Institute of Mathematics, EPFL MATH524

Chapter 1: Statistical foundations

Lecturer: Rajita Chandak *Spring 2025*

Here we provide an overview of basic probability theory and statistics that serves as a rough baseline upon which this course will be built. Students are expected to be familiar with the material presented here.

1 Probability theory

We start by review some basic probability concepts that will be used repeatedly throughout this course.

1.1 Random variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space (discrete or continuous).

Definition 1 (Random variable)**.** *A random variable is a real-valued function* $X : \Omega \to \mathbb{R}$

1.2 Properties of random variables

Given a random variable X on the probability space $(\Omega, \mathcal{F}, P)$, its distribution p is the probability measure on $\mathbb{R}$, the range of X, defined as

$$p(A) = \mathbb{P}(X \in A), \qquad \forall \text{ “nice” } A \subset \mathbb{R}.$$

where the collection of “nice” subsets has to do with measurability conditions. We will not worry too much about these constraints in this course. When the use of p may be unclear we will use p_X to emphasize dependence on the random variable. If μ is the probability measure on $\mathbb{R}$, we will use the notation $X \sim \mu$ to denote that X is distributed as μ.

The cummulative distribution function (CDF) of X is given by the function $F = F_X$, $F : \mathbb{R} \to [0, 1]$ where $F(a) = \mathbb{P}(X \leq a)$.

1.3 Comparing random variables

There are several different notions of equality of random variables:

1. Two R.V.s $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are identical if $X(\omega) = Y(\omega$ for all $\omega \in \Omega)$.
2. Two R.V.s $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are $\mathbb{P}$-almost surely equal if $\mathbb{P}(X = Y) := \mathbb{P}(\{\omega \in \Omega : X(\omega = Y(\omega))\}) = 1$.

3. Two discrete random variables X and Y are equal in distribution is for every $A \subset \mathbb{R}$, $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$. **Note:** This comparison also applies when X and Y are defined on two different probability spaces by equality of the two probability measures on all $A \subset \mathbb{R}$.

There are also many ways to order random variables.

1. Given two R.V.s $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$, X is said to be greater than Y if $X(\omega) \geq Y(\omega)$ for all $\omega \in \Omega$.
2. X is $\mathbb{P}$-almost surely greater than Y if $\mathbb{P}(X \geq Y) := \mathbb{P}(\{\omega \in \omega : X(\omega \geq Y(\omega))\}) = 1$.
3. X stochastically dominates Y (often denoted as $X \overset{d}{\geq} Y$) if $\mathbb{P}(X \geq a) \geq \mathbb{P}(Y \geq a)$.

1.4 Functions of random variables

The distribution of a random variable captures all "statistics" while ignoring details of particular realizations of the random variable. In practice, two distributions can be hard to compare. It is, therefore, useful to have a collection of numbers that describe salient features of the distribution that are easier to interpret. A few of the most common quantities used to understand probability distributions are listed here.

1. The expectation operator maps the random variable through the Lebesgue-Stieljes integral and can be written as

$$\mathbb{E}[g(X)] = \int g(x) \mathrm{d}F(x),$$

   where the integral becomes a sum when X is discrete. The limits of integration are determined by $\text{Range}(g(X)) \cap \mathbb{R}$.
2. The median is any number $a \in \mathbb{R}$ such that

$$\sum_{x \geq a} p(x) \geq \frac{1}{2} \quad \text{and} \quad \sum_{x \leq a} p(x) \geq \frac{1}{2}$$

3. The mode is defined as any number $c \in \mathbb{R}$ s.t. $p(c) \geq p(x)$ for all $x \in \mathbb{R}$.
4. The variance operator captures fluctuations of the random variable around the mean:

$$\mathbb{V}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2].$$

The expectation (mean) of a random variable tends to be a very useful quantity to work with. Some key properties of this operator are listed here:

1. (Non-negativity) If $X \geq 0$ *a.s.*, then $\mathbb{E}[X] \geq 0$.
2. (Linearity) The expectation operator is linear in the sense that $\mathbb{E}[aX + Y] = a\mathbb{E}[X] + \mathbb{E}[Y]$ whenever the RHS is well-defined.

3. (Monotonicity) If $X \leq Y a.s.$ and $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are well-defined $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
4. (Non-degeneracy) if $\mathbb{E}[|X|] = 0$ then, $X = 0 a.s.$
5. Given an event A, the indicator function $\mathbf{1}_A$ is defined as

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if} \omega \in A \\ 0 & \text{otherwise.} \end{cases}$$

   Then,

$$\mathbb{P}(A) = \mathbb{E}[\mathbf{1}_A].$$

   This statement is often useful since expectations tend to be easier to compute than probabilities.
6. (Functionals) If X has density f and $g(\cdot)$ is some function, $\mathbb{E}[g(X)] = \int g(x)f(x)\mathrm{d}x$.

