MATH357 - Statistics

Based on lectures from Winter 2025 by Prof. Abbas Khalili.
Notes by Louis Meunier

Contents

§1 Review of Probability

↪**Definition 1.1** (Measurable Space, Probability Space): We work with a set Ω = sample space = {outcomes}, and a σ-algebra $\mathcal{F}$, which is a collection of subsets of Ω containing Ω and closed under taking complements and countable unions. The tuple $(\Omega, \mathcal{F})$ is called *measurable space*.

We call a nonnegative function $P : \mathcal{F} \to \mathbb{R}$ defined on a measurable space a *probability function* if $P(\Omega) = 1$ and if $\{E_n\} \subseteq \mathcal{F}$ a disjoint collection of subsets of Ω, then $P\left(\bigcup_{n\geq 1} E_n\right) = \sum_{n\geq 1} P(E_n)$. We call the tuple $(\Omega, \mathcal{F}, P)$ a *probability space*.

↪**Definition 1.2** (Random Variables): Fix a probability space $(\Omega, \mathcal{F}, P)$. A Borel-measurable function $X : \Omega \to \mathbb{R}$ (namely, $X^{-1}(B) \in \mathcal{F}$ for every $B \in \mathfrak{B}(\mathbb{R})$) is called a *random variable* on $\mathcal{F}$.

- *Probability distribution*: X induces a probability distribution on $\mathfrak{B}(\mathbb{R})$ given by $P(X \in B)$
- *Cumulative distribution function (CDF)*:

$$F_X(x) := P(X \leq x).$$

Note that $F(-\infty) = 0, F(+\infty) = 1$ and F right-continuous.

We say X *discrete* if there exists a countable set $S := \{x_1, x_2, ...\} \subset \mathbb{R}$, called the *support* of X, such that $P(X \in S) = 1$. Putting $p_i := P(X = x_i)$, then $\{p_i : i \geq 1\}$ is called the *probability mass function* (PMF) of X, and the CDF of X is given by

$$P(X \leq x) = \sum_{i : x_i \leq x} p_i.$$

We say X *continuous* if there is a nonnegative function f, called the *probability distribution function* (PDF) of X such that $F(x) = \int_{-\infty}^{x} f(t) \, \mathrm{d}t$ for every $x \in \mathbb{R}$. Then,

- $\forall B \in \mathfrak{B}(\mathbb{R}), P(X \in B) = \int_B f(t) \, \mathrm{d}t$
- $F'(x) = f(x)$
- $\int_{-\infty}^{\infty} f(x) \, \mathrm{d}x = 1$

If $X : \Omega \to \mathbb{R}$ a random variable and $g : \mathbb{R} \to \mathbb{R}$ a Borel-measurable function, then $Y := g(X) : \Omega \to \mathbb{R}$ also a random variable.

↪**Definition 1.3** (Moments): Let X be a discrete/random random variable with pmf/pdf f and support S. Then, if $\sum_{x \in S} |x| f(x) / \int_S |x| f(x) \, \mathrm{d}x < \infty$, then we say the first moment/mean of X exists, and define

$$\mu_X = \mathbb{E}[X] = \begin{cases} \sum_{x \in S} x f(x) \\ \int_S x f(x) \, \mathrm{d}x \end{cases}.$$

Let $g : \mathbb{R} \to \mathbb{R}$ be a Borel-measurable function. Then, we have

$$\mathbb{E}[g(X)] = \begin{cases} \sum_{x \in S} g(x) f(x) \\ \int_S g(x) f(x) \end{cases}.$$

Taking $g(x) = |x|^k$ gives the so-called "kth absolute moments", and $g(x) = x^k$ gives the ordinary "kth moments". Notice that $\mathbb{E}[\cdot]$ linear in its argument.

For $k \geq 1$, if μ exists, define the central moments

$$\mu_k := \mathbb{E}\left[(X - \mu)^k\right],$$

where they exist.

↪**Definition 1.4** (Moment Generating Function (mgf)): If X a r.v., the mgf of X is given by

$$M(t) := \mathbb{E}\left[e^{tX}\right],$$

if it exists for $t \in (-h, h)$, $h > 0$. Then, $M^{(n)}(0) = \mathbb{E}[X^n]$.

↪**Definition 1.5** (Multiple Random Variable): $X = (X_1, ..., X_n) : \Omega \to \mathbb{R}^n$ a random vector if $X^{-1}(I) \in \mathcal{F}$ for every $I \in \mathfrak{B}_{\mathbb{R}^n}$. (It suffices to check for "rectangles" $I = (-\infty, a_1] \times \cdots \times (-\infty, a_n]$, as before.)

