Chapter 1

Probability

1.1 Basic concepts and results

A **random experiment** is when a set of all possible outcomes is known, but it is impossible to predict the actual outcome of the experiment. A **sample space**, denoted as Ω, contains all possible outcomes of the experiment. An **event** is a subset of Ω. We say that $A \subset \Omega$ has occrred if and only if the outcome of the experiment is an element of A. Formally, the family of events forms a σ-algebra of subsets of Ω that we denote by $\mathcal{A}$.

Note:

- $\Omega \in \mathcal{A}$
- $A \in \mathcal{A} \Rightarrow \bar{A} \in \bar{\mathcal{A}}$,where $\bar{A}$ indicates the compliment of A
- $A_1, A_2, \cdots \in \mathcal{A}$
- $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

1.1.1 Probability measures

Definition 1.1.1: Kolmogorov's axioms

- $P(A) \geq 0$
- $P(\Omega) = 1$
- If $A_i \bigcap A_j = \emptyset, i \neq j$, then $P(\cup_i A_i) = \sum_i P(A_i)$

Probability measure $P : \mathcal{A} \to \mathbb{R}$ satisfying Kolmogorov's axioms has the following properties:

- $P(\varnothing) = 0$
- $A \subset B \Rightarrow P(A) \leq P(B)$
- $0 \leq P(A) \leq 1$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(\bar{A}) = 1 - P(A)$
- $P(A - B) = P(A \cap \bar{B}) = P(A) - P(A \cap B)$

Definition 1.1.2: Conditional probability

If $P(B) > 0$,

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

We are re-evaluating the probability of A given the B space.

Let $\{A_1, A_2, \cdots\}$ denote a partition of $\Omega : \cup_i A_i = \Omega$; $A_i \cap A_j = \emptyset, i \neq j$. Meaning union makes up Ω and are mutually exclusive. Then if $P(A_i) > 0$ for all i

Theorem 1.1.1 Total probability theorem

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

$B = B \cap \Omega = B \cap [\cup_i A_i] = \cup_i (B \cap A_i)$ and $P(\cup_i B \cap A_i) = \sum_i P(B \cap A_i)$

Theorem 1.1.2 Bayes' theorem

If $P(B) > 0$

$$P(A_i|B) = \frac{P(B|A_j)P(A_j)}{\sum_i P(B|A_i)P(A_i)}$$

$$P(\underbrace{A_j}_{\text{explanation}} | \underbrace{B}_{\text{evidence}}) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\underbrace{P(B)}_{\text{substitute with total probability theorem}}}$$

1.1.2 Random variables

Definition 1.1.3: Random variable

Function defined in Ω and taking values in $\mathbb{R}$

$X : \Omega \to \mathbb{R}$

$\omega \mapsto X(\omega) = x$

A random variable induces a probability measure in $\mathbb{R}$ that we denote by P_X: if $B \subset \mathbb{R}$, $P_X(B) = P(A)$, where $A = X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$. Formally, there must be a σ-algebra of subsets of $\mathbb{R}$,$\mathcal{B}$, and we have to verify that for every set $B \in \mathcal{B}$ we have $X^{-1}(B) \in \mathcal{A}$. Typically, $\mathcal{B}$ is the so called Borel σ-algebra and it suffices to make sure that X satisfies $X^{-1}((-\infty, x]) \in \mathcal{A}$,$\forall x \in \mathbb{R}$.

Basically what it means is that we don't know if $X^{-1}(B) \in \mathcal{A}$ and for which B can I compute $P_X(B)$. If $X^{-1}(B) \in \mathcal{A}$ for B is in the Borel σ-algebra, then X is measurable.

Definition 1.1.4: Distribution function of a random variable

X: for all $x \in \mathbb{R}$

$F_X(x) = P_X((-\infty, x]) = P(X \leq x)$

It is suffice to know $F_X(\cdot)$ to be able to compute $P_X(B)$ for all $B \in \mathcal{B}$.

- For all $a > b$, $P(a < X \leq b) = F_X(b) - F_X(a)$
- $F_X(-\infty) = 0$; $F_X(\infty) = 1$

- F_X is right-continuous and non-decreasing
- The set of points at which F_X is discontinuous is either finite or countable (at most countable)

Definition 1.1.5: Discrete random variable

X is a discrete random variable if D_X is such that $P_X(D_X) = 1$

The probability mass function of X is defined as $f_X(x) = F_X(x) - \lim_{y \to x} F_X(y) = \begin{cases} P(X = x) & \text{if } x \in D_X \\ 0 & \text{otherwise} \end{cases}$

Any f satisfying the following is a probability mass function

- $f(x) \geq 0$ for all x
- $f(x) > 0$ iff $x \in D$, where $D \subset \mathbb{R}$ is finite or countable
- $\sum_{x \in D} f(x) = 1$

For any event $B \subset \mathbb{R}$, $P(X \in B) = \sum_{x \in B \cap D_X} f_X(x)$.

Note:

$F_X(x) = \sum_{y \leq x} f_X(y)$

$F_X(x) = P(X \leq x)$ cumulative distribution function
$\downarrow$
$f_X(x) = P(X = x)$ probability mass function
where $0 \leq f_X(x) \leq 1$

Discrete distribution include Bernoulli, binomial, Poisson, geometric, negative binomial, multinomial, hypergeometric, etc.

Definition 1.1.6: Continuous random variable

X is continuous if $P_X(D_X) = 0, D_X = \varnothing$ and if additionally there is f_X such that for all $x \in \mathbb{R}$

- $f_X(x) \geq 0 \rightarrow$ probability density function
- $F_X(x) = \int_{-\infty}^{+\infty} f(x)\, dx = 1$

At the points where F_X is differentiable, we have $F'_X(x) = f_X(x)$.

Any f satisfying the following conditions is a probability density function

- $f(x) \geq 0$ for all x
- $\int_{-\infty}^{+\infty} f(x)\, dx = 1$

Continuous distributions include uniform, exponential, gamma, chi-squared, normal. t-student, F-Snedcor, beta, Pareto, Weibull, log-normal, etc.

1.1.3 Functions of a random variable

Let X be a r.v. and $Y = h(X)$ where $h : \mathbb{R} \rightarrow \mathbb{R}$
In general, if $X = g(Y)$ with g invertible and differentiable, and X continuous, we have

$$f_Y(y) = |g'(y)|\, f_x(g(y))$$

Proof. $\dfrac{\partial F_X(x)}{\partial x} = f_X(x)$
Using chain rule: $(f \circ g)'(x) = [f(g(x))]' = f'(g(x))g'(x)$ ■

Definition 1.1.7: Expected value

Let $Y = h(X)$, a linear function.

The expected value of Y is defined by $E[Y] = \begin{cases} \sum_x h(x)\, f_X(x) & \text{if } X \text{ discrete} \\ \int_{-\infty}^{+\infty} h(x)\, f_X(x)\, dx & \text{if } X \text{ continuous} \end{cases}$

Formally, we must addtionally verify that the integral or series are absolutely convergent. $E[Y]$ may not exist.

There are two ways to compute $E[Y]$ with $Y = h(X)$, either use the definition above, or first obtain the distribution of Y and compute $E[Y] = \begin{cases} \sum_y y\, f_Y(y) & \text{if } Y \text{ discrete} \\ \int_{-\infty}^{+\infty} y\, f_Y(y)\, dy & \text{if } Y \text{ continuous} \end{cases}$. The two methods are equivalent.

Definition 1.1.8: Raw moment of oder k

$\mu_k' = E[X^k]$

Definition 1.1.9: Central moment of order k

$\mu_k = E[(X - \mu)^k]$, $\mu = E[X]$

Definition 1.1.10: Moment generating function

$M_X(s) = E[e^{sX}]$ whenever the expectation exists for s in a neighborhood of the origin.

- If $M_X(s)$ exists, then X has moments of all orders and $M^{(k)}(0) = E[X^k]$
- The moment generating function, when it exists, identifies the probability distribution

Some useful **properties**:

- $E[h_1(X) + h_2(X)] = E[h_1(X)] + E[h_2(X)]$
- If $c \in \mathbb{R}$, then $E[cX] = cE[X]$; $E[c] = c$
- If $c \in \mathbb{R}$, then $Var(cX + b) = c^2 Var(X)$
- $Var(X) = E[X^2] - (E[X])^2$
- $Var(X) \geq 0$; $Var(X) = 0 \Leftrightarrow P(X = c) = 1$ for some $c \in \mathbb{R}$

1.1.4 Bivariate random variables

$$(X, Y) : \Omega \to \mathbb{R}^2$$
$$\omega \mapsto (X(\omega), Y(\omega)) = (x, y)$$

If (X, Y) discrete, we define the joint probability mass function as $f(x, y) = P(X = x, Y = y)$. If $(X; Y)$ continuous, then there exists the joint probability density function, $f(x, y)$ such that for all $(x, y) \in \mathbb{R}^2$,

- $f(x, y) \geq 0$
- $F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, dv\, du$

Example 1.1.1

X = weight, Y = height $\Rightarrow$ Z = BMI

Definition 1.1.11: Marginal distributions

$$f_X(x) = \begin{cases} \sum_y f(x,y) & \text{if } (X,Y) \text{ discrete} \\ \int_{-\infty}^{+\infty} f(x,y)\,dy & \text{if } (X,Y) \text{ continuous} \end{cases}$$

Definition 1.1.12: Expectation of $Z = h(X,Y)$

$$E[Z] = \begin{cases} \sum_x \sum_y h(x,y)\, f(x,y) & \text{if } (X,Y) \text{ discrete} \\ \int_{-\infty}^{+\infty} h(x,y)\, f(x,y)\,dy\,dx & \text{if } (X,Y) \text{ continuous} \end{cases}$$

Definition 1.1.13: Conditional didstributions

$f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}$, y fixed: $f_Y(y) > 0$

function of x for every y where $f_Y(y) > 0$

Definition 1.1.14: Raw moment of order (r,s)

$\mu'_{(r,s)} = E[X^r\, Y^s]$

Definition 1.1.15: Central moment of order (r,s)

$\mu_{(r,s)} = E[(X-\mu_X)^r\,(Y-\mu_Y)^s]$

Definition 1.1.16: Covariance

$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \mu_{(1,1)}$

If x and y are positively associated $\longrightarrow \mathrm{Cov}(x,y) > 0 \longrightarrow$ If x is larger than its mean, then typically y is larger than its mean.

Some useful **properties**:

- $\mathrm{Cov}(X,Y) = E[X,Y] - E[X]E[Y]$
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- $\mathrm{Cov}(cX,Y) = c\mathrm{Cov}(X,Y), c \in \mathbb{R}$
- $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$
- $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\,\mathrm{Cov}(X,Y)$

Example 1.1.2 (Portfolio management)
$\mathrm{Cov}(x,y) < 0$
$\mathrm{Var}(x,y) < \mathrm{Var}(x) + \mathrm{Var}(y)$

Theorem 1.1.3 Law of iterated expectation
If $Z = h(X,Y)$ then $E[Z] = E_X[E[Z|X]]$

Theorem 1.1.4 Law of total variance
$\mathrm{Var}(Y) = \mathrm{Var}_X(E[Y|X]) + E_X[\mathrm{Var}(Y|X)]$

Other useful tricks:

- $E[h(X)\,Y \mid X = x] = h(x)\,E[Y \mid X = x]$
- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, E[Y|X])$

  Proof.

$$\begin{aligned}\mathrm{Cov}(X, E[Y|X]) &= E[X\,E[Y|X]] - E[X]\,E[E[Y|X]] \\ &= E[E[XY|X]] - E[X]\,E[Y] \\ &= E[XY] - E[X]\,E[Y] \\ &= \mathrm{Cov}(X, Y)\end{aligned}$$

  ■

1.1.5 Independence

Definition 1.1.17: Stochastic independence

X and Y are stochastically independent if and only if $\forall (x, y) \in \mathbb{R}^2,\ f(x, y) = f_X(x)\,f_Y(y)$

If X and Y are independent, then

- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

  Proof. $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2 \times \underbrace{\mathrm{Cov}(X, Y)}_{\to 0}$ ■

- $M_{X+Y}(s) = M_X(s)\,M_Y(s)$

  Proof. $M_{X+Y}(s) = E[e^{s(X+Y)}] = E[\underbrace{e^{sx}}_{u}\,\underbrace{e^{sy}}_{v}]$

  x and y independent stochastically $\Rightarrow$ u and v independent

  $M_{X+Y}(s) = E[e^{sx}]\,E[e^{sy}] = M_X(s)\,M_Y(s)$ ■

- $\mathrm{Cov}(X, Y) = 0$

  Proof. $\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = \underbrace{E[XY]}_{X,Y\text{uncorrelated}} - E[X]E[Y] = E[X]E[Y] - E[x]E[Y] = 0$ ■

- $E[X^r Y^s] = E[X^r]\,E[Y^s]$
- $E[Y \mid X = x] = E[Y]$; $E[X \mid Y = y] = E[X]$
- $f_{X|Y=y}(x) = f_X(x)$; $f_{Y|X=x}(y) = f_Y(y)$

  Proof. $f_{X|Y=y}(x) = \dfrac{f(x, y)}{f_Y(y)} = \dfrac{f_X(x) f_Y(y)}{f_Y(y)} = f_X(x)$ ■

Definition 1.1.18: Mean independence

Y is mean independent of X iff $E[Y \mid X = x]$ does not depend on x for all x.

Proof. $E[Y|X = x] = c$

$E[Y|X] = c \Rightarrow E[E[Y|X]] = c \Rightarrow E[Y] = c \rightarrow$ conditional is equal to marginal ■

Definition 1.1.19: Uncorrelatedness

X and Y are uncorrelated iff $\text{Cov}(X, Y) = 0$

Useful **results**:

- If X and Y are stochastically independent, then Y is mean-independent of X, and X is mean independent of Y.
- If Y is mean-independet of X, then X and Y are uncorrelated. The converse is not true.

  Proof. Y mean independence of $X \Rightarrow \text{Cov}(X, Y) = \text{Cov}(X, E[Y|X]) = \text{Cov}(X, c) = 0 \Rightarrow$ uncorrelated ■

- If Y is uncorrelated with X, then $E[XY] = E[X]E[Y]$
- If Y is mean-independent of X, then $E[X^k Y] = E[X^k]E[Y]$ for all k
- If Y and X are stochastically independent, then $E[X^k Y^r] = E[X^k]E[Y^r]$ for all k, r

Note:

stochastic independence $\Rightarrow$ mean independence $\Rightarrow$ uncorrelatedness

1.2 Convergence of sequences of random variables

If $\{X_n\}_{n=1}^{\infty}$ is a sequence of random variables and X is a random variable,

$$X_n : \underbrace{\Omega}_{\text{exists probability, } \sigma\text{-algebra}} \to \mathbb{R}$$

$$X_n \longrightarrow X \qquad \text{as } n \to +\infty$$

n can be population size, or can be the number of iterations for Monte Carlo simulation.

1.2.1 Notions of convergence of sequences

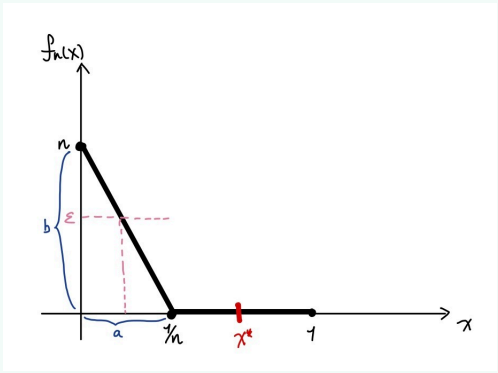
Notions of **convergence of sequences**: let $f_n, f : [0, 1] \to \mathbb{R}$

- Point wise convergence: $f_n(x) \to f(x)$ for all $x \in [0, 1]$
- Uniform convergence: $\sup_{x \in [0,1]} |f_n(x) - f(x)| \to 0$
- Convergence in L^P: $\int_0^1 |f_n(x) - f(x)|^P \, dx \to 0$
- Convergence in measure: $\mu(A_{n,\epsilon}) \to 0$ for all $\epsilon > 0$ where $A_{n,\epsilon} = \{x \in [0, 1] : |f_n(x) - f(x)| > \epsilon\}$

Example 1.2.1

$f_n : [0, 1] \to \mathbb{R}$

$$f_n(x) = \begin{cases} 0 & 1/n \le x \le 1 \\ n - n^2 x & 0 \le x < 1/n \end{cases}$$



As $n \to \infty$, a becomes smaller, b becomes bigger.

- Point wise convergence
  $\forall x \in [0,1]$
  $\forall x^* > 0, f_n(x^*) = 0 \quad \text{for } n > N \quad \text{except } f_n(0) = 0 \to \infty$

  $$\Rightarrow f_n(x) \to \begin{cases} 0 & \text{if } x \in [0,1] \\ \infty & \text{if } x = 0 \end{cases} \Rightarrow f_n \text{ is not converging pointwise to the null function.}$$

- Uniform convergence
  $\max |f_n(x)| = n \to +\infty \quad x \in [0,1] \Rightarrow f_n$ does not converge uniformly to the null function.

- Convergence in $L^1 \to P = 1$
  $$\int_0^1 |f_n(x)|\, dx = \frac{1}{2} = \underbrace{\frac{1}{n} \times n \times \frac{1}{2}}_{\text{area under the triangle}} \Rightarrow f_n \text{ does not converge in } L^1 \text{ to the null function.}$$

- Convergence in measure
  $A_{n,\epsilon} \subset [0, \frac{1}{n}]$
  $\mu(A_{n,\epsilon}) \leq \mu([0, \frac{1}{n}]) = \frac{1}{n} \to$ as $n \to \infty, \mu \to \infty \Rightarrow f_n$ converges to the null function in measure.

1.2.2 Convergence of random variables

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and X is a random variable, all defined in the same probability space $(\Omega, \mathcal{A}, P)$.

Definition 1.2.1: Almost surely convergence

X_n converges to X almost surely, or with probability 1, $X_n \xrightarrow{\text{a.s.}} X$, iff

$$P[\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}] = 1$$

Similar to pointwise convergence, no need for expectation.

Note:

$$\underbrace{P(\underbrace{X_n(\omega)}_{\text{set}} \to \underbrace{x(\omega)}_{\text{point}})}_{\text{set of which it happens has a probability of 1}} = 1$$

Definition 1.2.2: Convergence in the rth mean

X_n converges to X in the rth mean, $r \geq 1$, $X_n \xrightarrow{r} X$, iff

$$E[|X_n - X|^r] \to 0$$

Each point will be weighted with the same probability. Expectation is involved in this case.

Note:

When $r = 2$, it is the mean square convergence, often used for quality checking.

Definition 1.2.3: Convergence in probability

X_n converges in probability to X, $X_n \xrightarrow{P} X$, iff for all $\epsilon > 0$

$$P(|X_n - X| > \epsilon) \to 0$$

It is similar to measure in convergence. Often used to check for quality of estimator. Note that this is no longer a Lebesque measure, it is now a probability measure. $P\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\} \to 0$ as $n \to 0$.

Definition 1.2.4: Convergence in distribution

X_n converges in distribution to X, $X_n \xrightarrow{d} X$, iff

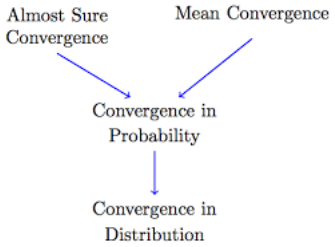
$$F_n(x) \to F_n$$

for all x continuity point of F, where $F(x) = P(X \leq x)$ and $F_n(x) = P(X_n \leq x)$

Has nothing to do with the random variable. Often used for hypothesis testing. It does not need the requirement that all points are defined in the same probability space $(\Omega, \mathcal{A}, P)$ as there is no ω in the density function.

Some useful **remarks**:

- Convergence in distribution is really about the convergence of the sequence of probability functions and not the random variables themselves.
- When defining convergence in the rth mean, it is assumed that the corresponding expected values exist: $E[|X_n|^r] < \infty$ and $E[|X|^r] < \infty$
- When $X_n \xrightarrow{1} X$, we say that X_n converges to X in mean; when $X_n \xrightarrow{2} X$, we say that X_n converges to X in quadratic mean.