1.5 Independence

Two events A and B are independent (with respect to the probability measure $\mathbb{P}$) if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

An infinite collection of events $\{A_i\}$ is independent if every finite sub-collection is independent. A collection of discrete r.v. $X_1, X_2, \ldots X_k$ are mutually independent if

$$\mathbb{P}(X_1, \ldots X_k) = \prod_{i=1}^{k} \mathbb{P}(X_i).$$

i.e., X_i are mutually independent iff their joint density is a product of its marginals.

If random variables are dependent, a useful measure of their dependence is the covariance.

$$\mathbb{C}\mathrm{ov}(X, Y) :== \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

The following properties of the covariance are important:

1. $\mathbb{C}\mathrm{ov}(X, X) = \mathbb{V}(X)$.
2. $\mathbb{C}\mathrm{ov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
3. If X and Y are independent, $\mathbb{C}\mathrm{ov}(X, Y) = 0$. The converse is not true.
4. Covariance is bi-linear. i.e.,

$$\mathbb{C}\mathrm{ov}(\sum_i X_i, \sum_j Y_j) = \sum_{i,j} \mathbb{C}\mathrm{ov}(X_i, Y_j)$$

1.6 Conditional expectation

We start by defining conditional probability of an event A given B.

Definition 2 (Conditional probability)**.** *Consider a probability space* $(\Omega, \mathcal{F}, P)$ *and* $A, B \in \mathcal{F}$ *with* $P(B) > 0$*. Then, the conditional probability is defined as*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

To better understand conditional probabilities, solve the following exercise:

Exercise 1. *The mapping* $\Omega \ni A \mapsto P(A|B) \in [0, 1]$ *defines a probability measure.*

In standard notation, the measure $P(\cdot|B)$ denotes the modified estimate of the probabilities of events conditioned on the "information" that event B occurred. Thus, one can also think of the original probability measure as a conditional probability on Ω, i.e.,, $P(B) = P(B|\Omega)$. Note that by definition, if A and B are independent, $P(A|B) = P(A)$. This means that the probability of A does not change when given any information about B.

The definition also implies that we can write $P(A \cap B) = P(A|B)P(B)$. Which leads to the fact that if we have a disjoint partition of Ω given by $\{B_i\}$ with $P(B_i) > 0$ for all i, we can write

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i).$$

This relation is also known as the *law of total probability*.

This notion of conditional probabilities can now be used to understand the fact that the expectation of a r.v. depends on the probability measure assigned to the sample space. A notion that becomes crucial in many different applications including variance reduction techniques and large deviation theory.

Suppose we know that event $A \in \mathcal{F}$ occurs and that $\mathbb{P}(A) > 0$. Then, we know that the original probability measure is in fact the conditional probability $\mathbb{P}(\cdot|A)$, which defines a new probability on Ω. We also know that the distribution is now altered from $p_X(\cdot) = \mathbb{P}(\cdot)$ to $p_{X|A}(\cdot) = \mathbb{P}(\cdot|A)$. The conditional expectation of X given event A is simply the expectation of X under this new conditional probability distribution. We define this formally below.

Definition 3 (Conditional expectation)**.** *The conditional expectation of* X *given* A *when* $\mathbb{P}(A) > 0$ *is defined by*

$$\mathbb{E}_{\mathbb{P}}[X|A] = \int x \mathrm{d}p_{X|A}.$$

Similar definitions and arguments can be made for conditioning on another r.v. instead of an event. In particular, we can show that the expectation of X conditioned on Y is only a function of Y.

Definition 4. $\mathbb{E}[X|Y] = h(Y)$ *where* h *is defined by*

$$h(y) = \mathbb{E}[X|Y = y] \ \forall y \in Range(Y)$$

It is important to realize here that $\mathbb{E}[X|Y = y]$ is a constant but $\mathbb{E}[X|Y]$ is a random variable (defined precisely by $h(Y)$). One can also think of $\mathbb{E}[X|Y]$ as the best approximation of X given Y.

Next, we list some basic properties of conditional expectations.

1. (Linearity) $\mathbb{E}[aX + bZ|Y] = a\mathbb{E}[X|Y] + b\mathbb{E}[Z|Y]$.
2. (Tower property) $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$, $\mathbb{E}[\mathbb{E}[X|Y]|g(Y)] = \mathbb{E}[X|g(Y)]$ and $\mathbb{E}[\mathbb{E}[X|g(Y)]|Y] = \mathbb{E}[X|g(Y)]$.
3. (Independence) If $X \perp\!\!\!\perp Y$, $\mathbb{E}[X|Y] = \mathbb{E}[X]$

Finally, we introduce the notion of conditional independence.