Let F be the CDF of X, and let $A \subseteq \{1, ..., n\}$, enumerating A by $\{i_1, ..., i_k\}$. Then, the CDF of the subvector $X_A = \left(X_{i_1}, ..., X_{i_k}\right)$ is given by

$$F_{X_A}\left(x_{i_1}, ..., x_{i_k}\right) = \lim_{\substack{x_{i_j} \to \infty, \\ i_j \in \mathcal{I} \setminus A}} F(x_1, ..., x_n).$$

In particular, the marginal distribution of X_i is given by

$$F_{X_i}(x) = \lim_{x_1, ..., x_{i-1}, x_{i+1}, ..., x_n \to +\infty} F(x_1, ..., x, ..., x_n).$$

Let $g : \mathbb{R}^n \to \mathbb{R}$ measurable. Then,

$$\mathbb{E}[g(X_1, ..., X_n)] = \begin{cases} \sum_{(x_1, ..., x_n)} g(x_1, ..., x_n) f(x_1, ..., x_n) \\ \int \cdots \int g(x_1, ..., x_n) f(x_1, ..., x_n) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n \end{cases}.$$

We have the notion of a joint mgf,

$$M(t_1, ..., t_n) = \mathbb{E}\left[e^{\sum_{i=1}^n t_i X_i}\right],$$

if it exists for $0 < \left(\sum_{i=1}^n t_i^2\right)^{\frac{1}{2}} < h$ for some $h > 0$. Notice that $M(0, ..., 0, t_i, 0, ..., 0)$ is equal to the mgf of X_i.

↪**Definition 1.6** (Conditional Probability): Let $(X_1, \ldots, X_n)$ a random vector. Let $\mathcal{I} = \{1, \ldots, n\}$ and A, B disjoint subsets of $\mathcal{I}$ with $k := |A|, h := |B|$. Write $X_A = \left(X_{i_1}, \ldots, X_{i_k}\right)^t$, similar for B. Then, the conditional probability of A given B is given by

$$f_{X_A|X_B}(x_a|x_b) := f_{X_A \mid X_B = x_B}(x_A) = \frac{f_{X_A,X_B}(x_a, x_b)}{f_{X_b}(x_b)},$$

provided the denominator is nonzero. Sometimes we have information about conditional probabilities but not the main probability function; we have that

$$f(x_1, \ldots, x_n) = f(x_1)f(x_2 \mid x_1)f(x_3|x_1, x_2)\cdots f(x_n \,|x_1, \ldots, x_{n-1}),$$

which follows from expanding the previous definition and observing the cancellation.

Let $X = (X_1, \ldots, X_n) \sim F$. We say $X_1, \ldots, X_n$ (mutually) independent and write $\coprod_{i=1}^n X_i$ if

$$F(x_1, \ldots, x_n) = \prod_{i=1}^n F_{X_i}(x_i),$$

where F_{X_i} the marginal cdf of X_i. Equivalently,

$$\begin{aligned}
\coprod_{i=1}^n X_i \Leftrightarrow f(x_1, \ldots, x_n) &= \prod_{i=1}^n f_{X_i}(x_i) \\
\Leftrightarrow P(X_1 \in B_1, \ldots, X_n \in B_n) &= \prod_{i=1}^n P(X_i \in B_i) \; \forall \, B_i \in \mathfrak{B}_{\mathbb{R}} \\
\Leftrightarrow M_X(t_1, \ldots, t_n) &= \prod_{i=1}^n M_{X_i}(t_i).
\end{aligned}$$

If X, Y are two random variables with cdfs F_X, F_Y such that $F_X(z) = F_Y(z)$ for every z, we say X, Y identically distributed and write $X \stackrel{d}{=} Y$ (note that X need not equal Y pointwise). If $X_1, \ldots, X_n$ a collection of random variables that are independent and identically distributed with common cdf F, we write $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} F$.

Further, define the covariance, correlation of two random variables X, Y respectively:

$$\mathrm{Cov}(X, Y) := \sigma_{X,Y} := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mu_X \mu_Y, \qquad \rho_{X,Y} := \frac{\sigma_{XY}}{\sigma_X \sigma_Y},$$

if $\mathbb{E}[|X - \mathbb{E}[X]| \, |Y - \mathbb{E}[Y]|] < \infty$.

↪**Theorem 1.1**: If $X_1, \ldots, X_n$ independent and $g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}$ borel-measurable functions, then $g_1(X_1), \ldots, g_n(X_n)$ also independent.

↪**Definition 1.7** (Conditional Expectation): Let X, Y be random variables and $g : \mathbb{R} \to \mathbb{R}$ a borel-measurable function. We define the following notions:

$$\mathbb{E}[g(X)|Y = y] = \begin{cases} \sum_{x \in S_X} g(x) f(x|y) \text{ discrete} \\ \int_{S_x} g(x) f(x|y) \, \mathrm{d}x \text{ cnts} \end{cases}.$$

$$\operatorname{Var}(X|Y = y) = \mathbb{E}\left[X^2|Y = y\right] - \mathbb{E}^2[X|Y = y].$$

↪**Theorem 1.2**: If $\mathbb{E}[g(X)]$ exists, then $\mathbb{E}[g(X)] = \mathbb{E}[\mathbb{E}[g(X)|Y]]$, where the first nested $\mathbb{E}$ is with respect to x, the second y.

