

# STAT24410 NOTES

ADEN CHEN

## CONTENTS

1. Probability	2
2. Joint Distribution	6
3. Statistical Inference	9
Appendix A: Common Distributions	16

- Last update: Tuesday 29<sup>th</sup> October, 2024.
- See [here](#) for the most recent version of this document.

## 1. PROBABILITY

## 1.1. CDF.

1.1.1. *Properties of CDF.*

- Nondecreasing.
- Right continuous.
- $\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1$ .

1.1.2. *Inverse of CDF.*

$$F^-(x) := \inf\{u : x \leq F(u)\}.$$

**Proposition 1.1.** *Let  $F$  be the cdf of  $X$ . If  $F$  is continuous and strictly increasing, then  $Y := F(X) \sim \text{Uniform}[0, 1]$ .*

**Proof.** For any  $y \in [0, 1]$ ,

$$\mathbb{P}(F(X) \leq y) = F(F^{-1}(y)) = y.$$

□

**Proposition 1.2.** *Let  $U \sim \text{Uniform}[0, 1]$  and  $X$  be the cdf of  $X$ . Then  $F^{-1}(U) \stackrel{\mathcal{D}}{=} X$ .*

**Proof.** For any  $x \in [0, 1]$ ,

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

□

*Remark 1.3.* This is useful for simulation.

**1.2. Transformations.** For  $Y := h(X)$ , if  $h$  is one-to-one and differentiable, then

$$f_Y(y) = f_X(h^{-1}(y)) \cdot \left| \frac{dh^{-1}(y)}{dy} \right|.$$

**1.3. Expectation.** For an r.v.  $X$ . We define

$$X^+ = \max\{X, 0\}, \quad X^- = \max\{-X, 0\}.$$

Note that  $X \equiv X^+ - X^-$ .

Since  $X^+$  is nonnegative,

$$\mathbb{E}(X^+) := \int_0^\infty x \, dF(x)$$

in the Riemann–Stieltjes sense, and similarly  $X^-$ .

**Definition 1.4.**  $X$  has expected value if at least one of  $\mathbb{E}(X^+)$  and  $\mathbb{E}(X^-)$  is finite, and when it does

$$\mathbb{E}(X) := \mathbb{E}(X^+) - \mathbb{E}(X^-).$$

**Definition 1.5.** We say  $Y$  **stochastically dominates**  $X$ ,  $Y \succeq X$ , if

$$\mathbb{P}(X > t) \leq \mathbb{P}(Y > t), \quad \forall t.$$

**Proposition 1.6.**

- *Linearity.*
- *If*

$$\int_{\mathbb{R}} |x| f(x) \, dx < \infty$$

*then*

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f(x) \, dx.$$

- *If  $X$  is stochastically dominated by  $Y$  then  $\mathbb{E}(X) \leq \mathbb{E}(Y)$ .*
- *If  $X$  and  $Y$  are independent, then  $\mathbb{E}(XY) = \mathbb{E}(X) \mathbb{E}(Y)$ .*

**Definition 1.7.** The **variance** of  $X$  is given by

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

**Proposition 1.8.**

- $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ .
- $\text{Var}(cX) = c^2 \text{Var}(X)$ .
- *If  $X$  and  $Y$  are independent, then  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ .*

**Proposition 1.9.** *If  $X \geq 0$  and there exists an at most countable subset  $S = \{x_1, x_2, \dots\}$  of isolated points such that  $F_X$  is continuously differentiable on  $[0, \infty) \setminus S$ , then*

$$\mathbb{E}(X) = \sum_{x \in S} x \mathbb{P}(X = x) + \int_0^\infty x F'_X(x) \, dx.$$

**1.4. Probability Inequalities.**

**Theorem 1.10** (Markov's Inequality). *If  $X \geq 0$  and  $c > 0$ , then*

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(X)}{c}.$$

*(Equality is attained when  $\mathbb{P}(X = 0 \text{ or } X = c) = 1$ .)*

**Proof.** Construct

$$Y := \begin{cases} c, & x \geq 0 \\ 0, & X < c. \end{cases}$$

Then  $Y \leq X$  and

$$\mathbb{E}(Y) = c \cdot \mathbb{P}(X \geq c) \leq \mathbb{E}(X).$$

□

**Theorem 1.11** (Chebychev's Inequality). *If  $c > 0$ , then*

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\mathbb{E}[(X - \mu)^2]}{c^2}$$