Definition 5 (Conditional independence)**.** *Given a probability space* $(\Omega, \mathcal{F}, P)$ *and an event* $A \in \mathcal{F}$ *with* $P(A) > 0$*, X and Y are said to be conditionally independent given A if for every Borel sets* B_1 *and* B_2 *in* $\mathbb{R}$*,*

$$P(X \in B_1, Y \in B_2|A) = P(X \in B_1|A)P(Y \in B_2|A).$$

Note that conditional independence is a natural extension of independence – it is independence with respect to the conditional probability measure instead of the original probability measure. This definition easily extends to multiple collections of random variables. Conditional independence with respect to a r.v. works in the same way as conditional independence with respect to sets. Proving the following results is a good exercise in understanding conditional independence.

Exercise 2. *Show that* X_1 *and* X_2 *are independent conditional on Y if for any nice bounded function* $f_1 : \mathbb{R} \mapsto \mathbb{R}$ *and* $f_2 : \mathbb{R} \mapsto \mathbb{R}$*,*

$$\mathbb{E}[f_1(X_1)f_2(X_2)|Y] = \mathbb{E}[f_1(X_1)|Y]\mathbb{E}[f_2(X_2)|Y].$$

In particular, X_1 *and* X_2 *are independent if and only if for every* f_1, f_2 *as above,*

$$\mathbb{E}[f_1(X_1)f_2(X_2)|Y] = \mathbb{E}[f_1(X_1)|Y]\mathbb{E}[f_2(X_2)|Y].$$

1.7 L^p spaces

Given a measurable function g on any σ-finite measure space $(\Omega, \mathcal{F}, \mu)$, for $p \geq 1$, the L^p norm of g, denotes as $\|g\|_p$ is defined as

$$\|g\|_p = \left(\int_\Omega |g|^p \mathrm{d}\mu \right)^{1/p}.$$

Note that the RHS is always well defined (but possibly equal to infinity) because it is an integral of a non-negative function with respect to a non-negative measure. The collection of measurable

functions whose L^p norm is finite is known as the L^p space. L^0 is often used to denote the collection of all measurable functions. For probability spaces, this notion is also known as the *moments* of a random variable. i.e., the p-th moment of a r.v. is defined as

$$\mathbb{E}[|X|^p] = \int |X|^p \mathrm{d}\mathbb{P} = \|X\|_p^p.$$

1.8 Inequalities

Some basic inequalities that are used frequently in analyzing random variables are stated in this section.

Lemma 1 (Markov's inequality)
If X is a nonnegative random variable, then

$$\forall a \in \mathbb{R}, \quad \mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Another more refined inequality provides bounds on deviations of X from its mean.

Lemma 2 (Paley-Zygmund inequality)
If X is nonnegative with $\mathbb{E}[X^2] \neq 0$, then, for any $0 \leq \alpha \leq 1$, we have

$$\mathbb{P}(X > \alpha\mathbb{E}[X]) \geq (1-\alpha)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

In particular,

$$\mathbb{P}(X \geq 0) \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

A simple application of Markov's inequality can provide a useful bound on the deviation of X from its mean.

Lemma 3 (Chebyshev's inequality)
For any $a > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| > a) \leq \frac{\mathbb{V}(X)}{a^2}.$$

The proof of Lemma 3 follows directly from Markov's inequality applied to $(X - \mathbb{E}[X])^2$.

The following inequality is a useful bound for when we only have access to marginal densities.

Lemma 4 (Cauchy-Schwarz inequality)
Suppose X and Y are s.t. $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$. Then,

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

Finally, a very useful inequality for relating the expectation of random variables to that of convex functions of random variables is given by Jensen's inequality.

Lemma 5 (Jensen's inequality)
If g is a convex function and X has well-defined expectation,

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

Lemma 6 (Holder's inequality)
For any two random variables X and Y,

$$\mathbb{E}[|XY] \leq \mathbb{E}[|X|^p]^{1/p}\mathbb{E}[|Y|^q]^{1/q}.$$

Furthermore,

$$\mathbb{E}[|X|^r] \leq (\mathbb{E}[|X|^s])^{\frac{r}{s}}. \tag{1}$$

In particular, if the s^{th} absolute moment of X is finite, then the r^{th} absolute moment is also finite.

Note that Cauchy-Schwarz is a particular case of Holder's inequality.

Exercise 3. *Show that (1) also follows from Jensen's inequality.*

1.9 Convergence

Let X_n be a sequence of r.v.s defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The three most common types of convergence of X_n to a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ are the following.

