)$ exists $\iff E[X^2] < \infty$.

(ii) $\text{Var}(X) = E[X^2] - (E[X])^2$.

(iii) $\text{Var}(X) = 0 \iff P(X = c) = 1$ for some $c \in \mathbb{R}$.

(iv) $\text{Var}(aX + b) = a^2 \text{Var}(X)$.

(v) $\min_{c \in \mathbb{R}} E[(X - c)^2] = \text{Var}(X) = E[(X - E[X])^2]$.

(vi) If X, Y are independent random variables, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

However, if X, Y are not independent, then

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

(vii) $\text{Var}(X) = \text{Var}(E[X|Y]) + E[\text{Var}(X|Y)] \geq \text{Var}(E[X|Y])$

Definition 1.14 (Moments of Distribution). Let X be a RV, $n \in \mathbb{N},\ \alpha \in \mathbb{R}$. If they exist, the following expectations have special names.

[1] Matthew Bernstein. 2017. URL `https://mbernste.github.io/files/notes/Variance.pdf`

(i) $E[X^n]$ is the n^{th} *raw* moment of X. A raw moment is a moment about the origin.

(ii) The n^{th} *central* moment of X is about the distribution's mean μ_x and is given by

$$m_n = E[(X-\mu_x)^n] = \int_{-\infty}^{\infty} (x-\mu_x)^n f_X(x)\ dx$$

Central moments are useful because they allow us to quantify properties of distributions in ways that are **location-invariant**. E.g. we may be interested in comparing the variability in height of adults versus children. We want to measure which group has greater variability while disregarding the heights of people in each group.

(iii) The n^{th} *standardized* moment is defined as the n^{th} central moment normalized by the standard deviation raised to the n^{th} power,

$$\bar{m}_n = \frac{m_n}{\sigma_n} = E\left[\left(\frac{X-\mu_x}{\sigma_x}\right)^n\right],$$

where m_n is defined as in (ii), and

$$\sigma_n = \sigma_x^n = \left(\sqrt{E[(X-\mu_x)^2]}\right)^n$$

(iv) A *sample* moment is an unbiased estimator of its respective raw, central, or standardized moment.

Recall that, given a statistical model, parameters summarize data for an entire population, while statistics summarize data from a sample of the population. We compute the former exactly using a statistical model and estimate it from data using the latter.

Example 1.15. For example, if we assume that $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ then the first raw moment is $E[X] = \mu_x$, and we estimate it with the sample mean.

Definition 1.16 (First 5 Moments).

1. The zeroth moment, m_0 represents the total mass of a distribution, and since probabilities are normalized quantities, it should always be equal to 1,

$$m_0 = \bar{m}_0 = E[(X-\mu_x)^0] = \int_{-\infty}^{\infty} (x-\mu_x)^0 f_X(x)\ dx = \int_{-\infty}^{\infty} f_X(x)\ dx = 1$$

2. The first *raw* moment, the **expectation** of X, is

$$\mu_1 = \mu_x = E[X] = \int_{-\infty}^{\infty} x f_X(x)\ dx,$$

it is the center of mass of a probability distribution. The first central and standardized moments are less interesting because they are always zeros.

3. The second *central* moment is called the **variance** of a random variable X, denoted $\mathrm{Var}(X)$,

$$m_2 = \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu_x)^2 f_X(x)\ dx$$

The second central moment increases quadratically as mass gets further away from the distribution's mean. In other words, variance captures how spread out a distribution is or its scale parameter. Points that are further away from the mean than others are penalized disproportionally. High variance means a wide distribution, which can loosely be thought of as a "more random" random variable.

4. The third *standardized* moment, called **skewness**, measures the relative size of the two tails of a distribution,

$$\bar{m}_3 = \mathbb{S}[X] = E[Z^3] = E\left[\left(\frac{X-\mu_x}{\sigma_x}\right)^3\right],$$

where Z is the *standard score* or *z-score.*

$$Z = \frac{X-\mu_x}{\sigma_x}$$

Skewness quantifies the relative size of the two tails, consider this: any data point less than a standard deviation from the mean results in a *standard score* less than 1; this is then raised to the third power, making the value even smaller. In other words, data points less than a standard deviation from the mean contribute very little to the final calculation of skewness. Since the cubic function preserves sign, if both tails are balanced, the skewness is zero. Otherwise, the skewness is positive for longer right tails and negative for longer left tails.

While a symmetric distribution always has a skewness of zero, the opposite claim is not always true: a distribution with zero skewness may be asymmetric. We'll see an example at the end of this section.