*for any  $\mu$ .*

**Proof.** Apply Markov's inequality to  $(X - \mu)^2$ .

□

**Theorem 1.12** (Chernoff's Inequality). *If  $c \in \mathbb{R}$  and  $t > 0$ , then*

$$\mathbb{P}(X \geq c) \leq e^{-tc} \mathbb{E}(e^{tX})$$

and

$$\mathbb{P}(X \leq c) \leq e^{tc} \mathbb{E}(e^{-tX}).$$

**Proof.** Apply Markov's inequality to  $e^{tX}$  and  $e^{-tX}$ .  $\square$

**Theorem 1.13** (Weak Law of Large Numbers). *Let  $X_1, X_2, \dots$  be i.i.d. with finite expectation  $\mu$  and variance  $\sigma^2$ . Then as  $n$  goes to infinity,  $\bar{X}_n \xrightarrow{P} \mu$ . That is*

$$\mathbb{P}\left[\left|\bar{X}_n - \mu\right| > \epsilon\right] \longrightarrow 0.$$

**Proof.** Note that  $\mathbb{E}(\bar{X}_n) = \mu$  and  $\text{Var}(\bar{X}_n) = \sigma^2/n$ . Chebyshev's gives

$$\mathbb{P}\left(\left|\bar{X}_n - \mu\right| < \epsilon\right) \leq \frac{\sigma^2}{n \cdot \epsilon^2} \longrightarrow 0$$

as  $n \rightarrow \infty$ .  $\square$

**Proposition 1.14** (Large Deviations). *Let  $X_1, X_2, \dots$  be i.i.d. with finite expectation  $\mu$  and variance  $\sigma^2$ . Let  $c > \mu$ . Then*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\bar{X}_n > c) = -\sup_t [tc - \kappa(t)],$$

where  $\kappa(t) = \log \mathbb{E}(e^{tX})$ .

We do not yet have the tools to prove that this is the limit, but we will use Chernoff's inequality to obtain a bound:

**Proof.** From Chernoff's inequality, for any  $t$  we have

$$\mathbb{P}(\bar{X}_n \geq c) = \mathbb{P}\left(\sum X_i \geq c \cdot n\right) \leq e^{-tnc} \mathbb{E}\left[e^{t(\sum X_i)}\right] = e^{-tnc+n\kappa(t)},$$

where  $\kappa(t) = \log \mathbb{E}(e^{tX})$ . Thus we have

$$\frac{1}{n} \log \mathbb{P}(\bar{X}_n \geq c) \leq -\sup_t [tc - \kappa(t)].$$

$\square$

*Remark 1.15.*

- $\mathbb{E}[e^{tX}]$  is the **moment generating function**.
- $\kappa(t)$  is the **cumulant generating function**.
- $\sup_t [tc - \kappa(t)]$  is the **Legendre Transform**.

**Definition 1.16.**  $X_n$  converges in distribution to  $X$ ,  $X_n \xrightarrow{\mathcal{D}} X$ , if

$$F_{X_n}(x) \longrightarrow F_X(x), \quad \forall x \in \mathbb{R}.$$

**Definition 1.17.** The **moment generating function** of  $X$  is

$$\begin{aligned} M : \mathbb{R} &\longrightarrow [0, \infty] \\ t &\longmapsto \mathbb{E}[e^{tX}]. \end{aligned}$$

**Proposition 1.18.** *Properties of the moment generating function:*

- $\mathbb{E}[X^m] = M_X^{(m)}(0)$  when Fubini grants

$$\mathbb{E}[e^{tX}] = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right] = \sum_{n=0}^{\infty} \frac{t^n \mathbb{E}(X^n)}{n!}.$$

- $M_{cX}(t) = M_X(ct)$ .
- If  $X$  and  $Y$  are independent, then

$$M_{X+Y}(t) = M_X(t) + M_Y(t).$$

- If  $X_1, X_2, \dots$  are i.i.d., then

$$M_{\sum X_i} = \prod M_{X_i}.$$

- $X_n \xrightarrow{\mathcal{D}} X$  if and only if  $M_{X_n} \rightarrow M_X$  in a neighborhood of 0.