Definition 6 (Convergence in distribution)**.** *Written as $X_n \xrightarrow{d} X$, $X_n \sim F_n$ converges in distribution to $X \sim F$ if $lim_{n\to\infty}F_n(x) = F(x)$ for all x for which F is continuous.*

Definition 7 (Convergence in probability)**.** *Written as $X_n \xrightarrow{p} X$, X_n converges in probability to X if $lim_{n\to\infty}\mathbb{P}(|X_n - X| > \varepsilon) = 0$ for all $\varepsilon > 0$.*

Definition 8 (Convergence almost surely (a.s.))**.** *Written as $X_n \xrightarrow{a.s.} X$, X_n converges almost surely to X if $\mathbb{P}(\{\omega \in \Omega : lim_{n\to\infty}X_n(\omega) = X(\omega)\}) = 1$.*

The following proposition outlines the relationship between each of these types of convergence.

Proposition 1 (Convergence of random variables)
Let X_n be a sequence of r.v.s and X an r.v. defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then,

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$$

None of the converse statements hold in general.

The weak and strong law of large numbers are two results that showcase the power of convergence in probability and almost-sure convergence. We first state the weak law of large numbers.

Theorem 1 (Weak Law of Large Numbers)
Suppose $\{X_i\}$ is a sequence of pairwise independent, identically distributed random variables with a finite mean μ. Let $S_n = \sum X_i$. Then, as $n \to \infty$, $(S_n - \mathbb{E}[S_n])/n$ converges to 0 in L^1, and therefore,

$$\frac{S_n - \mathbb{E}[S_n]}{n} \xrightarrow{p} 0,$$

and is equivalent to the statement that for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n - \mathbb{E}[S_n]}{n}\right| > \varepsilon\right) = 0$$

The identically distributed assumption in the WLLN statement can be relaxed to requiring that $\{X_i\}$ be uniformly integrable.

Now we state the strong version of the LLN.

Theorem 2 (Kolmogorov's Strong Law of Large Numbers)
Let $\{X_i\}$ be a sequence of i.i.d. r.v. defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $S_n = \sum X_i, n \in \mathbb{N}$. If $\mathbb{E}[|X_1|] < \infty$, then $\mathbb{P}$-almost surely,

$$\lim_{n\to\infty} \frac{S_n}{n} = \mathbb{E}[X_1]. \tag{2}$$

Conversely, if $\mathbb{P}(\limsup_n n^{-1}|S_n| < \infty) > 0$, then $\mathbb{E}[|X_1|] < \infty$ and (2) holds $\mathbb{P}$-almost surely.

A direct converse of first statement in Theorem 2 would mean that if $\mathbb{E}[|X_1|] = \infty$, then (2) does not hold, but this statement does not hold any meaning since the RHS of (2) is not well-defined in this case. This is why the second statement in the theorem is stated differently, implying that if $\mathbb{E}[|X_1|] = \infty$, then with positive probability S_n/n has no finite limit. i.e.,

$$\mathbb{E}[|X_1|] = \infty \Rightarrow \mathbb{P}(\limsup_n n^{-1}|S_n| < \infty) = 0.$$

Now, we state two popular forms of the central limit theorem (CLT).

Theorem 3 (Lindeberg-Feller CLT)
Let $\{X_i\}$ be a sequence of i.i.d. r.v. with $\mathbb{E}[X_1] = \mu$ and $\mathbb{V}(X_1) = \sigma^2 < \infty$. Then,

$$\frac{\sqrt{n}(X_i - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1).$$

Theorem 4 (Lyapunov's CLT)
Let $\{X_i\}$ be a sequence of independent r.v.s with $\mathbb{E}[X_i] = \mu_i$ and $\mathbb{V}(X_i) = \sigma_i^2 < \infty$ such that for some $\delta > 0$,

$$\frac{1}{s_n^{2+\delta}} \sum_i \mathbb{E}[|X_i - \mu_i|^{2+\delta}] \to 0 \quad \text{as } n \to \infty,$$

where $s_n^2 = \sum_i \sigma_i^2$. *Then,*

$$\frac{1}{s_n}\sum_i (X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0,1).$$

Note that in order to loosen the i.i.d. assumption from the Lindeberg-Feller CLT, the Lyapunov CLT requires slightly more than bounded second moment.

Finally, Slutsky's theorem allows for combining the LLN and CLT results.

Theorem 5 (Slutsky's Theorem)
Let X_n *and* Y_n *be sequences of r.v.s and* $c \in \mathbb{R}$. *Then,*

1. *If* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c$, *then* $X_n Y_n \xrightarrow{d} cX$.
2. *If* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c \neq 0$, *then* $X_n / Y_n \xrightarrow{d} X/c$.
3. *If* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c$, *then* $X_n + Y_n \xrightarrow{d} X + c$.

It is important to understand the convergence assumptions for Slutsky's results to hold. For example, it is **false** in general that $X_n \xrightarrow{d} X, Y_n \xrightarrow{d} Y \Rightarrow X_n + Y_n \xrightarrow{d} X + Y$ or $X_n Y_n \xrightarrow{d} XY$. However, both conclusions are true under the much stronger assumption that $(X_n, Y_n) \xrightarrow{d} (X, Y)$ or if X_n and Y_n are independent of each other.