Proof. **Convergence in mean implies convergence in probability**

$E[|X_n - X|] \to 0 \Rightarrow P(|X_n - X| > \epsilon) \to 0, \forall \epsilon > 0$

Using **Markov inequality**: $P(|y| > a) \leq \dfrac{|E[X_n - X|]}{\epsilon}$

$$0 \le \lim_{n\to\infty} P(|X_n - X| > \epsilon) \le \lim_{n\to\infty} \frac{\overbrace{E[|X_n - X|]}^{\to 0}}{\epsilon} = 0 \blacksquare$$

Proof. **Proof of convergence in probability implies convergence in distribution**

$X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} \Leftrightarrow P(|X_n - X| > \epsilon) \to 0 \Rightarrow P(X_n \le x) \to P(X \le x)\,, \forall x$

let $\epsilon > 0$,

$F_n(x) = P(X_n \le x)$

$F(x) = P(X \le x)$

Using the **total probability theorem**: $P(A) = P(A \cap B) + P(A \cap \bar{B}) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$

$$F_n(x) = P(\underbrace{X_n \le x}_{A}) = P(\underbrace{X_n \le x}_{A}, \underbrace{X \le x + \epsilon}_{B}) + P(\underbrace{X_n \le x}_{A}, \underbrace{X > x + \epsilon}_{\bar{B}}) \le F(x + \epsilon) * P(|X_n - x| > \epsilon)$$

$$F(x - \epsilon) - P(|X_n - X| > \epsilon) \le F_n(x) \le F(x + \epsilon) + \underbrace{P(|X_n - X| < \epsilon)}_{\to 0}$$

as $n \to \infty$, $\underbrace{F(x - \epsilon)}_{\xrightarrow{\epsilon\to 0} F(X)} \le \lim_{n\to\infty} F_n(x) \le \underbrace{F(x + \epsilon)}_{\xrightarrow{\epsilon\to 0} F(X)}$ $\blacksquare$

Some **converses**:

- If $X_n \xrightarrow{P} X$, then there exists $\{n_k\}_{k=1}^{+\infty}$ such that $X_{n_k} \xrightarrow{a.s.} X$ when $k \to +\infty$
- If $|X_n|^r$ is uniformly integratable, then $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{r} X$

Theorem 1.2.1 Skorokhod representation theorem

If $X_n \xrightarrow{d} X$ then there exists a probability space $(\Omega', \mathcal{A}', P')$ and r.v. $\{Y_n\}$ and Y, defined in Ω' such that

- $P'(Y_n \le y) = P(X_n \le y)$ and $P'(Y \le y) = P(X \le y)$ for all $y \in \mathbb{R}$. This means that X_n and Y_n are marginally equal in distribution, the same for X and Y.
- $Y_n \xrightarrow{a.s.} Y$

Other useful **results**:

- $X_n \xrightarrow{P} c \Leftrightarrow X_n \xrightarrow{d} c$, where $c \in \mathbb{R}$

Proof. $X_n \xrightarrow{d} x \Rightarrow X_n \xrightarrow{P} c \Leftrightarrow P(X_n \le x) \to \begin{cases} 0 & x < c \\ 1 & x < c \end{cases}$, not continuous at c

$P(|X_n - c| > \epsilon) \to 0,\ \forall \epsilon > 0$

$$\begin{aligned} P(|X_n - c| > \epsilon) &= P(X_n - c > \epsilon) + P(X_n - c < -\epsilon) \\ &= P(X_n > \epsilon + c) + P(X_n < c - \epsilon) \\ &= 1 - P(X_n \le \epsilon + c) + P(X_n < c - \epsilon) \\ &\le 1 - P(X_n \le \underbrace{\epsilon + c}_{>c}) + P(X_n \le \underbrace{c - \epsilon}_{<c}) \\ &\to 1 - 1 + 0 = 0 \end{aligned}$$

$\blacksquare$

- Since $E[(X_n - \theta)^2] = \text{Var}(X_n) + (E[X_n] - \theta)^2$ if $\text{Var}(X_n) \to 0$ and $E[X_n] \to \theta$. We have convergence in mean square to θ, and hence convergence in probability to θ.

Theorem 1.2.2 Continous mapping theorem

Let $h : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then

- $X_n \xrightarrow{a.s.} X \Rightarrow h(X_n) \xrightarrow{a.s.} h(X)$
- $X_n \xrightarrow{d} X \Rightarrow h(X_n) \xrightarrow{d} h(X)$
- $X_n \xrightarrow{P} X \Rightarrow h(X_n) \xrightarrow{P} h(X)$

Theorem 1.2.3 Slutsky theorem

Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables, X a random variable and c a real number. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then

- $X_n + Y_n \xrightarrow{d} X + c$
- $Y_n X_n \xrightarrow{d} cX$
- $X_n / Y_n \xrightarrow{d} X/c$ as long as $c \neq 0$

Wrong Concept 1.1: $X_n + Z_n \neq 2X$

Suppose that $X_n \xrightarrow{d} X$ where $X \sim N(0,1)$. Then with $Z_n = -X_n$ we have $Z_n \xrightarrow{d} X$. However, $X_n + Z_n = 0$, hence $X_n + Z_n$ does not converge in distribution to $2X$ as one might expect.

cdf of Z_n converges to cdf of X_n

$$\begin{aligned} Z_n \xrightarrow{d} X &\Leftrightarrow P(Z_n \leq z_n) \to \Phi(z_n),\ \forall z \in \mathbb{R} \\ &\Leftrightarrow P(-X_n \leq z) = P(X_n \geq -z) = 1 - P(X_n \leq -z) \\ &\to 1 - \Phi(-z) \\ \therefore Z_n &\xrightarrow{d} X \end{aligned}$$

This is why the Slutsky theorem is important, it showcases safe procedures.

Example 1.2.2 ($X_n \sim t(n) \Rightarrow X_n \xrightarrow{d} N(0,1)$ using Slutsky)

$X_n \sim t(n)$, $X_n = \dfrac{u_n}{\sqrt{\frac{v_n}{n}}}$

Assumptions: $\begin{cases} u_n \text{independent of } v_n \\ u_n \sim N(0,1) \\ v_n \sim \chi^2(n) \end{cases}$

What would be nice is to show that $\sqrt{\frac{v_n}{n}}$ converges to 1 then we can apply the Slutsky theorem.

Using the **mean square convergence**, we have

$$\text{Var}(\frac{v_n}{n}) = \frac{\text{Var}(v_n)}{n} = \frac{2n}{n^2} = \frac{2}{n} \to 0$$
$$E[\frac{v_n}{n}] = \frac{E[v_n]}{n} = \frac{n}{n} = 1 \to 1$$

We now have mean square convergence to 1.

Using the **Continuous mapping theorem**, we have

$$\frac{v_n}{n} \xrightarrow{P} 1 \Rightarrow \sqrt{\frac{v_n}{n}} \xrightarrow{P} 1$$

$\Rightarrow \frac{v_n}{n} \xrightarrow{2} 1$ and $\frac{v_n}{n} \xrightarrow{P} 1$

Now using the **Slutsky theorem**, we have

$$X_n = \frac{u_n}{\sqrt{\frac{v_n}{n}}} \xrightarrow{d} u_n \sim N(0,1)$$

1.3 Important asymptotic results

Theorem 1.3.1 Weak law of large numbers

Let $\{X_n\}_{n=1}^{+\infty}$ be a sequence of independent and identically distributed random variables, with $E[X_n] = \mu$ and $\text{Var}(X_n) = \sigma^2 < \infty$. Let also $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then we have that

$$\overline{X}_n \xrightarrow{P} \mu$$

Proof. Goal: $\overline{X}_n \xrightarrow{P} \mu \Rightarrow P(|\overline{X}_n - \mu| > \epsilon) \to 0$

Checking the validity of Chebychov's inequality,

$$E[\overline{X}_n] = E[\frac{1}{n}\sum_{i=1}^n X_i] = \frac{1}{n}\sum_{i=1}^n \underbrace{E[X_i]}_{\to \mu} = \frac{1}{n}\, n\, \mu = \mu$$

We can now apply the **Chebychov's inequality**: $P(\underbrace{|X-\mu|}_{\text{distance of distribution from its mean}} > \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2}$

$$P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\overbrace{\text{Var}(\overline{X}_n)}^{1}}{\epsilon^2} = \frac{\sigma^2}{n\,\epsilon^2} \to 0$$

$$1\text{: } \text{Var}(\overline{X}_n) = \frac{1}{n}\text{Var}(\sum_{i=1}^n X_i) \underbrace{=}_{\text{Var}(\sum)=\sum \text{Var}+\underbrace{2\text{Cov}}_{\text{iid}\to 0}} \frac{1}{n^2}\, n\, \sigma^2 = \frac{\sigma^2}{n} \blacksquare$$

Intuitively, the WLLN tell us that $\overline{X}_n$ becomes more and more concentrated around μ as n increases.

Theorem 1.3.2 Strong law of large numbers

Under the same conditions as above, we have

$$\overline{X}_n \xrightarrow{a.s.} \mu$$

Actually, it is only necessary to assume that $E[|X_i|] < +\infty$ for both laws to hold.

Theorem 1.3.3 Central limit theorem

Let $\{X_n\}_{n=1}^{+\infty}$ be a sequence of iid random variables possessing finite variance. Let $\mu = E[X_n]$ and $\sigma^2 = \text{Var}(X_n)$. Let also $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and

$$Z_n = \sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} = \frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\overline{X}_n - E[\overline{X}]}{\sqrt{\text{Var}(\overline{X})}} \xrightarrow{d} N(0,1)$$

Then we have

$$Z_n \xrightarrow{d} Z$$

where $Z \sim N(0,1)$

Proof. **Proof with assumption of existance of mgf**

Assume

1. $M_n(S) = E[e^{sX_n}]$ exists
2. $M_n(s) \longrightarrow M(s)$ for $s \in (-s_0, s_0)$

then $M(s) = E[e^{sX}] \Rightarrow X_n \xrightarrow{d} X$

Idea: X_n are iid r.v.
$E[e^{sX_n}] = M_{X_n}(s)$ exists for $s \in (-s_0, s_0) \Rightarrow Z_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$
Need to show $M_{Z_n}(s) \longrightarrow M_{N(0,1)}(s) = e^{s^2/2} \longrightarrow$ mgf of Z_n goes to $e^{s^2/2}$, the mgf of the normal distribution.

$$Z_n = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \underbrace{=}_{\text{Annex 1}} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} y_i \tag{1.1}$$

Annex 1:

$$Y_i = \frac{X_i - \mu}{\sigma}, \text{standardized version of the } X_i\text{'s}$$

$$\sum Y_i = \frac{\sum(X_i - \mu)}{\sigma} = \frac{\sum X_i - n\mu}{\sigma} = \frac{n\overline{X}_n - n\mu}{\sigma} = n\frac{\overline{X}_n - mu}{\sigma}$$

$$\frac{1}{\sqrt{n}}\sum Y_i = \frac{1}{\sqrt{n}} n \frac{\overline{X}_n - mu}{\sigma} = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$$

Using the moment generating function

$$\begin{aligned}
M_{Z_n}(s) &= E[e^{sZ_n}] = E[e^{s\frac{1}{\sqrt{n}}\sum Y_i}] \\
&= M_{\sum Y_i}(\frac{s}{\sqrt{n}}) \\
&= M_{Y_1}(\frac{s}{\sqrt{n}}) \times M_{Y_2}(\frac{s}{\sqrt{n}}) \times \cdots \times M_{Y_n}(\frac{s}{\sqrt{n}}) \longrightarrow \text{mgf of the sum of the variable is the product} \\
&= [M_Y(\frac{s}{\sqrt{n}})]^n \\
&= \sum_{k=0}^{2} M_Y^{(k)}(0)\frac{s^k}{k!} + \underbrace{r(s)}_{\frac{r(s)}{s^2}\to 0 \text{ as } s\to 0} \longrightarrow \text{Taylor's expansion of 2nd order, Annex 2} \\
&= 1 + \frac{s^2}{2!} + r(s)
\end{aligned} \tag{1.2}$$

Annex 2:

$$
\begin{aligned}
M_Y^{(k)}(0) &= E[Y^k] \\
M_Y^{(0)}(0) &= E[Y^0] = 1 \\
M_Y^{(1)}(0) &= E[Y^1] = 0 \\
M_Y^{(2)}(0) &= E[Y^2] = \frac{E[(x_i-\mu)^2]}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1
\end{aligned}
$$

Back to the moment generating function

$$
\begin{aligned}
M_{Z_n}(s) &= [M_Y(\frac{s}{\sqrt{n}})]^n \\
&= [1 + \frac{s^2/2}{n} + r(\frac{s}{\sqrt{n}})]^n \\
&= [1 + \frac{\frac{s^2}{2} + n\overbrace{r(s/\sqrt{n})}^{\to 0}}{n}]^n \underbrace{\to}_{\text{Annex 3}} e^{s^2/2}
\end{aligned}
\tag{1.3}
$$

Annex 3:

$$
[1 + \frac{u_n}{v_n}]^{v_n} \to e^c, u_n \to c, v_n \to \infty
$$

■

The CLT is often used to compute probabilities of the type $P(\overline{X}_n \leq x)$ approximating them by $\Phi(\sqrt{n}\frac{(x-\mu)}{\sigma})$ for sufficiently large n.

$$
\begin{aligned}
P(\overline{X}_n \leq x) &= P(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq \sqrt{n}\frac{x-\mu}{\sigma}) \\
&\approx \Phi(\sqrt{n}\frac{x-\mu}{\sigma}) \\
P(Z_n \leq z) &\to \Phi(z)
\end{aligned}
$$

Intuitively, the CLT tells us that the distribution of $\bar{X}_n$ is well approximated by a normal distribution for sufficiently large n as long as the variance is finite. Additionally, if the distribution of X_n is close to symmetric, then the rate of convergence is faster. Rate of convergence is related to the coefficient of symmetry, $\frac{E[(X-\mu)^3]}{(\text{Var}(X))^{3/2}} = \gamma_1$. If the distribution is symmetric, $\gamma_1 = 0$.

Theorem 1.3.4 Lévy's continuity theorem

Suppose that $\{X_n\}_{n=1}^{+\infty}$ is a sequence of random variables and let $M_n(s)$ denote the mgf of $X_n, n = 1, 2, \cdots$. Additionally assume that

$$
\lim_{n \to +\infty} M_n(s) = M(s)
$$

for s in a neighborhood of the origin, and that $M(\cdot)$ is the mgf of a random variable X.

In these circumstances,

$$
X_n \xrightarrow{d} X
$$

Example 1.3.1 (Application : Bernoulli)

$\{X_n\}_{n=1}^{+\infty}$ iid $B(1,\theta)$ where $\theta \in (0,1)$. By the **central limit theorem**,

$$
\sqrt{n}\frac{\overline{X}_n - \theta}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} N(0,1)
$$

On the other hand, the **WLLN** ensures that $\overline{X}_n \xrightarrow{d} \theta$.
By the **continuous mapping theorem**,

$$\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow{P} 1$$

and **Slutsky's theorem** allows us to conclude that

$$\sqrt{n}\,\frac{\overline{X}_n-\theta}{\sqrt{\theta(1-\theta)}}\,\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} = \sqrt{n}\,\frac{\overline{X}_n-\theta}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow{d} N(1,0)$$

which in practice means that, for large n

$$P\left(\frac{\overline{X}_n-\theta}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \le x\right) \approx \Phi(x)$$

Proof. $X_i \sim B(1,\theta), \quad E[x_i]=\theta \quad \mathrm{Var}(x_i)=\theta(1-\theta)$

By the **CLT**, $\sqrt{n}\,\underbrace{\frac{\overline{X}_n-\theta}{\sqrt{\theta(1-\theta)}}}_{\text{issue in the denominator}} \xrightarrow{d} N(0,1)$

$$\sqrt{n}\,\frac{\overline{X}_n-\theta}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow{d} N(0,1) = \underbrace{\sqrt{n}\,\frac{\overline{X}_n-\theta}{\sqrt{\theta(1-\theta)}}}_{\xrightarrow{d} N(0,1) \text{ by CLT}}\ \underbrace{\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}}_{\xrightarrow{P} 1, \text{ Annex 1}}$$

Annex 1:

- $\overline{X}_n \xrightarrow{P} E[X_i]=\theta$ by **WLLN**
- $\sqrt{\overline{X}_n(1-\overline{X}_n)} \to \sqrt{\theta(1-\theta)}$ by **continuous mapping theorem**

■

Example 1.3.2 (Application : $P(X \in A)$ using Simple Monte Carlo)

Notice that $P(X \in A) = E[Y]$ where $Y = I_A(X) = \begin{cases} 1 & , x \in A \\ 0 & , x \notin A \end{cases}$

Let $X_1, X_2, \cdots, X_M$ be iid r.v. with the same distribution as X, and $Y_i = I_A(X_i), i = 1, \ldots, M$. Then by **SLLN**,

$$\bar{Y}_M = \frac{1}{M}\sum_{i=1}^{M} Y_i \xrightarrow{a.s.} E[Y] = P(X \in A)$$

where M is the simulation length.

For sufficiently large M,

$$P(X \in A) \approx \frac{1}{M}\sum_{i=1}^{M} y_i = \frac{1}{M}\underbrace{\#\{i=1,\ldots,M : x_i \in A\}}_{\text{observed proportions of } x_i \in A}$$

Simple Monte Carlo allows us to replace the analytical knowledge of a probability distribution by a sufficiently large sample of iid draws from the distribution since almost all aspects of that probability distribution can be arbitrarily approximated using that sample.

Example 1.3.3 (Application : $f(a)$ for some $a \in \mathbb{R}$ using Simple Monte Carlo)

For a continous distribution with density f,

$$\begin{aligned} f(a) &= \lim_{\delta \to 0} \frac{F(a+\delta - F(a))}{\delta} \\ &= \frac{1}{\delta} \frac{1}{M} \#\{i = 1, \dots, M : a < x_i < a + \delta\} \end{aligned}$$

That is, the histogram of $x_1, \dots, x_M$ is an approximation to the density of X.

Theorem 1.3.5 Delta method

Let $\{X_n\}_{n=1}^{+\infty}$ be a sequence of r.v. such that $\forall \theta \in \Theta$

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

Let $\underbrace{\theta_0}_{\text{interior point of } \Theta} \in \overbrace{\Theta}^{\text{open set}}$ and g be a differentiable function such that $g'(\theta_0) \neq 0$. Then

$$\sqrt{n}(\underbrace{g(X_n)}_{\text{typically non-linear}} - g(\theta_0)) \xrightarrow{d} N(0, \sigma^2 [g'(\theta_0)]^2)$$

Proof. Using the 1st order Taylor expansion

$g(x) = g(\theta_0) + g'(\theta_0)(x - \theta_0) + r(x - \theta_0), \quad \frac{r(x-\theta_0)}{x - \theta_0} \to 0 \text{ as } x \to \theta_0$

$$\begin{aligned} g(x_n) - g(\theta_0) &= g'(\theta_0)(x_n - \theta_0) + r(x_n - \theta_0) \\ \sqrt{n}\,(g(x_n) - g(\theta_0)) &= \underbrace{\sqrt{n}\,(g(x_n) - g(\theta_0))}_{\text{Annex 1}} + \underbrace{\sqrt{n}\, r(x_n - \theta_0)}_{\text{Annex 2}} \\ \sqrt{n}\,(g(x_n) - g(\theta_0)) &= \underbrace{\sqrt{n}\,(g(x_n) - g(\theta_0))}_{\xrightarrow{d} N(0, [g'(\theta_0)]^2 \sigma^2)} + \underbrace{\sqrt{n}\, r(x_n - \theta_0)}_{\xrightarrow{P} 0} \end{aligned}$$

Annex 1:

$\sqrt{n}(x_n - \theta_0) \xrightarrow{d} N(0, \sigma^2)$

By **Slutsky's theorem**, $\underbrace{g'(\theta_o)}_{\text{constant}, \xrightarrow{P} g'(\theta_0)} \underbrace{\sqrt{n}(x_n - \theta_0)}_{\xrightarrow{d} T(\cdot)} \xrightarrow{d} g'(\theta_0) N(0, \sigma^2) = N(0 \times g'(\theta_0), [g'(\theta_0)]^2 \sigma^2)$

$\Rightarrow \sqrt{n}(x_n - \theta_0) \xrightarrow{d} N(0, [g'(\theta_0)]^2 \sigma^2)$

Annex 2:

Step 1:

$(T_n - \theta) = \underbrace{\frac{1}{a_n}}_{\xrightarrow{P} 0} \underbrace{a_n(T_n - \theta)}_{\xrightarrow{d} T(\cdot)} \xrightarrow{d} 0 \times T(\cdot) = 0 \Rightarrow T_n \xrightarrow{d} \theta \quad \underbrace{\Leftrightarrow}_{\text{this applies because } \theta \text{ is a constant}} \quad T_n \xrightarrow{P} \theta$

$\therefore \sqrt{n}(T_n - \theta) \xrightarrow{d} T \Rightarrow T_n \xrightarrow{P} \theta$

Step 2:

$\sqrt{n}(x_n - \theta_0) \xrightarrow{d} N(0, \sigma^2)$

By step 1, we can conclude that $x_n \xrightarrow{P} \theta_0 \quad X_n - \theta_0 \xrightarrow{P} 0$

We also know that $\frac{r(x)}{x} \to 0$

By **continuous mapping theorem**, $\frac{r(x_n-\theta_0)}{x_n-\theta_0} \xrightarrow{P} 0$

Step 3:

$$\sqrt{n}\, r(x_n-\theta_0) \Leftrightarrow \underbrace{\sqrt{n}(x_n-\theta_0)}_{\xrightarrow{d} N(0,\sigma^2)} \underbrace{\frac{r(x_n-\theta_0)}{x_n-\theta_0}}_{\xrightarrow{P} 0}$$

By **Slutsky's theorem**, $\sqrt{n}\, r(x_n-\theta_0) \xrightarrow{d} 0 \quad \underbrace{\Rightarrow}_{\text{true for constant}} \quad \sqrt{n} r(x_n-\theta_0) \xrightarrow{P} 0$ ■

Example 1.3.4 (Application : log-odds ratio)

Suppose that $X_1, \ldots, X_n$ are iid $B(1,\theta)$. Then the **CLT** ensures that

$$\sqrt{n}\frac{\bar{X}_n-\theta}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} N(0,1)$$

which is equivalent to $\sqrt{n}(\bar{X}_n-\theta) \xrightarrow{d} N(0,\theta(1-\theta))$.

- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n x_i$ representing the proportion of successes in the random sample
- θ representing the probability of success in the population

Now we are interested in the asymptotic distrbution of $Y_n = \ln \frac{\bar{X}_n}{1-\bar{X}_n}$ which is the empirical log odds of success, a non-linear function. With $g(x) = \ln \frac{x}{1-x}$, following $g'(x) = \frac{1}{x(1-x)}$. The **delta method** ensures that

$$\sqrt{n}(Y_n - \ln\frac{\theta}{1-\theta}) \xrightarrow{d} N(0, [\theta(1-\theta)]^{-1})$$

which is often written as

$$Y_n \overset{a}{\sim} N(\ln\frac{\theta}{1-\theta}, \frac{[\theta(1-\theta)]^{-1}}{n})$$

Proof. Asymptotic distribution of $T_n = \ln \frac{\bar{X}_n}{1-\bar{X}_n}$

$\sqrt{n}\frac{\bar{X}_n-\theta}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} N(0,1) \Leftrightarrow \sqrt{n}(\bar{X}_n-\theta) \xrightarrow{d} N(0,\theta(1-\theta))$

$$g(x) = \ln\frac{x}{1-x} \to g'(x) = \frac{(\frac{x}{1-x})}{(\frac{x}{1-x})} = \frac{\frac{1-x-(-1)x}{(1-x)^2}}{\frac{x}{1-x}} = \frac{1-x+x}{(1-x)^2}\frac{1-x}{x} = \frac{1}{x(1-x)}$$

Applying the **delta method**, $\sqrt{n}(T_n - g(\theta_0)) \xrightarrow{d} N(0, [g'(\theta_0)]^2\sigma^2)$

$$\Rightarrow \sqrt{n}(T_n - \ln\frac{\theta_0}{1-\theta_0}) \xrightarrow{d} N\left(0, \left(\frac{1}{\theta(1-\theta)}\right)^2 \theta_0(1-\theta_0)\right) \Leftrightarrow \sqrt{n}(T_n - \ln\frac{\theta_0}{1-\theta_0}) \xrightarrow{d} N(0, [\theta_0(1-\theta_0)]^{-1})$$ ■

Example 1.3.5 (Application : variance stabalizing)

Suppose $X_1, \ldots, X_n$ are $B(0,\theta)$. Then the **CLT** ensures that

$$\sqrt{n}\frac{\bar{X}_n-\theta}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} N(0,1)$$

Note that the asymptotic variance depends on the true value of θ, meaning that the variance, σ^2 is not fixed, thus giving us the motive to stabalize the variance. Our goal is to find a g such that $\sqrt{n}(g(\bar{X}_n) - g(\theta)) \xrightarrow{d} N(0,1)$, which is the same as solving for $g'(x) = \frac{1}{\sqrt{\theta(1-\theta)}}$.

$[g'(x)]^2\theta(1-\theta) = 1 \Leftrightarrow g'(x) = \frac{1}{\theta(1-\theta)} = \theta^{-1/2}(1-\theta)^{-1/2} \Rightarrow g(\theta) = 2\ \arcsin\sqrt{\theta}$

After this, the asymptotic distribution would be normal with a constant variance.

$$\sqrt{n}(2\ \arcsin\sqrt{\bar{X}_n} - 2\ \arcsin\sqrt{\theta}) \xrightarrow{d} N(0,1)$$

When can we apply this technique?
$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \ln(\mu))$
From **delta method**, $\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, [g'(\mu)]^2 ln(\mu))$
the variance stabalizing transformation g satisfies $[g'(x)]^2 \ln(\mu) = 1 \Leftrightarrow g'(\mu) = \frac{1}{\sqrt{\ln(\mu)}} \Rightarrow g(\mu) = \int_c^{\mu} \frac{1}{h(t)}\, dt$
c being some constant that ensures the integral exists, and with this c,
$\sqrt{n}(g(x_n) - g(\mu)) \xrightarrow{d} N(0,1)$

Chapter 2

Classical Statistical Model

2.1 Probability versus statistical inference

Probability theory begins with a completely specified model which we assume are correct and we compute the probabilities of certain events. For example,

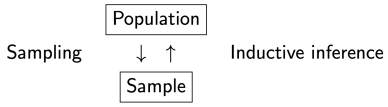
$$X_1, \ldots, X_n \overset{iid}{\sim} B(1, \theta)$$

$$T = \sum_{i=1}^{n} X_i \sim B(n, \theta)$$

$$P(T = t|\theta) = \binom{n}{t} \theta^t (1-\theta)^{n-t} \quad t = 0, 1, \ldots, n$$

with $n = 20 \rightarrow P(T = 10|\theta)$ is calculatable if we know θ. On the other hand, for **statistical inference**, we observe the realization of certain events, and using that information we try to infer the probabilistic model that governs the corresponding random experiment. For example, $T = 10 \rightarrow$ observed outcome. I want to use this information to make inference about θ.

Statistical data result from experiments conducted on a subset of a poopulation, the sample, and we try to extend the conclusions obtained to the whole population.



Inductive inference means that there is uncertainty regarding the resulting inference. If we are just drawing finite samples, then we cannot be certain the result is in fact representative of the entire population. The opposite would be **deductive inference** where it is of mathematics. No questions about the validity of the inference. If A holds $\rightarrow$ B definitely holds.

2.2 Model specification

To formalize the process of statistical inference. The characteristic of interest is modeled as a random variable X with cumulative distribution function (cdf) F, the statistical model. The model must be specified either through a **parametric model** where F is a known up to a finite dimensional parameter, e.g. X as normal with mean μ and variance σ^2, both unknown. Or a **nonparametric model** where F is specified in a nonparametric fashion, e.g. F is an element f the set of all continuous and symmetric distribution. Focusing on the parametric statistical model:

$$\mathcal{F} = \{F(\cdot|\theta) : \underbrace{\theta}_{\text{parameter}} \in \underbrace{\Theta}_{\text{parameter space}} \}$$

Example 2.2.1 (Application : daily return of financial asset)
We can propose a normal $\rightarrow \mathcal{F} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}$ or a gamma $\rightarrow \mathcal{F} = \{G(\alpha, \lambda) : \alpha, \lambda > 0\}$

Example 2.2.2 (Application : insurance policy)
If we are interested in the number of claims per year in an insurance policy, we can propose the Poisson $\rightarrow \mathcal{F} = \{Po(\lambda) : \lambda > 0\}$

The specificiation is important and results from many factors, namely based on the knowledge of the problem at hand, knowledge of previous studies, and knowledge of probability theory. The consequence of model misspecification is always negative but is smaller for larger samples.

2.2.1 Sampling

Random sampling means that the observed data are one of many possible data sets we could have obtained in the same circumstances. The set of n observations, $(x_1, \ldots, x_n)$ which we have observed is a realization of an n-dimensional random variables $(X_1, \ldots, X_n)$.

$$\begin{aligned} (X_1, \ldots, X_n) &\quad \text{Random sample} \\ (x_1, \ldots, x_n) &\quad \text{Observed sample} \end{aligned}$$

The **sample space** is a subset of $\mathbb{R}^n$ that contains the set of possible values for $x_1, \ldots, x_n$. We denoteit by $\mathcal{X}$.

Definition 2.2.1: IID random sampling

When the n random variables that compose the random sample are

- mutually independent $\rightarrow x_i \perp\!\!\!\perp x_j|\theta$
- identically distributed, with the same distribution as $X \rightarrow x_i \sim x_j|\theta$

we say that $(X_1, \ldots, X_n)$ constitutes an iid random sample of size n obtained from the population X. In notation, $X_1, \ldots, X_n|\theta \overset{iid}{\sim} X$, X follows the common distribution of all x_i's, $x_i \sim X$.

If $\mathcal{F} = \{f(\cdot|\theta) : \theta \in \Theta\}$ and $X_1, \ldots, X_n|\theta \overset{iid}{\sim} X$, then

$$\begin{aligned} F_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^{n} F_{X_i}(x_i|\theta) \qquad \text{by independence} \\ &= \prod_{i=1}^{n} F(x_i|\theta) \qquad \text{since } X_i \sim X \end{aligned}$$

and similarly for the probability density function

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Proof. **Poisson distribution**

$$\begin{aligned} f(x_1, \ldots, x_n|\lambda) &= \prod_{i=1}^{n} f(x_i|\lambda) \\ &= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\ &= P(x_1 = x_1, \ldots, x_n = x_n|\lambda) \\ &= e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \rightarrow \text{Annex 1} \end{aligned}$$

Annex 1:

- $e^a e^b = e^{a+b}$
- $a^x a^y = a^{x+y}$

■

2.3 Statistics

Definition 2.3.1: Statistic

A statistic is any function of $(X_1, \dots, X_n)$ that does not depend on unknown parameters.

Example 2.3.1 (Statistic)

In the context of a $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ and $\sigma > 0$ unknown.

Uni-dimensional statistics include

- $T = \sum_{i=1}^n X_i$
- $\bar{X} = \frac{1}{n} T$
- $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$

Bi-dimensional statistics include

- $(T, \sum_{i=1}^n X_i^2)$
- $(\bar{X}, S^2)$

Example 2.3.2 (Not statistic)

$$\sum_{i=1}^n (X_i - \mu)^2 \qquad \frac{1}{\sigma^2} \sum_{i=1}^n X_i^2$$

are not statistics becuase they depend on unknown parameters. If σ^2 is known, then $\frac{1}{\sigma^2} \sum_{i=1}^n X_i^2$ is a statistic.

Statistics operate a data reduction and are summaries of the information contained in the random sample. Statistics are random variables, as usual, it is important to distinguish between the random variable and its observed value.

population	random sample	observed sample
X	$(X_1, \dots, X_n)$	$(x_1, \dots, x_n)$
population mean	sample mean	mean of the sample
$\mu = E[X]$	$\bar{X} = \frac{1}{n} \sum_i X_i$	$\bar{x} = \frac{1}{n} \sum_i x_i$
population variance	sample variance	variance of the sample
$\sigma^2 = \mathrm{Var}(X)$	$S^2 = \frac{1}{n} \sum_i (X_i - \bar{X})^2$	$s^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$

Probability vs. Statistics vs. Data exploration

2.4 Sampling distribution

The sampling distribution of a statistic corresponds to its probability distribution: as $(X_1, \dots, X_n)$ varies according to its distribution, what is the resulting probabilistic behavior of $T(X_1, \dots, X_n)$. In classical inference, it is important to know the sampling distribution of statistics because that is necessary to evaluate the performance of statistical methodologies. The **objective** is to determine aspects of the sampling distribution of a statistic T knowing aspects of the probability distribution of the poopulation X.

There are different methods to obtain the sampling distribution of a statistic.

- **Change of variable**: If X is continuous,

$$F_T(t|\theta) = P(T \leq t|\theta) = \int_{A(t)} \sum_{i=1}^{n} f(x_i|\theta)\, dx_1 \dots dx_n$$

  where $A(t) = \{(x_1, \dots, x_n) \in \mathbb{R}^n : T(x_1, \dots, x_n) \leq t\}$. If X is discrete, replace integrals with sums.
- Determining the **moment generating function** of T
- Using **well-known properties** of the distribution of X
- **Asymptotic approximations** to the sampling distribution of certain statistics (from CLT and related results)
- Using **simulations**

Example 2.4.1 (Change of variable)
Let $T = \sum_{i=1}^{n} X_i$

- If $(X_1, \dots, X_n)$ is am iid random sample from a $Po(\lambda)$ population, since the sum of independent Poisson is still Poisson, we have $T \sim Po(n\lambda)$, hence $f_T(t|\lambda) = e^{-n\lambda} \frac{(n\lambda)^t}{t!}$, $t \in N_0$.
- If $(X_1, \dots, X_n)$ is an iid random sample from a $N(\mu, \sigma^2)$ population, then $T \sim N(n\mu, n\sigma^2)$.
- If $(X_1, \dots, X_n)$ is an iid random sample from a $B(1, \theta)$ population, then $T \sim B(n, \theta)$.

Example 2.4.2 (Monte Carlo simulation)

1. Draw N independent samples of size n from the distribution of X
2. For each of those samples, comptute the observed values of the statistic T
3. The N resulting numbers, $(t_1, \dots, t_n)$ constitute a sample of size N drawn from the sampling distribution of T

2.4.1 Sample distribution of the sample moments

Definition 2.4.1: Sample moments

Let $(X_1, \dots, X_n)$ be an iid random sample of size n from a population X. For $k \in \mathbb{N}$ we define the kth raw sample moment as

$$M'_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

and the kth central sample moment by

$$M_k = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^k$$

$\begin{cases} \mu'_k = E[x^k] \to k\text{th raw moment} \\ M'_k = \frac{1}{n} \sum x_i^k \to k\text{th sample raw moment} \end{cases}$ $\begin{cases} \mu_k = E[(x-\mu)^k] \to k\text{th central moment} \\ M_k = \frac{1}{n} \sum (x_i - \bar{x})^k \to k\text{th central sample moment} \end{cases}$

We want to observe how they behave in relation to each other.

Once again, it is important to distinguish the **sample moments**, M'_k and M_k, from the **population moments**, $\mu'_k = E[X^k]$ and $\mu_k = E[(X - E[X])^k]$, and the **observed sample moments**, $m'_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$ and $m_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k$.

Note:

Important special cases: $\bar{X} = M_1'$ and $S^2 = M_2$, the sample mean and the sample variance.

Theorem 2.4.1 Properties of the sample mean

If all the moments exist, then

$$\begin{aligned} E[\bar{X}] &= E[X] = \mu \\ Var(\bar{X}) &= \frac{Var(X)}{n} = \frac{\sigma^2}{n} \\ \mu_3 &= \frac{\mu_3}{n^2} \\ \mu_4 &= \frac{3\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3} \end{aligned}$$

Proof. $E[\bar{X}]$

$$E[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n} n\mu = \mu$$

■

Proof. $Var(\bar{X})$

$$\begin{aligned} Var(\bar{X}) &= E[(\bar{X} - \mu)^2] \\ &= E[(\frac{1}{n}\sum x_i - \mu)^2] \\ &= E[\frac{1}{n^2}(\sum (x_i - \mu))^2] \\ &= \frac{1}{n^2} E[\overbrace{(\sum \underbrace{a_i}_{=x_i-\mu})^2}^{\text{Annex 1}}] \\ &= \otimes \end{aligned}$$

Annex 1: $(\sum a_i)^2 = (\sum_i a_i)(\sum_j a_j) = \sum_i \sum_j a_i a_j = \sum_i a_i^2 + \sum_i \sum_{i \neq j} a_i a_j$ □

$$\begin{aligned} \otimes &= \frac{1}{n^2}[E[\sum_i a_i^2] + E[\sum_i \sum_{i \neq j} a_i a_j]] \\ &= \frac{1}{n^2}[\sum_i \underbrace{E[a_i^2]}_{E[(X-\mu)^2]=Var(X)=\sigma^2} + \sum_i \sum_{i \neq j} \underbrace{E[a_i a_j]}_{\text{Cov}(a_i,a_j)=0 \rightarrow \text{ Annex 2}}] \\ &= \frac{1}{n^2}(n\sigma^2 + 0) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Annex 2: Since $a_i \perp\!\!\!\perp a_j$, expectation of the product is the product of the expectation for independent variables. This implies that $E[(x_i - \mu)(x_j - \mu)] = E[(x_i - \mu)]E[(x_j - \mu)] = 0 \times 0 = 0$ ■

By the **Weak Law of Large Numbers**, $\bar{X} \xrightarrow{P} \mu$. The distribution is more and more concentrated around μ as n increases. The distribution of $\bar{X}$ is centered around μ and $\lim_{n \to +\infty} Var(\bar{X}) = 0$.

Proof. Asymmetry μ_3

$$\mu_3(\bar{X}) = E[(\bar{X} - \mu)^3] = \frac{1}{n^3} E[\underbrace{(\sum a_i)^3}_{\text{Annex 1}}] = \otimes$$

Annex 1:

$$
\begin{aligned}
(\sum a_i)^3 &= (\sum_i a_i)(\sum_j a_j)(\sum_k a_k) = \sum_i \sum_j \sum_k a_i a_j a_k \\
&= (\sum a_i)^2 (\sum_k a_k) = (\sum_i a_i^2 + \sum_i \sum_{i \neq j} a_i a_j)(\sum_k a_k) \\
&= \sum_i \sum_k a_i^2 a_k + \sum_i \sum_{i \neq j} \sum_k a_i a_j a_k \\
&= \sum_i a_i^3 + \underbrace{\sum_i \sum_{k \neq i} a_i^2 a_k}_{E[\cdot]=0} + \underbrace{\sum_i \sum_{j \neq i} \sum_{k \neq i} a_i a_j a_k}_{E[\cdot]=0} + \underbrace{\sum_i \sum_{j \neq i} a_i^2 a_j}_{E[\cdot]=0} \qquad \square
\end{aligned}
$$

$$
\otimes = \frac{1}{n^3} \sum_{i=1}^{n} E[(x_i - \mu)^3] = \frac{1}{n^3} n \mu_3 = \frac{\mu_3}{n^2} \xrightarrow{n \to \infty} 0
$$

■

As n goes to infinity, the distribution becomes symmetric. This evidence is compatible with the **Central Limit Theorem**. The related concept to asymmetry is **skewness**, $\gamma_1 = \frac{\mu_3}{\sigma^3}$. σ^3 is used to make γ_1 dimensionless, as in independent of the unit measurement of the x.

The commonly used concept related to μ_4 is **kurtosis** which is often denoted as $\gamma_2 = \frac{\mu_4}{\sigma^4}$. It is the indication of heavy tails. If $\gamma_2 > 3$, the distribution has a heavier tail than Gaussian. The excess kurtosis can also be used which is just $\gamma_2 - 3$, indicating heavier tail than Gaussian if bigger than 0.

Proof. Kurtosis γ_2

$$
\begin{aligned}
\gamma_2(\bar{X}) &= \frac{\mu_4(\bar{X})}{\underbrace{\sigma^4_{(\bar{X})}}_{\text{s.d. to the power of 4}}} = \frac{\frac{3\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}}{(\sqrt{\frac{\sigma^2}{n}})} \\
&= \frac{\frac{3\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}}{(\frac{(\sigma^2)^2}{n^2})} \overset{\mu_2 = \sigma^2}{=} \frac{\frac{3\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}}{(\frac{(\mu_2)^2}{n^2})} \\
&= 3 + \frac{1}{n} \underbrace{\frac{\mu_4 - 3\mu_2^2}{\mu_2^2}}_{\text{A}} \xrightarrow{n \to \infty} 3
\end{aligned}
$$

A: $\frac{\mu_4}{\mu_2^2} - \frac{3\mu_2^2}{\mu_2^2} = \frac{\mu_4}{\sigma^4} - 3 \to \gamma_2 - 3 \Rightarrow$ excess kurtosis.

■

Theorem 2.4.2 Properties of the sample variance

If all the moments exist,

$$
\begin{aligned}
E[S^2] &= \frac{n-1}{n} \sigma^2 \\
Var(S^2) &= \frac{\mu_4 - \mu_2^2}{n} - 2\frac{\mu_4 - 2\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3} \xrightarrow{n \to \infty} 0 \text{ , roughly centered around } S^2
\end{aligned}
$$

Since $E[S^2] = \frac{n-1}{n}\sigma^2 < \sigma^2$, always strictly smaller than the variance, we define the **bias-corrected sample variance**

$$
S'^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n}{n-1} S^2
$$

Notice that $M'_k = \frac{1}{n}\sum X_i^k$ which is the average of iid r.v. but $M_k = \frac{1}{n}\sum(X_i - \bar{X})^k$. We cannot use LLN or CLT to study M_k becuase $(X_i - \bar{X})$ and $(X_j - \bar{X})$ for $i \neq j$ are not iid as $\bar{X}$ depends on both X_i and X_j.

