# STAT24410 NOTES

ADEN CHEN

| Quarter | Instructor | Last Updated |
|---|---|---|
| 2024 Fall | Dan L. Nicolae | Thursday 2$^{\text{nd}}$ January, 2025 |

See here for the most recent version of this document.

Contents

# 1. Probability

## 1.1. The Cumulative Distribution Function.

**Proposition 1.1.** *Properties of the CDF:*

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

**Definition 1.2.** The **generalized inverse distribution function** is defined as

$$F^{-}(x) \coloneqq \inf\{u : x \le F(u)\}.$$

**Proposition 1.3.** *Let $F$ be the CDF of $X$. If $F$ is continuous and strictly increasing, then $Y \coloneqq F(X) \sim \mathrm{Uniform}[0,1]$.*

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

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

□

**Proposition 1.4.** *Let $U \sim \mathrm{Uniform}[0,1]$ and $F$ be the CDF of $X$. Then $F^{-1}(U) \sim F$.*

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

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

□

*Remark* 1.5. This is useful for simulation.

## 1.2. Transformations.

For $Y \coloneqq h(X)$, if $h$ is one-to-one and differentiable, then

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

## 1.3. Expectation.

For an random variable $X$. We define

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

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

Since $X^+$ is nonnegative, we may define

$$\mathrm{E}(X^+) \coloneqq \int_0^\infty x \,\mathrm{d}F(x)$$

in the Riemann–Stieltjes sense, and similarly $\mathrm{E}(X^-)$.

**Definition 1.6.** $X$ has expected value if at least one of $\mathrm{E}(X^+)$ and $\mathrm{E}(X^-)$ is finite, and when it does we define

$$\mathrm{E}(X) \coloneqq \mathrm{E}(X^+) - \mathrm{E}(X^-).$$

**Definition 1.7.** We say $Y$ **stochastically dominates** $X$, $Y \succeq X$, if for each $t$ we have $\mathbb{P}(X > t) \leq \mathbb{P}(Y > t)$.

**Proposition 1.8.** *Properties of* E*:*

- *Linearity.*
- *If*
$$\int_{\mathbb{R}} |x| f(x)\, \mathrm{d}x < \infty$$
*then*
$$\mathrm{E}(X) = \int_{\mathbb{R}} x f(x)\, \mathrm{d}x.$$
- *If $X$ is stochastically dominated by $Y$ then* $\mathrm{E}(X) \leq \mathrm{E}(Y)$.
- *If $X$ and $Y$ are independent, then* $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$.
- *(Hille)* E *commutes with closed (in particular, continuous) linear operators.*

**Definition 1.9.** The **variance** of $X$ is defined as
$$\mathrm{Var}(X) := \mathrm{E}[(X - \mathrm{E}(X))^2]$$

**Proposition 1.10.** *Properties of* Var*:*

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

**Proposition 1.11.** *If $X \geq 0$ and there exists an at most countable subset $S = \{x_1, x_2, \ldots\}$ of isolated points such that $F_X$ is continuously differentiable on $[0, \infty) \setminus S$, then*
$$\mathrm{E}(X) = \sum_{x \in S} x\mathbb{P}(X = x) + \int_0^\infty x F_X'(x)\, \mathrm{d}x.$$

1.4. **Probability Inequalities.**

**Theorem 1.12** (Markov's Inequality)**.** *If $X \geq 0$ and $c > 0$, then*
$$\mathbb{P}(X \geq c) \leq \frac{\mathrm{E}(X)}{c}.$$
*(Equality is attained when $\mathbb{P}(X = 0$ or $X = c) = 1$.)*

**Proof.** Construct
$$Y := c \cdot \mathbb{1}_{\{x \geq c\}}(X).$$
Then $Y \leq X$ and
$$\mathrm{E}(Y) = c \cdot \mathbb{P}(X \geq c) \leq \mathrm{E}(X).$$
□

**Theorem 1.13** (Chebychev's Inequality)**.** *If $c > 0$, then for any $\mu$ we have*

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

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

**Theorem 1.14** (Chernoff's Inequality)**.** *If $c \in \mathbb{R}$ and $t > 0$, then*

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

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

**Theorem 1.15** (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be iid with finite expectation $\mu$ and variance $\sigma^2$. Then as $n$ goes to infinity, $\overline{X}_n \xrightarrow{\mathrm{p}} \mu$. That is,*

$$\mathbb{P}\left[\left|\overline{X_n} - \mu\right| > \epsilon\right] \longrightarrow 0.$$

**Proof.** Note that $\mathrm{E}(\overline{X_n}) = \mu$ and $\mathrm{Var}(\overline{X_n}) = \sigma^2/n$. Chebyshev's gives

$$\mathbb{P}\left(\left|\overline{X_n} - \mu\right| > \epsilon\right) \leq \frac{\sigma^2}{n \cdot \epsilon^2} \longrightarrow 0.$$

□

**Proposition 1.16** (Large Deviations)**.** *Let $X_1, X_2, \ldots$ be iid with finite expectation $\mu$ and variance $\sigma^2$. Let $c > \mu$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\overline{X_n} > c) = -\sup_t [tc - \kappa(t)],$$

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

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

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

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

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

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

□

*Remark* 1.17.

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

**Definition 1.18.** A sequence of random variables $X_n$ **converges in distribution** to $X$, $X_n \xrightarrow{\mathscr{D}} X$, if their cdfs converge pointwise to the cdf of $X$. That is, if

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

**Definition 1.19.** The **moment generating function** of $X$ is

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

**Proposition 1.20.** *Properties of the moment generating function:*

- $\mathrm{E}[X^n] = M_X^{(n)}(0)$ *when Fubini grants so.*
$$\mathrm{E}\left[e^{tX}\right] = \mathrm{E}\left[\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right] = \sum_{n=0}^{\infty} \frac{t^n\,\mathrm{E}(X^n)}{n!}.$$
- $M_{cX}(t) = M_X(ct)$.
- *If $X$ and $Y$ are independent, then*
$$M_{X+Y}(t) = M_X(t) + M_Y(t).$$
- *If $X_1, X_2, \ldots$ are iid, then*
$$M_{\sum X_i} = \prod M_{X_i}.$$
- $X_n \xrightarrow{\mathscr{D}} X$ *if and only if* $M_{X_n} \to M_X$ *in a neighborhood of* $0$.

**Theorem 1.21** (Central Limit Theorem)**.** *If $X_1, X_2, \ldots$ are iid,* $\mathrm{E}(X_i) = \mu$, *and* $\mathrm{Var}(X_i) = \sigma^2$, *then*

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

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 $M_X'(0) = \mathrm{E}[X] = 0$ and $M_X''(0) = \mathrm{E}[X^2] = \sigma^2$, 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 $N(0, \sigma^2)$. $\square$

## 2. Joint Distribution

### 2.1. Random Vectors and Joint Distributions.

**Proposition 2.1.**

- $$F(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(x) \, \mathrm{d}x.$$
- *If $F$ is continuous and differentiable, then $X$ has density*
$$f(X) = \frac{\partial^n F(x)}{\partial x_1 \ldots \partial x_n}.$$
- *If $X_1, X_2, \ldots, X_n$ are independent, then*
$$F_X(x) = F_{X_1}(x_1) \ldots F_{X_n}(x_n).$$
- *If $F$ is differentiable, then*
$$f_X(x) = f_{X_1}(x_1) \ldots f_{X_n}(x_n),$$
*and conversely!*
- *If $X = (X_1, X_2, \ldots, X_n)$ has density $f_X$, then $X_I$ has density*
$$f_I(x_I) = \int_{\mathbb{R}^{n-|I|}} f\left(x_I, x_{S_n \setminus I}\right) \, \mathrm{d}x_{S_n \setminus I},$$
*where $S_n \coloneqq \{1, 2, \ldots, n\}$ are all the indices. Think "integrating out" the other variables.*