Theorem 6 (Delta method)
If $\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, *then*

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

for any function g *that is differentiable at* μ *and* $g'(\mu) \neq 0$.

1.10 Big-O and little-o notation

O notation in mathematics is used to bound sequences, however, the idea behind this notation is very similar to their application in computer science for computational complexity. There are two key terms here:

Definition 9 (*Big O*)**.** *Given two strictly positive sequences* a_n *and* b_n,

$$a_n = O(b_n) \iff \limsup_{n\to\infty} \frac{a_n}{b_n} \leq C,$$

for some $C > 0$. $a_n = O(1)$ *implies that* a_n *is bounded.*

This can be understood as a_n is *not larger than* b_n. In particular, if $a_n, b_n \to 0$, then this means that a_n does not decrease at a slower rate than b_n.

Definition 10 (*Little-o*)**.** *Given two strictly positive sequences a_n and b_n,*

$$a_n = o(b_n) \iff \limsup_{n\to\infty} \frac{a_n}{b_n} = 0.$$

$a_n = o(1)$ *implies that* $a_n \to 0$.

This means that a_n is *smaller than* b_n. If $a_n, b_n \to 0$, this implies that a_n goes to 0 faster than b_n.

It should be clear from the definitions that little-o implies big-O. The following properties of this notation should be clear:

Proposition 2
Given two strictly positive sequences $a_n, b_n \to 0$,

1. $CO(a_n) = O(a_n), Co(a_n) = o(a_n)$ *for any* $C \in \mathbb{R}$.
2. $O(a_n) + O(b_n) = O(a_n + b_n), o(a_n) + o(b_n) = o(a_n + b_n)$.
3. $O(a_n)O(b_n) = O(a_n b_n), o(a_n)o(b_n) = o(a_n b_n)$.
4. $O(a_n) + o(b_n) = O(a_n + b_n), O(a_n)o(b_n) = o(a_n b_n)$.
5. $o(1)O(a_n) = o(a_n)$.
6. $a_n^p = o(a_n^q)$ *for* $p > q \geq 0$.
7. $(a_n + b_n)^p = O(a_n^p + b_n^p)$

Both the big-O and little-o notation are purely deterministic, which is useful when we don't have to worry about randomness in the sequences. However, as is often the case in statistical theory, sequences may have inherent randomness for which this notation is not applicable. In that case, we introduce stochastic versions, identified typically by the p subscript.

Definition 11 (Little-o_p)**.** *Given a strictly positive non-random sequence a_n and random variable X_n,*

$$\begin{aligned} X_n = o_p(a_n) &\iff \frac{|X_n|}{a_n} \xrightarrow{p} 0 \\ &\iff \lim_{n\to\infty} \mathbb{P}(\frac{|X_n|}{a_n} > \varepsilon) = 0, \quad \forall \varepsilon > 0. \end{aligned}$$

To say $X_n \xrightarrow{p} 0$ *we write* $X_n = o_p(1)$

Definition 12 (Big-O_p)**.** *Given a strictly positive non-random sequence a_n and random variable X_n,*

$$\begin{aligned} X_n = O_p(a_n) &\iff \forall \varepsilon > 0, \exists C_\varepsilon > 0, n_o(\varepsilon) \in \mathbb{N} : \forall n \geq n_0(\varepsilon), \mathbb{P}\left(\frac{|X_n|}{a_n} > C_\varepsilon\right) < \varepsilon \\ &\iff \lim_{C\to 0} \limsup_{n\to\infty} \mathbb{P}\left(\frac{|X_n|}{a_n} > C\right) = 0. \end{aligned}$$

A similar list of properties to Proposition 2 can be established for the o_p and O_p notation.

Exercise 4. *Show that Chebyshev's inequality implies that* $X_n - \mathbb{E}[X_n] = O_p(\sqrt{\mathbb{V}(X_n)})$.

1.11 Analytic tools

In this section we state some common analytic results.

Theorem 7 (Taylor's theorem)
Let $f : \mathbb{R} \to \mathbb{R}$ *and* $x \in \mathbb{R}$. *Assume that* f *has* p *continuous derivatives in an interval* $(x - \delta, x + \delta)$ *for some* $\delta > 0$. *Then, for any* $|h| < \delta$,

$$f(x + h) = \sum_{j=0}^{p} \frac{f^{(j)}(x)h^j}{p!} + R_n,$$

where $R_n = o(h^p)$. *Note that* R_n *depends on* x. *The exact value of* R_n *can be determined if we additionally assume that the* $p + 1$ *derivative of* f *exists. In which case,*

$$R_n = \frac{f^{(p+1)}(\xi)h^{p+1}}{(p + 1)!} = o(h^p)$$

for some $\xi \in (x - \delta, x + \delta)$. *Furthermore, if* $f^{(p+1)}$ *is bounded, the error goes to 0 in* $(x - \delta, x + \delta)$.