Theorem 2.4.3 Properties of the bias-corrected sample variance

If all the moments exist,

$$E[S'^2] = \sigma^2$$
$$Var(S'^2) = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\mu_2^2)$$

Proof. $E[S'^2]$

$$S'^2 = \frac{nS^2}{n-1}$$
$$E[S'^2] = \frac{n}{n-1}E[S^2] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2$$

■

Theorem 2.4.4 Properties of central sample moments

If all the moments exist,

$$E[M_k] = \mu_k + O(\frac{1}{n})$$
$$Var(S'^2) = \frac{c}{n} + O(\frac{1}{n^2})$$

where c is a constant which involves central population moments of order $\leq 2k$.

The central sample moments have similar behavior as S^2.

$$a_n = O(b_n) \Leftrightarrow \frac{a_n}{b_n} \text{ is limited}$$
$$a_n = O(\frac{1}{n}) \Leftrightarrow \frac{a_n}{\frac{1}{n}} \text{ is limited} = na_n$$
$$a_n = \underbrace{\frac{1}{n}}_{\to 0}\underbrace{na_n}_{\text{limited}} \to 0$$
$$\Rightarrow O(\frac{1}{n}) \text{ sequence goes to 0 roughtly at the rate of } \frac{1}{n}$$

Theorem 2.4.5 Asymptotic distribution of $\bar{X}$

As long as $Var(X)$ is finite, we have as a direct consequence of the **Central limit theorem** that

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

This result is typically used in the form

$$P(\bar{X} \leq x) \approx \Phi\left(\sqrt{n}\frac{x-\mu}{\sigma}\right)$$

that is, $\bar{X} \overset{a}{\sim} N(\mu, \frac{\sigma^2}{n})$.

In general, unimodality and symmetry have a positive impact on the speed of convergence. If the distribution is already behaving like normal, the rate of convergence would be faster.

2.4.2 Order statistics

Definition 2.4.2: Order statistics

Let $(X_1, \ldots, X_n)$ be an iid random sample. The ith order statistic is denoted by $X_{(i)}$ and satisfies

$$\underbrace{X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}}_{\text{not iid}}$$

Order statistics are not iid because if we know $X_{(n)}$ is 10, then $X_{(1)}$ cannot be larger than 10. For notation purposes, $(Y_1, \ldots, Y_n) \equiv (X_{(1)}, \ldots, X_{(n)})$.

The order statistics have a joint pdf given by

$$g(y_1, y_2, \ldots, y_n) = n! \prod_{i=1}^{n} f(y_i) \qquad \text{if } y_1 < y2 < \ldots < y_n$$

If $u > v$, the joint pdf of (Y_u, Y_v) is

$$g_{u,v}(y,z) = \frac{n!}{(u-1)!(v-u-1)!(n-v)!} \times [F(y)]^{u-1}[F(z)-F(y)]^{v-u-1}[1-F(z)]^{n-v} f(y) f(z) \quad \text{if } y < z$$

The cdf of Y_v is

$$g_v(y) = \frac{n!}{(v-1)!(n-v)!}[F(y)]^{v-1}[1-F(y)]^{n-v} f(y) = G_v'(y)$$

and the pdf of Y_v is

$$G_v(y) = \sum_{j=v}^{n} \binom{n}{j} [F(y)]^j [1-F(y)]^{n-j} = P(Y_v \leq y)$$

Proof. The joint distribution of $G_{u,v}(y,z)$ follows a **multinomial distribution** with the following characteristics:

- n objects
- k categories
- The probability that each object belongs to category i is p_i, $i = 1, \ldots, k$ independently
- w_i is the number of objects among the n that belongs to class i, $i = 1, \ldots, k$

$$\begin{aligned} G_{u,v}(y,z) &= P(Y_u \leq y, Y_v \leq z) \\ g_{u,v}(y,z) &= \frac{\partial^2 G(y,z)}{\partial y \partial z} = \lim_{\delta_1 \to 0^+, \delta_2 \to 0^+} P(y < y_u \leq y + \delta_1, z < y_v \leq z + \delta_2) \end{aligned}$$

This is a multinomial with 5 classes:

$$\begin{cases} P_1 = F(y) \\ P_2 = P(y < x \leq y + \delta_1) = F(y + \delta_1) - F(y) \\ P_3 = F(z) - F(y + \delta_1) \\ P_4 = P(z < x \leq z + \delta_2) = F(z + \delta_2) - F(z) \\ P_5 = 1 - F(z + \delta_2) \end{cases}$$

with

$$\begin{cases} w_1 = u - 1 \\ w_2 = 1 \\ w_3 = v - u - 1 \longrightarrow \text{Annex 1} \\ w_4 = 1 \\ w_5 = n - v \end{cases}$$

Annex 1: $n = u - 1 + 1 + 1 + n - v + w_3 \Leftrightarrow n = u + 1 + n - v + w_3 \Leftrightarrow w_3 = n - u - 1 - n + v = v - u - 1$
where

$$P(y < Y_u \leq y + \delta_1, z < Y_v \leq z + \delta_2) = \frac{n!}{(u-1)!1!(v-u-1)!1!(n-v)!} P_1^{u-1} P_2^1 P_3^{v-u-1} P_4^1 P_5^{n-v}$$

thus,

$$\begin{aligned}
g_{u,v}(y,z) &= \lim_{\delta_1 \to 0^+, \delta_2 \to 0^+} \frac{n!}{(u-1)!(v-u-1)!(n-v)!} [F(y)]^{u-1} [F(y+\delta_1) - F(y)] \\
&\times [F(z) - F(y+\delta_1)]^{v-u-1} [F(z+\delta_2) - F(z)] [1 - F(z+\delta_2)]^{n-v} \\
&= \frac{n!}{(u-1)!(v-u-1)!(n-v)!} [F(y)]^{u-1} [F(z) - F(y)]^{v-u-1} [1 - F(z)]^{n-v} \\
&\times \underbrace{\lim_{\delta_1 \to 0^+} \frac{F(y+\delta_1) - F(y)}{\delta_1}}_{f(y)} \underbrace{\lim_{\delta_2 \to 0^+} \frac{F(z+\delta_2) - F(z)}{\delta_2}}_{f(z)} \\
&= \frac{n!}{(u-1)!(v-u-1)!(n-v)!} [F(y)]^{u-1} [F(z) - F(y)]^{v-u-1} [1 - F(z)]^{n-v} f(y) f(z)
\end{aligned}$$

■

Proof. The cdf of Y_v

$$\begin{aligned}
G_v(y) &= P(Y_v \leq y) = P(\text{at least } v \text{ of the } X_i\text{'s are } \leq y) \\
&= \underbrace{\sum_{j=v}^{n} P(\text{exactly } j \text{ of the } X_i\text{'s are } \leq y)}_{N_y = \#\{i : x_i \leq y\}} \\
&= \sum_{j=v}^{n} \binom{n}{j} [F(y)]^j [1 - F(y)]^{n-j} \quad \text{since } N_y \sim B(n, F(y))
\end{aligned}$$

The pdf is just the derivative of the cdf, thus

$$\begin{aligned}
g_v(y) = G_v'(y) &= \sum_{j=v}^{n} \binom{n}{j} \left[j[F(y)]^{j-1} f(y) [1 - F(y)]^{n-j} - [F(y)]^j (n-j) [1 - F(y)]^{n-j-1} (-f(y)) \right] \\
&= f(y) \sum_{j=v}^{n} \binom{n}{j} [F(y)]^{j-1} [1 - F(y)]^{n-j} [j + n - j - nj] \\
&= f(y) \sum_{j=v}^{n} \binom{n}{j} [F(y)]^{j-1} [1 - F(y)]^{n-j} [n - nj + j] \\
&= f(y) \sum_{j=v}^{n} \binom{n}{j} [F(y)]^{j-1} [1 - F(y)]^{n-j} [n(1-j) + j] \\
&= f(y) \sum_{j=v}^{n} \binom{n}{j} [F(y)]^{j-1} [1 - F(y)]^{n-j} [n - nj + j] \\
&= \frac{n!}{(v-1)!(n-v)!} [F(y)]^{v-1} [1 - F(y)]^{n-v} f(y)
\end{aligned}$$

■

Theorem 2.4.6 Important special cases : the maximum and the minimum

The pdf and the cdf of the minimum and the maximum are

$$G_1(y) = 1 - [1 - F(y)]^n \qquad g_1(y) = n f(y)[1 - F(y)]^{n-1}$$
$$G_n(y) = [F(y)]^n \qquad g_n(y) = n f(y)[F(y)]^{n-1}$$

with the joint cdf of

$$g_{1,n}(y, z) = n(n-1)[F(z) - F(y)]^{n-2} f(y) f(z) \qquad y < z$$

Proof. $G_n(y)$

$$\begin{aligned} G_n(y) &= P(Y_n \leq y) = P(X_{(n)} \leq y) \\ &= P(X_{(1)} \leq y, \ldots, X_{(n)} \leq y) \rightarrow \text{Annex 1} \\ &= P(X_{(1)} \leq y)\, P(X_{(2)} \leq y) \cdots P(X_{(n)} \leq y) \quad \text{because } x_i \perp\!\!\!\perp x_j, i \neq j \\ &= [F(y)]^n \quad \text{because iid} \end{aligned}$$

Annex 1: If $X_{(n)} \leq y \Leftrightarrow X_{(1)} \leq y, \ldots, X_{(n)} \leq y$

$g_n(y) = G_n'(y)$ if continuous

■

Proof. $G_1(y)$

The idea is that

$$X_{(1)} \leq y \Leftrightarrow \exists_i X_i \leq y$$

Thus we can write

$$\begin{aligned} G_1(y) &= P(X_{(1)} \leq y) = P(\exists_i X_i \leq y) \\ &= P(\overline{\forall_i X_i > y}) = 1 - P(\forall X_i > y) \\ &= 1 - P(X_{(1)} > y, \ldots, X_{(n)} > y) = 1 - P(X_{(1)} > y) \cdots P(X_{(n)} > y) \quad \text{because } x_i \perp\!\!\!\perp x_j, i \neq j \\ &= 1 - [1 - F(y)]^n \quad \text{because iid} \end{aligned}$$

■

Proof. $g_{1,n}(y, z)$

$$\begin{aligned} G_n(y) &= G_{1,n}(x, y) + P(x < X_i \leq y, \forall i) \\ &= G_{1,n}(x, y) + \underbrace{[F[(y) - F(x)]^n}_{P(a<X\leq b)=F(b)-F(a)} \quad \text{justified through } \textbf{total probability theorm} \\ \Leftrightarrow [F(y)]^n &= G_{1,n}(x, y) + [F(y) - F(x)]^n \\ \Leftrightarrow G_{1,n}(x, y) &= [F(y)]^n - [F(y) - F(x)]^n \,, x < y \end{aligned}$$

The cumulative distribution function is just the derivative relative to both variables, so

$$\begin{aligned} g_{1,n}(x, y) &= \frac{\partial^2}{\partial y \partial x} G_{1,n}(x, y) \\ &= \frac{\partial}{\partial y}[0 - n(-f(x))[F(y) - F(x)]^{n-1}] \\ &= nf(x)(n-1)f(y)[F(y) - F(x)]^{n-2} \,, n > 2 \end{aligned}$$

■

Example 2.4.3 (Pareto)

Pareto distribution is one of the fat tail distributions.

$$f(x) \propto x^{-(\theta+1)}, x > c \Leftrightarrow f(x) = bx^{-(\theta+1)}$$

To find b, we impose that the integral of the pdf is equal to 1.

$$\begin{aligned}
\int_c^{+\infty} bx^{-(\theta+1)}\, dx &= 1 \\
b\left[-\frac{1}{\theta}x^{-\theta}\right]_c^{+\infty} &= 1 \\
b\left(0 - \frac{c^{-\theta}}{-\theta}\right) &= 1 \\
b\frac{c^{-\theta}}{\theta} &= 1 \\
b &= \theta c^{\theta}
\end{aligned}$$

Polynomial decays slower than exponential decays. This means that the tails are heavier.

$$\begin{aligned}
F(x) = P(X \leq x) &= \int_c^x f(u)\, du \\
&= \int_c^x \theta c^{\theta} u^{-(\theta+1)}\, du \\
&= \theta c^{\theta}\left[-\frac{u^{-\theta}}{\theta}\right]_c^x \\
&= \theta c^{\theta}\left(\frac{x^{-\theta} - c^{-\theta}}{-\theta}\right) \\
&= c^{\theta}(c^{-\theta} - x^{-\theta}) \\
&= 1 - \left(\frac{c}{x}\right)^{\theta} \\
&= 1 - c^{\theta}x^{-\theta}, x > c
\end{aligned}$$

Now showing the distribution of the minimum order statistic.

$$\begin{aligned}
g_1(y) &= nf(y)[1 - F(y)]^{n-1} \\
&= n\theta c^{\theta} y^{-(\theta+1)}\left[c^{\theta}y^{-\theta}\right]^{n-1} \\
&= n\theta c^{\theta} y^{-(\theta+1)} c^{\theta(n-1)} y^{-\theta(n-1)} \\
&= n\theta c^{\theta n} y^{-(\theta n+1)}, y > c
\end{aligned}$$

This has the same structure as a Pareto distribution with parameter θn. Thus, the minimum of a sample of size n from a Pareto(θ) distribution is still a Pareto distribution with parameter θn.

$$X_{(1)} \sim Pa(c, \theta n)$$

However, the maximum does not have a simple form.

$$g_n(y) = n\theta c^{\theta} y^{-(\theta+1)}\left[1 - c^{\theta}y^{-\theta}\right]^{n-1}, y > c$$

The maximum does not belong to the Pareto family.

Example 2.4.4 (Exponential)

Let $(X_1, \ldots, X_n)$ be an iid random sample from an $Exp(\lambda)$ population. The pdf and cdf of the exponential distribution are

$$f(x) = \lambda e^{-\lambda x}, x > 0 \qquad F(x) = 1 - e^{-\lambda x}, x > 0$$

The pdf of the minimum order statistic is

$$\begin{aligned} g_1(y) &= nf(y)[1 - F(y)]^{n-1} \\ &= n\lambda e^{-\lambda y}\left[e^{-\lambda y}\right]^{n-1} \\ &= n\lambda e^{-\lambda y} e^{-\lambda n y + \lambda y} \\ &= n\lambda e^{-\lambda n y}, y > 0 \end{aligned}$$

This has the same structure as an exponential distribution with parameter λn. Thus, the minimum of a sample of size n from an $Exp(\lambda)$ distribution is still an exponential distribution with parameter λn.

$$X_{(1)} \sim Exp(\lambda n)$$

Also notice that

$$E[x] = \frac{1}{\lambda} \Rightarrow E[X_{(1)}] = \frac{1}{\lambda n} < \frac{1}{\lambda}$$

This makes sense as it is often used to model duration.
The pdf of the maximum order statistic is

$$\begin{aligned} g_n(y) &= nf(y)[F(y)]^{n-1} \\ &= n\lambda e^{-\lambda y}\left[1 - e^{-\lambda y}\right]^{n-1}, y > 0 \end{aligned}$$

The maximum does not belong to the exponential family.

Definition 2.4.3: Sample quantile

Let $p \in (0, 1)$ and $k = np$. Then the sample quantile of order p is Z_p such that

$$Z_p = \begin{cases} \frac{Y_k + Y_{k+1}}{2} & \text{if } k \text{ is an integer} \\ Y_{(\lceil k \rceil)} & \text{if } k \text{ is not an integer} \end{cases}$$

where $\lceil k \rceil$ is the integer part of k.

Theorem 2.4.7 Asymptotic distribution of the sample quantile of order p

Let Z_p be the sample quantile of order p of an iid random sample of size n, obtained from a continuous population with density f. Denoted by ξ_p the population quantile of order p. If f is continuous and positive at ξ_p, then

$$\sqrt{n} f(\xi_p) \frac{Z_p - \xi_p}{\sqrt{p(1-p)}} \xrightarrow{d} N(0, 1)$$

$$\xi_p : P(X \leq \xi_p) \geq p \quad \text{and} \quad P(X \geq \xi_p) \geq 1 - p$$

If X is continous, $F(\xi_p) = p$. It is actually quite difficult to prove the relationship between the population quantle ξ_p and the sample quantile Z_p.

Example 2.4.5 (Normal)

If $X \sim N(\mu, \sigma^2)$, then $\xi_{\frac{1}{2}} = \mu$. Then $f(\xi_{\frac{1}{2}}) = (2\pi\sigma^2)^{-\frac{1}{2}}$. Hence,

$$\sqrt{n}\frac{Z_{\frac{1}{2}} - \mu}{\sigma} \xrightarrow{d} N(0, \frac{\pi}{2}) \Leftrightarrow \sqrt{\frac{2n}{\pi\sigma^2}}(Z_{\frac{1}{2}} - \mu) \xrightarrow{d} N(0,1)$$

Theorem 2.4.8 Fisher-Tippett-Gnedenko Theorem

If $\exists a_n, b_n, b_n > 0$ then

$$\lim_{n\to\infty} P\left(\frac{X_{(n)} - a_n}{b_n} \leq x\right) = F(x)$$

and F is the cdf of a non-degenerate r.v., then F belongs to one of the following three families:

- **Gumbel** : $F(x) = e^{-e^{-(x-\mu)/\beta}}, x \in \mathbb{R}$, with $\mu \in \mathbb{R}$ and $\beta > 0$
- **Fréchet** : $F(x) = 0, x \leq \mu$ and $F(x) = e^{-((x-\mu)/\beta)^{-\alpha}}, x > \mu$ with $\mu \in \mathbb{R}$, $\beta > 0$ and $\alpha > 0$
- **Weibull** : $F(x) = e^{-((\mu-x)/\beta)^{\alpha}}, x < \mu$ and $F(x) = 1, x \geq \mu$ with $\mu \in \mathbb{R}$, $\beta > 0$ and $\alpha > 0$

One of the three extreme value distributions.

2.4.3 A few sampling distributions

Normal population

Let $(X_1, \ldots, X_n)$ be an iid random sample from a $N(\mu, \sigma^2)$ distribution.

- Distribution of the sample mean $\bar{X}$

  $\bar{X}$ is a linear combination of independent normals, hence it follows a normal distribution. We know that $E[\bar{X}] = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$. Thus,

  $$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{or} \quad \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \sim N(0,1)$$

  Unlike the CLT which is approximately $N(0,1)$, this is exact.

- Distribution of the bias-corrected sample variance S'^2

  Because $\frac{X_i - \mu}{\sigma} \sim N(0,1)$, we have $\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_1^2$. Thus, $\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_n^2$.

  Also $\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n [(X_i - \bar{X}) + (\bar{X} - \mu)]^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$. Hence,

  $$\sum_{i=1}^n (X_i - \mu)^2 = (n-1)\frac{S'^2}{\sigma^2} + n\frac{(\bar{X} - \mu)^2}{\sigma^2}$$

  Since $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, we have $n\frac{(\bar{X} - \mu)^2}{\sigma^2} \sim \chi_1^2$. Also, $\bar{X}$ and S'^2 are independent in the context of the normal distribution. Thus,

  $$(n-1)\frac{S'^2}{\sigma^2} = \sum_{i=1}^n (X_i - \mu)^2 - n\frac{(\bar{X} - \mu)^2}{\sigma^2} \sim \chi_{n-1}^2$$

  Therefore,

  $$\frac{(n-1)S'^2}{\sigma^2} \sim \chi_{n-1}^2$$

- Student ratio

When the population variance is unknown, we have the student ratio:

$$\frac{\bar{X}-\mu}{S'/\sqrt{n}} = \frac{\bar{X}-\mu}{S/\sqrt{n-1}} \sim t(n-1)$$

Proof. $\frac{N(0,1)}{\sqrt{\chi^2(n)/n}} \sim t(n)$, $N(0,1) \perp\!\!\!\perp \chi^2(n)$
We know that $\bar{X}$ and S'^2 are independent, thus

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S'^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S'/\sqrt{n}} \sim t(n-1)$$

and because $t(n) \xrightarrow{d} N(0,1)$, we have

$$\frac{\bar{X}-\mu}{S'/\sqrt{n}} \xrightarrow{d} N(0,1)$$

■

Two normal distributions

$$X_1 \sim N(\mu_1, \sigma_1^2)$$
$$X_2 \sim N(\mu_2, \sigma_2^2)$$

Two random samples, mutually independent, of size m and n respectively: $(X_{11}, \ldots, X_{1m})$ and $(X_{21}, \ldots, X_{2n})$.

