Mathematical Statistics

Davi Sales Barreira

May 26, 2022

Contents

List of Definitions

List of Theorems

Notes mostly based on Keener [2] and Shao [4]. Some concepts are taken from Gentle [1]. Unless stated otherwise, we always assume the probability space $(\Omega, \mathcal{F}, \mathsf{P})$, where $\mathcal{F}$ is the Borel σ-algebra.

1 Initial Definitions

1.1 Statistical Modelling

Let's start with some of the the main definitions. These notes have as prerequisites some knowledge of Measure Theory and Probability.

One can say that the goal of statistical analysis is to make inferences regarding a probability measure undergirding a data generating process. More formally, let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. The measure P is known as the *population* probability measure. A process generates data according to this probability space, and the goal is to infer the probability measure P.

On the definition of random sampling (skip)

This informal definition gives an intuition of what one is usually trying to accomplish, but it's not rigorous enough. A more precise framework of what statistical modelling is doing can be formulated as decision theory. Still, unfortunately, most books don't formalize concepts such as "samples" or "observed data", relying on our personal intuition, which I humbly consider mathematical malpractice. Hence, I'll take the liberty to define such things, thus, one should take them with care.

Definition 1.1 ($\circledast$ Sampling Process). Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. A function $f_s : \mathbb{N} \to \Omega$ is called a sampling process of this probability space if

$$\lim_{n \to +\infty} \frac{\sum_{i=1}^{n} \mathbb{1}_A(f_s(i))}{n} = \mathsf{P}(A), \quad \forall A \in \mathcal{F}. \tag{1}$$

Note that this definition implies, for example, that the first sample of an experiment is just $f_s(1)$. This definition is similar to what one gets with random number generators. The value $i \in \mathbb{N}$ is what is called the "seed", and the function f_s is the generator, which we are calling the sampler, or, sampling process.

These definition makes us pose questions such as:

- do these sampling processes exist for all/any probability space?;
- how does one constructs a sampling process function?.

The answer is actually simple. If our probability measures are discrete with finite support, one can construct sampler in the following way. Suppose that your Ω can either be 0 with probability p or 1. If $p \in \mathbb{Q}$, then $p = n/m$, and repeating infinitely the sequence $(0000...0111...1)$ with n zeros and $m-n$ ones, is a sampler for this measure P. Now, if $p \in \mathbb{I}$, then there is a sequence of rational numbers $q_k = \frac{n_k}{m_k} \to p$. The sampler with concatenated sequences of n_k zeros and $m_k - n_k$ ones samples measure P. The same argument can be used for any discrete measure P with finite support.

A much more complicated question is about independent sampling. A first challenge is how to define it, since the previous samplers we showed are clearly (intuitively) not random. For this, we'll use the notion of randomness created by Martin-Löf in 1966 [3], which we'll call ML-Randomness.

Definition 1.2 (⊛ Independent Sampling Process). Let $(\Omega, \mathcal{F}, P)$ be a probability space. A sampler $f_s : \mathbb{N} \to \Omega$ is called an independent sampler if the infinite sequence $f_s(\mathbb{N}) = (f(1), f(2), ...)$ is ML-Random.

ML-Randomness is not simple define, but it intuitively accomplishes what we'd expect in a random sequence. A sampler should be independent if we cannot predict the outcome of the next number in the sequence, which is not true if we actually know how the function is being generated. For example, the sampler $(01010101...)$ is a valid sampler for the Bernoulli distribution, but it's not ML-Random (i.e. independent) since we can easily predict each of it's outcome. Hence, the independent sampler should be, in a sense, impossible to compute, in other words, there should not be an "easy" formula to compute each outcome. All this intuition is captured by the ML-Randomness definition, but requires an understanding of Computation Theory, which I'll leave for another time.

Suppose you have measured the length of your table 10 times and you got the following measurements $\{1.10, 1.15, ..., 1.13, 1.09\}$. The question is then, what is the best estimate for the "actual" size of the table? This question is ill-posed, unless we mathematically define why an estimate would be better than another. For example, if you are measuring your table with the objective of finding out if it'll go through the door, perhaps it's better to be pessimistic and use the largest measurement as the size of the table. Or maybe you want to sell it on the internet, and you want to give a precise description for a potential buyer, thus, it might be better to use the average of all the measurements. Informally, the estimate of the table, be it the average, the maximum, the minimum, is what we call a statistic.

Definition 1.3 (Statistic - Shao [4] pg 100). Let X be a random variable. Hence, a measurable function $T(X)$ is called a **statistic** if the value of $T(X)$ is known if the value of X is known, i.e. a statistic is a function of the *observed data.*

Now, returning to the problem of estimating the size of the table. To properly answer the question of which estimate (statistic) to use, we have to formalize the decision process via *decision theory.* This is done in the following matter, first, we have to assign a *loss function* L that represents how much we are penalized by guessing the incorrect value for the size of the table. Suppose that the actual size is θ and the measurements are $x = (x_1, ..., x_n)$, hence $L(\theta, T(x))$ is our loss. This function will change depending of the problem at hand, for example, sometimes overestimating may be much worse than underestimating, and sometimes small errors may cause much less problems than larger error, thus, the loss function that you choose should properly model each case.

It might seem that we are all set to decide on our estimate, but there is still a loose thread. We need to somehow explain why the measurements are imprecise. For example, if you know that your ruler has some imprecisions, and the distribution of the error between a measurement and the actual size $(X_i - \theta)$ follows a Normal distribution with $N(0.1, 1)$. So, perhaps instead of using the average of the measurements, you should use the average minus 0.1. Note that we modelled each measurement as an independent random variable $X_i \sim N(\theta + 0.1, 1)$.