### 2.2. Transformations.

**Definition 2.2.** The **Jacobian** of $g : G \to H \subset \mathbb{R}^n$, where $G$ and $H$ are open, is given by
$$Jg(y) \coloneqq \det \left[ \frac{\partial g_i}{\partial y_j} \right].$$

**Proposition 2.3.** *If $X : \Omega \to H \subset \mathbb{R}^n$ and $h : H \to G \subset \mathbb{R}^n$, where $H$ and $G$ are open, are such that $h$ is one-to-one and differentiable and $h^{-1} : G \to H$ is differentiable. Then $Y \coloneqq h(X)$ has density*
$$f_Y(y) = \begin{cases} f_X(h^{-1}(y)) \cdot \left| Jh^{-1}(y) \right|, & y \in G \\ 0, & y \notin G. \end{cases}$$

**Definition 2.4.** The Gamma function is given by
$$\Gamma(\lambda) \coloneqq \int_0^\infty e^{-x} x^{\lambda - 1} \, \mathrm{d}x.$$

**Proposition 2.5.** *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 2.6.** We define the **conditional density** as

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

2.4. **Covariance and Correlation.**

**Definition 2.7.** The **covariance** of random variables $X$ and $Y$ is

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left((X-\mu_X)\cdot(Y-\mu_Y)\right).$$

Their **correlation** is given by

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

**Proposition 2.8.** *Properties:*

- $\mathrm{Var}(a+bX) = b^2\,\mathrm{Var}(X)$.
- $\mathrm{Cov}(a+bX, c+dY) = bd\,\mathrm{Cov}(X,Y)$.
- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$.
- *If $X$ and $Y$ are independent, then* $\mathrm{Cov}(X,Y) = 0$. *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* $\mathrm{Cov}(SZ, TZ) = 0$.

**Theorem 2.9.**

- *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 $\mathrm{E}(X|Y)$ is a random variable.

**Proposition 2.10.**

*(i) If $X$ and $Y$ are independent, then we have* $\mathrm{E}(X|Y) = \mathrm{E}(X)$ *with probability 1 .*
*(ii) Law of total expectation / Tower law:* $\mathrm{E}[\mathrm{E}[X|Y]] = \mathrm{E}[X]$.
*(iii) With probability* 1 *we have the following:*

$$\mathrm{E}[g(X)h(Y)|Y] = h(Y)\,\mathrm{E}(g(X)|Y), \quad \mathrm{E}[X|T(Y)] = \mathrm{E}[\mathrm{E}[X|T(Y)]|Y].$$

(iv) *Law of total variations: we have*

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

*where*

$$\mathrm{Var}(Y|X) \coloneqq \mathrm{E}(Y^2|X) - (\mathrm{E}(Y|X))^2.$$

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 $X \coloneqq Y$ if
$$U \leq \frac{c \cdot f(Y)}{h(Y)}$$
and return to (1) otherwise.

*Remark* 2.11.

- 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. Point Estimates

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

**Definition 3.2.**

- The **survival** function is defined as
$$\begin{aligned} S : [0,\infty) &\longrightarrow [0,1] \\ x &\longmapsto \mathbb{P}(T > x) = 1 - F_Y(x). \end{aligned}$$
- The **failure rate** function is defined as
$$h(x) \coloneqq \frac{f(x)}{S(x)}.$$

*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,
$$\log(1 - F(x)) = -\int_0^x h(t)\mathrm{d}t + C.$$
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 $T \sim \text{Exponential}(\lambda)$:
$$f(x) = \lambda \exp\left(-\int_0^x \lambda \mathrm{d}t\right) = \lambda \exp(-\lambda x), \quad \forall x > 0.$$
- 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 $T_1, T_2, \ldots \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$ and estimate $\lambda$.

*Remark* 3.5. Metrics to evaluate an estimator:
- Bias: $\mathrm{E}(\hat{\lambda}) - \lambda$.
- Variance: $\mathrm{Var}[\hat{\lambda}]$.
- Mean Squared Error: $\mathrm{MSE}[\hat{\lambda}] = \mathrm{E}[(\hat{\lambda} - \lambda)^2] = \text{Bias}^2 + \text{Variance}$.

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

$$\hat{\theta}_n \xrightarrow{\text{p}} \theta.$$

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

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

3.1.1. *Asymptotic Estimation.*

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

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

*Example* 3.8. Consider $T_n \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$.
- Since $\mathrm{E}(\overline{T}_n) = 1/\lambda$, we may use $\hat{\lambda} := 1/\overline{T}_n$ as an estimator for $\lambda$.
- Since
$$\mathrm{E}\left[\sum T_i^2/n\right] = \frac{2}{\lambda^2},$$
we may also use
$$\hat{\lambda}_2 = \sqrt{\frac{2n}{\sum T_i^2}}$$
as an estimator.

*Example* 3.9.
- Consider $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}[0, \theta]$. We have $\mathrm{E}[X] = \theta/2$.
$$\hat{\theta} := 2\hat{X}.$$
- Consider $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$. We have $\mathrm{E}[X] = \alpha/\beta$ and $\mathrm{E}[X^2] = \alpha/\beta^2 + (\alpha/\beta)^2$. Thus we solve
$$\frac{\hat{\alpha}}{\hat{\beta}} = \overline{X}, \quad \frac{\hat{\alpha}}{\hat{\beta}^2} + \frac{\hat{\alpha}^2}{\hat{\beta}^2} = \frac{\sum X_i^2}{n}.$$

The following theorems help us characterize these estimators.

**Theorem 3.10** (Continuous Mapping Theorem)**.**

*(i) if $X_n \xrightarrow{\text{p}} X$ and $g$ is continuous, then $g(X_n) \xrightarrow{\text{p}} g(X)$.*
*(ii) If $X_n \xrightarrow{\mathscr{D}} X$ and $g$ is continuous, then $g(X_n) \xrightarrow{\mathscr{D}} g(X)$.*

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

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

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

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

*and $g$ is continuously differentiable, then*

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

*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 $g'(\tilde{\theta}_n) \xrightarrow{\text{p}} g'(\theta)$ and $\sqrt{n}(X_n - \theta) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma^2)$, 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*

$$\overline{X} \xrightarrow{\text{p}} \mathrm{E}[X],$$

  *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 $\overline{T}_n$. Since

$$T_i \overset{\text{iid}}{\sim} \text{Exponential}(\lambda) = \text{Gamma}(1, \lambda),$$

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 $\hat{\lambda}_1 := 1/\overline{T}_n$ is biased for small $n$:

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

The estimator
$$\hat{\lambda}_3 \coloneqq \frac{n-1}{n}\hat{\lambda}_1,$$
then, is unbiased and has smaller variance.

*Remark* 3.15. This is a rare case. Oftentimes, we have instead a tread 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 $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} F_\theta$, where $\theta \in \mathbb{R}^k$ is a parameter for the distribution. Let $f(x, \theta)$[1] be the density or pmf of $F_\theta$. The **Likelihood** of $\theta$ given observations $X_1, X_2, \dots, X_n$ is
$$L(\theta) = L_n(\theta) \coloneqq \prod_{i=1}^{n} f(X_i, \theta).$$
The **maximum likelihood estimator** is the value at which $L$ obtains its maximum:
$$\hat{\theta} = \hat{\theta}_n \coloneqq \arg\max_{\theta} L(\theta).$$

*Remark* 3.17. It is often easier to work with the **log likelihood**:
$$\ell(\theta) = \ell_n(\theta) \coloneqq \log L(\theta).$$

*Remark* 3.18.