- Difference of the sample means

$$\bar{X_1} - \bar{X_2} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}) \Leftrightarrow \frac{(\bar{X_1} - \bar{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$$

  However, the previous result is limited if the population variances are not known. We can assume both variances are equal, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Then,

$$T = \frac{\frac{(\bar{X_1} - \bar{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{m} + \frac{1}{n}}}}{\sqrt{\frac{(m-1)S_1'^2 + (n-1)S_2'^2}{m+n-2}}} \sim t(m+n-2)$$

  Considering the equation as $T = \frac{U}{\sqrt{V/(m+n-2)}}$, $U\frac{1}{\sigma} \sim N(0,1)$ and $V\frac{1}{\sigma^2} \sim \chi^2(m+n-2)$.
  When the variances are unkown and different, if the sample sizes are large, Slutsky's theorem allows us to replace the population variances by the sample variances and obtain the same distribution in the limit. However, for small sample sizes, we can use the Welch-Satterthwaite approximation

$$\frac{(\bar{X_1} - \bar{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1'^2}{m} + \frac{S_2'^2}{n}}} \overset{a}{\sim} t(\nu)$$

  where ν is the largest integer that does not exceed

$$\nu = \frac{\left(\frac{S_1'^2}{m} + \frac{S_2'^2}{n}\right)^2}{\frac{(S_1'^2/m)^2}{m-1} + \frac{(S_2'^2/n)^2}{n-1}}$$

- Two sample variances

$$U = \frac{(m-1)S_1'^2}{\sigma_1^2} \sim \chi^2_{m-1}$$
$$V = \frac{(n-1)S_2'^2}{\sigma_2^2} \sim \chi^2_{n-1}$$

Since $U \perp\!\!\!\perp V$, we have

$$F = \frac{U/(m-1)}{V/(n-1)} = \underbrace{\frac{S_1'^2}{S_2'^2}}_{\text{ratio of the corrected sample variance}} \times \underbrace{\frac{\sigma_2^2}{\sigma_1^2}}_{\text{inverted ratio of the variances}} \sim F(m-1, n-1)$$

Bernoulli population

Consider there are two types of individuals in the population: the ones who posses a certain attribute and the one who do not. Let $(X_1, \ldots, X_n)$ be an iid random sample of size n from a $B(1, \theta)$ population, where θ is the proportion of successes in the population (known). It is useful to establish the sampling distribution of two statistics:

- The number of individuals in the sample who posses the attribute: $T = \sum_{i=1}^n X_i$
- The proportion of individuals in the sample who posses the attribute: $\bar{X} = \frac{T}{n}$

Clearly, $T \sim B(n, \theta)$, hence

- $P(T = t) = \binom{n}{t} \theta^t (1-\theta)^{n-t}, t = 0, 1, \ldots, n$
- $P(\bar{X} = z) = P(T = nz) = \binom{n}{nz} \theta^{nz} (1-\theta)^{n-nz}, z = \frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}$

Theorem 2.4.9 De Moivere-Laplace theorem

For large sample sizes,

$$\frac{T - n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow{d} N(0,1) \Leftrightarrow \frac{\bar{X} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \xrightarrow{d} N(0,1)$$

use when $n > 20, n\theta \geq 5, n\theta(1-\theta) \geq 5, 0.1 < \theta < 0.9$.

Combined with the continuity correction, we have

$$\begin{aligned} P(a \leq T \leq b) &\approx P\left(\frac{a - 0.5 - n\theta}{\sqrt{n\theta(1-\theta)}} \leq Z \leq \frac{b + 0.5 - n\theta}{\sqrt{n\theta(1-\theta)}}\right) \\ &= \Phi\left(\frac{b + 0.5 - n\theta}{\sqrt{n\theta(1-\theta)}}\right) - \Phi\left(\frac{a - 0.5 - n\theta}{\sqrt{n\theta(1-\theta)}}\right) \end{aligned}$$

with $a < b, a, b = 0, 1, \ldots, n$.

Proof. From the **CLT**, we get

$$\sqrt{n}\frac{\bar{X} - \theta}{\sqrt{\theta(1-\theta)}} \overset{a}{\sim} N(0,1)$$

Now going through some algebraic manipulations and applying the **CLT** again, we have

$$\sqrt{n}\frac{T - n\theta}{n\sqrt{\theta(1-\theta)}} = \frac{T - n\theta}{\sqrt{n}\sqrt{\theta(1-\theta)}} = \frac{T - n\theta}{\sqrt{n\theta(1-\theta)}} \overset{a}{\sim} N(0,1)$$

■

The motivation for the continuity correction is that the binomial distribution is discrete while the normal distribution is continuous. When approximating a discrete distribution with a continuous one, we need to adjust for this difference.

$$\begin{aligned} p(T = x) &= P\left(\frac{T - n\theta}{\sqrt{n\theta(1-\theta)}} = \frac{x - n\theta}{\sqrt{n\theta(1-\theta)}}\right) \overset{a}{\sim} N(0,1) \\ &\approx 0 \end{aligned}$$

Thus we need to consider the interval around x that captures the probability mass: $P(T = x) = P(x - 0.5 < T < x + 0.5)$.

Theorem 2.4.10 Law of rare events

Technically, $n \to \infty$ and $\theta \to 0$ such that $n\theta = \lambda$ is constant. Then,

$$T \overset{a}{\sim} Po(\lambda)$$

use for $n > 20$ and $\theta \notin (0.1, 0.9)$ and $n\theta < 5$.

Two Bernoulli populations

Consider two Bermoulli populations with success probabilities θ_1 and θ_2 respectively. We want to compare θ_1 and θ_2. $\theta_1 - \theta_2$ will be unknown and we want to make inference about this quantity through the statistics $\bar{X_1} - \bar{X_2}$, the difference between the sample proportions in two independent samples:

- $(X_{11}, \ldots, X_{1m}) \Rightarrow \bar{X_1} = \sum_{i=1}^{n} \frac{X_{1i}}{m}$
- $(X_{21}, \ldots, X_{2n}) \Rightarrow \bar{X_2} = \sum_{j=1}^{n} \frac{X_{2j}}{n}$

By the **De Moivere-Laplace theorem**, we have

$$\frac{\bar{X_1} - \theta_1}{\sqrt{\theta_1(1-\theta_1)/m}} \xrightarrow{d} N(0,1) \qquad \frac{\bar{X_2} - \theta_2}{\sqrt{\theta_2(1-\theta_2)/n}} \xrightarrow{d} N(0,1)$$

and using the independence between the two samples, we get

$$\frac{(\bar{X_1} - \bar{X_2}) - (\theta_1 - \theta_2)}{\sqrt{\frac{\theta_1(1-\theta_1)}{m} + \frac{\theta_2(1-\theta_2)}{n}}} \xrightarrow{d} N(0,1)$$

Proof.

$$\begin{aligned} X_m^* &\to N(0, \sigma^2) \\ Y_n^* &\to N(0, \delta^2) \end{aligned}$$

Then,

$$\begin{aligned} Z_{m,n} &= \frac{(X_m^* + Y_n^*) - (\mu - \theta)}{\underbrace{\sqrt{\frac{\sigma^2}{m} + \frac{\delta^2}{n}}}_{\sqrt{A}}} \\ &= \frac{(X_m - \mu) - (Y_n - \theta)}{\sqrt{A}} \\ &= \frac{X_m - \mu}{\sqrt{\sigma^2/m}} \times \underbrace{\frac{\sqrt{\sigma^2/m}}{\sqrt{A}}}_{\to\sqrt{c}} - \frac{Y_n - \theta}{\sqrt{\delta^2/n}} \times \underbrace{\frac{\sqrt{\delta^2/n}}{\sqrt{A}}}_{=\sqrt{1-\frac{\sigma^2/m}{A}} \to \sqrt{1-c}} \end{aligned}$$

Using the moment generating functions, we have

$$\begin{aligned}
M_{Z_{m,n}}(S) &= E\left[e^{SZ_{m,n}}\right] \\
&= E\left[e^{S\frac{X_m-\mu}{\sqrt{\sigma^2/m}}\frac{\sqrt{\sigma^2/m}}{\sqrt{A}} - S\frac{Y_n-\theta}{\sqrt{\delta^2/n}}\frac{\sqrt{\delta^2/n}}{\sqrt{A}}}\right] \\
&= M_{\frac{X_m-\mu}{\sqrt{\sigma^2/m}}}\left(S\frac{\sqrt{\sigma^2/m}}{\sqrt{A}}\right) M_{\frac{Y_n-\theta}{\sqrt{\delta^2/n}}}\left(S\frac{\sqrt{\delta^2/n}}{\sqrt{A}}\right) \\
&= \underbrace{M_{\sqrt{n}\frac{X_m-\mu}{\sqrt{\sigma^2/m}}}\left(S\frac{\sqrt{\sigma^2/m}}{\sqrt{A}}\right)}_{\to e^{S^2c/2}} \underbrace{M_{\sqrt{n}\frac{Y_n-\theta}{\sqrt{\delta^2/n}}}\left(S\frac{\sqrt{\delta^2/n}}{\sqrt{A}}\right)}_{\to e^{S^2(1-c)/2}} \\
&\to e^{S^2c/2+S^2/2-S^2c/2} = e^{S^2/2}
\end{aligned}$$

which is the mgf of a standard normal distribution.

Using the **Lévy theorem**, we conclude that $Z_{m,n} \xrightarrow{d} N(0,1)$.

■

Example 2.4.6 (Bernoulli : observation of large proportion of bad clients)
$\theta_1 = 0.05, \theta_2 = 0.06, m = 400, n = 500$

$$\begin{aligned}
P(\bar{X_1} - \bar{X_2} > 0) &= P\left(\frac{(\bar{X_1} - \bar{X_2}) - (\theta_1 - \theta_2)}{\sqrt{\frac{\theta_1(1-\theta_1)}{m} + \frac{\theta_2(1-\theta_2)}{n}}} > \frac{0 - (-0.01)}{\sqrt{\frac{0.05\times0.95}{400} + \frac{0.06\times0.94}{500}}}\right) \\
&= 1 - \Phi(0.66) \approx 0.2546
\end{aligned}$$

Note:
Care must be taken when extrapolating conclusions from the samples to the whole universe.

Gamma population

If $X \sim G(\alpha, \lambda)$, then for $\alpha, \lambda > 0$,

$$f(x|\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x > 0$$

where α is the shape parameter and λ is the rate parameter. Sometimes, the scale parameter $\beta = \frac{1}{\lambda}$ is used instead of the rate parameter.

Couple of special cases regarding the Gamma distribution:

- If $\alpha \in \mathbb{N}$, it is also known as the Erlang distribution.
- If $\alpha = 1$, it reduces to the exponential distribution with parameter λ.
- If $\alpha = \frac{n}{2}$ and $\lambda = \frac{1}{2}$, then $X \sim \chi^2_n$

Other special properties of the Gamma distribution:

- If $X_1 \sim G(\alpha_1, \lambda)$ is independent of $X_2 \sim G(\alpha_2, \lambda)$, then $X_1 + X_2 \sim G(\alpha_1 + \alpha_2, \lambda)$
- If $c > 0$ and $X \sim G(\alpha, \lambda)$, then $cX \sim G(\alpha, \frac{\lambda}{c})$
- If $X \sim G(\alpha, \lambda)$, then $2\lambda X \sim G(\alpha, \frac{1}{2}) = \chi^2_{2\alpha}$

Proof.

$$X \sim G(\alpha, \lambda)$$

Using the property $cX \sim G(\alpha, \frac{\lambda}{c})$ with $c = 2\lambda$, we have

$$2\lambda X \sim G\left(\alpha, \frac{\lambda}{2\lambda}\right) = G\left(\alpha, \frac{1}{2}\right)$$

With some algebraic manipulation and using the property $G(\frac{n}{2}, \frac{1}{2}) = \chi^2_n$, we get

$$G\left(\frac{2\alpha}{2}, \frac{1}{2}\right) = \chi^2_{2\alpha}$$

■

Another useful result is often used:

$$X \sim \chi^2_n \Rightarrow \sqrt{2X} - \sqrt{2n} \xrightarrow{d} N(0,1)$$

where the square root is used to make the χ^2_1 distribution more symmetric.

Proof. χ^2 are additive if independent, $X \sim \sum_{i=1}^{n} X_i$ where $X_i \sim \chi^2_1$ with $E[X_i] = 1$ and $Var(X_i) = 2$.

Using the **CLT**, we have

$$\frac{X - n\,E[X_i]}{\sqrt{n\,Var(X_i)}} \xrightarrow{d} N(0,1) \Leftrightarrow \frac{X - n}{\sqrt{2n}} \xrightarrow{d} N(0,1)$$

Now doing some algebraic manipulation,

$$\frac{\frac{X}{n} - 1}{\sqrt{2}\frac{\sqrt{n}}{n}} = \frac{\frac{X}{n} - 1}{\frac{\sqrt{2}}{\sqrt{n}}} = \sqrt{n}\frac{\frac{X}{n} - 1}{\sqrt{2}} \xrightarrow{d} N(0,1)$$

Using **Slutsky's**, we have

$$\sqrt{n}\left(\frac{X}{n} - 1\right) \xrightarrow{d} N(0,2)$$

Now applying the **Delta method** with $g(x) = \sqrt{x}$ and $g'(x) = \frac{1}{2}x^{-\frac{1}{2}}$,

$$\sqrt{n}\left(\sqrt{\frac{X}{n}} - 1\right) \xrightarrow{d} N\left(0, \underbrace{\left(\frac{1}{2}\right)^2}_{g'(1)} \underbrace{2}_{\sigma^2}\right) \Leftrightarrow \sqrt{X} - \sqrt{n} \xrightarrow{d} N\left(0, \frac{1}{2}\right)$$

Now dividing by $\sqrt{\frac{1}{2}}$ on both sides and using **Slutsky's** again, we get

$$\frac{\sqrt{X} - \sqrt{n}}{\sqrt{\frac{1}{2}}} = \sqrt{2X} - \sqrt{2n} \xrightarrow{d} N(0,1)$$

■

Another useful property of the Gamma distribution. Let $X_1, \ldots, X_n$ be a random sample of size n from a $G(\alpha, \lambda)$ population. Then,

$$\sum_{i=1}^{n} X_i \sim G(n\alpha, \lambda) \Leftrightarrow \bar{X} \sim G(n\alpha, n\lambda) \Leftrightarrow 2n\lambda\bar{X} \sim \chi^2_{2n\alpha}$$

Chapter 3

Sufficiency and Information

3.1 Sufficiency

Consider a parametric model $\mathcal{F} = \{f(\cdot|\theta) : \theta \in \Theta\}$ with $X_1, \dots, X_n$ iid random sample of size n extracted from X. The goal is to use the information contained in $X_1, \dots, X_n$ to produce inferential statements about the unknown parameter θ. When we compute statistics, i.e. functions of the random sample, we are summarizing the information contained in the random sample.

Open Question 3.1.1: When can we be assured that in the process we are not losing any relevant information about the parameter?

When all information that is lost is spurious, i.e. irrelevant for θ.

Example 3.1.1 (Poisson)

$X|\lambda \sim Po(\lambda)$ where $\lambda > 0$. We observe a random sample of size $n = 2$ and we know that the observed value of the statistic $T = \sum_{i=1}^{n} X_i = 31$. What can we say about the random sample X_1, X_2?

We want the possibility of the sample given all the information that I have.

$$\begin{aligned}
\frac{P(X_1 = x_1, X_2 = x_2|\lambda)}{P(T = 31|\lambda)} &= \frac{\frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}}{\frac{e^{-2\lambda}(2\lambda)^{31}}{31!}} \\
&= \frac{31!}{x_1!x_2!} \times \frac{\lambda^{x_1+x_2}}{\lambda^{31}} \times \left(\frac{1}{2}\right)^{31} \\
&= \frac{31!}{(31-x_2)!x_2!} \times \left(\frac{1}{2}\right)^{x_2} \left(1-\frac{1}{2}\right)^{31-x_2} \\
\Rightarrow \quad X_2|T = 31 &\sim B(31, \frac{1}{2})
\end{aligned}$$

We don't need to know λ.

Definition 3.1.1: Sufficient statistics

We say that a statistic T is sufficient for $\mathcal{F}$ or for θ, if the conditional distribution of the random sample given the observed value of T does not depend on the unkown parameter θ for all θ.

What is essentially means is that $f(x_1, \ldots, x_n|\theta, t)$ does not depend on θ.

$$\begin{aligned} f(x_1, \ldots, x_n|\theta, t) &= \frac{f(x_1, \ldots, x_n, t|\theta)}{f_T(t, \theta)} \\ &= \begin{cases} \frac{f(x_1, \ldots, x_n|\theta)}{f_T(t, \theta)} & \text{if } T(x_1, \ldots, x_n) = t \\ 0 \quad \text{otherwise} \end{cases} \end{aligned}$$

Example 3.1.2 (Definition : Bernoulli)

$X|\theta \sim B(1, \theta)$, and $X_1, \ldots, X_n|\theta \overset{iid}{\sim} B(1, \theta)$

Intuitively, the sequence of success and failure should not matter, only the proportion is important. Thus, $T = \sum_{i=1}^{n} X_i$ should be sufficient for θ.

$$\begin{aligned} f(x_1, \ldots, x_n|\theta, t) &= \frac{f(x_1, \ldots, x_n|\theta)}{f_T(t|\theta)} \quad \text{if } T(x_1, \ldots, x_n) = t, \sum x_i = t \\ &= \frac{\prod_{i=1}^{n} f(x_i|\theta)}{f_T(t|\theta)} \end{aligned}$$

Because $T|\theta \sim B(n, \theta)$

$$\begin{aligned} f(x_1, \ldots, x_n|\theta, t) &= \frac{\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \quad \text{if } \sum x_i = t \\ &= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{1}{\binom{n}{t}} \end{aligned}$$

The definition of sufficient statistic is not very useful to discover sufficient statistics. Thus we need to rely on another theorem.

Theorem 3.1.1 Halmos-Savage Factorization Criterion

A statistic T is sufficient for θ if and only if there are non-negative functions g and h such that

- g depends on θ and on the random sample exclusively through the observed value of T. Meaning it depends on parameter and statistics.
- h depends exclusively on the random sample.
- $f(x_1, \ldots, x_n|\theta) = g(T(x_1, \ldots, x_n); \theta) \times h(x_1, \ldots, x_n)$

Proof. T is sufficient for $\theta \Leftrightarrow$ There are non-negative function of g and h.

We first start from left to right.

If T is sufficient for θ, then

$$f(x_1, \ldots, x_n|t, \theta) = \frac{f(x_1, \ldots, x_n|\theta)}{f_T(t|\theta)} \quad \text{does not depend on } \theta$$

meaning that

$$f(x_1, \ldots, x_n|t) = \frac{f(x_1, \ldots, x_n|\theta)}{f_T(t|\theta)} \quad \text{if } T(x_1, \ldots, x_n) = t$$

so we get

$$f(x_1,\dots,x_n|\theta) = \underbrace{f_T(t|\theta)}_{g(t,\theta)}\underbrace{f(x_1,\dots,x_n|t)}_{h(x_1,\dots,x_n)}$$

Now going from right to left.

For X discrete, T is discrete. And for simplification, $\tilde{x} = (x_1,\dots,x_n)$.

$$\begin{aligned} f_T(t|\theta) = P(T=t|\theta) &= \sum_{\tilde{x}:T(\tilde{x})=t} f(\tilde{x}|\theta) \\ &= \sum_{\tilde{x}:T(\tilde{x}=t)} g(t;\theta)h(\tilde{x}) \\ &= g(t;\theta)\sum_{\tilde{x}:T(\tilde{x})=t} h(\tilde{x}) \end{aligned}$$

we now have

$$\begin{aligned} f(x_1,\dots,x_n|t,\theta) &= \frac{f(x_1,\dots,x_n|\theta)}{f_T(t|\theta)} \quad \text{if } T(\tilde{x}) = t \\ &= \frac{g(t,\theta)h(\tilde{x})}{g(t,\theta)\sum_{\tilde{x}:T(\tilde{x})=t} h(\tilde{x})} \\ &= \frac{h(\tilde{x})}{\sum_{\tilde{x}:T(\tilde{x})=t} h(\tilde{x})} \quad \text{does not depend on } \theta \end{aligned}$$

■

Example 3.1.3 (Factorization criterion : Poisson)
Let $X|\lambda \sim Po(\lambda)$

$$\begin{aligned} f(x_1,\dots,x_n|\lambda) = \prod_{i=1}^n f(x_1|\lambda) &= \prod_{i=1}^n e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} \\ &= e^{-n\lambda}\frac{\lambda^{\sum x_i}}{\prod x_i!} \\ &= \underbrace{e^{-n\lambda}\lambda^{\sum x_i}}_{g(\sum x_i;\lambda)}\underbrace{\frac{1}{\prod x_i!}}_{h(\tilde{x})} \end{aligned}$$

Example 3.1.4 (Factorization criterion : Uniform)
Let $X|\theta \sim U(0,\theta)$, $\theta > 0$.

$$\begin{aligned} f(x|\theta) &= \frac{1}{\theta} \quad , 0 < x < \theta \\ &= \frac{1}{\theta} I_{(0,\theta)}(x) \end{aligned}$$

with $I_A(x) = \begin{cases} 1 & , x \in A \\ 0 & , \text{otherwise} \end{cases}$

$$\begin{aligned} f(\tilde{x}|\theta) = \prod_{i=1}^n f(x_i|\theta) &= \prod_{i=1}^n \frac{1}{\theta} I_{(0,\theta)}(x_i) \\ &= \theta^{-n} I_{(0,\infty)}(x_{(1)})\, I_{(0,\theta)}(x_{(n)}) \\ &= \underbrace{\theta^{-n} I_{(0,\theta)}(x_{(n)})}_{g(x_{(n)})}\underbrace{I_{(0,\infty)}(x_{(1)})}_{h(\tilde{x})} \end{aligned}$$

Example 3.1.5 (Factorization criterion : Shifted exponential)

Consider the shifted exponential distribution $f(x|\lambda,\delta) = \lambda e^{-\lambda(x-\delta)} I_{[\delta,\infty)}(x)$

$$f(\tilde{x}|\lambda,\delta) = \prod_{i=1}^{n} f(x_i|\lambda,\delta) = \prod_{i=1}^{n} \lambda e^{-\lambda(x_i-\delta)} I_{[\delta,\infty)}(x_i)$$
$$= \underbrace{\lambda^n e^{-\lambda\sum x_i} e^{n\lambda\delta} I_{[\delta,\infty)}(x_{(1)})}_{g(\sum x_i, x_{(1)};\lambda,\delta)} \times \underbrace{1}_{h(\tilde{x})}$$

By the factorization criterion, $(\sum x_i, x_{(1)})$ is sufficient for (λ,δ).