↪**Theorem 1.3**: If $\mathbb{E}\left[X^2\right] < \infty$, then $\operatorname{Var}(X) = \operatorname{Var}(\mathbb{E}[X|Y]) + \mathbb{E}[\operatorname{Var}(X|Y)]$. In particular, $\operatorname{Var}(X) \geq \operatorname{Var}(\mathbb{E}[X|Y])$.

§2 Statistics

§2.1 Sample Distributions

↪**Definition 2.1** (Inference): We consider some population with some characteristic we wish to study. We can model this characteristic as a random variable $X \sim F$. In general, we don't have access to F, but wish to take samples from our population to make inferences about its properties.

(1) *Parametric inference:* in this setting, we assume we know the functional form of X up to some parameter, $\theta \in \Theta \subset \mathbb{R}^d$, where Θ our "parameter space". Namely, we know $X \sim F_\theta \in \mathcal{F} := \{F_\theta \mid \theta \in \Theta\}$.

(2) *Non-parametric inference:* in this setting we know noting about F itself, except perhaps that F continuous, discrete, etc.

Other types exist. We'll focus on these two.

↪**Definition 2.2** (Random Sample): Let $X_1, ..., X_n \overset{\text{iid}}{\sim} F$. Then $X_1, ..., X_n$ called a *random sample* of the population.

We also call X_i the "pre-experimental data" (to be observed) and x_i the "post-experimental data" (been observed).

↪**Definition 2.3** (Statistics): Let $X_1, ..., X_n \overset{\text{iid}}{\sim} F$ where X_i a d-dimensional random vector. Let

$$T : \underbrace{\mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d}_{n-\text{fold}} \to \mathbb{R}^k$$

be a borel-measurable function. Then, $T(X_1, ..., X_n)$ is called a *statistic*, provided it does not depend on any unknown.

⊛ **Example 2.1**: $\overline{X_n} := \frac{1}{n} \sum_{i=1}^{n} X_i$ (the "sample mean") and $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X_n}\right)^2$, (the "sample variance") are both typical statistics.

↪**Theorem 2.1**: Let $x_1, ..., x_n \in \mathbb{R}$, then

(a) $\operatorname{argmin}_{\alpha \in \mathbb{R}} \left\{\sum_{i=1}^{n} (x_i - \alpha)^2\right\} = \overline{x_n}$;

(b) $\sum_{i=1}^{n} (x_i - \overline{x_n})^2 = \sum_{i=1}^{n} \left(x_i^2\right) - n\overline{x_n}^2$;

(c) $\sum_{i=1}^{n} (x_i - \overline{x_n}) = 0$.

↪**Theorem 2.2**: Let $X_1, ..., X_n \overset{\text{iid}}{\sim} F$, and $g : \mathbb{R} \to \mathbb{R}$ borel-measurable such that $\operatorname{Var}(g(X)) < \infty$. Then,

(a) $\mathbb{E}\left[\sum_{i=1}^{n} g(X_i)\right] = n\mathbb{E}[g(X_1)]$;

(b) $\operatorname{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n \operatorname{Var}(X_1)$.

↪**Theorem 2.3**: Let $X_1, ..., X_n \overset{\text{iid}}{\sim} F$ with $\sigma^2 < \infty$, then

1. $\mathbb{E}\left[\overline{X_n}\right] = \mu$, $\operatorname{Var}\left(\overline{X_n}\right) = \frac{\sigma^2}{n}$, $\mathbb{E}\left[S_n^2\right] = \sigma^2$.
2. If $M_{X_1}(t)$ exists in some neighborhood of 0, then $M_{\overline{X_n}}(t) = M_{X_1}\left(\frac{t}{n}\right)^n$, where it exists.

↪**Theorem 2.4**: Let $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$. Then

1. $\overline{X_n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$;
2. $\overline{X_n}, S_n^2$ are independent;
3. $\frac{(n-1)S_n^2}{\sigma^2} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X_n}\right)^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

Remark 2.1:

2. actually holds iff the underlying distribution is normal.

PROOF. We prove 2. We first write S_n^2 as a function of $\left(X_2 - \overline{X}_n, ..., X_n - \overline{X}_n\right)$:

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

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

Then, it suffices to show that $\overline{X}_n$ and $\left(X_2 - \overline{X}_n, \ldots, X_n - \overline{X}_n\right)$ are independent.

Consider now the transformation

$$\begin{cases} y_1 &= \overline{x}_n \\ y_2 &= x_2 - \overline{x}_n \\ &\ddots \\ y_n &= x_n - \overline{x}_n \end{cases} \Rightarrow \begin{cases} x_1 &= y_1 - \sum_{i=2}^{n} y_i \\ x_2 &= y_2 + y_1 \\ &\ddots \\ x_n &= y_n + y_1 \end{cases},$$

which gives Jacobian

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

and so

$$\begin{aligned} f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) &= |J| \cdot f_{X_1,\ldots,X_n}(x_1(y_1,\ldots,y_n),\ldots,x_n(y_1,\ldots,y_n)) \\ &= n \cdot \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i(y_1,\ldots,y_n)-\mu)^2} \\ &\approx \underbrace{e^{-n\frac{(y_1-\mu)^2}{2\sigma^2}}}_{\text{only } y_1} \cdot \underbrace{e^{-\frac{1}{2\sigma^2}\left\{\left(\sum_{i=2}^{n} y_i\right)^2 + \sum_{i=2}^{n} y_i^2\right\}}}_{\text{no } y_1 \text{ dependence}}, \end{aligned}$$

and hence as the PDFs split, we conclude Y_1 independent of $Y_2, \ldots, Y_n$ and so $\overline{X}_n$ independent of $\left(X_2 - \overline{X}_n, \ldots, X_n - \overline{X}_n\right)$ and so in particular of any Borel-measurable function of this vector such as S_n^2, completing the proof.