Again, we return to the question "what estimate should you use?". Unfortunately, since our measurement are random, we cannot say for certain that the estimate you choose to use will actually be the optimal one for that specific day. But we can consider the *average risk*, based on the distribution of the error. Therefore, we have

$$E[L(\theta, T(X))] = \int L(\theta, T(X)) dP_x = \int L(\theta, T(X)) p_X(x) dx, \tag{2}$$

where the risk of using an estimate $T_1(X)$ is $E[L(\theta, T_1(X))]$. And with this, you can decide on which estimate to choose based on which one has the lowest risk.

Note that the above problem could be reinterpreted as trying to infer which Normal distribution $N(\theta+0.1, 1)$ better generated the sample in terms of minimizing the loss function. This problem would then be what we call *parametric*, since our inference is restricted to a parametric family of probability distributions.

Definition 1.4 (Parametric Family). $\mathcal{P}_\Theta := \{P_\theta \ : \ \theta \in \Theta\}$ is a parametric family, where for each $\theta \in \Theta$, P_θ is a probability measure in $(\Omega, \mathcal{F})$.

Definition 1.5 (Risk). Let $\mathcal{P}_\Theta := \{P_\theta \ : \ \theta \in \Theta\}$ be a parametric family, $T(X)$ a statistic, and $L(\theta, T)$ a loss function (e.g. $L(\theta, T) = |\theta - T(X)|^2$). The risk of T is

$$R(\theta, T) := E_\theta[L(\theta, T(X))]. \tag{3}$$

There are infinite parametric families, and the choice of an specific family will vary with the problem at hand. For example, if you are tasked with estimating a distribution where you know values vary only inside $[0, 1]$, it makes no sense to consider distributions such as Normal. Also, if you know your samples are discrete values, it makes no sense considering continuous distributions. One of the most useful families is the Exponential Family.

Definition 1.6 (Exponential Family). A parametric family $\mathcal{P}_\Theta$ dominated by a probability measure ν on $(\Omega, \mathcal{F})$ is called an Exponential Family if

$$\frac{dP_\theta}{d\nu} = p_\theta(x) = \exp\left\{ \left[\sum_i \eta_i(\theta) T_i(x) \right] - B(\theta) \right\} h(x), \quad \forall x \in \Omega. \tag{4}$$

Where each T_i is a statistic, η_i is a function from $\Theta \to \mathbb{R}$, h is a Borel measurable function on $(\Omega, \mathcal{F})$, and $B(\theta) = \log\{\int_\Omega h(x) \exp\{[\sum_i \eta_i(\theta) T_i(x)]\} d\nu(x)\}$, i.e. $B(\theta)$ normalizes the distribution $p_\theta(x)$ so that it integrates to 1.

The Exponential Family has a *canonical form*, which uses parameter η. Let

$$\eta(\theta) = (\eta_1(\theta), ..., \eta_d(\theta)) = \xi = (\xi_1, ..., \xi_d). \tag{5}$$

Define $\Xi := \{\xi : \xi \in \eta(\Theta)\}$, which is called the *natural parameter space.* Note that $\Xi \subset \mathbb{R}^d$.

The Exponential Family contains many known distributions, such as the Normal, Gamma, Exponential, Binomial, Beta, and more. Hence, for example, the family of all Normal distributions is a subfamily of the Exponential Family.

Definition 1.7 (Full rank - Shao [4]). A subfamily of the Exponential Family is called *full rank* if there is an open set contained in it's natural parameter space Ξ.

1.2 Sufficiency, Completeness and Minimality

Definition 1.8 (Sufficient Statistic). Suppose that X has a distribution family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Then, T is a sufficient statistic for $\mathcal{P}$ if for every t and θ_1,$\theta_2 \in \Theta$, then $P_{\theta_1}(X \mid T(X) = t) = P_{\theta_2}(X \mid T(X) = t)$, i.e., once we know the value of the statistic, the probability distribution of X is independent of the parameter θ.

Definition 1.9 (Factorization Theorem). Suppose that X has a distribution family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ dominated by ν (e.g. if ν is the Lebesgue measure, then each $P_\theta \in \mathcal{P}$ has a density function). Thus, a statistic T is sufficient if, and only if, there exist measurable functions $g_\theta \geq 0$ and $h \geq 0$ such that the density of a probability measure $P_\theta \in \mathcal{P}$ can be written as

$$p_\theta(x) = g_\theta(T(x))h(x), \quad \text{almost everywhere for } \nu. \tag{6}$$

Definition 1.10 (Minimal Sufficient). A statistic T is minimal sufficient for a family $\mathcal{P}$ if for every sufficient statistic S of $\mathcal{P}$, there is a measurable function ϕ such that $T = \phi(S)$ a.s. $\mathcal{P}$. In words, a minimal sufficient statistic can always be obtained from the other sufficient statistics.

For example, let $T(X) = \max X$ be the minimal sufficient. Hence, $S(X) = (S_1(X), S_2(X)) = (\min(X), \max(X))$ is another sufficient statistic that computes both the minimum and the maximum of the observed data. Note that $\phi(S(X)) = S_2(X) = \max(X) = T(X)$, hence, T is obtained from S.

Definition 1.11 (Completeness). A statistic $T(X)$ is complete for a family for $\mathcal{P}$ if for every $P \in \mathcal{P}$ and for any Borel f with $E[f(T)] = 0$, this implies that $f(T) = 0$ a.s. P.

Theorem 1.12. If T is complete and sufficient, then T is minimal sufficient.

Theorem 1.13. If an Exponential family is full rank, then T is minimal sufficient, where $T = (T_1, ..., T_n)$ for each T_i in equation (4).

Definition 1.14 (Ancillary). A statistic V is called ancillary with respect to a parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if its distribution does not depend on θ. This means that $V(X)$ provides no "information" on inferring θ.