Two of things to notice regarding sufficient statistics.

- There is always a sufficient statistic, it can be all the samples.
- Sufficient statistics are not unique.

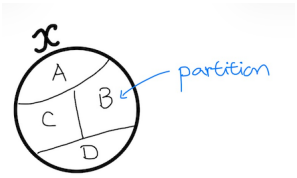
Example 3.1.6 (Un-uniqueness of sufficient statistics : Poisson)

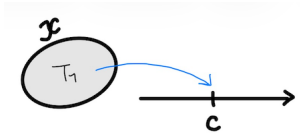
$X \sim Po(\lambda)$ where $t = \sum x_i$

- $f(\tilde{x}|\lambda) = e^{-n\lambda}\lambda^t \frac{1}{\prod x_i!} \Rightarrow T = \sum x_i$ is sufficient.
- $f(\tilde{x}|\lambda) = e^{-n\lambda}\lambda^{n\bar{x}} \frac{1}{\prod x_i!} \Rightarrow T = \bar{x}$ is sufficient.
- $f(\tilde{x}|\lambda) = e^{-n\lambda}\lambda^{T_1}\lambda^{T_2} \frac{1}{\prod x_i!} \Rightarrow (T_1, T_2)$ is sufficient where $T_1 = \sum_{i=1}^{n_1} x_i$ and $T_2 = \sum_{i=n_1+1}^{n} x_i$.

Another important concept is the **partition** induced in the sample space by a statistic.

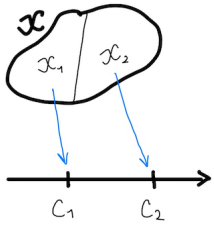
- Any statistics $T : \mathcal{X} \to \mathbb{R}^q$ induces a partition in the sample space $\mathcal{X}$ and the partitions do not intercept.





$$A + B + C + D = \mathcal{X} \Rightarrow T_1(\tilde{x}) = C \to \Pi = \{\mathcal{X}\}$$

If a statistic takes 2 values : $T_1(\tilde{x}) = \begin{cases} c_1 & , \tilde{x} \in \mathcal{X}_1 \\ c_2 & , \tilde{x} \in \mathcal{X}_2 \end{cases} \to \begin{cases} \mathcal{X}_1 \cup \mathcal{X}_2 = \mathcal{X} \\ \mathcal{X}_1 \cap \mathcal{X}_2 = \varnothing \\ \Pi = \{\mathcal{X}_1, \mathcal{X}_2\} \end{cases}$



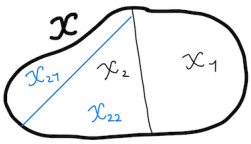
- The finer the partition induced by T in $\mathcal{X}$, the less information is lost; the smaller is the data reduction operated by T.
- A sufficient statistic operates a data reduction that does not involve loss of relevant information about the parameter; the partition it induces is also said to be sufficient.

Note:

When the partition is finer as it can be, $T(\tilde{x}) = \tilde{x}$, $\Pi = \cup_{\tilde{x}\in\mathcal{X}}\{\tilde{x}\}$, because no information will be lost.

- The notion of partition is more general than that of statistic; different statistics can induce the same partition, in which case they are said to be equivalent, i.e. they are one-to-one.
- If the partition induced by T is finer than the partition induced by S, then S is a function of T. In that case, if S is sufficient, so is T, that is

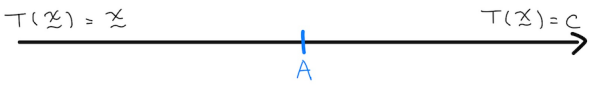
$$\left.\begin{array}{l} S = h(T) \\ S \text{ sufficient} \end{array}\right\} \Rightarrow T \text{ sufficient}$$



S is a function of T

$\begin{cases} S \to \{\mathcal{X}_1, \mathcal{X}_2\} \\ T \to \{\mathcal{X}_1, \mathcal{X}_{21}, \mathcal{X}_{22}\} \end{cases}$ If S is sufficient, then T is sufficient, because T is finer than S partition wise.

- However, if T is sufficient and $S = h(T)$, it is not a given that S is also sufficient unless S is injective, in which case, S and T are equivalent.
- We are interested in finding statistics which are sufficient but operates the least data reduction. In other words, statistics that induce the coarsest partition that is still sufficient.



Point A is what we want to find, the minimum sufficient statistic. We lose information as we are moving to the right hand side.

Definition 3.1.2: Minimal sufficient statistic

A statistic is said to be minimal sufficient for $\mathcal{F}$ if it is sufficient, and if S is any other sufficient statistic, then $T = h(S)$ for some h.

Consider a binary relation in $\mathcal{X}$ defined by $\begin{cases} \tilde{y} \in \mathcal{X} \\ \tilde{x} \in \mathcal{X} \end{cases}$, $\tilde{y}\, \boldsymbol{R}\, \tilde{x}$.

$$f(\tilde{y}|\theta) = c(\tilde{x}, \tilde{y}) f(\tilde{x}|\theta) \Leftrightarrow \underbrace{\frac{f(\tilde{y}|\theta)}{f(\tilde{x}|\theta)}}_{\text{likelihood ratio}} = c(\tilde{x}, \tilde{y})$$

where $c(\tilde{x}, \tilde{y}) > 0$ and does not depend on θ.

This binary relation is an equivalence relation, that is, it is

- symmetric: $\tilde{x}\, \boldsymbol{R}\, \tilde{y} \Leftrightarrow \tilde{y}\, \boldsymbol{R}\, \tilde{x}$
- reflexive: $\tilde{x}\, \boldsymbol{R}\, \tilde{x}$
- transitive: $\tilde{x}\, \boldsymbol{R}\, \tilde{y}$ and $\tilde{y}\, \boldsymbol{R}\, \tilde{z} \Rightarrow \tilde{x}\, \boldsymbol{R}\, \tilde{z}$

Hence, it induces a partition in $\mathcal{X}$ with parts

$$\Pi_x = \{y \in \mathcal{X} : y \, \boldsymbol{R} \, x\} \text{ for } x : f(x|\theta) > 0 \text{ for some } \theta \in \Theta$$
$$\Pi_0 = \{y \in \mathcal{X} : f(y|\theta) = 0 \quad \forall \theta \in \Theta\}$$

Proof. Transitive

Consider the case where

$$\tilde{x} \, \boldsymbol{R} \, \tilde{y} \Rightarrow f(\tilde{y}|\theta) = c_1 f(\tilde{x}|\theta)$$
$$\tilde{y} \, \boldsymbol{R} \, \tilde{z} \Rightarrow f(\tilde{y}|\theta) = c_2 f(\tilde{z}|\theta)$$

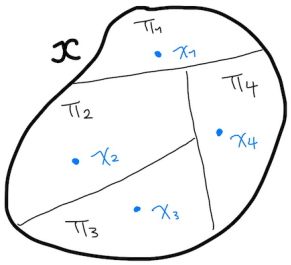
then we have

$$c_1 f(\tilde{x}|\theta) = c_2 f(\tilde{z}|\theta) \Rightarrow f(\tilde{x}|\theta) = \frac{c_2}{c_1} f(\tilde{z}|\theta) \Rightarrow \tilde{x} \, \boldsymbol{R} \, \tilde{z}$$

■

Theorem 3.1.2 Lehmann-Scheffe

The partition with parts Π_0 and $\{\Pi_x\}$ described above is minimal sufficient, and any statistic which induces it is minimal sufficient.



$$\Pi = \{\Pi_1, \Pi_2, \Pi_3, \Pi_4\}, G(\tilde{x}) = \tilde{x}_i \text{ if } \tilde{x} \in \Pi_i$$

Proof. Consider the statistic G that to each $x \in \mathcal{X}$ associates a representative of the element of the partition to which it belongs

$$x \in \mathcal{X} \mapsto x_\Pi \in \Pi_x = G(x)$$

Proof that G is sufficient

$$\tilde{x} \in \Pi_i \Rightarrow \tilde{x} \, \boldsymbol{R} \, \tilde{x}_i$$
$$f(\tilde{x}|\theta) = c(\tilde{x}_i, \tilde{x}) f(\tilde{x}|\theta) \underset{G(\tilde{x})=\tilde{x}_i}{=} \underbrace{c(\tilde{x}_i, \tilde{x})}_{g(\tilde{x})} \underbrace{f(G(\tilde{x}|\theta))}_{g(G;\theta)}$$
$$\Rightarrow G \text{ is sufficient}$$

Now **proof that G is a "grosser" partition**

$$U \text{ is sufficient}$$
$$\Pi_U = \{\Pi_x^*\}, \quad \left.\begin{matrix} \tilde{x} \in \mathcal{X} \\ \tilde{y} \in \Pi_x^* \end{matrix}\right\} \rightarrow \tilde{x} \in \Pi_i$$
$$\text{since } U \text{ is sufficient,}$$
$$f(\tilde{x}|\theta) = g(U(\tilde{x}); \theta) h(\tilde{x})$$
$$f(\tilde{y}|\theta) = g(U(\tilde{y}); \theta) h(\tilde{y}) = g(U(\tilde{x}); \theta) h(\tilde{y}) \Leftrightarrow$$
$$h(\tilde{x}) f(\tilde{y}|\theta) = \underbrace{h(\tilde{x}) g(U(\tilde{x}); \theta)}_{f(\tilde{x}|\theta)} h(\tilde{y}) \Leftrightarrow$$
$$f(\tilde{y}|\theta) = f(\tilde{x}|\theta) \frac{h(\tilde{y}|\theta)}{h(\tilde{x}|\theta)} \Rightarrow \tilde{y} \, \boldsymbol{R} \, \tilde{x} \Rightarrow U \text{ is a finer partition than } G$$
$$\therefore G \text{ is the minimal sufficient statistic}$$

■

Example 3.1.7 (Minimal sufficient statistic : Normal)

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population and let $\theta = (\mu, \sigma^2)$

$$\begin{aligned} f(\tilde{x}|\theta) &= \prod_{i=1}^{n} f(x_i|\mu\sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\} \\ &= (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}(\sum x_i^2 - 2n\bar{x}\mu + n\mu^2)\} \\ &= (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}\sum x_i^2\} \exp\{\frac{n\bar{x}\mu}{\sigma^2}\} \exp\{-\frac{1}{2\sigma^2}n\mu^2\} \end{aligned}$$

Now using the **Lehmann-Scheffe theorem**,

$$\begin{aligned} \frac{f(\tilde{y}|\theta)}{f(\tilde{x}|\theta)} &= \frac{(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}\sum y_i^2\} \exp\{\frac{n\bar{x}\mu}{\sigma^2}\} \exp\{-\frac{1}{2\sigma^2}n\mu^2\}}{(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}\sum x_i^2\} \exp\{\frac{n\bar{x}\mu}{\sigma^2}\} \exp\{-\frac{1}{2\sigma^2}n\mu^2\}} \\ &= \exp\{-\frac{1}{2\sigma^2}(\sum y_i^2 - \sum x_i^2)\} \exp\{\frac{n\mu}{\sigma^2}(\bar{y}-\bar{x})\} \quad \text{does not depend on } \theta \\ &\Leftrightarrow \begin{cases} \sum y_i^2 = \sum x_i^2 \\ \bar{y} = \bar{x} \end{cases} \end{aligned}$$

A minimal sufficient statistic is $(\bar{x}, \sum x_i^2)$ or $(\bar{x}, s^2)$

Example 3.1.8 (Minimal sufficient statistic : Exponential)

Let $X_1, \ldots, X_n$ be a random sample from a $Ex(\lambda)$ population and let $\theta = \lambda$

$$f(x|\lambda) = \lambda e^{-\lambda n}, n > 0$$

$$f(\tilde{x}|\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda n} = \lambda^n \exp\{-\lambda \sum x_i\}$$

$$\frac{f(\tilde{y}|\theta)}{f(\tilde{x}|\theta)} = \frac{\lambda^n \exp\{-\lambda \sum y_i\}}{\lambda^n \exp\{-\lambda \sum x_i\}} = \exp\{-\lambda(\sum y_i - \sum x_i)\}$$

Does not depend on λ if and only if $\sum y_i = \sum x_i \Leftrightarrow \bar{x} = \bar{y} \Rightarrow \bar{x}$ is minimal sufficient

3.2 Ancilarity and completeness

Definition 3.2.1: Ancillary statistic

A statistic T is said to be ancillary if its sampling distribution does not depend on the unknown parameter θ.

T is ancillary if $f_T(t|\theta)$ does not depend on θ.

Definition 3.2.2: Location-scale family of distributions

The location-scale family of distribution is composed by all the probability distributions such that the associated comulative distribution function is of the form

$$G(x|\delta,\lambda) = G\left(\frac{x-\lambda}{\delta}\right)$$

where G is a function that does not involve unknown parameters, $\lambda \in \mathbb{R}$ is the location parameter, and $\delta > 0$ is the scale parameter.

This family includes the location family (δ is known) and the scale family (λ is known).

Example 3.2.1 (Location-scale family : Normal)

Suppose that $X \sim N(\mu, \sigma^2)$

$$\frac{x-\mu}{\sigma} \sim N(0,1)$$

$$\begin{aligned} f(x|\mu,\sigma^2) &= P(X \leq x|\mu\sigma^2) \\ &= P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}|\mu,\sigma^2\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

where $G = \Phi, \lambda = \mu, \delta = \sigma$.

The normal family of distributions is a member of the location-scale family with location parameter μ and scale parameter δ.

Example 3.2.2 (Location-scale family : Uniform)

Suppose that $X \sim U(0,\theta)$, $\theta > 0$, then we have

$$f(x|\theta) = \frac{1}{\theta} I_{(0,\theta)}(x)$$

$$\begin{aligned} F(x|\theta) &= \frac{x}{\theta} \underbrace{I_{(0,\theta)}(x)}_{0<x<\theta \Leftrightarrow 0<\frac{x}{\theta}<1} \\ &= \frac{x}{\theta} I_{(0,1)}(\frac{x}{\theta}) \\ &= G\left(\frac{x}{\theta}\right) \end{aligned}$$

where $G(y) = y\, I_{(0,1)}(y)$

$U(0,\theta)$ belongs to the scale family with scale parameter $\delta = \theta$ and $\lambda = 0$.

Couple **remarks** regarding the location-scale family

- The distribution of X is part of the location-scale family with location parameter λ and scale parameter δ if and only if the distribution of $\frac{(x-\lambda)}{\delta}$ does not depend on unknown parameters.
- If the distribution of X is a member of the location-scale family with location paratmer λ and scale parameter δ then

any statistic which is a function of $X_1, \ldots, X_n$ only through the vector $\left(\frac{X_i-\lambda}{\delta}, i = 1, \ldots, n\right)$ is ancillary.

$$T = T(X_1, \ldots, X_n) = H\left(\frac{X_1-\lambda}{\delta}, \ldots, \frac{X_n-\lambda}{\delta}\right), \quad T \text{ is ancillary}$$

Proof. X belongs to the location-scale family

$$\Leftrightarrow F(X) = P(X \le x) = G\left(\frac{X-\lambda}{\delta}\right)$$

$$\Rightarrow P\left(\frac{X-\lambda}{\delta} \le x\right) = P(X \le \delta x + \lambda) = G\left(\frac{\delta x + \lambda - \lambda}{\delta}\right) = G(x) \rightarrow \text{ doesn't depend on } (\lambda, \delta)$$

Let G be the CDF of $Y = \frac{X-\lambda}{\delta} \Leftrightarrow X = \delta Y + \lambda$

$$P(X \le x) = P(\delta Y + \lambda \le x) = P\left(Y \le \frac{x-\lambda}{\delta}\right) = G\left(\frac{x-\lambda}{\delta}\right)$$

■

Example 3.2.3 (Location-scale and statistic : Uniform)
Let $X|\theta \sim U(\theta - \frac{1}{2}, \theta + \frac{1}{2}), \theta \in \mathbb{R}$

$$f(x|\theta) = \frac{1}{1} \underbrace{I_{(\theta-\frac{1}{2}, \theta+\frac{1}{2})}(x)}_{\theta-\frac{1}{2}<x<\theta+\frac{1}{2} \Leftrightarrow -\frac{1}{2}<x-\theta<\frac{1}{2}} = I_{(-\frac{1}{2}, \frac{1}{2})}(x-\theta) = G(x-\theta)$$

where $G(y) = I_{(-\frac{1}{2}, \frac{1}{2})}(y)$

The distribution of X belongs to the location family with $\lambda = \theta$ and $\delta = 1$. The distribution of $x - \theta$ does not depend on unknwon parameters.

$$R = X_{(n)} - X_{(1)} = (X_{(n)} - \theta) - (X_{(1)} - \theta) = H(X_1 - \theta_1, \ldots, X_n - \theta_n)$$

- It would be natural to expect a minimal sufficient statistic and an ancillary statistic to be indepenendent. However, that is not the case in general.

Example 3.2.4 (Minimal sufficient statistic and ancillary statistic : Uniform)
Let $X|\theta \sim U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$

$$\frac{f(\tilde{y}|\theta)}{f(\tilde{x}|\theta)} = \frac{I_{(y_{(n)}-\frac{1}{2}, \infty)}(\theta) I_{(-\infty, y_{(1)}+\frac{1}{2})}(\theta)}{I_{(x_{(n)}-\frac{1}{2}, \infty)}(\theta) I_{(-\infty, x_{(1)}+\frac{1}{2})}(\theta)}$$

does not depend on θ if and only if $\begin{cases} x_{(n)} = y_{(n)} \\ x_{(1)} = y_{(1)} \end{cases} \Rightarrow (x_{(1)}, x_{(n)})$ is minimal sufficient, $(x_{(n)} - x_{(1)}, x_{(1)})$ is minimal sufficient.

$(R, X_{(1)})$ is minimal sufficient but contains an element that is ancillary.

Definition 3.2.3: Complete statistic

A statistic T is said to be complete if and only if

$$E[h(T)|\theta] = 0 \ \forall \theta \in \Theta \Rightarrow h(T) \equiv 0$$

When T is complete and $h_1(T)$ and $h_2(T)$ are two functions of T that have the same expected value, then it must be the case that they are the same, $h_1(T) = h_2(T)$.

Proof.

$$
\begin{aligned}
&E_\theta[h_1(T) - h_2(T)] = 0 \quad \forall \theta \in \Theta \\
&\Downarrow \\
&h_1(T) - h_2(T) = 0 \\
&\Updownarrow \\
&h_1(T) = h_2(T) \quad \text{must be the same}
\end{aligned}
$$

■

Theorem 3.2.1

Any statistic that is sufficient and complete is minimal sufficient.