**Theorem 1.19** (Central Limit Theorem). *If  $X_1, X_2, \dots$  are i.i.d.,  $\mathbb{E}(X_i) = \mu$ , and  $\text{Var}(X_i) = \sigma^2$ , then*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2).$$

The following proof works only when we have enough regularity; it is meant to provide a certain intuition (the general proof needs complex analysis):

**Proof.** We assume  $\mu = 0$  and consider the mgf.

$$M_{\sum X_i/\sqrt{n}}(t) = M_{\sum X_i}\left(\frac{t}{\sqrt{n}}\right) = \left[M_{X_i}\left(\frac{t}{\sqrt{n}}\right)\right]^n.$$

We obtain an approximation though Taylor:

$$M_X\left(\frac{t}{\sqrt{n}}\right) \approx M_X(0) + \frac{t}{\sqrt{n}}M_X'(0) + \frac{t^2}{n}M_X''(0)$$

Noting that  $M_X'(0) = \mathbb{E}[X] = 0$  and  $M_X''(0) = \mathbb{E}[X^2] = \sigma^2$ , we have

$$M_{\sum X_i/\sqrt{n}}(t) \approx \left[1 + \frac{t^2\sigma^2}{n}\right]^n \longrightarrow e^{t^2\sigma^2}.$$

The last term is precisely the mgf of  $N(0, \sigma^2)$ . □

## 2. JOINT DISTRIBUTION

## 2.1. Random Vectors and Joint Distributions.

**Proposition 2.1.**

•

$$F(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(x) \, dx.$$

- If  $F$  is continuous and differentiable, then  $X$  has density

$$f(X) = \frac{\partial^n F(x)}{\partial x_1 \cdots \partial x_n}.$$

- If  $X_1, X_2, \dots, X_n$  are independent, then

$$F_X(x) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$

- If  $F$  is differentiable, then

$$f_X(x) = f_{X_1}(x_1) \cdots f_{X_n}(x_n),$$

and conversely!

- If  $X = (X_1, X_2, \dots, X_n)$  has density  $f_X$ , then  $X_I$  has density

$$f_I(x_I) = \int_{\mathbb{R}^{n-|I|}} f(x_I, x_{S_n \setminus I}) \, dx_{S_n \setminus I},$$

where  $S_n := \{1, 2, \dots, n\}$  are all the indices. Think “integrating out” the other variables.

## 2.2. Transformations.

**Definition 2.2.** The **Jacobian** of  $g : G \rightarrow H \subset \mathbb{R}^n$ , where  $G$  and  $H$  are open, is given by

$$J_g(y) := \det \left[ \frac{\partial g_i}{\partial y_j} \right].$$

If  $X : \Omega \rightarrow H \subset \mathbb{R}^n$  and  $h : H \rightarrow G \subset \mathbb{R}^n$ , where  $H$  and  $F$  are open, are such that  $h$  is one-to-one and differentiable and  $h^{-1} : G \rightarrow H$  is differentiable. Then  $Y := h(X)$  has density

$$f_Y(y) = \begin{cases} f_X(h^{-1}(y)) \cdot |Jh^{-1}(y)|, & y \in G \\ 0, & y \notin G. \end{cases}$$

**Definition 2.3.** The Gamma function is given by

$$\Gamma(\lambda) := \int_0^\infty e^{-x} x^{\lambda-1} \, dx.$$

**Proposition 2.4.** *Properties:*

- $\Gamma(1) = 1$ .
- $\Gamma(1/2) = \sqrt{\pi}$ .
- $\Gamma(x+1) = x\Gamma(x)$ .
- $\Gamma(n) = (n-1)!$  for any  $n \in \mathbb{N}$ .

**2.3. Conditional distribution.** The continuous case:

**Definition 2.5.** We define the **conditional density** as

$$f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

**2.4. Covariance and Correlation.**

**Definition 2.6.** The **covariance** of random variables  $X$  and  $Y$  is

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mu_X) \cdot (Y - \mu_Y)).$$

Their **correlation** is given by

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

**Proposition 2.7.** *Properties:*

- $\text{Var}(a + bX) = b^2 \text{Var}(X)$ .
- $\text{Cov}(a + bX, c + dY) = bd \text{Cov}(X, Y)$ .
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$ .
- If  $X$  and  $Y$  are independent, then  $\text{Cov}(X, Y) = 0$ . But the converse is not true. For example, if  $Z \sim N(0, 1)$ , and  $S$  and  $T$  are random signs (1 or -1), then  $\text{Cov}(SZ, TZ) = 0$ .