- Might be non-unique. Consider $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(\theta, \theta+1)$.
- 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:
$$\max_{\mu,\sigma^2} L(\mu, \sigma^2) \geq \max_{\sigma} L(X_1, \sigma^2) \geq \frac{1}{2^n}\left[\frac{1}{\sqrt{2\pi}\sigma}\right]\prod_{k=1}^{n} e^{-X_1^2/2}.$$

[1] Some also write $f_\theta(x)$ or $f(x|\theta)$.

## 3.3. Likelihood Theory.

**Definition 3.19.** The **score function** is defined as

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

*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*

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

*Equivalently,*

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

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

(a) $\mathrm{E}_\theta\left[\dot{\ell}_n(\theta)\right] = 0.$

(b) $-\mathrm{E}_\theta\left[\ddot{\ell}_n(\theta)\right] = \mathrm{E}\left[\dot{\ell}_n(\theta)^2\right].$

**Definition 3.23 (Fisher Information).**

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

That is,

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

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,*

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

*Remark* 3.26. The Cramér–Rao lower bound is attained if and only if

$$\mathrm{Corr}(\hat{\theta}(X), \dot{\ell}(X)) = 1.$$

By Cauchy-Schwarz inequality, this is equivalent to $\hat{\theta}(X)$ and $\dot{\ell}(X)$ being linearly related random variables. That is,

$$\dot{\ell}(\theta) = \alpha(\theta)\hat{\theta}(X) + \beta(\theta)$$

for functions $\alpha$ and $\beta$ that do not depend on $X$.

**Proposition 3.27.** *Under the regularity conditions in the Cramér–Rao inequality, there exists an unbiased estimator $\hat{\theta}(X)$ of $\theta$ whose variance attains the Cramér–Rao lower bound if and only if the score can be expressed in the form*

$$\dot{\ell}(\theta) = I(\theta)\left\{\hat{\theta}(X) - \theta\right\},$$

*or, equivalently, if and only if the function*

$$\frac{\dot{\ell}(\theta)}{I(\theta)} + \theta$$

*does not depend on $\theta$.*

**Theorem 3.28** (Fisher)**.** *Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} f(x|\theta_0)$, with $f$ satisfying certain smoothness conditions. As $n \to \infty$, we have*

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

*and*

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

*Remark* 3.29. One may also think

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

**Proposition 3.30.** *Assumptions:*

- *Common support.*
- *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 3.31.** *If $\hat{\theta}_n$ is an MLE of $\theta$, then $\hat{\tau}_n \coloneqq g(\hat{\theta}_n)$ is an MLE of $g(\theta)$.*

3.4. **Jensen Inequality.**

**Theorem 3.32.** *If $f : \mathbb{R} \to \mathbb{R}$ is convex and $X$ is a random variable such that* $\mathrm{E}|X| < \infty$*, then*

$$f(\mathrm{E}\,X) \leq \mathrm{E}\,f(X).$$

**Proof.** From the convexity of $f$ we know $f(x) \geq f(y) + f'(y)(x-y)$ for any $x$ and $y$. Setting $y = \mu =: \mathrm{E}\, X$ gives
$$f(X) \geq f(\mu) + f'(\mu)(X-\mu), \quad \forall x, y.$$
Taking expectation on both sides gives the desired result. □

3.4.1. *Applications of Jensen Inequality.*

- If $f$ is concave, then $f(\mathrm{E}\, X) \geq \mathrm{E}\, f(X)$.
- The convex function $x \mapsto x^2$ and the concave function $x \mapsto \log x$ give two special cases:
$$(\mathrm{E}\, X)^2 \leq \mathrm{E}\, X^2, \quad \log \mathrm{E}\, X \geq \mathrm{E} \log X.$$
- If $x_1, x_2, \ldots, x_n > 0$ and $p_i \geq 0$ such that $\sum p_i = 1$, then
$$\prod x_i^{p_i} \leq \sum p_i x_i.$$

  *Remark* 3.33. When $p_i = 1/n$, this result reduces to the geometric mean-arithmetic mean inequality.

  **Proof.** Let $X$ be a discrete variable such that $\mathbb{P}(X = x_i) = p_i$. Then
$$\sum p_i \log x_i = \mathrm{E} \log X \leq \log \mathrm{E}\, X \leq \sum p_i x_i.$$
  Taking exp on both sides gives the desired result. □
- **Hölder's inequality**: If $X, Y \geq 0$ are random variables and $p, q > 0$ are such that $1/p + 1/q = 1$, then
$$\mathrm{E}(XY) \leq (\mathrm{E}\, X^p)^{1/p} \cdot (\mathrm{E}\, Y^q)^{1/q}.$$

  **Proof.** If $\mathrm{E}\, X^p = \mathrm{E}\, X^q = 1$, then
$$XY = (X^p)^{1/p}(Y^q)^{1/q} \leq \frac{1}{p}X^p + \frac{1}{q}X^q,$$
  where the last inequality follows from the previous result. Taking expectation on both sides then gives $\mathrm{E}[XY] \leq \mathrm{E}\, X^p\, \mathrm{E}\, Y^q$.

  For the general case, normalize by setting
$$\tilde{X} := \frac{X}{(\mathrm{E}\, X)^{1/p}}, \quad \tilde{Y} := \frac{Y}{(\mathrm{E}\, Y)^{1/q}}.$$
  □
- **Cauchy Inequality**: Taking $p = q = 2$ in Hölder gives
$$\mathrm{E}|XY| \leq \sqrt{\mathrm{E}\, X^2}\sqrt{\mathrm{E}\, Y^2}.$$
- The consistency of likelihood.

### 3.5. **Multivariate Normal.**

**Definition 3.34.** The random vector $X = (X_1, X_2, \dots, X_k)$ is said to follow a **multivariate normal distribution** if for each $a \in \mathbb{R}^k$, $a^\intercal x$ is normal. We write

- $\mu = \operatorname{E} X \in \mathbb{R}^k$.
- $\Sigma = \operatorname{Var}(X) = \operatorname{E}\left[(X-\mu)(X-\mu)^\intercal\right] \in \mathbb{R}^{2k}$.

**Proposition 3.35.**

- *If $\Sigma$ is positive definite, then $X$ has density*
$$f(X) = \frac{1}{(2\pi)^{k/2}\det(\Sigma)} \exp\left(-\frac{1}{2}(X-\mu)^\intercal \Sigma^{-1}(X-\mu)\right).$$
- *If $(X_1, X_2)$ is bivariate normal and* $\operatorname{Cov}(X_1, X_2) = 0$*, 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*
$$V = a + BU \sim N_p(a + B\mu, B\Sigma B^\intercal).$$

## 4. Confidence Intervals

**Definition 4.1.** Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F_\theta$. Confidence intervals (CIs) are probabilistic statements on data of the form

$$\mathbb{P}_\theta\left[A(X_1, \ldots, X_n) \le \theta \le B(X_1, \ldots, X_n)\right] = \alpha.$$

The interval

$$[A(X_1, \ldots, X_n), B(X_1, \ldots, X_n)]$$

is called a $\alpha \cdot 100\%$ **confidence interval**.