5. The fourth *standardized* moment, **kurtosis**, measures the combined weight of the tails relative to the distribution. If either or both tails increases, the kurtosis will increase.

$$\bar{m}_4 = \mathbb{K}[X] = \frac{\mu_4}{\sigma_4} = E\left[\left(\frac{X-\mu_x}{\sigma_x}\right)^4\right]$$

Unlike skewness's cubic term which preserves sign, kurtosis's even power means that the metric is always positive and that long tails on either side dominate the calculation. Just as we saw with skewness, kurtosis's fourth power means that standard scores less than 1—again, data near the peak of the distribution—only marginally contribute to the total calculation. In other words, kurtosis measures tailedness, not peakedness.

Definition 1.17 (Moment Generating Function). Let X be a random variable. The moment generating function (MGF) of X is given as

$$M_X(s) = E[e^{sX}] = \begin{cases} \int_{-\infty}^{\infty} e^{sX} f_X(x)\, dx, & \text{if } X \text{ has a density.} \\ \sum_x e^{sX} f_X(x), & \text{if } X \text{ is discrete.} \end{cases}$$

provided that $M_X(s)$ exists in a neighbourhood $(-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$ such that it is finite and defined.

In other words, the moment-generating function of X is the expectation of the random variable e^{sX}.

To see why it is called a "moment-generating function", note that

$$\frac{d^k}{dt^k} M_X(t) = E\left[\frac{d^k}{dt^k} e^{tX}\right] = E[X^k e^{tX}],$$

thus,

$$\left.\frac{d^k}{dt^k} M_X(t)\right|_{t=0} = E[X^k].$$

In other words, the k^{th} derivative of the MGF evaluated at $t = 0$ is the $k^t h$ moment. This also means that the MGF's Taylor series expansion,

$$E[e^{tX}] = E\left[\sum_{k=1}^{\infty} \frac{1}{k!} t^k X^k\right] = \sum_{k=1}^{\infty} \frac{1}{k!} t^k E[X^k]$$

is really an infinite sum of weighted raw moments.

Definition 1.18 (Characteristic Function). Let X be a RV, the **Characteristic function** of X is given by $\phi_X(t) = E[e^{itX}]$ for $t \in \mathbb{R}$. It is always defined.

1.1 *Multiple Random Variables*

Definition 1.19 (Random Vector). A **Random Vector** $X = (X_1, ..., X_n)$ is a measurable function from $(\Omega, \mathcal{A}, P)$ to $\mathbb{R}^n$.

Definition 1.20 (Multivariate Distribution Function). A **Multivariate distribution function** $F : \mathbb{R}^n \to [0, 1]$ is defined by

$$F(X_1, ..., X_n) = P(X_1 < x_1, ..., X_n < x_n)$$

Definition 1.21 (Independence). The RV X_1, X_2 are called **independent** if their joint distribution function $F(X_1, X_2)$ is of the form

$$F_{(X_1,X_2)}(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2) \quad \forall x_1, x_2 \in \mathbb{R}$$

The variables X_i, $i \in I$ where I is an arbitrary index set, are called **independent** if for any $k \in \mathbb{N}$ and any collection $i_1, ..., i_k \in I$, then

$$F_{(X_{i_1},...,X_{i_k})}(x_{i_1}, ..., x_{i_k}) = \prod_{i=1}^{k} F_{X_{ij}}(x_{ij}), \quad (x_{i_1}, ..., x_{i_k}) \in \mathbb{R}^k$$

Corollary 1.21.1. Suppose $X_1, ..., X_n$ are random variables and $g_1, ..., g_n$ are (Borel) measurable functions, $g_i : \mathbb{R} \to \mathbb{R}$. Then, if $X_1, ..., X_n$ are independent, $g_1(X_1), ..., g_n(X_n)$ are also independent.

Definition 1.22 (Independently and Identically Distributed RVs). A sequence of RVs $X_1, X_2, X_3, ...$ is called independent and identically distributed (**iid**) if $\{X_i\}_{i\in\mathbb{N}}$ are independent and X_i has the same distribution for all i.

Notation : if X_1, X_2 are independent, we write $X_1 \perp\!\!\!\perp X_2$

Remark 1.23. If $X_1, ..., X_n$ are independent and $M_{X_i}(t)$ exists for $|t| < \varepsilon \forall i \in \{1, ..., n\}$, then

$$M_X(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

Definition 1.24. Let $(X_1, ..., X_n) = X$, $g : \mathbb{R}^n \to \mathbb{R}$ a borel-measurable function, then $g(X)$ is a random variable. The **Expectation** is

$$E[g(X_1, ..., X_n)] = \begin{cases} \sum_{X_1,...,X_n} g(x_1, ..., x_n) f_X(x_1, ..., x_n), & \text{if } X \text{ is discrete} \\ \int \ldots \int_{-\infty}^{\infty} g(x_1, ..., x_n) f_X(x_1, ..., x_n)\, dX_1 \ldots dX_n, & \text{if } X \text{ has density.} \end{cases}$$

Definition 1.25 (Moments and Central Moments). Let (X, Y) be a random vector, then

(i) The **moment** of (X, Y) of order $(j + k)$ is

$$E[X^j Y^k],\ j, k \in \mathbb{N}.$$

(ii) The **central moment** (X, Y) of order $(j + k)$ is

$$E[|X - E[X]|^j |Y - E[Y]|^k],\ j, k \in \mathbb{N}.$$

Provided all expectations above are finite.