For 3, note that

$$\begin{aligned} V := \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\left(X_i - \overline{X}_n\right) - \left(\mu - \overline{X}_n\right)\right)^2 \\ &= \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}{\sigma^2} + \frac{n\left(\overline{X}_n - \mu\right)^2}{\sigma^2} =: W_1 + W_2. \end{aligned}$$

The first part, W_1, of this summation is just $(n-1)\frac{S_n^2}{\sigma^2}$, a function of S_n^2, and the second, W_2, is a function of $\overline{X}_n$. By what we've just shown in the previous part, these two are independent. In addition, $V \sim \chi^2_{(n)}$ and

$$W_2 = \frac{n\left(\overline{X}_n - \mu\right)^2}{\sigma^2} = \left(\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \sim \chi^2_{(1)},$$

since the inner random variable is a standard normal. Then, since W_1, W_2 independent, $M_V(t) = M_{W_1}(t) M_{W_2}(t)$, so for $t < \frac{1}{2}$,

$$M_{W_1}(t) = \frac{M_V(t)}{M_{W_2}(t)} = \frac{(1-2t)^{-\frac{n}{2}}}{(1-2t)^{-\frac{1}{2}}} = (1-2t)^{-\frac{(n-1)}{2}},$$

hence $W_1 \sim \chi^2_{(n-1)}$. ■

↪**Proposition 2.1**: Let $X \sim t(\nu)$, the Student t-distribution i.e

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

then

- $\mathrm{Var}(X) = \frac{\nu}{\nu-2}$ for $\nu > 2$
- If $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{(\nu)}$ are independent random variables, then $T = \frac{Z}{\sqrt{V/\nu}} \sim t(\nu)$.

↪**Theorem 2.5**: Let $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then,

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

Remark 2.2: By combing CLT and Slutsky's Theorem, T asymptotes to $\mathcal{N}(0,1)$, but this gives a general distribution. Note that for large n, $t(n-1)$ approximately normal too.

PROOF. Notice that

$$W_1 := \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1), \qquad W_2 := \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

are independent, and

$$T = \frac{W_1}{\sqrt{W_2/(n-1)}}$$

so by the previous proposition $T \sim t(n-1)$. ■

↪**Proposition 2.2**: Given $U \sim \chi^2_{(m)}, V \sim \chi^2_{(n)}$ independent, then $F = \frac{U/m}{V/n} \sim F(m,n)$. If $T \sim t(\nu)$, $T^2 \sim F(1,\nu)$.

↪**Theorem 2.6**: Let $X_1, \ldots, X_m \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$ be mutually independent random samples. Then,

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

Proof. We have that $U = \frac{(m-1)S_m^2}{\sigma_1^2} \sim \chi^2_{(m-1)}$ and $V = \frac{(n-1)S_n^2}{\sigma_2^2}$ are independent so by the previous proposition

$$F = \frac{U/(m-1)}{V/(n-1)} \sim F(m-1, n-1).$$

■

§2.2 Order Statistics

↪**Definition 2.4**: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$. Then, the *order statistics* are

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)},$$

where $X_{(i)}$ the ith largest of $X_1, \ldots, X_n$.

↪**Definition 2.5** (Related Functions of Order Statistcs): The *sample range* is defined

$$R_n := X_{(n)} - X_{(1)}.$$

The *sample median* is defined

$$M(X_1, \ldots, X_n) := \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ odd} \\ \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+1}{2}\right)}}{2} & \text{if } n \text{ even.} \end{cases}$$

↪**Theorem 2.7** (Distribution of Max, Min): Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F, f$.

(Discrete)

(a) $P(X_{(1)} = x) = [1 - F(x^-)]^n - [1 - F(x)]^n$

(b) $P(X_{(n)} = y) = [F(y)]^n - [F(y^-)]^n$

(Continuous)

(c) $F_{X_{(1)}}(x) = P(X_{(1)} \leq x) = 1 - [1 - F(x)]^n, \qquad f_{X_{(1)}}(x) = n \cdot f(x)[1 - F(x)]^{n-1}$

(d) $F_{X_{(n)}}(y) = [F(y)]^n, \qquad f_{X_{(n)}}(y) = n \cdot f(y)[F(y)]^{n-1}$

