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 $S := \{x_1, x_2, \ldots\} \subset \mathbb{R}$, 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

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

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 $X \overset{d}{=} Y$ (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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$.

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:

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

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

↪**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 $\mathrm{Var}(X) = \mathrm{Var}(\mathbb{E}[X|Y]) + \mathbb{E}[\mathrm{Var}(X|Y)]$. In particular, $\mathrm{Var}(X) \geq \mathrm{Var}(\mathbb{E}[X|Y])$.

§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 $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. 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 $\left(X_2 - \overline{X}_n, \dots, X_n - \overline{X}_n\right)$ 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

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

and hence as the PDFs split, we conclude Y_1 independent of $Y_2, \dots, Y_n$ and so $\overline{X}_n$ independent of $\left(X_2 - \overline{X}_n, \dots, X_n - \overline{X}_n\right)$ 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 $X_1, ..., X_m \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_1, ..., Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$ 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 $X_1, ..., X_n \overset{\text{iid}}{\sim} F$. Then, the *order statistics* are

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

where $X_{(i)}$ the ith largest of $X_1, ..., X_n$. The *sample range* is defined

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