Theorem 1.26 (Holder Inequality). Let $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, if $E[|X|^p], E[|Y|^q] < \infty$, then

$$E[|XY|] \leq (E[|X|^p])^{\frac{1}{p}} (E[|Y|^q])^{\frac{1}{q}}$$

For the case $p = q = 2$, we have the Cauchy-Schwartz inequality: $E[|XY|] \leq \sqrt{E[X^2]}\sqrt{E[Y^2]}$.

Definition 1.27 (Covariance). Let (X, Y) be a random vector. Then $E[(XE[X])(YE[Y])]$ is called the covariance of (X, Y), denoted $\text{Cov}(X, Y)$ provided it exists.

Lemma 1.28 (Properties of Covariance).

(i) $\text{Cov}(X, Y)$ exists if $E[X^2], E[Y^2] < \infty$.

(ii) $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$

(iii) $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$

(iv) $\mathrm{Cov}(a, X) = E[aX] - aE[X]\ \forall a \in \mathbb{R}$

Definition 1.29 (Pearson Correlation Coefficent). Let (X, Y) be a random vector, and $\mathrm{Cov}(X, Y)$ exist, then

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

is called the **Pearson Correlation Coefficient.**

Lemma 1.30 (Properties of Correlation).

(i) $|\rho(X, Y)| \leq 1$

(ii) If X, Y are independent, then $\rho(X, Y) = 0$

(iii) $\rho(-X, Y) = -\rho(X, Y)$

(iv) $\rho(X, Y) = \rho(Y, X)$

(v) $\rho = \pm 1 \iff X = aY + b$ almost surely

Definition 1.31 (Expectation of Random Vectors). Let $X = (X_1, ..., X_n)^T$ be a random vector such that $E[X_i] < \infty\ \forall i$, then

$$E[(X_1, ..., X_n)] = E[X] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix}$$

Theorem 1.32. Let X a random vector as above, then for any $a = (a_1, ..., a_n)^T \in \mathbb{R}^n$,

$$E[a^T X] = E[a_1 X_1 + ... + a_n X_n] = a^T E[X]$$

Theorem 1.33. If X is a random vector as above, and $X_1, ..., X_n$ **Are Independent**, then

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

Theorem 1.34. Let $X = (X_1, ..., X_n)^T$ be a random vector such that $E[X_i] < \infty$, then for any $a = (a_1, ..., a_n) \in \mathbb{R}^n$,

$$\mathrm{Var}(a_1 X_1 + ... + a_n X_n) = \mathrm{Var}(a^T X) = a^T X a$$

1.2 *Limit Theorems*

Definition 1.35 (Converging in Probability). Let $\{X_n\}$ be a sequence of RVs defined on some *probability space* (Ω, S, P). We say that a sequence $\{X_n\}$ **converges in probability** to the RV X if for every $\varepsilon > 0$,

$$P\{|X_n - X| > \varepsilon\} \to 0 \quad \text{as} \quad n \to \infty,$$

and we write

$$X_n \xrightarrow{P} X.$$

Definition 1.36 (Converging Almost Surely). Let $X_1, X_2, \dots$ an arbitrary sequence of random variables, and X a RV. Then we say $X_n \to X$ **almost surely** if and only if $P\{\omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1$, in which case we write

$$X_n \overset{a.s.}{\to} X.$$

Definition 1.37 (Converging in Distribution (in Law)). Let $\{X_n\}$ be a collection of random variables and X be a RV, then X_n converges in law or in distribution to X as $n \to \infty$ if for any continuity point x of F_X,

$$F_{X_n}(x) \overset{n\to\infty}{\longrightarrow} F_X(x),$$

in other words, we have pointwise convergence of the distribution function, and we write

$$X_n \overset{\mathcal{D}}{\to} X \quad \text{or} \quad X_n \rightsquigarrow X$$

Theorem 1.38 (Central Limit Theorem). Consider the proper rescaling $\sqrt{n}(\overline{X_n} - \mu)$, where W_n is a sequence of *iid* RVs such that $E[X_1] = \mu < \infty$, and $\text{Var}(X_1) = \sigma^2 < \infty$. By the **Weak Law of Large Numbers**, we have $\overline{X_n} \overset{P}{\to} \mu$. Then,

$$\sqrt{n}\left(\frac{\overline{X_n} - \mu}{\sigma}\right) \overset{\mathcal{D}}{\to} \mathcal{N}(0,1),$$

i.e. converges in distribution.