Proof. (a) Notice

$$P(X_{(1)} = x) = P(X_{(1)} \leq x) - P(X_{(1)} < x).$$

We have

$$\begin{aligned} P\big(X_{(1)} \le x\big) &= 1 - P\big(X_{(1)} > x\big) \\ &= 1 - P(X_1 > x, X_2 > x, \dots, X_n > x) \\ &= 1 - P(X_1 > x)P(x_2 > x)\cdots P(X_n > x) \\ &= 1 - [1 - F(x)]^n, \end{aligned}$$

and similarly

$$P\big(X_{(1)} < x\big) = 1 - P\big(X_{(1)} \ge x\big) = 1 - [1 - F(x^-)]^n,$$

where $F(x^-) = \lim_{z \to x^-} F(z)$. So in all,

$$P\big(X_{(1)} = x\big) = [1 - F(x^-)]^n - [1 - F(x)]^n.$$

(b) is very similar. For (c), we have

$$\begin{aligned} P\big(X_{(1)} \le x\big) &= 1 - P\big(X_{(1)} > x\big) \\ &= 1 - P(X_1 > x, \dots, X_n > x) \\ &= 1 - [1 - F(x)]^n. \end{aligned}$$

(d) is similar. ■

↪**Theorem 2.8** (Distribution of jth Order Statistics): Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} F, f$.

(Discrete) Suppose the X_i's take values in $S_x = \{x_1, x_2, \dots\}$ and put $p_i = P(X_i)$. Then,

$$F_{X_{(j)}}(x_i) = P\big(X_{(j)}(x_i) \le x_i\big) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k},$$

where $P_i = P(X_i \le x_i) = \sum_{\ell=1}^{i} p_\ell$.

(Continuous)

$$F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} F^k(x)[1 - F(x)]^{n-k},$$

so

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f(x)[F(x)]^{j-1}[1 - F(x)]^{n-j}.$$

Proof. For discrete, we have

$$P\big(X_{(j)}(x_i) \le x_i\big) = P(\text{at least } j \text{ out of } X_1, \dots, X_n \le x_i).$$

Then,

$$P(\text{at least } j \text{ out of } X_1, \dots, X_n \le x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}.$$

Continuous is similar. ■

§2.3 Large Sample/Asymptotic Theory

↪**Definition 2.6** (Convergence in Probability): We say $T_n = T(X_1, \ldots, X_n)$ converges *in probability* to θ $T_n \xrightarrow{P} \theta$ as $n \to \infty$ if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|T_n - \theta| > \varepsilon) = 0.$$

↪**Definition 2.7** (Convergence in Distribution): Find a positive sequence $\{r_n\}$ with $r_n \to \infty$ such that

$$r_n(T_n - \theta) \xrightarrow{d} T,$$

where T a random variable.

↪**Theorem 2.9** (Slutsky's): Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} a$ for some $a \in \mathbb{R}$. Then,

$$X_n + Y_n \xrightarrow{d} X + a$$

$$X_n Y_n \xrightarrow{d} aX,$$

and if $a \neq 0$,

$$\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{a}.$$

↪**Theorem 2.10** (Continuous Mapping Theorem (CMT)): Suppose $X_n \xrightarrow{P} X$ and g is continuous on the set C such that $P(X \in C) = 1$. Then,

$$g(X_n) \xrightarrow{P} g(X).$$

⊛ **Example 2.2**: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ with $\mu = \mathbb{E}[X_i], \sigma^2 = \mathrm{Var}(X_i) < \infty$. Then,

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

since we may rewrite

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

The numerator $\xrightarrow{d} \mathcal{N}(0,1)$ by CLT. $S_n^2 \xrightarrow{P} \sigma^2$, so the denominator goes to 1 in probability.

↪**Definition 2.8** (Big $\mathcal{O}$, Little o Notation): Let $\{a_n\}, \{b_n\} \subseteq \mathbb{R}$ real sequences.

- We say $a_n = \mathcal{O}(b_n)$ if $\exists\, 0 < c \in \mathbb{R}$ and $N \in \mathbb{N}$ such that $|\frac{a_n}{b}| \leq c$ for every $n \geq N$.
- We say $a_n = o(b_n)$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$.

↪**Definition 2.9** (Big $\mathcal{O}_p$, Little o_p Notation): Let $\{X_n\}, \{Y_n\}$ sequences of random variables.

- We say $X_n = \mathcal{O}_p(1)$ if $\forall\, \varepsilon > 0$ there is a $N_\varepsilon \in \mathbb{N}$ and $C_\varepsilon \in \mathbb{R}$ such that
$$P(|X_n| > C_\varepsilon) < \varepsilon$$
for every $n > N_\varepsilon$.
  - We say $X_n = \mathcal{O}_p(Y_n)$ if $X_n/Y_n = \mathcal{O}_p(1)$.
- We say $X_n = o_p(1)$ if $X_n \xrightarrow{P} 0$.
  - We say $X_n = o_p(Y_n)$ if $X_n/Y_n = o_p(1)$.

↪**Proposition 2.3**: If $X_n \xrightarrow{d} X$, then $X_n = \mathcal{O}_p(1)$.