Theorem 1.15 (Basu). If T is complete and sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, and V is ancillary, then T and V are independent under P_θ for any $\theta \in \Theta$.

1.3 Statistics vs. Estimators

Note that in books like Keener [2], the author will sometimes refer to something as a statistic, and sometimes as an estimator. In terms of the mathematical definition, a statistic and an estimator are the same thing, i.e. measurable functions of the data. Informally, for parametric families, one usually uses the term statistic when the function $T(X)$ is used to estimate the parameter θ of the distribution and estimator for a function $\delta(X)$ that is trying to estimate a function $g(\theta)$ of the parameter.

For example, suppose that we have a family of Normal distributions with $N(\theta, 1)$, that is, we don't know the mean, but the variance is known. Suppose that we collected n independent samples, thus, a statistic would be $T(X_1, ..., X_n) = \sum_{i=1}^{n} \frac{x_i}{n}$, the sample average. Note that the goal here is to $T(X_1, ..., X_n) \approx \theta$, hence, a statistic.

Now, suppose that we are instead interested in estimating not the mean (θ), but actually the squared mean (θ^2). In this case, the function $\delta(X_1, ..., X_n) = \left(\sum_{i=1}^{n} \frac{x_i}{n}\right)^2$ would be an estimator of $g(\theta) = \theta^2$.

Theorem 1.16 (Rao-Blackwell). Let T be a sufficient statistic for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, let δ be an estimator for $g(\theta)$, and define $\eta(T) = E[\delta(X) \mid T]$. If $\theta \in \Theta, R(\theta, \delta) < +\infty$, and $L(\theta, \cdot)$ is convex, then

$$R(\theta, \eta) \leq R(\theta, \delta). \tag{7}$$

What this theorem is implying is that if you want to estimate a function of your parameter θ, if there exists a sufficient statistic T, then the best estimator, in terms of the risk, can be written as a function of T.

Definition 1.17 (Unbiased Estimator). An estimator $\delta(X)$ for $g(\theta)$ is unbiased if

$$E_\theta \delta(X) = g(\theta), \ \forall \theta \in \Theta. \tag{8}$$

Definition 1.18 (UMVU). An unbiased estimator δ^* is uniformly minimum variance (UMVU) if

$$Var_\theta(\delta^*) \leq V_\theta(\delta), \ \forall \theta \in \Theta, \tag{9}$$

for any other unbiased estimator δ.

Note that if we consider the squared loss function for $g(\theta)$ as $L(\theta, \delta) = |g(\theta), \delta|$, and then

$$R(\theta, \delta) = E_\theta[(\delta - g(\theta))^2] = Var_\theta(\delta) + b^2(\theta, \delta), \tag{10}$$

where $b(\theta, \delta) = E_\theta[\delta(X) - g(\theta)]$ is the bias. Hence, if our estimator is unbiased, the risk is equivalent to the variance, and the UMVU estimator is the one with least risk among all unbiased estimators.

1.4 Fisher Information and Cramér-Rao Inequality

Definition 1.19 (Fisher Information). Let $\mathcal{P} = P_\theta : \theta \in \Theta$ be dominated by a measure ν, such that each P_θ has a density function p_θ. The Fisher Information is given by

$$I(\theta) := E_\theta \left[\left(\frac{\partial \log p_\theta(X)}{\partial \theta} \right) \left(\frac{\partial \log p_\theta(X)}{\partial \theta} \right)^T \right]. \tag{11}$$

Note that for $\theta \in \mathbb{R}^n$, $I(\theta)$ is an $n \times n$ matrix, where

$$I(\theta)_{i,j} := E_\theta \left[\frac{\partial \log p_\theta(X)}{\partial \theta_i} \frac{\partial \log p_\theta(X)}{\partial \theta_j} \right]. \tag{12}$$

Note that for the existence of the Fisher Information, the following regularity condition must be satisfied ([1]):

- The parameter space $\Theta \subset \mathbb{R}^n$ is convex and contains an open set;
- For any x in the support and $\theta \in \Theta^\circ$ (interior of Θ), $\frac{\partial p_\theta(x)}{\partial \theta}$ and $\frac{\partial^2 p_\theta(x)}{\partial \theta^2}$ exist and are finite, and $\frac{\partial^2 p_\theta(x)}{\partial \theta^2}$ is continuous in θ;

- The support is independent of θ, hence $P_\theta \in \mathcal{P}$ all share the same support.

Proposition 1.20 (Fisher Information Equivalent Form). Let $I(\theta)$ be the Fisher Information and suppose that the regularity conditions hold. Then,

$$E\left[\frac{\partial \log p_\theta(X)}{\partial \theta}\right] = 0, \tag{13}$$

and

$$I(\theta) = E_\theta\left[\left(\frac{\partial \log p_\theta(X)}{\partial \theta}\right)\left(\frac{\partial \log p_\theta(X)}{\partial \theta}\right)^T\right] = -E_\theta\left[\frac{\partial^2 \log p_\theta(X)}{\partial \theta^2}\right]. \tag{14}$$

Proof. Let's assume that P_θ has a density p_θ (e.g. P_θ is dominated by the Lebesgue measure). From the regularity conditions, we can pass the derivative to the inside of the integral,

$$\begin{aligned} E\left[\frac{\partial \log p_\theta(X)}{\partial \theta}\right] &= \int \frac{\partial \log p_\theta(X)}{\partial \theta} p_\theta(x)dx \\ &= \int \frac{1}{p_\theta(X)}\frac{\partial p_\theta(X)}{\partial \theta} p_\theta(x)dx \\ &= \frac{\partial}{\partial \theta}\int p_\theta(x)dx = \frac{\partial}{\partial \theta}1 = 0. \end{aligned}$$