*Remark* 4.2. We are typically interested in $\alpha = 0.95$ or $\alpha = 0.99$.

*Remark* 4.3.

- The probabilistic statement concerns the interval ends, not $\theta$, which is fixed. The interval ends are random variables.
- Interpretation (frequentest): the long run frequency of the CI covering $\theta$ is $\alpha$.

**Definition 4.4.** The $\alpha$ quantile of $X \sim F$, $q_\alpha$, is such that

$$\mathbb{P}[X \le q_\alpha] = \alpha.$$

*Example* 4.5. Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(\lambda) = \text{Gamma}(1, \lambda)$. Then $\sum X_i \sim \text{Gamma}(n, \lambda)$ and $\lambda \sum X_i \sim \text{Gamma}(n, 1)$. Note that the distribution of $\lambda \sum X_i$ does not depend on $\lambda$. We then have

$$\left[\frac{q_{0.025}}{\sum X_i}, \frac{q_{0.975}}{\sum X_i}\right],$$

where $q$ refers to the quantile of $\text{Gamma}(n, 1)$, is a 95% CI.

**Definition 4.6.** Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F_\theta$. The function

$$g(X_1, \ldots, X_n, \theta)$$

is called a **pivot** if its distribution does not depend on $\theta$.

*Remark* 4.7. One may use the distribution of the pivot $g(X_1, \ldots, X_n, \theta) \sim F^*$ to build CIs. Let $L$ and $U$ be the $(1-\alpha)/2$ and $1-(1-\alpha)/2$ quantiles for $F^*$. Then

$$\alpha = \mathbb{P}\left[L \le g(X_1, \ldots, X_n, \theta) \le U\right] = \mathbb{P}\left[\theta \in S(X_1, \ldots, X_n, L, U)\right]$$

for some set $S$. If $S$ is an interval, it is a CI.

**Theorem 4.8.** *Let* $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. *Let*

$$\overline{X_n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad S^2 \coloneqq \frac{1}{n-1} \sum (X_i - \overline{X})^2.$$

*Then*

$$\sqrt{n} \cdot \frac{\overline{X} - \mu}{S} \sim t_{n-1}, \quad (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

*Remark* 4.9. Thus $\sqrt{n} \cdot \frac{\overline{X}-\mu}{S}$ is a pivot estimator for $\mu$ and $(n-1)\frac{S^2}{\sigma^2}$ is a pivot estimator for $\sigma$.

*Remark* 4.10. We may use the central limit theorem and the above results to obtain approximate CIs for large samples. Let $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} F$ with $\mathrm{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$. The central limit theorem gives

$$\sqrt{n} \cdot \frac{\overline{X} - \mu}{\sigma} \approx \mathcal{N}(0, 1).$$

Thus

$$\left[\overline{X} - q_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + q_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

where $q$ is the quantiles on $\mathcal{N}(0, 1)$ contains $\mu$ with probability $\alpha$. We can approximate $\sigma$ using $S$ to obtain the following CI:

$$\left[\overline{X} - q_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + q_{\alpha/2}\frac{S}{\sqrt{n}}\right].$$

Note that we used two approximations: central limit theorem and using $S$ to approximate $\sigma$.

*Remark* 4.11. For a MLE $\hat{\theta}$, we can use the following two results to construct approximate CIs:

$$\sqrt{n}(\hat{\theta} - \theta) \approx \mathcal{N}\left(0, \frac{1}{I(\theta)}\right), \sqrt{nI(\theta)}(\hat{\theta} - \theta) \approx \mathcal{N}(0, 1).$$

*Remark* 4.12. The above cases fail, however, if either the distribution of the pivot or the variance of the estimators is unknown.

## 5. The Bootstrap

**Definition 5.1.** Let $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} F$. The **empirical distribution function (edf)**, $\hat{F}_n$, is the CDF that puts probability $1/n$ at each $X_i$.

$$\hat{F}_n(x) \coloneqq \frac{1}{n} \sum \mathbb{1}_{\{X_i \leq x\}}.$$

*Remark* 5.2. Note that $\mathbb{1}_{\{X_i \leq x\}} \sim \text{Bernoulli}(F(x))$. This gives the following properties:

**Proposition 5.3.**

- *$\hat{F}(x)$ is an unbiased estimator for $F(x)$:*
$$\mathrm{E}[\hat{F}(x)] = F(x).$$
- *$\hat{F}(x)$ has variance:*
$$\mathrm{Var}(\hat{F}(x)) = \frac{F(x)(1 - F(x))}{n}.$$
- *By the law of large numbers,*
$$\hat{F}(x) \xrightarrow{\text{p}} F(x).$$
*Moreover, $\hat{F}_n(x) \to F(x)$ uniformly. That is:*

**Theorem 5.4** (Glivenko-Cantelli). *If $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} F$, then as $n \to \infty$ we have*

$$\sup_x \left| \hat{F}_n(x) - F(x) \right| \longrightarrow 0.$$

*Remark* 5.5. For variable $\theta \coloneqq T(F)$, we can thus construct estimator $\hat{T} \coloneqq T(\hat{F})$.

*Example* 5.6. For $T = \int x \, \mathrm{d}F(x)$, $\theta$ is the mean. For $T = \int (x - \mu)^2 \, \mathrm{d}F(x)$, $\theta$ is the variance. For $T = F^{-1}(1/2)$, $\theta$ is the median.

*Remark* 5.7. Let $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} F$ and $T_n \coloneqq g(X_1, \dots, X_n)$. We want to find $\mathrm{Var}(T_n)$. If it is possible to sample from $F$, then we may repeated the following procedure

- Take repeated samples of size $n$.
- Calculate $T_n$ for each sample.

to obtain $k$ samples of $T_n$, $T_{n,1}, \dots, T_{n,k}$. We may use

$$\frac{1}{k} \sum \left( T_{n,j} - \overline{T}_n \right)^2$$

as an estimator for the variance of $T_n$, $\mathrm{Var}_F(T_n)$.

*Remark* 5.8. If we cannot directly sample from $F$, we may use $\hat{F}$ as an approximation. That is, given a sample of size $n$, we sample repeatedly with replacement $k$ samples also of size $n$ from the given sample, and calculate the statistic of interest for each sample to estimate the distribution of $T_n$. This procedure is called **bootstrapping**, and each sample is called a **bootstrap sample**.

## 6. Hypothesis Testing

We want to test whether a set of given data is generated by a certain data generating model.

The idea: we use a certain distance between the ecdf and the theoretical cdf in the density space as a test statistic.

*Example* 6.1. Given $X_i \overset{\text{iid}}{\sim} F$, we want to test if $F$ is the cdf of a normal distribution. Test statistic:

- **Kolmogorov–Smirnov**: $S := \sup_x \left|F(x) - \hat{F}(x)\right|$.
- Quantiles: e.g., compare $Q_3 - Q_1$ with $X_{(\lfloor 3N/4 \rfloor)} - X_{(\lfloor N/4 \rfloor)}$.
- **Shapiro-wilk**:
$$W := \frac{\left(\sum a_i x_{(i)}\right)^2}{\sum (x - \overline{x})^2}.$$

### 6.1. Hypothesis Testing for Parametric Models.

Let $X_i \overset{\text{iid}}{\sim} F_\theta$ with $\theta \in \Omega$. The **null hypothesis**:
$$H_0 : \theta \in \Omega_0 \subset \Omega.$$
The **alternative hypothesis**:
$$H_A : \theta \in \Omega_1.$$
We often have $\Omega_1 = \Omega \setminus \Omega_0$.