↪**Proposition 2.4** (The Delta Method (First Order)): Let $\sqrt{n}(X_n - \mu) \xrightarrow{d} V$ and g a real-valued function such that g' exists at $x = \mu$ and $g'(\mu) \neq 0$. Then,
$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} g'(\mu)V.$$
In particular, if $V \sim \mathcal{N}(0, \sigma^2)$ then
$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} \mathcal{N}\left(0, g'(\mu)^2\sigma^2\right).$$

Proof. Taylor expanding the LHS,
$$\sqrt{n}\{g(X_n) - g(\mu)\} = g'(\mu)\sqrt{n}(X_n - \mu) + o_p(1) \to g'(\mu)V.$$
∎

↪**Proposition 2.5** (The Delta Method (Second Order)): Suppose $\sqrt{n}(X - n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ and $g'(\mu) = 0$ but $g''(\mu) \neq 0$. Then,
$$n\{g(X_n) - g(\mu)\} \xrightarrow{d} \sigma^2 \frac{g''(\mu)}{2} \cdot \chi^2_{(1)}.$$

Proof.
$$g(X_n) - g(\mu) = \frac{g''(\mu)}{2}(X_n - \mu)^2 + o_p(1),$$
so

$$n(g(X_n) - g(\mu)) = \sigma^2 \frac{g''(\mu)}{2}\left[\frac{\sqrt{n}(X_n - \mu)}{\sigma}\right]^2 + o_p(1).$$

The bracketed term converges in distribution to $\mathcal{N}(0, 1)$ and the $o_p(1)$ term converges in probability to zero, so the proposition follows by applying Slutsky's Theorem. ■

⊛ **Example 2.3**: $\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ by CLT. Letting $g(x) = x^2$, and assuming $\mu \neq 0$, then

$$\sqrt{n}\left(\overline{X}_n^2 - \mu^2\right) \to \mathcal{N}(0, 4\mu^2\sigma^2),$$

by the first-order delta method.

↪**Proposition 2.6**: Let $X_1, ..., X_n \overset{\text{iid}}{\sim} F$, and denote the ECDF $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \leq x)$. Then,
1. $\mathbb{E}[F_n(x)] = F(x)$;
2. $\text{Var}\,(F_n(x)) = \frac{1}{n}F(x)(1 - F(x))$;
3. $nF_n(x) = \sum_{i=1}^n \mathbb{1}(X_i \leq x) \sim \text{Bin}(n, F(x))$;
4. $\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1-F(x))}} \xrightarrow{d} \mathcal{N}(0, 1)$.
5. $F_n(x) \xrightarrow{P} F(x)$.
6. $P(|F_n(x) - F(x)| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$, by Hoeffing's Inequality.
7. $\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = \|F_n - F\|_\infty \xrightarrow{\text{a.s.}} 0$, by the Glivenko-Cantelli Theorem.
8. $P(\|F_n - F\|_\infty > \varepsilon) \leq C\varepsilon e^{-2n\varepsilon^2}$ for some constant C (Dvoretzky-Kiefer-Wolfowitz Theorem).

Remark 2.3: The constant in 8. was shown to be 2 by Massart.

§2.4 Parametric Inference

↪**Definition 2.10** (Point Estimator): Let $X_1, ..., X_n$ a random sample. A *point estimator* $\hat{\theta} := \hat{\theta}(X_1, ..., X_n)$ is an estimator of a parameter θ if it is a statistic.

⊛ **Example 2.4**: Let X be a random variable denoting whether a randomly selected electronic chip is operational or not, i.e. $X = \begin{cases} 1 \text{ operational} \\ 0 \text{ else} \end{cases}$, supposing $X \sim \text{Ber}(\theta)$, then $0 < \theta < 1$ is the probability a randomly selected chip is operational. Let $X_1, ..., X_n \overset{\text{iid}}{\sim} \text{Ber}(\theta)$. Then,

$$\mathcal{F} = \{\text{Ber}(\theta) : 0 < \theta < 1\}, \qquad \Theta = (0, 1).$$

Then, possible estimators are $\overline{X}_n$, $\frac{X_1 + X_2}{2}$, just X_2, etc.

↪**Definition 2.11** (Bias): An estimator $\hat{\theta}_n$ is an *unbiased* estimator of θ if

$$\mathbb{E}_\theta\left[\hat{\theta}_n\right] = \theta, \qquad \forall\, \theta \in \Theta,$$

where the expected value is taken with respect to the distribution of $\hat{\theta}_n$ (and thus depends on the distribution F_θ).

Generally, the *bias* of an estimator $\hat{\theta}_n$ is defined

$$\operatorname{Bias}\left(\hat{\theta}_n\right) := \mathbb{E}_\theta\left[\hat{\theta}_n\right] - \theta, \qquad \theta \in \Theta.$$

If $\hat{\theta}_n$ unbiased, then $\operatorname{Bias}\left(\hat{\theta}_n\right) = 0$.

⊛ **Example 2.5**: For instance, recalling the previous example,

$$\mathbb{E}_\theta\left[\overline{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta[X_i] = \frac{1}{n}n\theta = \theta,$$

so $\overline{X}_n$ unbiased. Also,

$$\mathbb{E}_\theta[X_1] = \theta,$$

so just X_1 also unbiased, as is $\frac{X_1+X_2}{2}$.