Now, for the second part, let's do for the 1-D case for clarity

$$\frac{\partial^2 \log p_\theta(x)}{\partial \theta^2} = \frac{\partial^2 p_\theta(x)}{\partial \theta^2}\frac{1}{p_\theta(x)} - \frac{\partial p_\theta(x)}{\partial \theta}\frac{1}{p_\theta(x)^2} = \frac{\partial^2 p_\theta(x)}{\partial \theta^2}\frac{1}{p_\theta(x)} - \left(\frac{\partial \log p_\theta(x)}{\partial \theta}\right)^2.$$

Note that, due to the regularity conditions, we can exchange the derivative and the integral, hence

$$E[\frac{\partial^2 p_\theta(x)}{\partial \theta^2}\frac{1}{p_\theta(x)}] = \int \frac{\partial^2 p_\theta(x)}{\partial \theta^2}\frac{1}{p_\theta(x)} p_\theta(x)dx = \frac{\partial^2}{\partial \theta^2}\int p_\theta(x)dx = 0.$$

We conclude that

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

□

Theorem 1.21. (Crámer-Rao) Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a dominated family, and suppose that the regularity conditions for the Fisher Information are met. For an unbiased estimator δ of $g(\theta)$, i.e. $E_\theta[\delta(X)] = g(\theta)$, with $g(\theta)$ differentiable, then

$$Var_\theta(\delta) \geq g'(\theta)^T [I(\theta)]^{-1} g'(\theta) \tag{15}$$

where $I(\theta)$ is the Fisher Information. The equation above is for the general case where $\theta \in \mathbb{R}^n$. If $\theta \in \mathbb{R}$, then

$$Var_\theta(\delta) \geq \frac{[g'(\theta)]^2}{I(\theta)}. \tag{16}$$

Corollary 1.22. If X and Y are independent random variables with densities $p_\theta(x)$ and $q_\theta(y)$, and the Fisher Information regularity conditions are met, then

$$I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta). \tag{17}$$

2 Large Sample Theory

This section focuses on the different types of convergence.

2.1 Convergence of Random Variables

It's common to say that a sequence $x_n \to x$ if, and only if,$|x_n - x| \to 0$. When we are talking about a sequence of random variables, this notion starts to become more delicate. Remember that a random variable X_n is just a function from Ω to some space E which is usually $\mathbb{R}$. The one possible notion of convergence between random variables would be that $X_n : \Omega \to \mathbb{R}$ converges to $X : \Omega\mathbb{R}$ if for every $\omega \in \Omega$, we have that $|X_n(\omega) - X(\omega)| \to 0$. This is what is commonly known as everywhere convergence.

Now, we know that there are events in which the probability of them occurring is zero. Hence, we can modify the idea of convergence to something a bit weaker, for example, we can say that X_n converges to X almost everywhere if they converge everywhere except in a set of probability zero, i.e. X_n converges everywhere to X for all $\omega \in A \subset \Omega$ such that $P(A) = 1$, so it only does not converge to $\omega \in A^c$ where $P(A^c) = 0$.

Definition 2.1 (Almost Sure Convergence). A sequence of random variables Y_n converge almost surely to Y if there exists a set $M \subset \Omega$, such that $P(M) = 0$ and for every $\omega \in \Omega \setminus M$

$$|Y_n(\omega) \to Y(\omega)| \to 0. \tag{18}$$

An equivalent and more compact way of writing this is

$$\lim_{n \to \infty} P(Y_n \to Y) = 1. \tag{19}$$

Surprisingly, this notion of convergence does not correspond to a topology, i.e. almost sure convergence is not topological convergence.

Given $p \geq 1$, the $L^p(P)$ space consists of the equivalent class of functions such that $|X|^p$ is integrable, i.e.

$$\int_\Omega |X|^p dP < +\infty. \tag{20}$$

In this space, we have the following norm:

$$||X||_{L^p} := (E|X|^p)^{1/p}. \tag{21}$$

Theorem 2.2 (Convergence L^p implies L^q in finite space). If the measure space is finite (e.g. probability space), then convergence in $L^p(P)$ implies convergence in $L^q(P)$ if $p \geq q$.

Proof. Use Jensen's Inequality

$$(E|X|^q)^{\frac{p}{q}} \leq E|X|^{\frac{qp}{q}} = E|X|^p.$$

Now, just take the p-th root

$$((E|X|^q)^{\frac{p}{q}})^{1/p} = (E|X|^q)^{1/q} \leq (E|X|^p)^{1/p}.$$

□

Definition 2.3 (Convergence in Probability). Let X_n be a sequence of random variables. We say that X_n converges in probability to a random variable X if for every $\varepsilon > 0$

$$\lim_{n \to +\infty} P(|X_n - X| \geq \varepsilon) = 0. \tag{22}$$

In this case, we write $X_n \to_p X$.

Proposition 2.4. If $E[(X_n - X)^2] \to 0$, then $X_n \to_p X$.

Proof. Use Chebyshev's inequality. $\square$

Theorem 2.5 (Weak Law of Large Numbers). Let X_n be an i.i.d sequence of random variables with $Var(X_i) = \sigma^2 < +\infty$. Then

$$\bar{X}_n := \frac{X_1 + ... + X_n}{n} \to_p EX_1. \tag{23}$$

Note that there exists a stronger form in which convergence in probability occurs even with infinite variance.

Proof.

$$Var(\bar{X}_n) = \frac{1}{n^2}(Var(X_1) + ... + Var(X_n)) = \frac{\sigma^2}{n}.$$

Then, apply Chebyshev to $P(|\bar{X}_n - EX_1| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0$. $\square$

Proposition 2.6. If f is continuous at c and $X_n \to_p c$, then $f(X_n) \to_p f(c)$.