*Remark* 6.2. Note a certain asymmetry: we usually know a lot more about $H_0$ (the "status quo") than $H_1$.

**Definition 6.3.** Let $S$ be the set of all possible values for $X = (X_1, \dots, X_n)$. The values for which we do not reject $H_0$, $S_0$, is called the **acceptance region**. The values for which we reject $H_0$, $S_1$, is called the **rejection region**. Note that we require $S = S_0 \cup S_1$.

**Definition 6.4.** $T = T(X)$ is called a **test statistic** if
$$S_1 = \{x : T(x) \in R_1\}$$
for some $R_1 \subset \mathbb{R}$.

**Definition 6.5.** A **type I error**, or a false positive, is the rejection of the null hypothesis when it is actually true. A **type II error**, or a false negative, is the failure to reject a null hypothesis that is actually false.

**Definition 6.6.** The function
$$\pi : \Omega \longrightarrow [0, 1], \quad \pi(\theta) := \mathbb{P}_\theta(x \in S_1)$$
is called the **power function**.

*Remark* 6.7. Note we can represent type I errors as $\pi(\theta)$ with $\theta \in \Omega_0$; and type II errors as $1 - \pi(\theta)$ with $\theta \in \Omega_1$. Ideally, we want $\pi$ to be small on $\Omega_0$ and large on $\Omega_1$. We often find $S_1$ such that $\pi$ is low on $\Omega_0$ and hope for the best for $\Omega_1$.

**Definition 6.8.** The **size** of the test is $\sup_{\theta \in \Omega_0} \pi(\theta)$.

**Definition 6.9.** A test is a **level $\alpha$ test** if it has size $\leq \alpha$.

*Remark* 6.10. For convenience of calculating size, we often want either simple $H_0$ such that $\theta = \theta_0$, or the power function to be constant on $\Omega_0$.

*Example* 6.11. Let $X_i \overset{\text{iid}}{\sim} F$ such that $\mathrm{E}[X_i] = \mu$ with known variance $\mathrm{Var}[X_i] = \sigma^2$. Let

$$H_0 : \mu = \mu_0, \quad H_A : \mu > \mu_0.$$

Under $H_0$, the CLT gives

$$T(X) \coloneqq \sqrt{n} \cdot \frac{\overline{X} - \mu_0}{\sigma} \approx \mathcal{N}(0, 1).$$

Then, we may set the rejection region by picking $c$ such that

$$\mathbb{P}_\mu \left(\{T(X) \geq c\}\right) = \alpha.$$

*Example* 6.12. Same set up as above, with

$$H_0 : \mu = \mu_0, \quad H_A : \mu \neq \mu_0.$$

We may set

$$S_1 \coloneqq \{X : |T(X)| > c_2\}$$

to be such that $\mathbb{P}_\mu(X \in S_1) \approx \alpha$.

*Remark* 6.13. If $\sigma$ is unknown, we may use the fact that under $H_0$,

$$\sqrt{n} \cdot \frac{\overline{X} - \mu_0}{S} \sim t_{n-1}.$$

6.2. ***p*-value.**

**Definition 6.14.** The ***p*-value** is the smallest level $\alpha$ for which we reject $H_0$ with the observed data.

**Proposition 6.15.** *If under $H_0$, $T \sim F$, then $p = \mathbb{P}(T \geq T_{obs})$. Moreover,* $F(p) \sim \mathrm{Uniform}[0, 1]$.

## 7. Likelihood Ratio Test

Let $H_0$ and $H_1$ be simple hypotheses (that is, are of the form $\theta = \theta_i$). We may define the test statistic using the Likelihood ratio

$$LR(X) \coloneqq \frac{L(\theta_0|X)}{L(\theta_1|X)} = \frac{\prod f(X_i|\theta_0)}{\prod f(X_i|\theta_1)}$$

and the rejection region as

$$S_1 \coloneqq \{X : LR(X) \leq c\}.$$

We know that this test is the most powerful test (with fixed level) with the following result:

**Theorem 7.1** (Neyman-Pearson Lemma)**.** *With $LR$ and $S_1$ as above, if the type I and type II errors are $\alpha$ and $\beta$, then any other test with $\alpha$ type I error has a type II error larger than $\beta$.*

More generally, we have the following:

**Definition 7.2.** For $H_0 : \theta \in \Omega_0$ and $H_A : \theta \in \Omega_1$, we can define the **likelihood ratio** as follows:

$$\Lambda(X) \coloneqq \frac{\sup_{\theta\in\Omega_0} L(\theta|X)}{\sup_{\theta\in\Omega} L(\theta|X)}.$$

**Theorem 7.3.** *If $\Omega \subset \mathbb{R}^p$ is open and $\Omega_0$ is obtained by fixing $k$ coordinates of $\theta$ and if the assumptions of the MLE hold, then under $H_0$ we have*

$$-2\log\Lambda(X) \xrightarrow{\mathscr{D}} \chi^2_k.$$

*Remark* 7.4. More generally,

$$-2\log\Lambda(X) \xrightarrow{\mathscr{D}} \chi^2_{\dim H_A - \dim H_0}.$$

The below proof is meant to provide a certain intuition. It deals only with the case $p = k = 1$.

**Proof.** Let $p = k = 1$ and $H_0 : \theta = \theta_0$. Let $\hat{\theta}$ be the MLE. We have

$$\Lambda(X) = \frac{L(\theta_0|X)}{L(\hat{\theta}|X)}$$

and thus

$$-2\log\Lambda(X) = 2(\ell(\hat{\theta}) - \ell(\theta_0)).$$

Taylor gives

$$\ell(\theta_0) \approx \ell(\hat{\theta}) + \dot{\ell}(\hat{\theta})(\theta_0 - \hat{\theta}) + \frac{1}{2}\ddot{\ell}(\hat{\theta})(\theta_0 - \hat{\theta})^2.$$

Under certain regularities we have $\dot{\ell}(\hat{\theta}) = 0$. Thus rearranging gives

$$2\left[\ell(\hat{\theta}) - \ell(\theta_0)\right] = \left[\sqrt{n}\left(\hat{\theta} - \theta_0\right)\right]^2 \left[-\frac{1}{n}\ddot{\ell}(\hat{\theta})\right].$$

We complete the proof by noting that under $H_0$, by Fisher's theorem we have $\sqrt(\hat{\theta} - \theta_0) \xrightarrow{\mathscr{D}} \mathcal{N}(0, 1/I(\theta_0))$, and by the law of large numbers we have $-1/n \cdot \ddot{\ell}(\hat{\theta}) \xrightarrow{\mathrm{p}} I(\theta_0)$. □

7.1. **Hypothesis Testing and Confidence Intervals.** Let $X_i \overset{\text{iid}}{\sim} F_\theta$ and $A(\theta_0)$ be the acceptance region for a test with $H_0 : \theta = \theta_0$ at level $\alpha$. We have that

**Proposition 7.5.** *Let $S(X) = \{\theta_0, X \in A(\theta_0)\}$ be the set of parameters rejected by data $X$. Then $S(X)$ is a $1 - \alpha$ confidence set.*

**Proof.**

$$P_\theta(\theta \in S(X)) = \mathbb{P}_\theta(X \in A(\theta)) \geq 1 - \alpha.$$

□

The converse of the above theorem is also true:

**Proposition 7.6.** *Let $S(X)$ be a $1 - \alpha$ confidence set and define*

$$A(\theta_0) \coloneqq \{X : \theta_0 \in S(X)\}.$$

*Then $A(\theta_0)$ is the acceptance region of a level $\alpha$ test of $H_0 : \theta = \theta_0$ and $H_A : \theta \neq \theta_0$.*

**Proof.**

$$\mathbb{P}_{\theta_0}(X \notin A(\theta_0)) = \mathbb{P}_{\theta_0}\left(\theta_0 \notin S(X)\right) \leq \alpha.$$

□

## 8. Multiple Testing

| | Not rejected | rejected | |
|---|---|---|---|
| $H_0$ true | $U$ | $V$ | $m_0$ |
| $H_A$ true | $T$ | $S$ | $m - m_0$ |
| | $m - R$ | $R$ | $m$ |

- $R$ is the number of discoveries, $V$ the number of false discoveries.
- We want $U$ and $S$ to be large, and $V$ and $T$ to be small.