**Theorem 2.8.**

- If  $(X, Y)$  has density  $f$ , then  $X|Y$  has density

$$\frac{f(x, y)}{f_Y(y)}.$$

- If  $(X, Y)$  has a pmf, then  $X|Y$  is discrete with pmf

$$\frac{p(x, y)}{p_Y(y)}.$$

Note that  $E(X|Y = y)$  is a number, and  $\mathbb{E}(X|Y)$  is a random variable.

**Proposition 2.9.**

- (i) If  $X$  and  $Y$  are independent, then

$$\mathbb{E}(X|Y) = \mathbb{E}(X) \quad \text{with probability 1.}$$

- (ii) Law of total expectation / Tower law:

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

- (iii)

$$\mathbb{E}[g(X)h(Y)|Y] = h(Y) \mathbb{E}(g(X)|Y) \quad \text{with probability 1.}$$

And

$$\mathbb{E}[X|T(Y)] = \mathbb{E}[\mathbb{E}[X|T(Y)|Y]] \quad \text{with probability 1.}$$

(iv) *Law of total variations*

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

where

$$\text{Var}(Y|X) := \mathbb{E}(Y^2|X) - (\mathbb{E}(Y|X))^2.$$

**2.5. Rejection Sampling.** If for some constant  $c$  we have

$$h(x) \geq c \cdot f(x), \quad \forall x,$$

then we can obtain a sample from distribution with density  $f$  using samples from distribution with density  $h$  using **rejection sampling**:

- (1) Sample  $Y$  from  $g$  and  $U$  from Uniform(0, 1), with  $Y$  and  $U$  independent.
- (2) Set  $X := Y$  if

$$U \leq \frac{c \cdot f(Y)}{h(Y)}$$

and return to (1) otherwise.

*Remark 2.10.*

- Think sampling on the area under  $f$  (as a subset of the area under  $g$ ).
- Rejection sampling can also be used if

$$f(x) = \frac{g(x)}{N},$$

where  $N$  is an unknown constant (e.g., an integral with numerical approximations but no closed form solutions). We need only find  $h$  such that

$$h(x) \geq cN \cdot g(x).$$

Think

$$h(x) \gg g(x).$$



## 3. STATISTICAL INFERENCE

*Example 3.1.* Modeling lifetime  $T : \Omega \rightarrow [0, \infty)$ .

**Definition 3.2.**

- The **survival** function is defined as

$$S : [0, \infty) \longrightarrow [0, 1]$$

$$x \longmapsto \mathbb{P}(T > x) = 1 - F_Y(x).$$

- The **failure rate** function is defined as

$$h(x) := \frac{f(x)}{S(x)}.$$

*Remark 3.3.*

$$\mathbb{P}(T \leq x + \Delta x | T > x) = \frac{\mathbb{P}[x < T \leq x + \Delta x]}{\mathbb{P}[T > x]} = \frac{F(x + \Delta x) - F(x)}{S(x)} \approx \Delta x \cdot \frac{f(x)}{S(x)} = \Delta x \cdot h(x).$$

Think of an increasing failure rate as “aging.”

Given  $h$  we can recover  $f$ :

$$h(x) = \frac{f(x)}{1 - F(x)} = -\frac{\partial}{\partial x} \log(1 - F(x)).$$

So,

$$\log(1 - F(x)) = -\int_0^x h(t) dt + C.$$

Since  $F(0) = 0$  we know  $C = 0$ . We have

$$s(x) = \exp\left(-\int_0^x h\right)$$

and

$$f(x) = h(x) \exp\left(-\int_0^x h\right).$$

*Example 3.4.*

- If  $h(x) = \lambda$  is a constant function, we have  $T \sim \text{Exponential}(\lambda)$ :

$$f(x) = \lambda \exp\left(-\int_0^x \lambda dt\right) = \lambda \exp(-\lambda x), \quad \forall x > 0.$$

- If  $h(x) = \alpha + \beta x$  with  $\alpha, \beta > 0$ , then  $T$  follows the Gompertz distribution.
- If  $h(x) = \lambda \beta x^{\beta-1}$ , then  $T$  follows the Weibull distribution.

**3.1. Estimating parameters.** We next assume  $T_1, T_2, \dots \stackrel{\text{iid}}{\sim} \text{Exponential}(\lambda)$  and estimate  $\lambda$ .