Proof. Since f is continuous at c, for any $\varepsilon > 0$, there is $\delta_\varepsilon > 0$, such that $|x-c| < \delta_\varepsilon \implies |f(x)-c| < \varepsilon$. Thus if $\omega_0 \in \{\omega \in \Omega : |X_n(\omega)-c| < \delta_\varepsilon\}$, then $X_n(\omega_0) = x_0 \in B_{\delta_\varepsilon}(c)$, hence $|f(X_n(\omega_0)) - f(c)| < \varepsilon$, implying that

$$\begin{aligned}\{\omega \in \Omega : |X_n(\omega) - c| < \delta_\varepsilon\} \subset \{\omega \in \Omega : |f(X_n(\omega)) - f(c)| < \varepsilon\} \implies \\ P(|X_n - c| < \delta_\varepsilon) \leq P(|f(X_n) - f(c)| < \varepsilon) \implies \\ P(|f(X_n) - f(c)| \geq \varepsilon) \leq P(|X_n - c| \geq \delta_\varepsilon).\end{aligned}$$

$\square$

Theorem 2.7 (Strong Law of Large Numbers). If $X_1, X_2...$ are i.i.d. with finite mean $\mu = EX_i$, and if $\overline{X}_n = (X_1 + ... + X_n)/n$, then

$$\overline{X}_n \to_{a.s} \mu. \tag{24}$$

2.2 Weak Convergence / Convergence in Distribution

Definition 2.8 (Weak convergence or Convergence in distribution). Let $X_n : \Omega \to E$ be a sequence of random variables. We say that X_n converges weakly, or in distribution, to $X : \Omega \to E$ if

$$\lim_{n \to +\infty} E[f(X_n)] = E[f(X)], \quad \forall f : E \to \mathbb{R} \in C_b. \tag{25}$$

We denote weak convergence by $X_n \rightharpoonup X$. Also, remember that C_b is the set of bounded continuous functions[1].

Theorem 2.9 (Convergence in distribution alternative definition). A sequence X_n converges weakly to X, if, and only if,

$$F_n(x) \to F(x), \quad \text{for all } x \text{ where } F \text{ is continous,} \tag{26}$$

where F_n is the cumulative distribution of X_n and F is the cumulative distribution of X.

Corollary 2.10. If g is a continuous function and $Y_n \rightharpoonup Y$, then

$$g(Y_n) \rightharpoonup g(Y). \tag{27}$$

Theorem 2.11 (Central Limit Theorem). Let $X_1, \dots$ be i.i.d. with $EX_i = \mu$ and $VarX_i = \sigma^2$, both finite. For $\overline{X}_n = (X_1 + \dots + X_n)/n$, then

$$(\overline{X}_n - \mu)\sqrt{n} \rightharpoonup N(0, \sigma^2) \iff \frac{(\overline{X}_n - \mu)}{\frac{\sigma}{\sqrt{n}}} \rightharpoonup N(0, 1). \tag{28}$$

Equivalently,

$$\overline{X}_n \rightharpoonup N(\mu, \frac{\sigma^2}{n}). \tag{29}$$

This means that the sample average converges to the real average as a normal distribution with decreasing variance.

The CLT does not say how fast the convergence happens. This is given in the following result.

Theorem 2.12 (Berry-Esseen). Let $X_1, \dots$ be i.i.d. with $EX_i = \mu$ and $VarX_i = \sigma^2$, both finite. For $\overline{X}_n = (X_1 + \dots + X_n)/n$, then

$$|\frac{\overline{X}_n - \mu}{\sqrt{n}} - \Phi(x/\sigma)| \leq \frac{3E[|X_1 - \mu|^3]}{\sigma^3\sqrt{n}} \tag{30}$$

Theorem 2.13 (Delta Method). Let $X_1, \dots$ be i.i.d. with $EX_i = \mu$ and $VarX_i = \sigma^2$, both finite. For $\overline{X}_n = (X_1 + \dots + X_n)/n$, and a function f differentiable at μ, then

$$\sqrt(n)(f(\overline{X}_n) - f(\mu)) \rightharpoonup N(0, [f'(\mu)]^2\sigma^2). \tag{31}$$

[1] The set of functions f are usually called *test functions*, since we can think of them as testing if a sequence random variables is indeed converging to X, i.e. *"to see if* $X_n \rightharpoonup X$, *just take each* f *in* C_b *and evaluate if* $E[f(X_n)] \to E[f(X)]$*; if the sequence passes this test for all functions* f*, then indeed it converges weakly to* X*"*.

2.3 Hierarchy of Convergences

Note that the notions of convergence presented are such that a.s. convergence is stronger than convergence in probability which is stronger than convergence in distribution. The result below illustrates this.

Theorem 2.14 (Convergence Hierarchy). If $X_n \to_{a.s.} X \implies X_n \to_p X$. And, $X_n \to_p X \implies X_n \rightharpoonup X$.

Proof. First, if $X_n \to_{a.s.} X$, then

$$0 = P(\limsup_{n\to\infty} |X_n - X| \geq \varepsilon) \geq \limsup_{n\to\infty} P(|X_n - X| \geq \varepsilon) \geq 0.$$

Hence $X_n \to_p X$. Now, if $X_n \to_p X$, for $f \in C_b(\Omega)$, we have $f(X_n) \to_p f(X)$[2], which implies $E|f(X_n) - f(X)| \to 0$[3]. Thus,

$$|Ef(X_n) - Ef(X)| \leq E|f(X_n) - f(X)|.$$

□

Proposition 2.15. If $X_n \rightharpoonup X$ and $P(X = c) = 1$ for some $c \in \mathbb{R}$, then $X_n \to_p X$.

The following two theorems present some useful properties when we have convergence of random variables.