The error rates can be measured by the following:

- Family-wise error rate (FWER): $P[V > 0]$.
- Per-family error rate (PFER): $\mathrm{E}\, V$.
- False discovery rate (FDR): $\mathrm{E}[V/R]$.

We discuss first the case of controlling FWER:

**Proposition 8.1.** *For $m$ independent tests, to obtain $\mathbb{P}[V] < \alpha$ for some $\alpha > 0$, we may reject when $p < \gamma$ for*

$$\gamma \coloneqq 1 - (1 - \alpha)^{1/m}.$$

*This is the **Sidak** correction.*

**Proof.** Noting that under $H_0$ we have $p \sim \mathrm{Uniform}[0, 1]$, we obtain

$$\mathbb{P}[V > 0] = \mathbb{P}[p_{(1)} < \gamma] = 1 - \mathbb{P}[p_{(1)} \geq \gamma] = 1 - (1 - \gamma)^m.$$

For

$$\gamma \coloneqq 1 - (1 - \alpha)^{1/m},$$

we get $1 - (1 - \gamma)^m < \alpha$. □

*Remark* 8.2. Using the approximation $\exp x \approx 1 + x$ for small $x$, we have

$$1 - (1 - \alpha)^{1/m} \approx 1 - e^{-\alpha/m} \approx 1 - \left(1 - \frac{\alpha}{m}\right) = \frac{\alpha}{m}.$$

Thus we may also set $\gamma \coloneqq \alpha/m$. This is the **Bonferroni** correction.

To control FDR (and the PFER), we use the following result:

**Algorithm 8.3** (Benjamini Hochberg procedure, 1985)**.**

- *Sort all p-values in ascending order.* $p_{(1)} \leq \cdots \leq p_{(m)}$.
- *Find the largest $j$ such that*

$$p_{(j)} \leq \frac{\alpha j}{m}.$$

- *Reject the tests with the $j$ smallest $p$-values.*

**Proof.** Let $N \subset \{1, 2, \ldots, m\}$ be the indices of the tests when $H_0$ is true. Note that $|N| = m_0$. Define
$$\alpha_r \coloneqq \frac{\alpha r}{m}, \quad \forall r = 1, 2, \ldots, m.$$
Note that $\alpha_R$ is the $p$-value threshold. We have
$$\mathrm{E}\left[\frac{V}{R}\right] = \mathrm{E}\left[\frac{1}{R}\sum_{k\in N} \mathbb{1}_{\{p_k \le \alpha_R\}}\right] = \sum_{k\in N}\sum_{r=1}^{m} \frac{1}{r}\mathbb{P}\left[p_r \le \alpha_R, R = r\right].$$
Now, define $R_k$ to be the number of false discoveries when doing the BH prodcedure at $\alpha$ with the $k$th $p$-value $p_k$ replaced by 0. Note that
$$\mathbb{P}\left[p_k \le \alpha_R, R = r\right] = \mathbb{P}\left[p_k \le \alpha_r, R_k = r\right] = \mathbb{P}\left[p_k \le \alpha_r\right] \cdot \mathbb{P}\left[R_k = r\right].$$
We thus have
$$\begin{aligned}
\mathrm{E}\left[\frac{V}{R}\right] &= \sum_{k\in N}\sum_{r=1}^{m} \frac{1}{r}\alpha_r \cdot \mathbb{P}[R_k = r] \\
&= \frac{\alpha}{m}\sum_{k\in N}\sum_{r=1}^{R} \mathbb{P}[R_k = r] \\
&= \frac{\alpha m_0}{m} \le \alpha.
\end{aligned}$$
Note that this is a very conservative estimate if $m_0 \ll m$. $\square$

If, on the other hand, we want to find the FDR for the rejection region $[0, \gamma]$, we may note that
$$\frac{V}{R} \approx \frac{m_0 \cdot \gamma}{R}.$$
Thus we need only estimate $m_0$. To do so we note that for $\lambda \in [0, 1]$ we have
$$\#\text{ of } p\text{-values } > \lambda \approx m_0(1 - \lambda).$$
Note however that there is a bias-variance trade-off: for small $\lambda$ this estimator is more biases, since it might include $p$-values generated by $H_A$; for large $\lambda$, on the other hand, fewer tests will have $p$-values larger than $\lambda$, and the estimator has more noise.

## 9. Bayesian Statistic

Recall Bayes' formula:

$$\mathbb{P}\left[A|B\right] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$$

and its generalization: If $H_1, \ldots, H_k$ is a partition of the sample space $\Omega$ and $D$ is an event with $\mathbb{P}[D] > 0$, then

$$P[H_i|D] = \frac{\mathbb{P}[H_i]\mathbb{P}[D|H_i]}{\sum_j \mathbb{P}[H_j]\mathbb{P}[D|H_j]}.$$

We call

- $\mathbb{P}[H_i]$ the **prior**, probabilities
- $\mathbb{P}[D|H_i]$ the **likelihood**,
- $\mathbb{P}[H_i|D]$ the **posterior** probabilities.

In hypothesis testing, we view $\theta$ as a random variable. Given a prior $f(\theta)$ and a model for data $f(X|\theta)$, we can then obtain the posterior $f(\theta|X)$ by

$$f(\theta|X) := \frac{f(X|\theta)f(\theta)}{f(X)},$$

where $f(X)$ is defined as

$$f(X) := \int f(X|\theta)f(\theta)\,\mathrm{d}\theta$$

so that $f(\theta|X)$ is a valid density function.

### 9.1. Credible Intervals.

- Equal tailed $1-\alpha$ credible interval:
$$\left[F^{-1}_{\theta|X}(\alpha/2), F^{-1}_{\theta|X}(1-\alpha/2)\right].$$
- High posterior density interval
$$I = \{\theta : f(\theta|X) \geq c\} \quad \text{s.t.} \quad \mathbb{P}_{\theta|X}(I) = 1-\alpha.$$

*Remark* 9.1.

- In credible intervals, $\theta$ is the random variable, not the interval ends.
- The high posterior density interval might be the union of several intervals.
- The interval lengths of high posterior density intervals are always no longer than those of the corresponding equal tailed credible intervals.

9.2. **Hypothesis Testing.** We use

$$\frac{\mathbb{P}_{\theta|X}(\theta \in \Omega_0)}{\mathbb{P}_{\theta|X}(\theta \in \Omega_1)}$$

*Example* 9.2. Let $\theta$ be the probability of obtaining heads. Let $X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ and let $X \coloneqq \sum X_i$.