⊛ **Example 2.6**: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_\theta$, $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$, $\mu = \mathbb{E}[X_i], \sigma^2 \operatorname{Var}(X_i)$. Then, $\hat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ an unbiased estimator of μ. Let $\hat{\sigma}_n^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$, then recalling $\mathbb{E}\left[\hat{\sigma}_n^2\right] = \sigma^2$, this is an unbiased estimator of σ^2. However, changing the constant term, to get

$$\hat{\sigma}_n^{*2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2,$$

is biased, with

$$\mathbb{E}_\theta\left[\hat{\sigma}_n^{*2}\right] = \frac{n-1}{n}\sigma^2,$$

so

$$\operatorname{Bias}\left(\hat{\sigma}_n^{*2}\right) = -\frac{\sigma^2}{n} < 0,$$

i.e. $\hat{\sigma}_n^{*2}$ underestimates the true parameter on average. Of course, in the limit it becomes 0.

⊛ **Example 2.7**: Let $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{U}(0, \theta)$, $\theta > 0$, $\Theta = (0, \infty)$. Recall $\mathbb{E}_\theta[X_i] = \frac{\theta}{2}$. Consider

$$\hat{\theta}_{n,1} := 2X_3, \qquad \hat{\theta}_{n,2} := 2\overline{X}_n, \qquad \hat{\theta}_{n,3} := X_{(n)}.$$

Then, $\mathbb{E}\left[\hat{\theta}_{n,i}\right] = \theta$ for $i = 1, 2$ and $\frac{n}{n+1}\theta$ for $i = 3$. Hence, we can scale the last one, $\hat{\theta}_{n,4} := \frac{n+1}{n}\hat{\theta}_{n,3}$, to get an unbiased estimator.

↪**Definition 2.12** (Mean-Squared Error): The *Mean-Squared Error* (MSE) of an estimator is defined

$$\begin{aligned}
\mathrm{MSE}_\theta\left(\hat{\theta}_n\right) &:= \mathbb{E}_\theta\left[\left(\hat{\theta}_n - \theta\right)^2\right] \\
&= \mathbb{E}_\theta\left[\left(\left(\hat{\theta}_n - \mathbb{E}_\theta\left[\hat{\theta}_n\right]\right) + \left(\mathbb{E}_\theta\left[\hat{\theta}_n\right] - \theta\right)\right)^2\right] \\
&= \mathrm{Var}_\theta\left(\hat{\theta}_n\right) + \left[\mathrm{Bias}\left(\hat{\theta}_n\right)\right]^2.
\end{aligned}$$

Remark that if $\mathbb{E}_\theta\left[\hat{\theta}_n\right] = \theta$, i.e. $\hat{\theta}_n$ unbiased, then $\mathrm{MSE}_\theta\left(\hat{\theta}_n\right) = \mathrm{Var}_\theta\left(\hat{\theta}_n\right)$.

↪**Definition 2.13** (Consistency): We say an estimator $\hat{\theta}_n$ of θ is *consistent* if $\hat{\theta}_n \overset{P}{\to} \theta$ as $n \to \infty$.

Remark 2.4: There are many ways of establishing consistency; by direct definition of convergence in probability, the WLLN (maybe continuous mapping theorem), or checking if $\mathbb{E}_\theta\left[\hat{\theta}_n\right] \to \theta$ (if this happens we say $\hat{\theta}_n$ "asymptotically unbiased") and $\mathrm{Var}_\theta\left(\hat{\theta}_n\right) \to 0$ as $n \to \infty$, for in this case by Chebyshev's Inequality we have consistency.

⊛ **Example 2.8**: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_\theta$.

1. $\hat{\mu}_n := \overline{X}_n \xrightarrow{P} \mu$ by WLLN, and $S_n^2 \xrightarrow{P} \sigma^2$ similarly.
2. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{U}(0, \theta)$, then $\mathbb{E}[X_i] = \frac{\theta}{2}$. Note that $\hat{\theta}_{n,1} = 2\overline{X}_n$ and $\hat{\theta}_{n,2} = \frac{n+1}{n} X_{(n)}$ are both unbiased estimators of θ, and both are consistent. To see the second one, we have that for any $\varepsilon > 0$,

$$\begin{aligned} P\left(|X_{(n)} - \theta| > \varepsilon\right) &= P\left(\theta - X_{(n)} > \varepsilon\right) \\ &= P\left(X_{(n)} < \theta - \varepsilon\right) \\ &= \left(\frac{\theta - \varepsilon}{\theta}\right)^n \to 0 \text{ as } n \to \infty. \end{aligned}$$

We have too that

$$\mathrm{MSE}_\theta\left(\hat{\theta}_{n,1}\right) = \mathrm{Var}_\theta\left(\hat{\theta}_{n,1}\right) = 4\mathrm{Var}_\theta\left(\overline{X}_n\right) = \frac{4}{n}\,\mathrm{Var}(X_i) = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

Also

$$\begin{aligned} \mathrm{MSE}_\theta\left(\hat{\theta}_{n,2}\right) = \mathrm{Var}_\theta\left(\hat{\theta}_{n,2}\right) &= \left(\frac{n+1}{n}\right)^2 \mathrm{Var}\left(X_{(n)}\right) \\ = \cdots &= \frac{\theta^2}{n(n+2)} = \frac{\theta^2}{3n} \cdot \frac{3}{n+2} \le \mathrm{MSE}_\theta\left(\hat{\theta}_{n,1}\right) \; \forall\, n \ge 1. \end{aligned}$$

We will focus on the class of unbiased estimators of a real-valued parameter, $\tau(\theta)$, $\tau : \Theta \to \mathbb{R}$.