Theorem 2.16 (Continuous Mapping Theorem). Let $f : \mathbb{R} \to \mathbb{R}$ a Borel measurable function such that $P(X \in D_f) = 0$, where D_f is the set of discontinuities of f. Then,

(i) $X_n \to_{a.s.} X \implies f(X_n) \to_{a.s.} f(X)$,

(ii) $X_n \to_p X \implies f(X_n) \to_p f(X)$,

(iii) $X_n \rightharpoonup X \implies f(X_n) \rightharpoonup f(X)$.

Theorem 2.17 (Slutsky's Theorem). Let (X_n) and (Y_n) be a sequence of random variables, and $X_n \rightharpoonup X$, $Y_n \to_p c$ for some $c \in \mathbb{R}$, then

(i) $X_n + Y_n \rightharpoonup X + c$,

[2] Prove this.

[3] Prove this.

(ii) $X_n Y_n \rightharpoonup cX$,

(iii) $\frac{X_n}{Y_n} \rightharpoonup \frac{X}{c}$, provided that $c \neq 0$.

Note the the previous theorems allow us to prove the following.

Example 2.1. Let $X_1, ..., X_n$ i.i.d. with finite mean μ and variance σ^2. Then,

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \rightharpoonup N(0, 1), \tag{32}$$

where $S := \sum_{i=1}^{n} \frac{(\overline{X} - X_i)^2}{\sqrt{n-1}}$, i.e. if we sample enough times, using the sample standard deviation as an approximation of the real variance, we converge to a standard Normal. To prove this, note that $S/\sqrt{n} \to_p \sigma^2/\sqrt{n}$, and use Slutsky.

3 Maximum Likelihood Estimation

3.1 Formalization of the MLE

Definition 3.1 (Likelihood). Let X be a random variable in the probability space $(\Omega, \mathcal{B}, P)$, with density function p_θ with respect to a measure (e.g. Lebesgue), and a parameter θ that characterizes the distribution. The **likelihood** function is $\ell_x(\theta) = p_\theta(x)$, sometimes abbreviated as $\ell(\theta)$.

In other words, the likelihood function is the same as the density function, but the variable is θ instead of x. Another way to think of the likelihood is as if $\ell_x(\theta) = p(\theta \mid x)$, that is the conditional density.

Definition 3.2 (Maximum Likelihood Estimator). Given an observed data x, the maximum likelihood estimator (MLE) is

$$\hat{\theta} \in \arg\max_{\theta \in \Theta} \ell_x(\theta). \tag{33}$$

It can be thought as the following. Suppose that you threw a coin 10 times, and observed 5 heads and 5 tails. The density function is then given by the Binomial distribution:

$$p_\theta(X = 5) = \binom{10}{5} \theta^5 (1 - \theta)^5, \tag{34}$$

where θ is the parameter that represents the coin bias. Now, the MLE is the θ such that $p_\theta(X = 5)$ is maximized, which, in this case, is 0.5.

If $X_1, X_2, ..., X_n$ are i.i.d. with density $p_\theta, \theta \in \Theta$, the likelihood function will be the product of each density. Thus, since taking the logarithm won't change the maximization parameter $\hat{\theta}$, it's common to talk about the log-likelihood as shown below

$$\log \ell(\theta) = \log \prod_{i=1}^{n} p_\theta(X_i) = \sum_{i=1}^{n} \log p_\theta(X_i). \tag{35}$$

A way to quantify how similar are two probability distributions is given by what is called the Kullback-Leibler (KL) Divergence.

Definition 3.3 (Kullback-Leibler Divergence). Let μ, ν be probability measures in $(\Omega, \mathcal{B})$, then

$$\mathrm{KL}(\mu, \nu) = E_\mu \left[\log \frac{\mu}{\nu}\right] = \int_\Omega \log \frac{d\mu}{d\nu} d\mu = \int_\Omega \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} d\nu \tag{36}$$

where $\frac{d\mu}{d\nu}$ is the Radon-Nikodym derivative and we consider $\lim_{x \to 0^+} x \log x = 0$ (note that this is necessary for KL to be well defined). Note that if $\mu, \nu \ll \lambda$, then for $p = \frac{d\mu}{d\nu}$ and $q = \frac{d\nu}{d\mu}$, then

$$\mathrm{KL}(\mu, \nu) = \int_\Omega p \log \frac{p}{q} d\lambda. \tag{37}$$

Definition 3.4 (Consistent Estimators). A sequence of estimators δ_n is consistent for $g(\theta)$ in a parametric family $\mathcal{P}$, if for any $\theta \in \Theta$

$$\delta_n \to_{P_\theta} g(\theta). \tag{38}$$

Note that the likelihood is a random function, since it's a function that varies according to the value obtained from observing the random variable $X : \Omega \to \mathbb{R}^n$. Let's formalize this notion of a random function.

Definition 3.5 (Continuous Random Function Indexed on X). Let $K \subset \mathbb{R}^n$, and $X : \Omega \to \mathbb{R}^p$ a random variable.

$$W_X(t) = h(t, X), \ t \in K, \tag{39}$$

where $h : K \times \mathbb{R}^n \to \mathbb{R}^p$ and h is continuous in t for every x.

Note that W_X is a random function, i.e. for every $\omega \in \Omega$ we get a value of $X(\omega) = x$, which gives a different function $W_x(t) \in C(K)$. Figure 1 illustrates the process of generating the indexed random function.

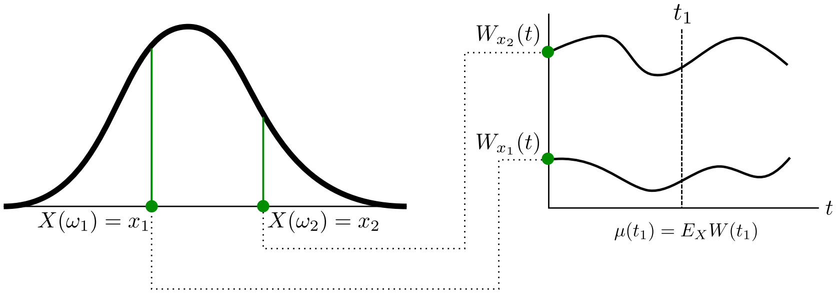

Figure 1: Example of random function W_X.