*Remark 3.5.* Metrics to evaluate an estimator:

- Bias:  $\mathbb{E}(\hat{\lambda}) - \lambda$ .
- Variance:  $\text{Var}[\hat{\lambda}]$ .
- Mean Squared Error:  $\text{MSE}[\hat{\lambda}] = \mathbb{E}[(\hat{\lambda} - \lambda)^2] = \text{Bias}^2 + \text{Variance}$ .

**Definition 3.6.** An estimator  $\hat{\theta}_n$  of  $\theta$  is said to be **consistent** if

$$\hat{\theta}_n \xrightarrow{P} \theta.$$

That is, if for any  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) = 0.$$

### 3.1.1. Asymptotic Estimation.

**Definition 3.7** (Method of Moments). Let  $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} F$  with  $n$  parameters. To estimate the parameters, we equate  $n$  (usually the first  $n$ ) theoretical moments to the  $n$  corresponding sample moments:

$$\mathbb{E}[X^k] = \frac{1}{n} \sum X_i^k, \quad 1 \leq k \leq n.$$

*Example 3.8.* Consider  $T_n \stackrel{\text{iid}}{\sim} \text{Exponential}(\lambda)$ .

- Since  $\mathbb{E}(\bar{T}_n) = 1/\lambda$ , we may use  $\hat{\lambda} := 1/\bar{T}_n$  as an estimator for  $\lambda$ .
- Since

$$\mathbb{E}\left[\sum T_i^2/n\right] = \frac{2}{\lambda^2},$$

we may also use

$$\hat{\lambda}_2 = \sqrt{\frac{2n}{\sum T_i^2}}$$

as an estimator.

*Example 3.9.*

- Consider  $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Uniform}[0, \theta]$ . We have  $\mathbb{E}[X] = \theta/2$ .  

$$\hat{\theta} := 2\hat{X}.$$
- Consider  $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$ . We have  $\mathbb{E}[X] = \alpha/\beta$  and  $\mathbb{E}[X^2] = \alpha/\beta^2 + (\alpha/\beta)^2$ . Thus we solve

$$\frac{\hat{\alpha}}{\hat{\beta}} = \bar{X}, \quad \frac{\hat{\alpha}}{\hat{\beta}^2} + \frac{\hat{\alpha}^2}{\hat{\beta}^2} = \frac{\sum X_i^2}{n}.$$

The following theorems help us characterize these estimators.

**Theorem 3.10** (Continuous mapping theorem).

- (i) if  $X_n \xrightarrow{P} X$  and  $g$  is continuous, then  $g(X_n) \xrightarrow{P} g(X)$ .
- (ii) If  $X_n \xrightarrow{\mathcal{D}} X$  and  $g$  is continuous, then  $g(X_n) \xrightarrow{\mathcal{D}} g(X)$ .

**Lemma 3.11** (Slutsky). If  $X_n \xrightarrow{\mathcal{D}} X$  and  $Y_n \xrightarrow{P} c$ , where  $c$  is a constant, then

$$X_n + Y_n \xrightarrow{\mathcal{D}} X + c, \quad X_n Y_n \xrightarrow{\mathcal{D}} cX.$$

**Theorem 3.12** (Delta Method). *If  $X_n$  is such that*

$$\sqrt{n}(X_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

*and  $g$  is continuously differentiable, then*

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2 [g'(\theta)]^2).$$

*Remark 3.13.* Intuition: We can write

$$\sqrt{n}(g(X_n) - g(\theta)) = g'(\tilde{\theta}_n) \sqrt{n}(X_n - \theta), \quad \tilde{\theta}_n \in (x_n, \theta).$$

We know that  $g'(\tilde{\theta}_n) \xrightarrow{P} g'(\theta)$  and  $\sqrt{n}(X_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$ , so Slutsky's gives the desired result.