2 Sample Moments and Their Distributions

2.1 Random Sampling

Definition 2.1 (Random Sample). Let X be a *random variable* (RV) with *Distribution Function* (DF), and let $X_1, \dots, X_n$ be *iid* RVs with common DF F. Then the collection $X_1, \dots, X_n$ is known as a **Random Sample** of size n from the DF F.

Definition 2.2 (Statistic). Let $X_1, \dots, X_n$ be n independent observations on an RV X and let $f : \mathbb{R}^n \to \mathbb{R}^k$ a *Borel-measurable* function. Then the RV $f(X_1, \dots, X_n)$ is called a **Statistic**, provided that it is not a function of any unknown parameters.

If $X_1, \dots, X_n$ is a random sample from F, their joint distribution is given by

$$F^*(X_1, \dots, X_n) = \prod_{i=1}^{n} F(x_i)$$

Definition 2.3 (Some common statistics). Let $X_1, ..., X_n$ be a random sample from a DF F, then the statistic

$$\overline{X} := n^{-1} S_n = \sum_{i=1}^{n} \frac{X_i}{n}$$

is called the **Sample Mean** and

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

is called the **Sample Variance**, with S the **Sample Standard Deviation.**

It should be remembered that sample statistics $(\overline{X}, S^2)$ are random variables, while population parameters (μ, σ^2) are fixed constants that are unknown most often than not.

2.2 *Sample Characteristics and Their Distributions*

In this section, we consider some commonly used sample characteristics and their distributions.

Definition 2.4 (Sample/Empirical Distribution Function). Let

$$F_n^*(x) = n^{-1} \sum_{i=1}^{n} \varepsilon(x - X_i),$$

where

$$\varepsilon(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Then $nF_n^*(x)$ is the number of X_k's that are $\leq x$, and is called the **Sample/Empirical Distribution Function**.

We note that $0 \leq F_n^*(x) \leq 1 \ \forall x$, is right continuous, non-decreasing, and $F_n^*(-\infty) = 0, \ F_n^*(\infty) = 1$, hence is a *Distribution Function.*

Theorem 2.5. The RV $F_n^*(x)$ has the *probability function*

$$P\left(F_n^*(x) = \frac{j}{n}\right) = \binom{n}{j} (F(x))^j (1 - F(x))^{n-j}, \quad j = 0, 1, ..., n$$

with mean

$$E[F_n^*(x)] = F(x)$$

and variance

$$\text{Var}(F_n^*(x)) = \frac{F(x)(1 - F(x))}{n}$$

Corollary 2.5.1. For each $x \in \mathbb{R}^n$, $F_n^*(x) \xrightarrow{P} F(x)$ as $n \to \infty$.

Corollary 2.5.2. For each $x \in \mathbb{R}$,

$$\frac{\sqrt{n}(F_n^*(x) - F(x))}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{L} Z \quad \text{as} \quad n \to \infty$$

Z is $\mathcal{N}(0, 1)$.

Proposition 2.6 (Chebychev's Inequality).

$$P\{|F_n^*(x) - F(x)| > \varepsilon\} \leq \frac{\text{Var}(F_n^*(x))}{\varepsilon^2}$$

or equivalently,

$$P\{(T_n - \theta)^2 \geq \epsilon^2\} \leq \frac{E(T_n - \theta)^2}{\epsilon^2}$$

We have convergence in probability of the empirical distribution function to the true distribution function, $F_n^*(x) \overset{p}{\to} F(x)$, using the Weak Law of Large Numbers (WLLN). But, we note that it can also be directly computed it using Chebychev's inequality:

$$P\{|F_n^*(x) - F(x)| > \varepsilon\} \leq \frac{\text{Var}(F_n^*(x))}{\varepsilon^2} = \frac{F(x)(1 - F(x))}{n\varepsilon^2} \to 0 \text{ as } n \to \infty$$

Theorem 2.7 (Glivenko-Cantelli). $F_n^*(x)$ converges uniformly to $F(x)$, that is, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\sup_{-\infty<x<\infty} |F_n^*(x) - F(x)| > \varepsilon\right) = 0$$