2.4.1 Uniformly Minimum Variance Unbiased Estimators (UMVUE)

↪**Definition 2.14** (UMVUE): Let $\boldsymbol{X} = (X_1, \ldots, X_n)^t$ be a random variable with a joint pdf/pmf given by

$$p_\theta(\boldsymbol{x}) = p_\theta(x_1, \ldots, x_n),$$

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

1. $\mathbb{E}_\theta[\mathrm{T}(\boldsymbol{X})] = \tau(\theta)$ for every $\theta \in \Theta$;
2. for any other unbiased estimator $\mathrm{T}^*(\boldsymbol{X})$ of $\tau(\theta)$, we have

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

↪**Proposition 2.7** (Cramér-Rau Lower Bound): We define in the case $d = 1$ ($\Theta \subseteq \mathbb{R}$) for convenience. Assume that

(1) the family $\{p_\theta : \theta \in \Theta\}$ has a common support $S = \{\boldsymbol{x} \in \mathbb{R}^n : p_\theta(\boldsymbol{x}) > 0\}$ that does not depend on θ;

(2) for $\boldsymbol{x} \in S$, $\theta \in \Theta$, $\frac{\mathrm{d}}{\mathrm{d}\theta} \log p_\theta(\boldsymbol{x}) < \infty$;

(3) for any statistic $h(\boldsymbol{x})$ with $\mathbb{E}_\theta[|h(\boldsymbol{x})|] < \infty$ for every $\theta \in \Theta$, we have

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int_S h(\boldsymbol{x}) p_\theta(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_S h(\boldsymbol{x}) \frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x},$$

whenever the right-hand side is finite.

Let $\mathrm{T}(\boldsymbol{X})$ be such that $\mathrm{Var}_\theta(\mathrm{T}(\boldsymbol{X})) < \infty$ and $\mathbb{E}_\theta[\mathrm{T}(\boldsymbol{X})] = \tau(\theta)$ for every every $\theta \in \Theta$. Then if $0 < \mathbb{E}_\theta\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log(p_\theta(\boldsymbol{x}))\right)^2\right] < \infty$ for every $\theta \in \Theta$, then the Cramér-Rao Lower Bound (CRLB) holds:

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

Remark 2.5: The quantity

$$I(\theta) := \mathbb{E}_\theta\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log(p_\theta(\boldsymbol{x}))\right)^2\right]$$

is called the *Fisher information* contained in $\boldsymbol{X}$ about θ.

PROOF. Note that $\tau(\theta) = \mathbb{E}_\theta[\mathrm{T}(\boldsymbol{X})]$ implies

$$\begin{aligned}
\tau'(\theta) &= \frac{\mathrm{d}}{\mathrm{d}\theta} \mathbb{E}[\mathrm{T}(\boldsymbol{X})] \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\left[\int_S \mathrm{T}(\boldsymbol{x}) p_\theta(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}\right] \\
\text{by ass. 2, 3} \qquad &= \int_S \mathrm{T}(\boldsymbol{x}) \frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\
&= \int_S \mathrm{T}(\boldsymbol{x}) \frac{\mathrm{d}}{\mathrm{d}\theta}[\log p_\theta(\boldsymbol{x})] p_\theta(\boldsymbol{x}) \, \mathrm{d}x \\
&= \mathbb{E}_\theta\left[\mathrm{T}(\boldsymbol{X}) \frac{\mathrm{d}}{\mathrm{d}\theta} \log p_\theta(\boldsymbol{X})\right], \qquad \forall\, \theta \in \Theta. \qquad \text{(I)}
\end{aligned}$$

On the other hand, by (3) with $h \equiv 1$, then

$$0 = \int_S \frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_S \left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(\boldsymbol{x})\right] p_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \forall\,\theta \in \Theta$$

$$\Rightarrow \mathbb{E}_\theta\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(\boldsymbol{X})\right] = 0. \qquad \text{(II)}$$

Combining (I) and (II),

$$\tau'(\theta) = \mathrm{Cov}_\theta\left(\mathrm{T}(\boldsymbol{X}), \frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(\boldsymbol{x})\right),$$

since $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, but the second of these terms vanishes by (II). Thus,

$$\left[\tau'(\theta)^2\right] = \mathrm{Cov}^2_\theta\left(\mathrm{T}(\boldsymbol{x}), \frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(\boldsymbol{X})\right)$$

■