Proof. Assume T is sufficient and complete.

$$
\begin{aligned}
&\exists T_1 \quad T_1 = g(T), \quad T_1 \text{ sufficient} \\
&h(T) = T - \underbrace{E[T|T_1]}_{\text{function of } T_1 \text{ because } T_1 = g(T)}
\end{aligned}
$$

Conditional expectation: $E[X|Y = y] = \int_X x f_{X|Y=y}(x)\, dx$ which is a function of y. This expectation could be a function of the parameter, need to make sure that it's not the case for $h(T)$ to be a statistic.

$$
E[T|T_1 = t_1] = \int_x T(\tilde{x}) f_{\tilde{x}|T_1=t_1,\theta}(\tilde{x})\, d\tilde{x}
$$

Since T_1 is sufficient, by definition, $f_{\tilde{x}|T_1=t_1,\theta}(\tilde{x})$ does not depend on $\theta \Rightarrow E[T|T_1]$ does not depend on θ. $\Rightarrow h(T)$ is a statistic.

$$
\begin{aligned}
E_\theta[h(T)] &= E_\theta[T] - \underbrace{E_\theta[E[T|T_1]]}_{\text{law of iterated expectation}} \\
&= E_\theta[T] - E_\theta[T] = 0
\end{aligned}
$$

Since T is complete, $h(T) = 0 \Leftrightarrow T = E[T|T_1] = f(T_1)$ ■

Wrong Concept 3.1: Converse

Let $X|\theta \sim U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$

$R = X_{(n)} - X_{(1)}$ which is ancillary $\rightarrow$ distribution does not depend on θ

$T(R, X_{(1)})$ is minimal sufficient but R has no information about θ so should throw away, but if you throw away then it is not minimal sufficient.

$$
h(T) = R - \underbrace{E[R]}_{=c \text{ because } R \text{ is ancillary}}
$$

$E[h(T)] = 0$ but $h(T) \neq 0 \Rightarrow T$ is not complete.

Turns out that sufficient and complete statistics operate a data reduction that is more effective than that operated by minimal sufficient statistics that are not complete. This is the subject of the Theorem of Basu.

Theorem 3.2.2 Basu's Theorem

Let T be a sufficient and complete statistic. Then T is independent of any ancillary statistic.

Proof. Assume

- T is sufficient and complete
- U is ancillary
- $P(U \in A|T) = h_A(T)$ because conditioning on T

$h_A(T)$ is a statistic and T is sufficient so does not depend on θ.

$$\begin{aligned} P(U \in A|T = t) &= \int_x 1_A(U(\tilde{x})) f_{\tilde{x}|T=t,\theta}(\tilde{x})\, d\tilde{x} \\ &= \int_{\{\tilde{x}:U(\tilde{x})\in A\}} U(\tilde{x}) f_{\tilde{x}|T=t,\theta}(\tilde{x})\, d\tilde{x} \quad \text{where } T = t, \theta \text{ does not depend on } \theta \text{ because } T \text{ sufficient} \end{aligned}$$

but $E[h_A(T)]$ could depend on θ but does not because

$$h_A(T) = P(U \in A|T) = E[1_A(U)|T] \longrightarrow \text{Bernoulli}$$

$$E[h_A(T)] = E[E[1_A(U)|T]] = E[1_A(U)] = \underbrace{P(U \in A)}_{\text{marginal probability}} = c \quad \text{does not depend on } \theta, \text{ so a constant}$$

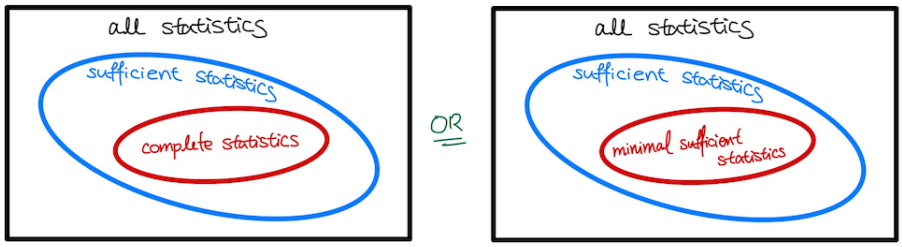
$$h_A(T) - c \quad \text{still a statistic because (statistic - constant) = statistic}$$

$$E[h_A(T) - c] = 0 \overset{T \text{ complete}}{\Longrightarrow} h_A(T) = c \overset{A \text{ arbitrary}}{\Longleftrightarrow} P(U \in A|T) = P(U \in A)$$

$\Rightarrow U$ and T are independent. ■

A couple **remarks**:

- A complete and sufficient statistic does not contain any ancillary information, that is, the data reduction operated by this type of statistics is more effective than that opearted by minimal sufficient statistics which are not complete.
- Either all the minimal sufficient statistics are complete or there are no sufficient and complete statistics.



3.3 Exponential family

Definition 3.3.1: Exponential family of distribution

We say that a random vector X is distributed according to a member of the k-parametric exponential family if its pdf or pmf can be expressed in the form

$$f(x|\theta) = c(\theta)\, h(x)\, \exp\left[\sum_{j=1}^{k} \omega_j(\theta)\, R_j(x)\right]$$

with support $\{x : f(x|\theta) > 0\}$ independent of $\theta = (\theta_1, \ldots, \theta_k)$. Also, $c(\theta) \geq 0, h(x) \geq 0$, and $R_j(x)$ are scalar functions of x.

The **canonical form** of a distribution which belongs to the k-parameter exponential family is obtained through the so-called natural parametrization, $\alpha_j = \omega_j(\theta)$, $j = 1, \ldots, k$

$$f(x|\alpha) = d(\alpha)\, h(x)\, \exp\left[\sum_{j=1}^{k} \alpha_j\, R_j(x)\right]$$

where $\alpha = (\alpha_1, \ldots, \alpha_k) \in A$ is called the natural parameter and A being called the natural parameteric space.

Example 3.3.1 (Exponential family : Normal)

$$\begin{aligned} f(X|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(X-\mu)^2\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(X^2 - 2\mu X - \mu^2)\right\} \\ &= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-\mu X}{2\sigma^2}\right\}}_{c(\theta)} \exp\left\{\underbrace{\frac{-1}{2\sigma^2}}_{\omega_1(\theta)} \underbrace{X^2}_{R_1(X)} + \underbrace{\frac{\mu}{\sigma^2}}_{\omega_2(\theta)} \underbrace{X}_{R_2(X)}\right\} \end{aligned}$$

$\Rightarrow$ The Normal distribution belongs to the 2-parameters exponential family.

Natural parameterization of the Normal distribution is

$$\alpha_1 = \frac{-1}{2\sigma^2} \quad \alpha_2 = \frac{\mu}{\sigma^2}$$

$(\alpha_1, \alpha_2) \in \mathbb{R} \times \mathbb{R} \rightarrow$ the natural parameter space

Example 3.3.2 (Exponential family : Poisson)

$$\begin{aligned} f(x|\theta) &= e^{-\theta}\frac{\theta^x}{x!} = e^{-\theta}\frac{1}{x!}\exp(\ln\theta^x) \\ &= \underbrace{e^{-\theta}}_{c(\theta)}\ \underbrace{\frac{1}{x!}}_{h(x)}\ \exp\left(\underbrace{\ln\theta}_{\omega_1(\theta)}\ \underbrace{x}_{R_1(x)}\right) \end{aligned}$$

$\Rightarrow$ The Poisson distribution is a member of the 1-parameter exponential family.

Theorem 3.3.1

The k-parameter exponential family structure is preserved under iid random sampling and there is a k-dimentional sufficient statistic regardless of the sample size.

Proof. $X_1, \ldots, X_n$ random sample from a model that is a member of the k-parameter exponential family.

$$f(\tilde{x}|\tilde{\theta}) = \prod_{i=1}^{n} f(x_i|\tilde{\theta}) = \prod_{i=1}^{n} c(\tilde{\theta})h(x_i)\exp(\sum_{j=1}^{k} w_j(\theta)R_j(x_i))$$

$$= [c(\tilde{\theta})]^n \prod_{i=1}^{n} h(x_i)\exp(\sum_{i=1}^{n}\sum_{j=1}^{k} w_j(\theta)R_j(x_i))$$

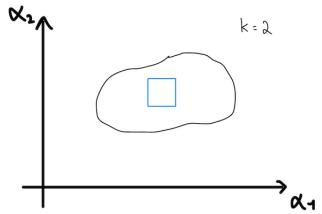
$$= c^*(\theta)h^*(\tilde{x})\exp(\underbrace{\sum_{j=1}^{k} w_j(\theta)}_{w_j(\theta)}\underbrace{\sum_{i=1}^{n} R_j(x_i)}_{R^*(\tilde{x})=T_j(\tilde{x})})$$

$$= \underbrace{c^*(\theta)\exp(\sum_{j=1}^{k} w_j(\theta)T_j(\tilde{x}))}_{g(\tilde{T};\theta)}\underbrace{h^*(\tilde{x})}_{h^*(\tilde{x})}$$

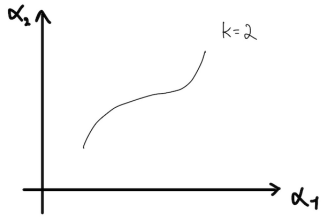
$T = (T_1, \ldots, T_k)$ is a sufficient statistic by the factorization criterion. ■

Theorem 3.3.2

The sufficient statistic $T = (T_1(X), \ldots, T_k(X))$ just described is complete if the natural parametric space A contains an open subset of $\mathbb{R}^k$



$\Rightarrow$ Contains an open subset of $\mathbb{R}^k$

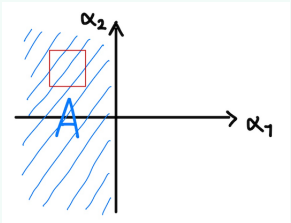


$\Rightarrow$ Doesn't contain an open subset of $\mathbb{R}^k$

Example 3.3.3 (Sufficient, complete, and minimal sufficient : Normal)

$$\alpha_1 = \frac{-1}{\sigma^2} \in \mathbb{R}_-$$

$$\alpha_2 = \frac{\mu}{\sigma^2} \in \mathbb{R}$$



$\Rightarrow (T_1, T_2)$ is complete it was already sufficient $\Rightarrow (T_1, T_2)$ is sufficient and complete $\Rightarrow$ minimal sufficient

Example 3.3.4 (Basu : $\bar{X}, S^2$ independet in Normal)

Begin by assuming that the variance is known, $\sigma^2 = \sigma_0^2$

$N(\mu, \sigma_0^2)$ belongs to the 1-parameter exponential family

$$\exp\left\{ \underbrace{\frac{-1}{2\sigma_0^2}x^2}_{(\text{all known})x^2} + \frac{\mu}{\sigma_0^2}x \right\}$$

$$\alpha_1 = \frac{\mu}{\sigma_0^2} \quad T_1 = \sum x_i$$

$\Rightarrow T_1$ is sufficient and complete $\Rightarrow \bar{X}$ is sufficient and complete

And since σ^2 is known, the Normal distribution is also part of the location family with parameter μ

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}((x_i - \mu)(\bar{x} - \mu))^2 = \frac{1}{n}\sum_{i=1}^{n}[(x_i - \mu) - \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)]^2 = (\cdots)$$

Essentially, $S^2 = H(x_i - \mu,\ i = 1, \ldots, n) \Rightarrow S^2$ is ancillary

$$P(\bar{X} \leq t, S^2 \leq s|\mu, \sigma_0^2) = P(\bar{X} \leq t|\mu, \sigma_0^2)P(S^2 \leq s|\mu, \sigma_0^2) \quad \forall \sigma_0^2 \rightarrow \text{ arbitrary}$$

$\Rightarrow \bar{X}$ and S^2 are independent in the Normal 2-parameter model

Claim 3.3.1

When we restrict a model, sufficiency is maintained, completeness not always.

Example 3.3.5 (Restricing a model : Normal)

Suppose that $X \sim N(\mu, \sigma^2)$ but consider a sub-family $\mu = \sigma^2$

Clearly, $T = (S^2, \bar{X})$ is still sufficient for this sub-family. However, the natural parameter space no longer contains an open subset of $\mathbb{R}^2$ so that the previous theorem (3.3.2) does not guarantee the completeness of T.

It is easy to see that T is in this setting not complete as $\dfrac{\bar{X} - nS^2}{(n-1)}$ is a function of T which has zero expected value but is obviously not null.

3.4 Fisher information

Definition 3.4.1: Likelihood function

When we observe the sample $(x_1, \ldots, x_n)$, the observed value of the random sample $X_1, \ldots, X_n$ with joint pdf or pmf function $f(x_1, \ldots, x_n)$, the corresponding likelihood function is a function of θ given by

$$L(\theta|x_1, \ldots, x_n) = f(x_1, \ldots, x_n|\theta)$$

Essentially it shows how likely it is to observe the sample given the parameter. And since $x_1, \ldots, x_n$ is fixed, it's usual to write $L(\theta)$ instead of $L(\theta|x_1, \ldots, x_n)$. Note that L is a function of θ and not of $x_1, \ldots, x_n$. If $X_1, \ldots, X_n$ **is an iid**

random sample, we have

$$L(\theta|x_1,\ldots,x_n) = f(x_1,\ldots,x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

The likelihood function is not a probability mass or density function. Hence it has no natural scale associated. In fact, it is more approriately defined as

$$L(\theta|x_1\ldots,x_n) \propto f(x_1,\ldots,x_n|\theta)$$

that is, up to a multiplying constant. This multiplying constant can depend on $x_1,\ldots,x_n$ but not on θ.

Example 3.4.1 (Proportional : Poisson)

$$\begin{aligned} f(x) &\propto \frac{\lambda^x}{x!} = c\frac{\lambda^x}{x!} \\ \sum_{x=0}^{\infty} = 1 &\Leftrightarrow \sum_{x=0}^{\infty} c\frac{\lambda^x}{x!} = 1 \\ &\Leftrightarrow c\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = 1 \\ &\Leftrightarrow ce^{\lambda} = 1 \\ &\Leftrightarrow c = e^{-\lambda} \end{aligned}$$

Anything that is not associated with the sample can be taken out.

What bears meaning are ratios like

$$\frac{L(\theta'|x_1,\ldots,x_n)}{L(\theta^*|x_1,\ldots,x_n)}$$

which measures how likely is θ' compared to θ^* having observed thet data $x_1,\ldots,x_n$.

Note:

The ratio has a meaning, but individual values at differnet λ doesn't mean anything.

With this definition, the likelihood function depends on the data only through the observed value of a **sufficient statistic**.

$$f(\tilde{x}|\theta) = g(T(\tilde{x});\theta)h(\tilde{x}) \Rightarrow L(\theta|\tilde{x}) \propto g(\tilde{x}|\theta) \propto g(T(\tilde{x});\theta)$$

Also note that

$$\frac{\partial \ln(f(x|\theta))}{\partial\theta} = \frac{\partial \ln(L(\theta|x))}{\partial\theta}$$

Proof.

$$\begin{aligned} L(\theta|x) &= cf(x|\theta) \\ \ln(L(\theta|x)) &= \ln c + \ln(f(x|\theta)) \\ \frac{\partial \ln(f(x|\theta))}{\partial\theta} &= \frac{\partial \ln(L(\theta|x))}{\partial\theta} \end{aligned}$$

Need to differentiate regarding to θ. ■

Let us consider the case $\Theta \in \mathbb{R}$. We have **regularity conditions:**

- C1: Θ is an open interval of $\mathbb{R}$
- C2: The set $\{x : f(x|\theta) > 0\}$, i.e. the support of $f(\cdot|\theta)$ does not depend on θ
- C3: The function $f(x|\theta), x \in \mathcal{X}, \theta \in \Theta$, is differentiable in θ for all x
- C4: We have $0 < E_\theta\left[\frac{\partial \ln f(X|\theta)}{\partial\theta}\right]^2 < \infty$ for all θ

- C5: It is ok to permute the symbols $\frac{\partial}{\partial\theta}$ and $\int dx$

Definition 3.4.2: Score function

Having observed the sample $x_1, \ldots, x_n$, the score function measures the relative variation of the likelihood function as a function of θ

$$S(\theta|x_1, \ldots, x_n) = \frac{L'(\theta|x_1, \ldots, x_n)}{L(\theta|x_1, \ldots, x_n)} = \frac{\partial \ln L(\theta|x_1, \ldots, x_n)}{\partial\theta}$$

For an iid random sample, we have $S(\theta|x_1, \ldots, x_n) = \sum_{i=1}^{n} S(\theta|x_i)$ where $S(\theta|x_i)$ is the score function associated with the ith observation.

Proof.

$$\begin{aligned}
L(\theta|x_1, \ldots, x_n) &\propto \prod_{i=1}^{n} f(x_i|\theta) \Rightarrow \\
\ln L(\theta|x_1, \ldots, x_n) &= c + \sum_{i=1}^{n} \ln f(x_i|\theta) \\
&= c^* + \sum_{i=1}^{n} \ln L(\theta|x_i) \\
\frac{\ln L(\theta|x_1, \ldots, x_n)}{\partial\theta} &= \sum_{i=1}^{n} \frac{\ln L(\theta|x_i)}{\partial\theta} = \sum_{i=1}^{n} S(\theta|x_i)
\end{aligned}$$

■

Theorem 3.4.1

Under the regularity conditions, we have that for all $\theta \in \Theta$

$$E_\theta[S(\theta|X_1, \ldots, X_n)] = 0$$

Proof. The idea is that if 1 is zero, the sum is also zero $\rightarrow E[S(\theta|x_i)] = 0$

$$\begin{aligned}
E[S(\theta|x_i)] &= \int_x \frac{\partial \ln f(x|\theta)}{\partial\theta} f(x|\theta)\, dx \\
&= \int_x \frac{\frac{\partial f(x|\theta)}{\partial\theta}}{f(x|\theta)} f(x|\theta)\, dx \\
&= \int_x \frac{\partial f(x|\theta)}{\partial\theta}\, dx \\
&= \frac{\partial}{\partial\theta} \underbrace{\int_x f(x|\theta)\, dx}_{=1} \\
&= \frac{\partial}{\partial\theta} 1 = 0
\end{aligned}$$

■

Definition 3.4.3: Fisher information

The Fisher information about θ contained in $X_1, \ldots, X_n$ is defined by

$$I_{(X_1,\ldots,X_n)}(\theta) = E_\theta\{[S(\theta|X_1, \ldots, X_n)]^2\}$$

Properties under the regularity condition:

- $I_{(X_1,\ldots,X_n)}(\theta) = Var[S(\theta|X_1, \ldots, X_n)]$

Proof. $Var(x) = E[x^2] - E^2[x]$ since $E[x] = 0 \Rightarrow Var(x) = E[x^2]$ with x being the score function. ■

- If $X = (X_1, X_2)$ with X_1 and X_2 independent $\forall\theta$, then

$$I_X(\theta) = \sum_{i=1}^{2} I_{X_i}(\theta)$$

Proof.

$$\begin{aligned}
f(\tilde{x}_1 m \tilde{x}_2|\theta) &= f(\tilde{x}_1|\theta) f(\tilde{x}_2|\theta) \\
\ln f(\tilde{x}_1 m \tilde{x}_2|\theta) &= \ln f(\tilde{x}_1|\theta) + \ln f(\tilde{x}_2|\theta) \\
\frac{\partial \ln f(\tilde{x}_1 m \tilde{x}_2|\theta)}{\partial\theta} &= \frac{\partial \ln f(\tilde{x}_1|\theta)}{\partial\theta} + \frac{\partial \ln f(\tilde{x}_2|\theta)}{\partial\theta} \\
S(\theta|\tilde{x}_1, \tilde{x}_2) &= S(\theta|\tilde{x}_1) + S(\theta|\tilde{x}_2) \\
I_{\tilde{x}_1,\tilde{x}_2}(\theta) &= Var(S(\theta|\tilde{x}_1, \tilde{x}_2)) = Var(S(\theta|\tilde{x}_1) + S(\theta|\tilde{x}_2)) \\
&= Var(S(\theta|\tilde{x}_1) + S(\theta|\tilde{x}_2)) = I_{\tilde{x}_1}(\theta) + I_{\tilde{x}_2}(\theta)
\end{aligned}$$

■

- Consequently, in the case of an iid random sample, $I_{(X-1,\ldots,X_n)}(\theta) = nI_{X_i}(\theta)$
- Useful formula to compute I

$$I_{(X_1,\ldots,X_n)}(\theta) = E[S^2(\theta|X)] = -E\left[\frac{\partial S(\theta|X)}{\partial\theta}\right] = -E\left[\frac{\partial^2 \ln L(\theta|X)}{\partial\theta^2}\right]$$

Theorem 3.4.2

If $T(X_1, \ldots, X_n)$ is a statistic, and under regularity conditions,

$$I_{(X_1,\ldots,X_n)}(\theta) \geq I_T(\theta)$$

and the equality holds if and only if T is sufficient for θ.

This makes sense as statistics contains all information needed but still performs data reduction.

Theorem 3.4.3

Consider the alternative parametrization $\theta = g(\phi)$ with g differentiable. Then the Fisher information about ϕ in X, $I_X^*(\phi)$ satisfies

$$I_X^*(\phi) = I_X(g(\phi))[g'(\phi)]^2$$

with $I_X(\theta)$ representing the Fisher about θ in X.