Definition 3.6 (Mean of Random Function). Let W be a continuous random function in K. The mean of this random function is

$$\mu(t) = EW(t),\ t \in K. \tag{40}$$

Lemma 3.7. If $E||W||_\infty < +\infty$, the μ is continuous, where

$$||W||_\infty := \sup_{t \in K} ||W(t)||. \tag{41}$$

These definitions are mostly used in order to rigorously define the likelihood and the MLE, which is a point in the likelihood function, more specifically, it's the t^* that maximizes the likelihood.

3.2 Consistency of the MLE

Theorem 3.8 (Compact MLE consistency). Let $W(\omega) := \log\left(\frac{f_\omega(X)}{f_\theta(X)}\right)$. If Θ is compact, $E_\theta||W||_\infty < +\infty$, $f_\omega(x)$ is a continuous function of ω a.e. x, and $P_\omega \neq P_\theta$ for all $\omega \neq \theta$, then under P_θ, $\hat{\theta}_n \to_p \theta$, where $\hat{\theta}_n$ is the maximum likelihood for $X_1, ..., X_n$ i.i.d with distribution P_θ.

This theorem is implying that if the parameter space is compact and the likelihood functions are continuous, then the MLE is consistent.

The next theorem also shows the consistency of the MLE, but for the case where the parameter space is not limited.

Theorem 3.9 (MLE consistency). Let $W(\omega) := \log\left(\frac{f_\omega(X)}{f_\theta(X)}\right)$. If $\Theta = \mathbb{R}^p$, $f_\omega(x)$ is a continuous function of ω for a.e x, $P_\omega \neq P_\theta$ for all $\omega \neq \theta$ and $f_\omega(x) \to 0$ as $|\omega| \to +\infty$. If $E_\theta \|\mathbb{1}_C W\|_\infty < +\infty$ for every compact $C \subset \mathbb{R}^p$, and if $E_\theta \sup_{|\omega|>a} W(\omega) < +\infty$ for some $a > 0$, then under P_θ, $\hat{\theta}_n \to_p \theta$.

Both theorems show that the MLE is a consistent estimator, but they say nothing about the rate of convergence. This is shown in the following theorem.

Theorem 3.10 (Rate of Convergence MLE). Assume:

1. $X_1, X_2, \ldots$ are i.i.d with density p_θ, for $\theta \in \Theta \subset \mathbb{R}$;
2. The set $A = \{x : p_\theta(x) > 0\}$ is independent of θ;
3. For every $x \in A$, $\partial^2 p_\theta(x)/\partial\theta^2$ exists and is continuous in θ;
4. Let $W(\theta) = \log p_\theta(X)$. The Fisher information $I(\theta)$ for a single observation exists, is finite, and $I(\theta) = E_\theta W'(\theta)^2 = -E_\theta W''(\theta)$;
5. For every θ in the interior of Θ, there exists $\varepsilon > 0$ such that $E_\theta \|\mathbb{1}_{[\theta-\varepsilon,\theta+\varepsilon]}\|_\infty < +\infty$;
6. The MLE $\hat{\theta}_n$ is consistent.

Then for any $\theta \in \Theta^\circ$,

$$\frac{\hat{\theta}_n - \theta}{\sqrt{n}} \rightharpoonup N\left(0, \frac{1}{I(\theta)}\right). \tag{42}$$

3.3 Confidence Intervals

Definition 3.11 (Confidence Interval). If δ_0, δ_1 are statistics, then the random interval (δ_0, δ_1) is called a $(1-\alpha)$ confidence interval for $g(\theta)$ if

$$P_\theta(g(\theta) \in (\delta_0, \delta_1)) \geq 1 - \alpha, \tag{43}$$

for all $\theta \in \Theta$. A set $S = S(X)$ is called a $1 - \alpha$ confidence region for $g(\theta)$ if

$$P_\theta(g(\theta) \in S) \geq 1 - \alpha, \tag{44}$$

for all $\theta \in \Theta$. Note that the confidence intervals vary with the observed data, hence, it is itself a random variable.

Definition 3.12 (Pivots). A variable X is a pivot if the distribution of X does not depend on the unknown parameters. They are usually useful for the construction of confidence intervals.

Example 3.1. Let $X_1, ...X_n$ i.i.d from $N(\mu, \sigma^2)$, thus

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$
$$V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

are pivots. Note that the distributions are not dependent on the parameters, but the functions are. Let's construct the confidence intervals such that $P_\theta(\mu \in (\delta_0, \delta_1)) = 1 - \alpha$. To do this, note that $T = \frac{Z}{\sqrt{(V/(n-1))}}$ is pivotal also, with t-distribution of $n-1$ degrees of freedom.

$$(\delta_0, \delta_1) := \left(\overline{X}_n - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \overline{X}_n + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right), \tag{45}$$

where $t_{\alpha/2,n-1}$ is the $\alpha/2$ percentile of the t-distribution with $n-1$ degrees of freedom, i.e.

$$P(T \geq t_{\alpha/2,n-1}) = \int_{t_{\alpha/2,n-1}}^{\infty} p_{T_{n-1}}(x)d(x) = \frac{\alpha}{2} \tag{46}$$

4 Hypothesis Testing

4.1 Likelihood ratio

Two hypothesis H_0 (null hypothesis) and H_1 (alternative hypothesis) consists of subsets of the parameter space Θ. In other words, if $H_0 \subset \Theta$ is the true hypothesis, it means that $X \sim P_\theta$ such that $\theta \in H_0$, and similarly for the alternative hypothesis. Hence, Hypothesis testing consists in defining a methodology in order to decide which hypothesis should be "picked".