Theorem 2.8 (Transforming Density Functions). Let X be a RV with PDF $f(x)$, and y a transformation function, then $y(X)$ is a derived random variable. Denote the inverse of the transformation $y(x)$ by $x(y)$. Then, the PDF of $y(X)$ is

$$g(y) = f(x(y)) \left|\frac{dx}{dy}\right|$$

Proof. Suppose X is a random variable whose probability density function is $f(x)$. By definition,

$$P(a \leq X < b) = \int_a^b f(x)\, dx$$

Any function of a random variable is itself a random variable and, if y is taken as some transformation function, $y(X)$ will be a derived random variable. Let $Y = y(X)$, and notice that if $X = a$, then $Y = y(a)$, hence

$$P(y(a) \leq Y < y(b)) = P(a \leq X < b) = \int_a^b f(x)\, dx = \int_{f(a)}^{f(b)} f(x(y)) \frac{dx}{dy}\, dy.$$

Notice that the right-hand integrand $f(x(y))\, \frac{dx}{dy}$ is expressed wholly in terms of y, denoting it by $g(y)$, we get

$$P(y(a) \leq Y < y(b)) = \int_{f(a)}^{f(b)} g(y)\, dy.$$

This demonstrates that $g(y)$ is the probability density function associated with Y. Now, if $\frac{dx}{dy}$ were to change sign, there would be values of x for which $y(x)$ would be multivalued, i.e. it couldn't be a PDF. Noting this, we conclude that the derived PDF should be written as

$$g(y) = f(x(y)) \left| \frac{dx}{dy} \right| .$$

□

3 Theory of point estimation

3.1 Consistency and Bias

Let X be an *random variable* defined on a probability space (ω, S, P). Suppose that the *distribution function* F of X depends on a certain number of parameters, and suppose further that the functional form of F is known except perhaps for a finite number of these parameters. Let $\theta = (\theta_1, \theta_2, ..., \theta_k)$ be the unknown parameter associated with F.

Definition 3.1 (Parameter). A **parameter** is any quantity of a statistical population that summarizes or describes an aspect of the population, such as a mean or a standard deviation. If a population exactly follows a known and defined distribution, for example the normal distribution, then a small set of parameters can be measured which completely describes the population.

A **parameter** is to a *population* as a **statistic** is to a *sample*; that is to say, a parameter describes the true value calculated from the full population (such as the population mean), whereas a statistic is an estimated measurement of the parameter based on a sample (such as the sample mean).

Definition 3.2 (Parameter Space). The set of all admissible values of the parameters of a distribution function F is called the **parameter space**, and is denoted by Θ.

Definition 3.3 (Parametric Family). A **parametric family** is a family of objects whose differences depend only on the chosen values for a set of *parameters*.

Example 3.4. For example, the probability density function F_X of a random variable X may depend on a parameter θ. In that case, the function may be denoted $f_X(\cdot\,;\theta)$ to indicate the dependence on the **parameter** θ. θ is not a formal argument of the function as it is considered to be fixed. However, each different value of the parameter gives a different probability density function. Then the **parametric family** of densities is the set of functions $\{f_X(\cdot\,;\theta) \mid \theta \in \Theta\}$, where Θ

denotes the **parameter space**. As an example, the normal distribution is a family of similarly-shaped distributions parametrized by their mean and their variance.

Let $X := X_1, ..., X_n$ be a r.v. with DF F_θ, where $\theta = (\theta_1, \theta_2, ..., \theta_k)$ is a vector of unknown parameters, $\theta \in \Theta$. Let $\varphi : \Theta \to \mathbb{R}$, in this section, we explore the problem of approximating $\varphi(\theta)$ on the basis of the observed value x of X.

Definition 3.5 (Point Estimator). Let $(X_1, ..., X_n)$ be a *random sample* of size n from the *random variable* F_θ. A statistic $\delta(X)$ is said to be a **point estimator** of φ if $\delta : \mathcal{X} \to \Theta$, where $\mathcal{X}$ is the set of possible values of X.

The problem of point estimation is to find an estimator δ for the unknown parametric function $\varphi(\theta)$ that has some nice properties. We define the following properties.

Definition 3.6 (Unbiasedness). A *point estimate* $\delta(X)$ of a *parameter* θ is called unbiased if $E_\theta[\delta] = \theta$, where

$$E_\theta[\delta] = \int t \, dF_{\delta,\theta}(t)$$

Definition 3.7 (Bias). Let θ be a RV, and $\hat{\theta}$ an estimator of θ. Then the **bias** of $\hat{\theta}$ relative to θ is defined as

$$\text{Bias}_\theta(\hat{\theta}) = E[\hat{\theta}] - \theta$$

An estimator is said to be **unbiased** if its bias is equal to zero for all values of parameter θ, or equivalently, if the expected value of the estimator matches that of the parameter.