We can now characterize estimators obtained from the method of moments:

**Proposition 3.14.**

- *Non-uniqueness: we can obtain multiple estimators.*
- *Consistency: Law of Large Numbers gives*

$$\bar{X} \xrightarrow{P} \mathbb{E}[X],$$

*and the continuous mapping theorem then gives consistency (under certain conditions).*

- *Asymptotic normality: the Delta Method gives normality (under certain conditions).*

3.1.2. *Estimators for Smaller  $n$ .* We can obtain the exact distribution of  $\bar{T}_n$ . Since

$$T \stackrel{\text{iid}}{\sim} \text{Exponential}(\lambda) = \text{Gamma}(1, \lambda),$$

we know by the properties of gamma distributions that

$$\sum T_i \sim \text{Gamma}(n, \lambda).$$

Again by properties of gamma distributions, we know that the estimator  $\hat{\lambda}_1 := 1/\bar{T}_n$  is biased for small  $n$ :

$$\mathbb{E}[\hat{\lambda}_1] = n \cdot \mathbb{E} \left[ \frac{1}{\sum T_i} \right] = \frac{n\lambda}{n-1}.$$

The estimator

$$\hat{\lambda}_3 := \frac{n-1}{n} \hat{\lambda}_1,$$

then, is unbiased and has smaller variance.

*Remark 3.15.* This is a rare case. Oftentimes, we have instead a trade off between bias and variance.

**3.2. Maximum Likelihood Estimator.** The above may be summed up as the following steps:

- Estimators
- Evaluations
- Distribution for estimators (which allows for the construction of probabilistic statements)

Maximum Likelihood estimator accomplishes all the above in a streamlined fashion.

**Definition 3.16.** Let  $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} F_\theta$ , where  $\theta \in \mathbb{R}^k$  is a parameter for the distribution. Let  $f(x, \theta)$ <sup>1</sup> be the density or pmf of  $F_\theta$ . The **Likelihood** of  $\theta$  given observations  $X_1, X_2, \dots, X_n$  is

$$L(\theta) = L_n(\theta) := \prod_{i=1}^n f(X_i, \theta).$$

The **maximum likelihood estimator** is the value at which  $L$  obtains its maximum:

$$\hat{\theta} = \hat{\theta}_n := \arg \max_{\theta} L(\theta).$$

*Remark 3.17.* It is often easier to work with the **log likelihood**:

$$\ell(\theta) = \ell_n(\theta) := \log L(\theta).$$

*Remark 3.18.*

- Might be non-unique. Consider  $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Uniform}(\theta, \theta + 1)$ .
- Might not exist. Consider  $X_1, X_2, \dots, X_n$  iid with density

$$f(x, \mu, \sigma^2) = \frac{1}{2} \left[ \frac{1}{\sqrt{2\pi}} e^{-x^2/2} + \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right].$$

Think  $X \sim \mathcal{N}(0, 1)$  with probability 1/2 and  $X \sim \mathcal{N}(\mu, \sigma^2)$  with probability 1/2. The likelihood function is unbounded:

$$\max_{\mu, \sigma^2} L(\mu, \sigma^2) \geq \max_{\sigma} L(X_1, \sigma^2) \geq \frac{1}{2^n} \left[ \frac{1}{\sqrt{2\pi}\sigma} \right] \prod_{k=1}^n e^{-X_k^2/2}.$$

### 3.3. Likelihood Theory.

**Definition 3.19 (Score Function).**

$$\dot{\ell}_n(\theta) := \frac{\partial}{\partial \theta} \ell_n(\theta) = \sum_{i=1}^n \frac{\frac{\partial f}{\partial \theta}(x_i, \theta)}{f(x_i, \theta)} = \sum_{i=1}^n \frac{f'(x_i, \theta)}{f(x_i, \theta)}.$$

*Remark 3.20.* We find the MLE by setting the score function to 0.

**Proposition 3.21.** If  $f(x, \theta)$  has common support, that is, if  $\{x : f(x, \theta) > 0\}$  does not depend on  $\theta$ , then

$$\mathbb{E}_{\theta_0} \left[ \frac{L_n(\theta)}{L_n(\theta_0)} \right] = 1.$$

<sup>1</sup>Some also write  $f_\theta(x)$  or  $f(x|\theta)$ .

Equivalently,

$$\mathbb{E} [\exp (\ell_n(\theta) - \ell_n(\theta_0))] = 1.$$

**Proposition 3.22.** *If the density functions are smooth, then*

- (a)  $\mathbb{E}_\theta [\dot{\ell}_n(\theta)] = 0.$
- (b)  $-\mathbb{E}_\theta [\ddot{\ell}_n(\theta)] = \mathbb{E} [\dot{\ell}_n(\theta)^2].$

**Definition 3.23 (Fisher Information).**

$$I(\theta) := \mathbb{E}_\theta [\dot{\ell}(\theta)^2] = \mathbb{E}_\theta [-\ddot{\ell}(\theta)].$$

That is,

$$I(\theta) := \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta} \log f(X, \theta) \right)^2 \right] = -\mathbb{E} \left[ \frac{\partial^2}{\partial \theta^2} \log f(X, \theta) \right],$$

where the expectation is taken with respect to  $X \sim f(x, \theta)$ .