It's possible to reformulate this problem in our decision theory framework. This would be equivalent to establishing a loss function

4.2 Generalized Likelihood Ratio

The Pearson-Neymman Lemma proves that the Likelihood Ratio is the best for the case of simple hypothesis (i.e. $H_0 = \{\theta_0\}$ and $H_1 = \{\theta_1\}$). Yet, this test may be extended for more general hypothesis by using the $\sup_{\theta \in H_0}$ or the MLE (when the maximum exists). Thus, the Generalized Likelihood Ratio test uses

$$\lambda(\theta) := \frac{\sup_{\theta \in H_1} \ell(\theta)}{\sup_{\theta \in H_0} \ell(\theta)}. \tag{47}$$

5 Bootstrap

6 Inequalities Galore

This section is a collection of useful inequalities related to Statistics. Here are inequalities that are not directly related to statistics, and are more "auxiliary", for example, the famous Cramér-Rao inequality is presented in Section 1.4.

Theorem 6.1 (Jensen's Inequality). Let $(\Omega, \mathcal{F}, P)$ be a probability space, $f : \mathbb{R} \to \mathbb{R}$ a convex function and $X : \Omega \to \mathbb{R}$ a random. Then

$$f(EX) \leq Ef(X). \tag{48}$$

If f is strictly convex, then the inequality is strict unless X is almost surely constant.

Theorem 6.2 (Cauchy-Schwarz). The famous Cauchy-Schwarz inequality has the following form

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \sum_{i=1}^{n} a_i^2 \sum_{j=1}^{n} b_j^2. \tag{49}$$

We can extend this for vectors and inner products. Let $\boldsymbol{u}, \boldsymbol{v}$ be vectors in $\mathbb{R}^n$, then

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle \leq \sqrt{\langle \boldsymbol{u}, \boldsymbol{u} \rangle}\sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle} = ||\boldsymbol{u}|| \cdot ||\boldsymbol{v}|| \tag{50}$$

.

Theorem 6.3 (Covariance Inequality). Let X and Y be random variables, then

$$\mathsf{Cov}(X, Y)^2 \leq \mathsf{Var}(X)\mathsf{Var}(Y). \tag{51}$$

Proof. If the random variables have finite variance, then $E[XY]$ is an inner product, hence by Cauchy-Schwarz

$$E[XY]^2 \leq E[X^2]E[Y^2]$$

Note that for $W = X - E[X]$ and $Z = Y - E[Y]$ we have

$$E[WZ]^2 = \mathsf{Cov}(X, Y)^2 \leq E[W^2]E[Z^2] = \mathsf{Var}(X)\mathsf{Var}(Y),$$

and we conclude the proof of the desired inequality. □

Theorem 6.4 (General form of Chebyshev's Inequality). Let X be a random variable and $g(X) > 0$ be a non-decreasing function on $\mathbb{R}$. Then, for any x,

$$P(X \geq x) \leq \frac{Eg(X)}{g(x)}. \tag{52}$$

Note that if $Y = |X - EX|$ and $g(Y) = Y^2$, we get the standard Chebyshev's Inequality

$$P(Y \geq x) = P(|X - EX| \geq x) \leq \frac{E|X - EX|^2}{x^2}. \tag{53}$$

7 Distributions Zoo

This section contains a collection of important distributions and some of their properties and derivations.

Normal Distribution For $X \sim N(\mu, \sigma^2)$ we have the density function

$$p(x) = \frac{1}{\sqrt{\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{54}$$

A $N(0, 1)$ is called standard Normal distribution. From a standard Normal, if $X \sim N(0, 1)$, for $Z = \sigma X + \mu$ then $Z \sim N(\mu, \sigma^2)$.

χ^2 Distribution The χ^2_k is the sum of the squares of k independent random variables with standard Normal distribution. The density function is

$$p(x) = \frac{x^{\frac{k}{2}-1} e^{-x/2}}{2^{k/2}\Gamma(\frac{k}{2})} \mathbb{1}_{(0,+\infty)}(x). \tag{55}$$

Gamma Distribution The Gamma distribution is a generalization of the χ^2 distribution, i.e. a χ^2_1 is a Gamma with 1 degree of freedom.

F-Distribution The F-Distribution can be obtained as the reason between two $\chi^2_{k_1}$ and $\chi^2_{k_2}$ random variables, i.e. $F(k_1, k_2) = \frac{X/k_1}{Y/k_2}$ where $X \sim \chi^2_{k_1}$ and $Y \sim \chi^2_{k_2}$.

t-Distribution

Let $Z \sim N(0, 1)$, $V \sim \chi^2_{n-1}$, then

$$T = \frac{Z}{\sqrt{\frac{V}{n-1}}} \sim t_{n-1}. \tag{56}$$

The density function for the distribution is

$$p_T(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\nu/2)(1 + \frac{x^2}{\nu})^{(\nu+1)/2}}, \tag{57}$$

where $\nu = n - 1$. Also, remember that

$$\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} dx, \; z > 0. \tag{58}$$

Which implies that $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$.

References

[1] James Gentle. *Theory of Statistics*. 2020. URL https://mason.gmu.edu/~jgentle/books/MathStat.pdf.

[2] R.W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer New York, 2010. ISBN 9780387938394. URL https://books.google.com.br/books?id=aVJmcega44cC.

[3] Per Martin-Löf. The definition of random sequences. *Information and control*, 9(6):602–619, 1966.

[4] Jun Shao. *Mathematical Statistics*. Springer-Verlag New York Inc, 2nd edition, 2003.