Definition 3.8 (Consistency). Let $X_1, X_2, ..., X_n$ be a sequence of *iid* r.v. with common DF F_θ, $\theta \in \Theta$ a sequence of point estimators $T(X_1, ..., X_n) = T_n$ is called

(i) **weakly consistent** for θ if $T_n \xrightarrow{p} \theta$

(ii) **strongly consistent** for θ if $T_n \xrightarrow{a.s.} \theta$

Theorem 3.9 (Markov's Inequality). If X is a nonnegative random variable and $a > 0$, then the probability that X is at least a is at most the expectation of X divided by a:

$$P(X > a) \leq \frac{\mathbb{E}[X]}{a}$$

Proposition 3.10 (Chernoff's Inequality). Suppose $X_1, ..., X_n$ are *iid* RV, let $X = \sum_{i=1}^n X_i$. We can apply Markov's inequality to get the following inequality

$$P(X \geq a) \leq \frac{\prod_{i=1}^n \mathbb{E}[e^{tX_i}]}{e^{ta}}$$

Proof. Apply the transformation $x \mapsto e^{tx}$

$$
\begin{aligned}
P(X \geq a) \leq P(e^{tX} \geq e^{ta}) \overset{(3.9)}{\leq} \frac{\mathbb{E}[e^{tX}]}{e^{ta}} &= \frac{\mathbb{E}[e^{t(X_1+X_2+\ldots+X_n)}]}{e^{ta}} \\
&= \frac{\mathbb{E}[\prod_{i=1}^{n} e^{t(X_i)}]}{e^{ta}} \\
&= \frac{\prod_{i=1}^{n} \mathbb{E}[e^{t(X_i)}]}{e^{ta}}
\end{aligned}
$$

□

Recall that if X, Y are independent, then $E[XY] = E[X]E[Y]$, and in particular, for any measurable functions g, f, $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$.

Definition 3.11 (Martingales).

Theorem 3.12 (Kolmogorov's Strong Law of Large Numbers). Let $X_1, X_2, \ldots$ be *iid* RV with common law $\mathcal{L}(X)$(common cumulative distribution function F). Then

$$\frac{S_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{a.s.} \mu = E[X] \text{ as } n \to \infty \iff E[X] < \infty$$

Theorem 3.13. If T_n is a sequence of estimators for θ such that $E[T_n] \to \theta$ and $\text{Var}(T_n) \to 0$ as $n \to \infty$, then T_n is a *consistent* estimator of θ.

3.2 *Invariance*

Definition 3.14 (Invariance Under Group). Let $\mathcal{G}$ be a group of Borel-measurable functions, with the group operation being composition of functions. A family of probability distributions is said to be **invariant under a group** $\mathcal{G}$ if $\forall g \in \mathcal{G}, \forall \theta \in \Theta$, we can find a unique $\theta^* \in \Theta$ such that the distribution of $g(X)$ is given by P_{θ^*} whenever X has the distribution P_θ.

3.3 *Sufficient Statistics*

Definition 3.15 (Family of Distributions). This terminology of "families" tends to be used when studying classes $\mathcal{C}_Y$ of functions into a set Y or "maps." Given a domain X, a family $\mathcal{F}$ of maps on X parameterized by some set Θ (the "parameters") is a function[2]

$$\mathcal{F} : X \times \Theta \to Y$$

for which

(i) for each $\theta \in \Theta$, the function $\mathcal{F}_\theta : X \to Y$ given by $\mathcal{F}_\theta(x) = \mathcal{F}(x, \theta)$ is in $\mathcal{C}_Y$ and

(ii) $\mathcal{F}$ itself has certain "nice" properties.

[2] Stack Exchange. Definition of family of a distribution, Dec 2017. URL `https://stats.stackexchange.com/questions/320746/definition-of-family-of-a-distribution`

The idea is that we want to vary functions from X to Y in a "smooth" or controlled manner. Property (i) means that each θ designates such a function, while the details of property (ii) will capture the sense in which a "small" change in θ induces a sufficiently "small" change in $\mathcal{F}_\theta$.

For statistical applications, $\mathcal{C}_Y$ is the set of all distributions on $\mathbb{R}$ (or, in practice, on $\mathbb{R}^n$ for some n, but to keep the exposition simple I will focus on $n = 1$). We may identify it with the set of all non-decreasing càdlàg functions $\mathbb{R} \to [0,1]$ where the closure of their range includes both 0 and 1: these are the cumulative distribution functions, or simply distribution functions. Thus, $X = \mathbb{R}$ and $Y = [0,1]$.

A family of distributions is any subset of $\mathcal{C}_Y$. Another name for a family is statistical model. It consists of all distributions that we suppose govern our observations, but we do not otherwise know which distribution is the actual one.

(i) A family can be empty.

(ii) $\mathcal{C}_Y$ itself is a family.

(iii) A family may consist of a single distribution or just a finite number of them.