Theorem 8 (Monotone convergence theorem)
Suppose $\{f_n\}_{n\in\mathbb{N}}$ *is a sequence of non-negative measurable functions s.t.* $f_1 \leq f_2 \leq \ldots \leq f_n$ *and* $\lim_{n\to\infty} f_n = f$. *Then,* $\lim_{n\to\infty} \int f_n = \int f$.

The monotone convergence theorem in its essence provides the conditions under which the order of integrals and limits can be swapped. This theorem directly implies the results of the next lemma and theorem.

Lemma 7 (Fatou's lemma)
Suppose $\{f_n\}_{n\in\mathbb{N}}$ *is a sequence of non-negative measurable functions. The following bound holds:*

$$\liminf_{n\to\infty} \int f_n \geq \int \liminf_{n\to\infty} f_n.$$

Theorem 9 (Dominated convergence theorem)
Suppose $\{f_n\}_{n\in\mathbb{N}}$ *is a sequence of non-negative measurable functions s.t.* $\lim_{n\to\infty} f_n = f$. *If there exists a measurable function* g *s.t.* $\int g < \infty$ *and* $|f_n| \leq g$ *for all* n, *then,* f_n, f *are integrable and*

$$\lim_{n\to\infty} \int f_n = \int f.$$

Theorem 10 (Continuous mapping theorem)
Let $\{X_n\}, X$ *be some* S*-valued random variables. Suppose there exists a function* $g : S \mapsto \tilde{S}$ *that is continuous. Suppose* $X_n \to X$ *in* S, *then* $g(X_n) \to g(X)$ *in* $\tilde{S}$.

For our purposes, we will only look at the continuous mapping theorem (CMT) for real-valued random variables and so we can generally take $S = \tilde{S} = \mathbb{R}$ (or some subset of the reals). Note that the CMT holds for all the types of convergence we consider (in distribution, in probability and almost surely).

Definition 13 (Exchangeability)**.** *A sequence of random variables $\{X_n\}_{n\in A}$ is exchangeable if and only if for every finite permutation σ, the permuted sequence $\{X_{\sigma(n)}\}_{n\in A}$ is equal in distribution to the original sequence.*

Note that any sequence of i.i.d. random variables is exchangeable by this definition.

Theorem 11 (Radon-Nikodym theorem)
Suppose $(\Omega, \mathcal{F})$ is a measurable space and μ, ν are two σ-finite measures on this space s.t. $\mu \ll \nu$ (μ is absolutely continuous with respect to ν). Then, there exists a unique $\mathcal{F}$-measurable function $f : (\Omega, \mathcal{F}) \mapsto ([0, \infty], \mathcal{B}([0, \infty]))$ such that

$$\nu(A) = \int_A f \mathrm{d}\mu$$

*for all $A \in \mathcal{F}$. This function f is known as the **Radon-Nikodym derivative** of ν with respect to μ or the **density** of ν w.r.t. μ and is denoted as*

$$f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}.$$

1.12 Additional references

Durrett (2009) is a classical textbook that covers probability theory with a plethora of exercises that will help build intuition. For a more detailed review of the topics covered in this handout, see the lecture notes by Amir Dembo (Dembo, 2023). For an introduction to basic analysis see Rudin et al. (1964).

2 Statistical principles

A *statistical model* on the population P is postulated to make analysis possible.

Definition 14 (Parametric family)**.** *A set of probability measures P_θ on $\{\Omega, \mathcal{F}\}$ indexed by $\theta \in \Theta$ is a parametric family if and only if $\Theta \subset \mathbb{R}^d$ for some fixed $d \in \mathbb{N}^+$ and each P_θ is a known probability measure when θ is known. We call Θ the parameter space and d the dimension.*

A parametric model, then, assumes that the population P is determined by a parametric family. The family is considered *identifiable* iff $\theta_1 \neq \theta_2 \implies P_{\theta_1} \neq P_{\theta_2}$. There are two common parametric families that are often using in statistical theory.

Definition 15 (Exponential families)**.** *A parametric family such that*

$$f_X(x|\theta) = h(x) \exp(\eta(\theta)T(x) - A(\theta)),$$

where T, h, η and A are known functions and h is non-negative.