**Remark 3.24.** Intuitively, the more variation there is in the density functions  $f(x, \theta)$  as we vary  $\theta$ , the more information we can get from data. Fisher information measures the variation in density functions by looking at its derivative.

**Theorem 3.25 (Cramér–Rao Inequality).** *Let  $T(X_n)$  be any unbiased estimator for  $g(\theta)$ . Then*

$$\text{Var}[T(X_n)] \geq \frac{[g'(\theta)]^2}{nI(\theta)}.$$

**Theorem 3.26 (Fisher).** *Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x|\theta_0)$ , with  $f$  satisfying certain smoothness conditions. As  $n \rightarrow \infty$ , we have*

$$\sqrt{nI(\theta_0)} \cdot (\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

and

$$\sqrt{nI(\hat{\theta})} \cdot (\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

**Remark 3.27.** One may also think

$$\hat{\theta} \approx \mathcal{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right).$$

**Proposition 3.28.** *Assumptions:*

- *Common support:*  $\{x : f(x, \theta) > 0\}$  does not depend on  $x$ .
- *Smoothness of densities.*
- *Distinct densities:* if  $\theta_1 \neq \theta_2$  then  $f(x, \theta_1) \neq f(x, \theta_2)$ .

*Properties of maximal likelihood estimators under the above assumptions:*

- *consistency,*
- *asymptotic normality,*
- *has known and optimal asymptotic variance (efficiency). That is, it attains the Cramér–Rao bound.*
- *Invariance in the following sense:*

**Theorem 3.29.** *If  $\hat{\theta}_n$  is an MLE of  $\theta$ , then  $\hat{\tau}_n := g(\hat{\theta}_n)$  is an MLE of  $g(\theta)$ .*

### 3.4. Jensen Inequality.

**Theorem 3.30.** *If  $f : \mathbb{R} \rightarrow \mathbb{R}$  is convex and  $X$  is a random variable such that  $\mathbb{E}|X| < \infty$ , then*

$$f(\mathbb{E} X) \leq \mathbb{E} f(X).$$

**Proof.** From the convexity of  $f$  we know  $f(x) \geq f(y) + f'(y)(x - y)$  for any  $x$  and  $y$ . Setting  $y = \mu =: \mathbb{E} X$  gives

$$f(X) \geq f(\mu) + f'(\mu)(X - \mu), \quad \forall x, y.$$

Taking expectation on both sides gives the desired result.  $\square$

#### 3.4.1. Applications of Jensen Inequality.

- If  $f$  is concave, then  $f(\mathbb{E} X) \geq \mathbb{E} f(X)$ .
- The convex function  $x \mapsto x^2$  and the concave function  $x \mapsto \log x$  give two special cases:

$$(\mathbb{E} X)^2 \leq \mathbb{E} X^2, \quad \log \mathbb{E} X \geq \mathbb{E} \log X.$$

- If  $x_1, x_2, \dots, x_n > 0$  and  $p_i \geq 0$  such that  $\sum p_i = 1$ , then

$$\prod x_i^{p_i} \leq \sum p_i x_i.$$

*Remark 3.31.* When  $p_i = 1/n$ , this result reduces to the geometric mean-arithmetic mean equality.

**Proof.** Let  $X$  be a discrete variable such that  $\mathbb{P}(X = x_i) = p_i$ . Then

$$\sum p_i \log x_i = \mathbb{E} \log X \leq \log \mathbb{E} X \leq \sum p_i x_i.$$

Taking exp on both sides gives the desired result.  $\square$

- **Hölder's inequality:** If  $X, Y \geq 0$  are random variables and  $p, q > 0$  are such that  $1/p + 1/q = 1$ , then

$$\mathbb{E}(XY) \leq (\mathbb{E} X^p)^{1/p} \cdot (\mathbb{E} Y^q)^{1/q}.$$

**Proof.** If  $\mathbb{E} X^p = \mathbb{E} Y^q = 1$ , then

$$XY = (X^p)^{1/p} (Y^q)^{1/q} \leq \frac{1}{p} X^p + \frac{1}{q} Y^q,$$

where the last inequality follows from the previous result. Taking expectation on both sides then gives  $\mathbb{E}[XY] \leq \mathbb{E} X^p \mathbb{E} Y^q$ .