Definition 3.16 (Sufficient). Let $X = (X_1, ..., X_n)$ be a sample from $\{F_\theta : \theta \in \Theta\}$. A statistic $T = T(X)$ is sufficient for θ or for the family of distributions $\{F_\theta : \theta \in \Theta\}$ if and only if the conditional distribution of X, given $T = t$, does not depend on θ, except perhaps on a null set.

In particular, a statistic is sufficient for a family of probability distributions if the sample from which it is calculated gives no additional information than the statistic, as to which of those probability distributions is the sampling distribution.

The above definition is not constructive since it requires that we first guess a statistic T and then check to see whether T is sufficient. Moreover, the procedure for checking that T is sufficient is quite time consuming. We now give a criterion for determining sufficient statistics.

Theorem 3.17 (Factorization Criterion). Let $X = (X_1, ..., X_n)$ be discrete RVs with PMF $p_\theta(x_1, ..., x_n),\ \ \theta \in \Theta$. Then $T(X_1, ..., X_n)$ is sufficient for θ if and only if we can write

$$p_\theta(x_1, ..., x_n) = h(x_1, ..., x_n) g_\theta(T(x_1, ..., x_n)),$$

where h is a nonnegative function of $x_1, ..., x_n$ only and does not depend on θ, and g_θ is a nonnegative nonconstant function of θ and $T(x_1, ..., x_n)$ only. The statistic $T(X_1, ..., X_n)$ and parameter θ may be multidimensional.

Remark 3.18. Theorem 3.17 also holds for the continuous case and, indeed, for quite arbitrary families of distributions.

3.4 *Maximum Likelihood*

The principle of maximum likelihood essentially assumes that the sample is representative of the population and chooses as the estimator that value of the parameter which maximizes the PDF $f_\theta(x)$.

Definition 3.19 (Likelihood function). The **likelihood function** (often simply called the *likelihood*) is the *joint probability distribution* of observed data viewed as a function of the parameters of a statistical model.

Let $(X_1, ..., X_n)$ be a random vector with PDF $f_\theta(x_1, ..., x_n)$, $\theta \in \Theta$. The likelihood function is defined as

$$L(\theta; x_1, ..., x_n) = f_\theta(x_1, ..., x_n),$$

considered as a function of θ. If θ is a multiple parameter, and $X_1, ..., X_n$ are *iid*, with PDF $f_\theta(x)$, the likelihood function is then

$$L(\theta; x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

Definition 3.20 (log Likelihood). It is convenient to work with the logarithm of the likelihood, since log is a monotone function,

$$\log L(\hat{\theta}; x_1, ..., x_n) = \sup_{\theta \in \Theta} \log L(\theta; x_1, ..., x_n)$$

Proposition 3.21. Let Θ be an open subset of $\mathbb{R}^k$, and suppose that $f_\theta(x)$ is a positive, differentiable function of θ, if a supremum $\hat{\theta}$ exists, it must satisfy the likelihood equations,

We say "equations" because there is one equation for each j.

$$\frac{\partial \log L(\hat{\theta}; x_1, ..., x_n)}{\partial \theta_j} = 0, \quad j = 1, 2, ..., k, \quad \theta = (\theta_1, ..., \theta_k)$$

Any nontrivial root of the likelihood equations is called an MLE in the loose sense. A parameter value that provides the absolute maximum of the likelihood function is called an MLE in the strict sense or, simply, an MLE.

Definition 3.22 (Maximum Likelihood Estimator (MLE)). The principle of maximum likelihood estimation consists of choosing an estimator of θ, $\hat{\theta}(x)$, that maximizes $L(\theta; x_1, ..., x_n)$. In other words, we want to find a mapping $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^k$ such that

$$L(\hat{\theta}; x_1, ..., x_n) = \sup_{\theta \in \Theta} \{L(\theta : x_1, ..., x_n)\}$$

If such a $\hat{\theta}$ exists, we call it a **maximum likelihood estimator**, or MLE

3.5 Completeness

The concept of sufficiency is used frequently with another concept, called completeness, which we now define.

Definition 3.23 (Completeness). Let $\{f_\theta \mid \theta \in \Theta\}$ be a family of PDFs (or PMFs). We say that this family is complete if

$$\mathbb{E}_\theta[g(X)] = 0 \quad \forall \theta \in \Theta$$

implies that

$$P_\theta\{g(X) = 0\} = 1 \quad \forall \theta \in \Theta$$

Definition 3.24 (Complete Estimator). A statistic $T(X)$ is said to be complete if the family of distributions of T is complete.

That is, if X is a RV whose probability distribution belongs to a parametric model P_θ parametrized by θ, and T is a statistic, then T is said to be complete for the distribution of X if, for every measurable function g,

$$\text{if } \mathbb{E}_\theta(g(T)) = 0 \text{ for all } \theta \text{ then } P_\theta(g(T) = 0) = 1 \text{ for all } \theta.$$

X will usually be a multiple RV. The family of distributions of T is obtained from the family of distributions of $X_1, X_2, ..., X_n$ by the usual transformation discussed in 1.1.