Definition 16 (Location-scale families)**.** *A family of distributions such that location and scale transformations generate distributions that belong to the family of distributions. Alternatively, it is a family of distributions where each member of the family can be obtained by a location and/or scale transformation of some canonical member of the family. i.e., for some $V \in \mathbb{R}$ and $M \in \mathbb{R}^+$*

$$\{P_{(\mu,\sigma)} : \mu \in V, \sigma \in M\}$$

where

$$P_{(\mu,\sigma)}(A) = P(\sigma^{-1/2}(A - \mu)), \qquad A \in \mathcal{F}$$

Definition 17 (Nonparametric family)**.** *A family of probability measures is non-parametric if there does not exist any Θ such that the definition 14 is satisfied.*

Non-parametric families can have minimal assumptions like $\{f_X : \mathbb{E}[X] < \infty\}$.

Let us suppose that X is now some realization from an unknown population P.

A measurable function $T(X)$ is called a *statistic* if $T(X)$ is a known operation on X. We can construct many statistics in accordance with this definition. For example, $T(X) = X$ is a statistic that does not transform the data in any way. This is a rather trivial statistic that will likely not be very useful for making inferences on P. More often, T is chosen such that it maps to a space that has smaller dimensionality than the original data X such that it captures some meaningful information in X that can then be used to make inference on the population P. Importantly, $\sigma(T(X)) \subset \sigma(X)$, and so T provides a reduction of the σ-field. Note that $T(X)$ is a random variable since if X is unknown, T may also be unknown, even if it is a known function. This problem of identifying the distribution of T without strong assumptions on the distribution of X is a large part of statistical theory.

Definition 18 (Order statistics)**.** *$X = (X_1, \ldots, X_n)$ that are i.i.d. from some distribution P. Let $X_{(i)}$ be the $i-th$ smallest value of $\{X_i\}$. Then $X_{(1)}, \ldots, X_{(n)}$ are called the ordered statistics.*

2.1 Sufficiency and completeness

When we consider T as a reduction of the σ-field, it is important to consider whether this reduction results in the loss of any information about P.

Definition 19 (Sufficiency)**.** *$X \sim P$. $T(X)$ is sufficient for P (or $\theta \in \Theta$ when considering a parametric family) if and only if $F_X(x|T)$ is known. i.e., the conditional distribution of X given T does not depend on P or θ.*

Note that the concept of sufficient statistics will depend on the family of distributions. In particular if T is sufficient for $P \in \mathcal{P}$, then it is also sufficient for $P \in \mathcal{P}_0 \subset \mathcal{P}$, but not necessarily for $P \in \mathcal{P}_1 \supset \mathcal{P}$.

The following theorem showcases the power of identifying sufficient statistics.

Theorem 12 (Factorization theorem)
Suppose $\{X_i\}$ *are samples from some* $P \in \mathcal{P}$*. Then,* T *is sufficient for* P *if and only if the pdf can be factored into two components:*

$$f_X(x) = g_P(T(x))h(x).$$

The pdf of X *then depends on* P *only through a function of* T*.*

Definition 20 (Minimal sufficiency)**.** *Suppose* T *is a sufficient statistic for* P*.* T *is a minimal sufficient statistic if and only if for any other sufficient statistic* S*, there exists a measurable function* g *such that* $T = g(S)$*.*

If T and S are both minimally sufficient, then there must exist a one-to-one measurable function g that maps S to T and vice-versa.

Definition 21 (Ancillary statistic)**.** *A statistic* $V(X)$ *is ancillary if its distribution does not depend on* P*. It is first-order ancillary if* $\mathbb{E}[V(X)] \perp\!\!\!\perp P$*.*

A trivial ancillary statistic is any constant $V(X) = c \; \forall X$. If V is a non-trivial ancillary statistic, then $\sigma(V(X)) \subset \sigma(X)$ is a non-trivial σ-field that does not contain any information about P.

Definition 22 (Completeness)**.** *A statistics* $T(X)$ *is complete if for an Borel function* f*,* $\mathbb{E}[f(T(X))] = 0$ *for all* $P \in \mathcal{P}$ *implies that* $f(T) = 0 a.s.$

If T is complete and $S = g(T)$, then S is also complete.

Lemma 8 (Lehmann-Scheffe)
A complete and sufficient statistic is minimal sufficient.

Note that the converse is not true in general. i.e., a minimal sufficient statistic need not be complete.

Theorem 13 (Basu's Theorem)
Suppose V *and* T *are statistics of* X*. If* V *is ancillary and* T *minimally sufficient, then* $V \perp\!\!\!\perp T$ *w.r.t any* $P \in \mathcal{P}$*.*

3 Statistical decision theory

Now, we turn to reviewing some foundational statistical decision theory.

A *statistical decision* is a conclusion drawn about P after observing X. A statistic T is then the *decision rule* that tells us what this conclusion will be when X is observed. This decision must be constructed or chosen based on some criteria that determines the performance of this decision. The criterion typically used to evaluate the decision is the *loss function* L. The *risk*, $R()$, is the average loss for T over the realizations X. Common loss functions include (i) squared error, (ii) absolute error and (iii) 0-1 error. The explicit form of each of these loss functions will depend largely on the assumptions we make about P and the problem we are trying to solve. We will look at each of these more carefully later in the course.