- Case 1: Let the prior be $\theta \sim \text{Uniform}[0,1]$. Then
$$f_{\theta|X} \propto \theta^X (1-\theta)^{n-X}, \quad 0 < \theta < 1$$
and we have
$$f_{\theta|X} \sim \text{Beta}(X+1, n-X+1).$$
We have the posterior mean $(X+1)/(n-X+1)$, which we may think of this as the frequentest estimator with two extra flips, one heads and one tails.
- Case 2: Let the prior be $\theta \sim \text{Beta}(\alpha, \beta)$. We have
$$f_{\theta|X} \propto \theta^{\alpha+X-1}(1-\theta)^{\beta+n-X-1}.$$
So
$$f(\theta|X) \sim \text{Beta}(\alpha + X, \beta + n - X).$$
Note that the posterior mean
$$\frac{X+\alpha}{n+\alpha+\beta} = \frac{\alpha}{\alpha+\beta}\frac{\alpha+\beta}{\alpha+\beta+n} + \frac{X}{n}\frac{n}{\alpha+\beta+n}$$
is a convex combination of the prior mean $\alpha/(\alpha+\beta)$ and the data mean $X/n$, and converges to the data mean as $n \to \infty$.

*Remark* 9.3. In case two, we have a family of distribution which when updated results in posterior distributions in the same family. Prior distributions like this are called **conjugate priors**.

*Example* 9.4. Travel to a city; saw a train with number $T$. Suppose trains are numbered $1 \ldots N$. What do we know about $N$? Frequentist's solution: MoM gives $\overline{N} = 2T - 1$. Bayesian: let's assume the prior distribution

$$\theta(N) \propto 1/N.$$

Note that this an **improper prior** since it does not have a density. We have then that

$$\Theta(N|T) \xrightarrow{\text{p}} \frac{1}{N^2}, \quad N \geq T$$

and

$$\mathbb{P}[N \geq x|T] = \frac{\sum_{n \geq x} \frac{1}{n^2}}{\sum_{n \geq 1} \frac{1}{n^2}} \approx \frac{\int_x^\infty \frac{1}{y^2}\,\mathrm{d}y}{\int_1^\infty \frac{1}{y^2}\,\mathrm{d}y} = \frac{T}{x}.$$

*Remark* 9.5.

- $\mathbb{P}[N \geq 2T|T] \approx 1/2$. So the posterior median is $\approx 2T$.
- The posterior mean is $\infty$.

*Example* 9.6 (Exponential Rate). Let $X_i \overset{\text{iid}}{\sim}$ Exponential$(\lambda)$ and let $\lambda \sim$ Gamma$(\alpha, \beta)$. We have

$$f(\lambda|X) \propto f(\lambda)f(X|\lambda) = \lambda^{\alpha-1}e^{-\lambda\beta}\prod_{i=1}^{n}\lambda e^{-\lambda X_i}$$
$$= \lambda^{\alpha+n-1}e^{-\lambda(\beta+\sum X_i)}.$$

Thus

$$f(\lambda|X) \sim \text{Gamma}\left(n+\alpha, \beta + \sum X_i\right).$$

We have the posterior mean

$$\frac{n+\alpha}{\beta+\sum X_i} = \frac{1+\frac{\alpha}{n}}{\overline{X}+\frac{\beta}{n}}$$

Recall that the MLE is $\overline{X}$.

*Example* 9.7 (Normal Mean). Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with known $\sigma^2$. Let $\mu \sim \mathcal{N}(\mu_0, \nu^2)$. We have then the posterior

$$\mu|X \sim \mathcal{N}\left(\frac{\frac{\mu_0}{\nu^2}+\frac{n\overline{X}}{\sigma^2}}{\frac{1}{\nu^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\nu^2}+\frac{n}{\sigma^2}}\right).$$

Note that the posterior is a weighted average of the prior mean and the sample mean. One may think of the weights ($n/\sigma^2$ and $1/\nu^2$) as the information contained in the data.

9.3. **Selecting Prior Distributions.** Most critical (and criticized) part of Bayesian statistics.

For discrete and finite sample space $\Omega$, we may use past experience to determine a prior. When $\Omega$ is an interval, we may discretize it and use the above method.

9.3.1. *Conjugate Priors.*

**Definition 9.8.** A family $\mathcal{F}$ of distributions is said to be **closed under sampling** from a model $f(X|\theta)$ if for each prior $f \in \mathcal{F}$, the posterior $f(\theta|X) \in \mathcal{F}$.

| $\mathcal{F}$ | $f(X\|\theta)$ |
|---|---|
| Beta | Bernoulli or Binomial |
| Gamma | Exponential |
| $\mathcal{N}$ | $\mathcal{N}$ (with fixed variance) |
| Gamma | Poisson |

*Example* 9.9.

9.3.2. *Uninformative Priors.*

- If $\Omega$ is discrete and finite, we may use the uniform prior.
- If $\Omega$ is an interval, we may use the uniform prior. Note that the uniform prior is not invariant under reparameterization.
- $\Omega = \mathbb{R}$. Flat (improper) prior.

**Definition 9.10.** If $X_i \overset{\text{iid}}{\sim} f(X|\theta)$ with fisher information $I(\theta)$. The **Jeffreys** prior is defined as

$$\pi_J(\theta) \propto \sqrt{I(\theta)}.$$

**Theorem 9.11.** *The Jeffreys prior is invariant under reparameterization.*

## 10. Statistical Decision Theory

We have $X_i \overset{\text{iid}}{\sim} f(X|\theta)$ with $\theta \in \Omega$. $\pi(\theta)$ is a prior on $\theta$. We are interested in estimating $\theta$. A "decision" is an estimator $\hat{\theta} = \hat{\theta}(X_i, \ldots, X_n)$.

**Definition 10.1.** A **loss function** is a function $L : \Omega \times \Omega \to [0, \infty)$.

*Example* 10.2. Common loss functions:

- Squared error loss: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$.
- Absolute error loss: $L(\theta, \hat{\theta}) = \left|\theta - \hat{\theta}\right|$.
- Zero-one loss: $L(\theta, \hat{\theta}) = \mathbb{1}_{\{\theta \neq \hat{\theta}\}}$.

**Definition 10.3.** The **frequentist risk** of $\hat{\theta}$ is

$$R(\theta, \hat{\theta}) \coloneqq \mathrm{E}_\theta\, L(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}) f(X|\theta)\, \mathrm{d}X.$$

*Remark* 10.4.

- The risk function does not depend on the data. It is a "pre-data" measure of performance.
- If $L$ is squared loss, then $R(\theta, \hat{\theta}) = \mathrm{MSE}(\hat{\theta})$.
- $R(\theta, \hat{\theta})$ is a function of $\theta$ — which $\theta$ to choose to make the comparison?

*Example* 10.5 (Two estimators with the same risk). Suppose

$$X_1, X_2 \overset{\text{iid}}{\sim} \begin{cases} \theta - 1, & \text{with probability } \frac{1}{2}, \\ \theta + 1, & \text{with probability } \frac{1}{2}. \end{cases}$$

Consider the estimators $\hat{\theta}_1 = (X_1 + X_2)/2$ and $\hat{\theta}_2 = X_1 + 1$. Using the zero-one loss function, we have

$$R(\theta, \hat{\theta}_1) = \mathbb{P}_\theta\,[X_1 = X_2] = \frac{1}{2},$$
$$R(\theta, \hat{\theta}_2) = \mathbb{P}_\theta\,[X_1 = \theta + 1] = \frac{1}{2}.$$

We cannot compare $\hat{\theta}_1$ and $\hat{\theta}_2$.