$g(T)$ can't depend on θ

Definition 3.25 (Exponential Family). If there exists real-valued functions $Q_1, ..., Q_k, D$ defined on Θ and Borel-measurable functions $T_1, ..., T_k, S$ on $\mathbb{R}^n$ such that

$$f_\theta(x) = \exp\left(\sum_{i=1}^{k} Q_i(\theta)T_i(x) + D(\theta) + S(x)\right) \tag{1}$$

we say that the family $\{f_\theta, \theta \in \Theta\}$ is a k-parameter **exponential family**.

Theorem 3.26. Let $\{f_\theta : \theta \in \Theta\}$ be a $k-$parameter exponential family as in (1), where $\mathbb{R}^k \supseteq \theta = (\theta_1, ..., \theta_k) \in \Theta$, $T_1, ..., T_k, S$ real valued functions on $\mathbb{R}^n$, $T = (T_1, ..., T_k)$, and $x = (x_1, ..., x_n)$, $k \leq n$. Let $Q = (Q_1, ..., Q_k)$, and suppose that the range of Q contains an open set in $\mathbb{R}^k$. Then,

$$T = (T_1(X), ..., T_k(X))$$

is a complete sufficient statistic.

3.6 UMVUE

Theorem 3.27 (Lehmann Sheffé). Let $\vec{X} = X_1, X_2, \ldots, X_n$ be a random sample from a distribution that has p.d.f (or p.m.f in the discrete case) $f(x : \theta)$ where $\theta \in \Omega$ is a parameter in the parameter space. Suppose $Y = u(\vec{X})$ is a sufficient statistic for θ, and let $\{f_Y(y : \theta) : \theta \in \Omega\}$ be a complete family. If $\varphi : \mathrm{E}[\varphi(Y)] = \theta$ then $\varphi(Y)$ is the unique MVUE of θ.

3.7 *Lower Bound for the Variance of an Unbiased Estimate*

Definition 3.28 (Score). Formally, the partial derivative with respect to θ of the natural logarithm of the likelihood function is called the **score**.

Definition 3.29 (Fisher Information). Let $X \sim f_\theta(x)$. The Fisher information is defined to be the variance of the score:

$$I(X) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \ln f_\theta(X) \right)^2 \middle| \theta \right] = \int_{\mathbb{R}} \left(\frac{\partial}{\partial\theta} \ln f_\theta(x) \right)^2 f_\theta(x)\, dx$$

3.8 *Review of Estimator Properties*

Unbiased. An estimator is unbiased if, on average, it produces parameter estimates that are equal to the true values of the parameters being estimated. Mathematically, this can be expressed as $E(\hat{\theta}) = \theta$, where $E(\hat{\theta})$ is the expected value of the estimator and θ is the true parameter value.

Consistent. Consistency refers to the property that as the sample size increases, the estimator converges in probability to the true parameter value. In other words, for any small positive number ϵ, $\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$, where n is the sample size.

PROVING CONSISTENCY. The easiest way to show convergence in probability/consistency is to invoke Chebyshev's Inequality, which states:

$P((T_n - \theta)^2 \geq \epsilon^2) \leq \frac{E[(T_n-\theta)^2]}{\epsilon^2}$.

Thus,

$P(|T_n - \theta| \geq \epsilon) = P((T_n - \theta)^2 \geq \epsilon^2) \leq \frac{E[(T_n-\theta)^2]}{\epsilon^2}$.

And so you need to show that $E[(T_n - \theta)^2]$ goes to 0 as $n \to \infty$.

Remark 3.30. The above requires that the estimator is at least **asymptotically unbiased**.

Sufficient. A statistic is considered sufficient if it contains all the information about the parameter that is available in the sample. In other words, no other statistic calculated from the same sample provides additional information about the parameter. This property is crucial in reducing data dimensionality while retaining essential information for parameter estimation.

PROVING SUFFICIENCY. Showing sufficiency is straightforward with the use of the *Factorization Criterion* (3.17).

Complete. Completeness is a property related to the ability of a statistic to detect all possible variations in the underlying distribution. A statistic is considered complete if, for any function g such that $E[g(T)] = 0$ for all values of the parameter, the only solution is that the probability of $g(T) = 0$ is 1. Here, T is the statistic in question.

References

Matthew Bernstein. 2017. URL `https://mbernste.github.io/files/notes/Variance.pdf`.

Stack Exchange. Definition of family of a distribution, Dec 2017. URL `https://stats.stackexchange.com/questions/320746/definition-of-family-of-a-distribution`.