Definition 23 (Admissibility)**.** *Suppose $\mathcal{T}$ is a class of decision rules. A rule $T \in \mathcal{T}$ is admissible if and only if there does not exist any $S \in \mathcal{T}$ that has lower risk. i.e.,*

$$R(T) \leq R(S) \; \forall \; S \in \mathcal{T}.$$

The idea of admissibility is to eliminate some decision rules from $\mathcal{T}$ to simplify the decision making process.

Often to further narrow to set of rules to consider, the class $\mathcal{T}$ can be restricted to contain only rules that also satisfy some additional desirable properties (e.g. linearity, unbiasedness, moment constraints, continuity, invariance, etc.).

Based on the goal, T can be called by many names. In point estimation, T is the estimator for a certain parameter of the distribution P.

3.1 Hypothesis testing

Hypothesis testing addresses the problem of deciding whether a given statement about P (usually identifying the class of distributions that contains the true DGP) is true, provided we have some information in the form of X. A general formulation of a hypothesis test takes the following form:

Given a null hypothesis H_0 and an alternative hypothesis H_1, the test statistic T is used to identify which hypothesis can be supported based on whether or not it falls into the critical (rejection) region. Typical formulation of the two hypothesis looks like

$$H_0 : P \in P_0 \qquad H_1 : P \in P_1.$$

We use the following example to illustrate the construction of a hypothesis test and its test statistic.

Example 1 (One-sided hypothesis testing)**.** *Suppose $\{X_i\} \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$. Where μ is unknown but $\sigma^2 < \infty$ is known. Consider the one sided hypothesis test:*

$$H_0 : \mu \leq \mu_0 \qquad H_1 : \mu > \mu_0,$$

where μ_0 is some fixed constant. It is straightforward to show that the sample mean $\bar{X}_n$ is sufficient for $\mu \in \mathbb{R}$. As a result we consider the class of tests: $T_c(X) = \mathbf{1}_{[c,\infty)}(\bar{X}_n)$. This means that the null H_0 is rejected if $\bar{X}_n > c$ where c is some constant that is chosen based on the normality assumption as follows:

$$\alpha_c(\mu) = \mathbb{P}(T_c(X) = 1) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right).$$

Since the normal CDF Φ is an increasing function,

$$\sum_{P \in P_0} \alpha_c(\mu) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu_0)}{\sigma}\right).$$

It is also true that

$$\sum_{P\in P_1}(1-\alpha_c(\mu)) = \Phi\left(\frac{\sqrt{n}(c-\mu_0)}{\sigma}\right).$$

If we set some significance level α, the most effective test is to choose c such that

$$\alpha = \sup_{P\in P_0}\alpha_c(\mu),$$

wherein, c must satisfy

$$1-\Phi\left(\frac{\sqrt{n}(c-\mu_0)}{\sigma}\right) = \alpha.$$

Solving for c, we get

$$c = \frac{\sigma z_{1-\alpha}}{\sqrt{n}} + \mu_0,$$

where $z_a = \Phi^{-1}(a)$.

Any hypothesis testing problem may have one of the following: where the trade-off between type

	H_0 Accepted	H_1 Rejected
H_0 true	✓	Type I error
H_1 true	Type II error	✓

I and type II errors needs to be balanced. One common method in balancing the two errors is to employ the Neyman approach which sets the critical value to be such that

$$\mathbb{P}(\text{Type I error}) = \mathbb{P}(\text{reject } H_0) \leq \alpha.$$

This is precisely the way in which c was chosen in Example 1. This formulation naturally leads to an important function, the **power function**.

Definition 24 (Power function)**.** *For a critical region C, the power function $\beta : \Theta \to [0,1]$ is given by*

$$\beta(\theta) = \mathbb{P}(\textit{reject } H_0) = \mathbb{P}(X \in C) = \mathbb{P}(T(X) > c)$$

for all $\theta \in \Theta$.

The power function then classifies both errors as

$$\mathbb{P}(\text{type I error}) = \beta(\theta),\ \theta \in \Theta_0$$

and

$$\mathbb{P}(\text{type II error}) = 1-\beta(\theta),\ \theta \in \Theta_1,$$

where Θ_0 and Θ_1 define the parameter regions covered by the null and alternative hypothesis, respectively.

A "good" hypothesis test will minimize the probability of making either type I or type II errors.

References

Dembo, A. (2023). Lecture notes in probability theory.

Durrett, R. (2009). *Elementary probability for applications*. Cambridge university press.

Rudin, W. et al. (1964). *Principles of mathematical analysis*, volume 3. McGraw-hill New York.