For the general case, normalize by setting

$$\tilde{X} := \frac{X}{(\mathbb{E} X^p)^{1/p}}, \quad \tilde{Y} := \frac{Y}{(\mathbb{E} Y^q)^{1/q}}.$$

$\square$

- **Cauchy Inequality:** Taking  $p = q = 2$  in Hölder gives

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

- The consistency of likelihood.

### 3.5. Multivariate Normal.

**Definition 3.32.** The random vector  $X = (X_1, X_2, \dots, X_k)$  is said to follow a **multivariate normal distribution** if for each  $a \in \mathbb{R}^k$ ,  $a^\top X$  is normal. We write

- $\mu = \mathbb{E} X \in \mathbb{R}^k$ .
- $\Sigma = \text{Var}(X) = \mathbb{E} [(X - \mu)(X - \mu)^\top] \in \mathbb{R}^{2k}$ .

**Proposition 3.33.** • If  $\Sigma$  is positive definite, then  $X$  has density

$$f(X) = \frac{1}{(2\pi)^{k/2} \det(\Sigma)} \exp \left( -\frac{1}{2} (X - \mu)^\top \Sigma^{-1} (X - \mu) \right).$$

- If  $(X_1, X_2)$  is bivariate normal and  $\text{Cov}(X_1, X_2) = 0$ , then  $X_1$  and  $X_2$  are independent.
- If  $U \sim N_k(\mu, \Sigma)$ ,  $a \in \mathbb{R}^p$ , and  $B$  is a  $p \times k$  matrix, then

$$V = a + BU \sim N_p(a + B\mu, B\Sigma B^\top).$$

## APPENDIX A: COMMON DISTRIBUTIONS

Distribution	Support	PMF	Mean	Variance
Binomial( $n, p$ )	$\{0, 1, 2, \dots, n\}$	$\binom{n}{x} p^x (1-p)^{n-x}$	$np$	$np(1-p)$
Geometric( $p$ )	$\{1, 2, 3, \dots\}$	$(1-p)^{x-1} p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson( $\lambda$ )	$\{0, 1, 2, \dots\}$	$\frac{\lambda^x e^{-\lambda}}{x!}$	$\lambda$	$\lambda$

TABLE 1. Key Properties of Discrete Distributions

Distribution	Support	PDF	Mean	Variance
Uniform( $a, b$ )	$[a, b]$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathcal{N}(\mu, \sigma^2)$	$(-\infty, \infty)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mu$	$\sigma^2$
Exponential( $\lambda$ )	$[0, \infty)$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma( $\alpha, \beta$ )	$(0, \infty)$	$\frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Beta( $\alpha, \beta$ )	$(0, 1)$	$\frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

TABLE 2. Key Properties of Continuous Distributions

**Proposition 3.34.** *Properties of Exponential distribution:*

(i) The “memoryless” property:

$$\mathbb{P}(T \leq x+y | T > x) = \mathbb{P}(T \leq y).$$

(ii)  $\text{Exponential}(\lambda) = \text{Gamma}(1, \lambda)$ .

**Proposition 3.35.** *Properties of Gamma distribution:*

(i) If  $X_i \stackrel{\text{iid}}{\sim} \text{Gamma}(\alpha_i, \beta)$  for  $i = 1, 2, \dots, N$ , then

$$\sum X_i \sim \text{Gamma}\left(\sum \alpha_i, \beta\right).$$

(ii) If  $X \sim \text{Gamma}(\alpha, \beta)$  and  $\alpha > 1$ , then

$$\mathbb{E}[1/X] = \frac{\beta}{\alpha-1}.$$



**Proof.**

(i) Note that

$$\mathbb{E} \left[ e^{tX_i} \right] = \left( 1 - \frac{t}{\beta} \right)^{-\alpha_i}, \quad \forall t < \beta.$$

We then have

$$M_{\sum X_i}(t) = \prod M_{X_i}(t) = \left( 1 - \frac{t}{\beta} \right)^{-\sum \alpha_i}.$$

(ii) We have

$$\mathbb{E}[1/X] = \int_0^\infty \frac{1}{x} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} e^{-\beta x} dx,$$

which we can integrate by reducing to the  $\Gamma$  function.

□