*Example* 10.6. Suppose $X \sim \mathcal{N}(\theta, 1)$ with $\theta \in \Omega = (0, 10)$. Consider estimators $\hat{\theta}_1 = X$ and $\hat{\theta}_2 = 5$. Using the squared error loss, we have

$$R(\theta, \hat{\theta}_1) = \mathrm{E}_\theta[(\theta - X)^2] = \operatorname{Var} X = 1,$$
$$R(\theta, \hat{\theta}_2) = \mathrm{E}_\theta[(\theta - 5)^2] = (5 - \theta)^2.$$

Neither estimator is uniformly better.

### 10.1. **Comparing Estimators.**

10.1.1. *The Frequentist Approach.*

**Definition 10.7.** The **maximum risk** of an estimator $\hat{\theta}$ is

$$\overline{R}(\hat{\theta}) = \max_{\theta} R(\theta, \hat{\theta}).$$

**Definition 10.8.** A **minimax estimator** is an estimator $\hat{\theta}$ such that

$$\overline{R}(\hat{\theta}) = \inf_{\tilde{\theta}} \overline{R}(\tilde{\theta}).$$

10.1.2. *The Bayes Approach.*

**Definition 10.9.** With $\pi$ as the prior, the **Bayes risk** is given by

$$r(\hat{\theta}) = \int R(\theta, \hat{\theta})\pi(\theta)\, \mathrm{d}\theta.$$

**Definition 10.10.** A **Bayes estimator** (associated with the loss function $L$ and prior $\pi$) is an estimator $\hat{\theta}$ such that

$$r(\hat{\theta}) = \min_{\tilde{\theta}} r(\tilde{\theta}).$$

*Example* 10.11. Let $X \sim \text{Binomial}(n, \theta)$ with prior $\theta \sim \text{Uniform}(0, 1)$ and squared error loss. Consider the estimators

$$\hat{\theta}_1 = \hat{\theta}_{MLE} = X/n, \quad \hat{\theta}_2^{\alpha,\beta} = \frac{\alpha + X}{\alpha + \beta + n} \quad (\alpha, \beta > 0).$$

(Note that $\hat{\theta}_2^{\alpha,\beta}$ is the posterior mean with a $\text{Beta}(\alpha, \beta)$ prior.) We have

$$R(\theta, \hat{\theta}_1) = \mathrm{E}[(\hat{\theta}_1 - \theta)^2] = \mathrm{Var}[\hat{\theta}_1] = \frac{\theta(1-\theta)}{n}$$

and

$$R(\theta, \hat{\theta}_2^{\alpha,\beta}) = \frac{n\theta(1-\theta) + [\alpha - (\alpha+\beta)\theta]^2}{(\alpha+\beta+n)^2}.$$

Note that

$$R(\theta, \hat{\theta}_2^{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}}) = \frac{n}{4(n + \sqrt{n})^2}$$

is constant as a function of $\theta$.

- We first compare the maximum risk:

  $$\overline{R}(\hat{\theta}_1) = \frac{1}{4n}, \quad \overline{R}(\hat{\theta}_2^{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}}) = \frac{1}{4n + 8\sqrt{n} + 4}.$$

  So $\hat{\theta}_2^{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}}$ is better estimator.

- Next, we compare the Bayes risk:
$$r(\hat{\theta}_1) = \int_0^1 \frac{\theta(1-\theta)}{n} 1 \, \mathrm{d}\theta = \frac{1}{6n}$$
and
$$r(\hat{\theta}_2^{\frac{\sqrt{n}}{2},\frac{\sqrt{n}}{2}}) = \frac{n}{4(n+\sqrt{n})^2}.$$
So $\hat{\theta}_1$ is the batter estimator for large $n$.

Note that $\hat{\theta}_2^{1,1}$ is the Bayes estimator with risk
$$r(\hat{\theta}_2^{1,1}) = \frac{1}{6(n+2)}.$$

**Definition 10.12.** The **posterior risk** is the risk calculated using the posterior distribution:
$$r(\hat{\theta}|X) = \int L(\theta, \hat{\theta}) f(\theta|X) \, \mathrm{d}\theta.$$

**Theorem 10.13.** *Let $\hat{\theta} = \hat{\theta}(X)$ be the value that minimizes the posterior risk $r(\hat{\theta}|X)$. Then $\hat{\theta}$ is the Bayes estimator.*

**Proof.** Note that
$$\begin{aligned} r(\hat{\theta}) &= \int R(\theta, \hat{\theta}) \pi(\theta) \, \mathrm{d}\theta \\ &= \int \int L(\theta, \hat{\theta}) f(X|\theta) \, \mathrm{d}x \; \pi(\theta) \, \mathrm{d}\theta \\ &= \int \int L(\theta, \hat{\theta}) f(X|\theta) \pi(\theta) \, \mathrm{d}\theta \, \mathrm{d}x \\ &= \int \int L(\theta, \hat{\theta}) \pi(\theta|X) f(X) \, \mathrm{d}\theta \, \mathrm{d}x \\ &= \int r(\hat{\theta}|X) f(X) \, \mathrm{d}x. \end{aligned}$$

□

**Theorem 10.14** (Bayes estimators)**.**
- *If L is squared error loss, then the Bayes estimator is the posterior mean.*
- *If L is absolute error loss, then the Bayes estimator is the posterior median.*
- *If L is zero-one loss, then the Bayes estimator is the posterior mode.*

We prove only the first statement.

**Proof.** Note: if $X$ is a random variable with mean $\mu$, then

$$\begin{aligned} \mathrm{E}[(X-c)^2] &= \mathrm{E}[(X-\mu)^2 + 2(X-\mu)(\mu-c) + (\mu-c)^2] \\ &= \mathrm{Var}[X] + (\mu-c)^2 \end{aligned}$$

is minimized at $c = \mu$.

The posterior risk is

$$r(\hat{\theta}|X) = \int (\theta - \hat{\theta})^2 f(\theta|X)\, \mathrm{d}\theta.$$

We think of $X = \theta$ and $c = \hat{\theta}$. Thus $r(\hat{\theta}|X)$ is minimized at $\hat{\theta} = \mathrm{E}[\theta|X]$. $\square$

## 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

### 10.2. **Properties of the normal distribution.**

**Proposition 10.15.** *Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$.*

- $\mathrm{E}[S^2] = \sigma^2$.
- *$\overline{X}$ and $S^2$ are independent.*
- *Let $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Let*

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad S^2 := \frac{1}{n-1} \sum (X_i - \overline{X})^2.$$

*Then*

$$\sqrt{n} \cdot \frac{\overline{X} - \mu}{S} \sim t_{n-1}, \quad (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

### 10.3. **Properties of the exponential distribution.**

**Proposition 10.16.**

*(i) The "memoryless" property:*

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

*(ii)* Exponential$(\lambda)$ = Gamma$(1, \lambda)$.

### 10.4. **Properties of the gamma distribution.**

**Proposition 10.17.**

*(i) If* $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_i, \beta)$ *for* $i = 1, 2, \ldots, N$, *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*

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

*(iii) If* $X \sim \text{Gamma}(\alpha, \beta)$, *then*

$$\beta X \sim \text{Gamma}\left(\alpha, 1\right).$$

**Proof.**

(i) Note that

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

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

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

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

□

10.5. **Properties of the gamma distribution.**

**Proposition 10.18.**

- Beta$(1, 1)$ = Uniform$(0, 1)$.
- *If* $X \sim \text{Beta}(\alpha, \beta)$, *then*

$$\mathrm{E}[X] = \frac{\alpha}{\alpha + \beta}, \quad \mathrm{Var}[X] = \frac{\alpha}{\alpha + \beta}\frac{\beta}{\alpha + \beta}\frac{1}{\alpha + \beta + 1}.$$