Wrong Concept 3.2: Poisson alternative parametrization

$$\begin{aligned}
&\phi = \frac{1}{\lambda} \Leftrightarrow \lambda = \frac{1}{\phi}, \quad I_X(\lambda) = \frac{1}{\lambda} \\
&I_X^*(\phi) = ? \\
&I_X^*(\phi) = I_X(\frac{1}{\phi}) = \lambda \Rightarrow \text{WRONG}
\end{aligned}$$

Example 3.4.2 (Alternative parametrization : Poisson)

Through brute force computation,

$$\begin{aligned} f(x|\phi) &= e^{-\frac{1}{\phi}} \frac{(\frac{1}{\phi})^x}{x!} \\ \ln f(x|\phi) &= -\frac{1}{\phi} - x \ln \phi + c \\ \frac{\partial f}{\partial \phi} &= \frac{1}{\phi^2} - \frac{x}{\phi} \\ \frac{\partial^2 \ln f}{\partial \phi^2} &= -2\phi^{-3} + \frac{x}{\phi^2} \\ I_X^*(\phi) &= -E[-2\phi^{-3} + \frac{x}{\phi^2}] = \frac{2}{\phi^3} - \frac{\frac{1}{\phi}}{\phi^2} = \frac{2}{\phi^3} - \frac{1}{\phi^3} = \frac{1}{\phi^3} \end{aligned}$$

Through the theorem,

$$\begin{aligned} \phi = \frac{1}{\lambda} \Leftrightarrow \lambda = \frac{1}{\phi} = g(\phi) \\ g'(\phi) = -\frac{1}{\phi^2} \\ I_X^*(\phi) = I_X(\frac{1}{\phi})[g'9\phi]^2 = \phi \frac{1}{\phi^4} = \frac{1}{\phi^3} \end{aligned}$$

The ideas the properties that we have described for the case $\Theta \in \mathbb{R}$ can be naturally extended to the multiparametric case $\Theta \subset \mathbb{R}^k$. Only difference is for the alternative parameterization

$$I_X^*(\phi) = J' I_X(g(\phi)) J$$

where J is a matrix of order k whose ith line is

$$\left(\frac{\partial g_i(\phi)}{\partial \phi_1}, \cdots, \frac{\partial g_i(\phi)}{\partial \phi_k} \right)$$

Chapter 4

Parametric Point Estimation

The context of the problem:

- Parametric statistical model for $\bar{X}, \mathcal{F} = \{f(\cdot|\theta) : \theta \in \Theta\}$
- $X_1, \dots, X_n$ are iid random sample of size n extracted from X
- Samples space is denoted by $\mathcal{X}$

The problem itself:

- To produce a **point estimate of** θ, that is, select an application $x \in \mathcal{X} \mapsto T(x) \in \Theta$ that to each observed sample associates a value for θ
- The application T is a statistic which we call an **estimator** for θ and the observed value of $T(x)$ is called the **estimate**
- We may be interested in estimating a function of θ, $\tau(\theta)$

4.1 Optimality criteria

In frequentist statistics, the quality of an estimator is assessed by looking at the population of estimates it produces, that is, at its sampling distribution. We are **evalauting the estimator** and not the estimate.

Note:

$E[\bar{X}] = \lambda \longrightarrow$ pre-experimental.
$\bar{X} \longrightarrow$ post-experimental, not so intersted in this accuracy.

4.1.1 Unbiasedness

Definition 4.1.1: Unbiased estimator

An estimator T is said to be an unbiased estimator of $\tau(\theta)$ if and only if

$$\forall \theta \in \Theta \quad E_\theta[T] = \tau(\theta)$$

In practice, this means that if we use T to estimate $\tau(\theta)$ a very large number of times, then the average of the estimates will be close to $\tau(\theta)$ no matter the true value of θ.

Proof. $E[T(\tilde{x})] = \tau(\theta)$

$$\begin{cases} \tilde{x}_1 \longrightarrow T(\tilde{x}_1) \\ \vdots \\ \tilde{x}_N \longrightarrow T(\tilde{x}_N) \end{cases} \Rightarrow \frac{1}{N} \sum_{i=1}^{N} T(\tilde{x}_i) \overset{LLN}{=} \tau(\theta)$$

■

The quantity $b(T) = E[T|\theta] - \tau(\theta)$ is known as the **bias of the estimator of** $\tau(\theta)$. An estimator which is not unbiased is said to be biased. There are cases where an unbiased estimator cannot be found. In general, the sample mean is an unbiased estimator of the population mean as long as it exists. The bias-corrected variance is an unbiased estimator of the population variance as long as it exists. By restricting the class of interesting estimators to the class of unbiased estimators, we may miss intersting estimators.

4.1.2 Most efficient estimation

Definition 4.1.2: More efficient estimator

Let T and T^* be two **unbiased** estimators of $\tau(\theta)$. We say that T is more efficient than T^* in the estimation of $\tau(\theta)$ if

$$Var_\theta(T) \leq Var_\theta(T^*), \quad \forall \theta \in \Theta$$

Theorem 4.1.1 Cramer-Rao inequality

Consider a statistical model satisfying the **regularity conditions** with $\Theta \subset \mathbb{R}$, and let $\tau(\theta)$ be a differentiable function. Let T be an **unbiased estimator** of $\tau(\theta)$ with finite variance. Additionally, assume that $\forall \theta \in \Theta$

$$E_\theta[T(X-1,\dots,X_n)S(\theta|X_1,\dots,X_n)] < +\infty$$
$$E_\theta[T(X-1,\dots,X_n)S(\theta|X_1,\dots,X_n)] = \tau'(\theta) \rightarrow (1)$$

in which case we say that T is a **regular estimator**. Then,

$$Var_\theta(T) \geq \frac{[\tau'(\theta)]^2}{nI_X(\theta)} \rightarrow \text{Cramer-Rao lower bound (CRLB)}$$

Proof. $(1) : E_\theta[T(X-1,\dots,X_n)S(\theta|X_1,\dots,X_n)] = \tau'(\theta)$

$$E[TS] = \int_x T(\tilde{x}) \frac{\partial \ln f(\tilde{x}|\theta)}{\partial \theta} f(\tilde{x}|\theta)\, d\tilde{x} = \int_x T(\tilde{x}) \frac{\frac{\partial f(\tilde{x}|\theta)}{\partial \theta}}{f(\tilde{x}|\theta)} f(\tilde{x}|\theta)\, d\tilde{x}$$
$$= \int_x \frac{\partial T(\tilde{x}) f(\tilde{x}|\theta)}{\partial \theta}\, d\tilde{x} = \frac{\partial}{\partial \theta} \int_x T(\tilde{x}) f(\tilde{x}|\theta)\, d\tilde{x} = \frac{\partial}{\partial \theta} E[T] = \tau'(\theta)$$

■

Proof. Cramer-Rao inequality

$$Cov(T,S) = E[TS] - E[T] \underbrace{E[S]}_{=0,\text{ regularity condition}} = E[TS] = \tau'(\theta)$$

Using Cauchy-Swartz inequality,

$$[\rho(T,S)]^2 \leq 1 \Leftrightarrow$$
$$\left[\frac{Cov(T,S)}{\sqrt{Var(T)Var(S)}}\right]^2 \leq 1 \Leftrightarrow \frac{[\tau'(\theta)]^2}{Var(T)Var(S)} \leq 1 \Leftrightarrow$$
$$Var(T) \geq \frac{[\tau'(\theta)]^2}{\underbrace{Var(S)}_{=I_{\tilde{x}}(\theta)\text{ regularity condition}}} = \frac{[\tau'(\theta)]^2}{I_{\tilde{x}}(\theta)}$$

■

The Cramer-Rao lower bound is only meaningful under the regularity conditions. Even when then regularity conditions are satisfied, there might not exist an estimator whose variance equals the CRLB.

Definition 4.1.3: Efficiency

The ratio between the CRLB and the variance of an **unbiased estimator** of $\tau(\theta)$ is known as efficiency

$$e(T) = \frac{\text{CRLB}}{Var_\theta(T)}$$

If the regularity conditions are satisfied, $0 \leq e(T) \leq 1$. And if T is an **unbiased** estimator of $\tau(\theta)$ and $e(T) = 1$, then T is known as the **most efficient estimator** of $\tau(\theta)$. There is also the notion of **asymptotic efficiency** $\lim_{n\to\infty} e(T)$ and also of asymptotically most efficient estimators.

Example 4.1.1 (Most efficient estimator : Poisson)

$X_1, \ldots, X_n \sim Po(\lambda)$

$$\tau(\theta) = \lambda \Rightarrow \tau'(\theta) = 1$$
$$Var(T) \geq \frac{1}{I_x(\lambda)} = \frac{\lambda}{n}$$
$$Var(\bar{X}) = \frac{\lambda}{n} = \text{CRLB} \Rightarrow \bar{X} \text{ is the most efficient estimator of } \lambda$$

Another case,

$$\tau(\theta) = e^{-\lambda} \Rightarrow \tau'(\theta) = -e^{-\lambda}$$
$$Var(W) \geq \frac{e^{-2\lambda}}{n/\lambda} = \frac{\lambda e^{-2\lambda}}{n}$$

$\Rightarrow$ If I find the estimator with this variance, then I've found the most efficient estimator

Corollary 4.1.2 CR theorem: Existance of most efficient estimator

Let T be a **regular and unbiased** estimator of $\tau(\theta)$. Then T is the most efficient estimator of $\tau(\theta)$ if and only if there exists $a(\theta)$ such that

$$S(\theta|x_1, \ldots, x_n) = a(\theta)[T(x_1, \ldots, x_n) - \tau(\theta)]$$

Proof. Begin with the condition of the Cauchy-Swartz inequality

$$T = a + bS \Leftrightarrow S = a + bT$$

Since $E[S] = 0$,

$$0 = a + b\tau(\theta) \Leftrightarrow a = -b\tau(\theta) \Rightarrow S = -b\tau(\theta) + bT \Leftrightarrow S = b(T - \tau(\theta))$$

■

Corollary 4.1.3 Sufficient & 1-parameter exponential family

The CRLB in the estimation of $\tau(\theta)$ is attained by an estimator T if and only if T is a **sufficient** statistic in the **one-parameter exponential family** with density

$$f(x) = h(x)c(\theta)\exp[Q(\theta)T(x)]$$

where $c(\theta) = \int a(\theta)\tau(\theta)\, d\theta$ and $Q(\theta) = \int a(\theta)\, d\theta$.

Proof.

$$S(\theta|\tilde{x}) = \frac{\partial \ln f(\tilde{x}|\theta)}{\partial \theta} = a(\theta)[T - \tau(\theta)]$$
$$\Rightarrow \ln f(\tilde{x}|\theta) = A(\theta)T - B(\theta) + c(\tilde{x})$$
$$\Rightarrow f(\tilde{x}|\theta) = \underbrace{e^{c(\tilde{x})}}_{h(\tilde{x})} \overbrace{\underbrace{e^{-B(\theta)}}_{c(\theta)} \exp\{\underbrace{A(\theta)}_{\omega_1(\theta)} \underbrace{T(\tilde{x})}_{R_1(\tilde{x})}\}}^{g(T;\theta)}$$

■

If there is a most efficient estimator for $\tau(\theta)$ then T must be suffcient for θ. There aren't most efficient estimators in models that do not admist one-dimentional sufficient statistics.

Example 4.1.2 (Most efficient estimator : Bernoulli)
$X \sim B(1, \theta)$

$$f(\tilde{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$
$$\ln f(\tilde{x}|\theta) = \sum x_i \ln\theta + (n - \sum x_i)\ln(1-\theta)$$
$$S(\theta|\tilde{x}) = \frac{\partial \ln f(\tilde{x}|\theta)}{\partial\theta} = \frac{\sum x_i}{\theta} + (n - \sum x_i)\frac{-1}{1-\theta} = \cdots = \frac{1}{\theta(1-\theta)}(\sum x_i - n\theta) = \frac{n}{\theta(1-\theta)}(\bar{X} - \theta)$$
$\Rightarrow \bar{X}$ is the most efficient estimator of θ

$$E[\bar{X}] = \theta = \tau(\theta)$$
$$Var(\bar{X}) = \frac{[\tau'(\theta)]^2}{nI_X(\theta)} = \frac{1}{nI_X(\theta)}$$
$$\text{since } Var(\bar{X}) = \frac{\theta(1-\theta)}{n} = \frac{1}{nI_X(\theta)} \Rightarrow nI_X(\theta) = \frac{1}{\theta(1-\theta)}$$

Example 4.1.3 (Most efficient estimator : Exponential)
$X \sim Ex(\lambda)$

$$f(x|\lambda) = \lambda e^{-\lambda x}, \quad x > 0$$
$$f(\tilde{x}|\theta) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}$$
$$\ln f(\tilde{x}|\lambda) = n\ln\lambda - \lambda \sum x_i$$
$$S(\lambda|\tilde{x}) = \frac{\partial f(\tilde{x}|\lambda)}{\partial\lambda} = \frac{n}{\lambda} - \sum x_i = \underbrace{-n}_{a(\lambda)}\Big(\underbrace{\bar{X}}_{T} - \underbrace{\frac{1}{\lambda}}_{\tau(\lambda)}\Big)$$

$\bar{X}$ is the most efficient estimator of $\tau(\lambda) = \frac{1}{\lambda} \rightarrow \tau'(\lambda) = -\frac{1}{\lambda^2}$

$$E[\bar{X}] = \frac{1}{\lambda} = \tau(\lambda)$$
$$Var(\bar{X}) = \frac{[\tau'(\theta)]^2}{nI_X(\lambda)} = \frac{\frac{1}{\lambda^4}}{nI_X(\lambda)} \Leftrightarrow$$
$$I_X(\lambda) = \frac{1}{\lambda^2}$$

Once we determine that T is the most efficient estimator of $\tau(\theta) \Rightarrow E[T] = \tau(\theta); Var(T) = \frac{[\tau'(\theta)]^2}{nI_X(\theta)}$ which may allow us to determine $I_X(\lambda)$ if we know $Var(T)$ or vice versa.

4.1.3 Uniformly minimum-variance unbiased estimation

Definition 4.1.4: Uniformly minimum-variance unbiased estimator (UMVUE)

Let T be an **unbiased** estimator of $\tau(\theta)$. If for any other unbiased estimator of $\tau(\theta)$, W, we have

$$Var(T|\theta) \leq Var(W|\theta), \quad \forall \theta \in \Theta$$

then T is the so called uniformly minimum-variance unbiased estimator of $\tau(\theta)$, or UMVUE.

Theorem 4.1.4 Rao-Blackwell

Let T be a **sufficient** statistic for θ and U an **unbiased** estimator of $\tau(\theta)$. Then $E[U|T]$ is an **unbiased** estimator of $\tau(\theta)$ with variance that is never superior to that of U. The two variances coincide if and only if U is a function of T.

Note:

The process of computing $E[U|T]$ is called Rao-Blackwellization.

Proof. U unbiased estimator of $\tau(\theta)$, T sufficient statistic.

$$E[U|T] \text{ is a statistic}$$

$$E[U|T=t] = \int_x U(\tilde{x}) \underbrace{f(\tilde{x}|t)}_{\text{doesn't depend on } \theta \text{ because } T \text{ sufficient}} dx$$

$$E[E[U|T]] = E[U] = \tau(\theta) \Rightarrow E[U|T] \text{ unbiased estimator of } \tau(\theta)$$

$$Var(U) = \underbrace{E[Var(U|T)]}_{\geq 0} + Var(E[U|T])$$

$$Var(U) \geq Var(E[U|T])$$

$$Var(U) = Var(E[U|T]) \Leftrightarrow E[Var(U|T)] = 0 \Leftrightarrow U \text{ is a function of } T$$

■

Example 4.1.4 (Rao-Blackwell : Normal)

$X_1, \ldots, X_n \sim n(\mu, 1), \quad \tau(\theta) = \mu^2 + 1$

We know that $U = \frac{\sum X_i^2}{n} = S^2 + \bar{X}^2$ is an unbaised estimator of $\tau(\mu)$

$$E[U] = \mu^2 + 1$$

$$\bar{X} \text{ is sufficient for } \mu$$

$$E[U|\bar{X}] = E[S^2 + \bar{X}^2|\bar{X}] = E[S^2|\bar{X}] + E[\bar{X}^2|\bar{X}]$$

$$\bar{X} \perp\!\!\!\perp S^2 \Rightarrow E[S^2|\bar{X}] = E[S^2] = \frac{n-1}{n} \times 1$$

$$\text{Also, } E[\bar{X}^2|\bar{X}] = \bar{X}^2$$

$$\Rightarrow E[U|\bar{X}] = \frac{n-1}{n} + \bar{X}^2$$

The UMVUE should be a **function of a sufficient statistic**. When we Rao-Blackwellize an unbiased estimator, we do not obtain necessarily the UMVUE because there is always the possibility that if we had started with another unbaised estimator we could haveobtained a smaller variance. That cannot happen if we **start with a sufficient and complete**

statistic: Let $g(T) = E[U|T]$ and $g^*(T)$ another unbiased estimator which is also a function of T. The **completeness** of T implies that $g(T) = g^*(T)$.

Theorem 4.1.5 Lehmann-Scheffé

If the statistical model admits a **sufficient and complete** statistic T and there is at least an **unbiased** estimator of $\tau(\theta)$ then there is an UMVUE for $\tau(\theta)$ that is unique and a function of T.

Startegies to find UMVUEs in models admitting sufficient and complete statistic:

1. Obtain an unbiased estimator and Rao-Blackwellize it using the sufficient and complete statistic.
   - Find $U : E[U] = \tau(\theta)$
   - Find T sufficient and complete $\rightarrow$ compute $E[U|T]$
2. Directly identify an unbiased estimator that is a function of a complete and sufficient statistic.
   - Find T sufficient and complete
   - Find $g(T)$ such that $E[g(T)] = \tau(\theta)$

Example 4.1.5 (Lehmann-Scheffé : Poisson)

Conditions:

- $X_1 \ldots, X_n|\theta \sim Po(\theta)$
- $Po(\theta)$ is a member of the 1-parameter exponential family with $\alpha = \ln\theta \in \mathbb{R}$
- $T = \sum x_i$ is sufficient and complete

Using method 1:

$$U = I_{\{0\}}(X_1) \Rightarrow E[U] = P(X_1 = 0) = e^{-\theta}$$

$$E[U|T=t] = E[I_{\{0\}}(X_1)|T=t] = P(X_1 = 0|T=t) = \frac{P(X-1=0; \sum_{i=2}^n X_1 = t)}{P(T=t)} = \frac{P(X_1=0)P(\sum_{i=2}^n X_1 = t)}{P(T=t)}$$

$$\left[\begin{cases} \sum_{i=2}^n x_i \sim Po((n-1)\theta) \\ T \sim Po(n\theta) \end{cases}\right]$$

$$= \frac{e^{-\theta}\frac{e^{-(n-1)\theta}[(n-1)\theta]^t}{t!}}{e^{-n\theta}\frac{(n\theta)^t}{t!}} = \left(\frac{n-1}{n}\right)^t \rightarrow \text{UMVUE}$$

Using method 2:

$$E[g(T)] = \sum_{t=0}^{\infty} g(T)e^{-n\theta}\frac{(n\theta)^t}{t!} = e^{\theta}, \quad \forall\theta > 0 \Leftrightarrow$$

$$\sum_{t=0}^{\infty} g(T)\frac{(n\theta)^t}{t!} = e^{n\theta}e^{-\theta}, \quad \forall\theta > 0 \Leftrightarrow$$

$$\sum_{t=0}^{\infty} g(T)n^t\frac{\theta^t}{t!} = e^{(n-1)\theta}, \quad \forall\theta > 0 \Leftrightarrow (1)$$

$$\sum_{t=0}^{\infty} g(T)n^t\frac{\theta^t}{t!} = \sum_{t=0}^{\infty}(n-1)^t\frac{\theta^t}{t!}, \quad \forall\theta > 0 \Rightarrow$$

$$g(t)n^t = (n-1)^t \Leftrightarrow g(t) = \left(\frac{n-1}{n}\right)^t \rightarrow \text{UMVUE}$$

(1):

$$e^x = \sum_{j=0}^{\infty} \frac{x^j}{j!}$$

$$e^{(n-1)\theta} = \sum_{t=0}^{\infty} \frac{[(n-1)\theta]^t}{t!} = \sum_{t=0}^{\infty} (n-1)^t \frac{\theta^t}{t!}$$

Relating efficiency and UMVU. If the model is **regular** and there exists a **most efficient estimator** for $\tau(\theta)$, then this estimator is necessarily the **UMVUE**. The variance of the UMVUE will not necessarily be equal to the CRLB. If the model is regular, there may not exist most efficient estimators, in this case the UMVUE will have a variance strictly larger than the CRLB. The **UMVU criterion does not depend on regularity conditions**, and hence, if theses are not met, the variance of the UMVUE may be smaller than the CRLB defined as if the model were regular.

4.1.4 Mean square error

This is used to compare unbiased estimators in terms of the **dispersion** of their distribution around $\tau(\theta)$.

Definition 4.1.5: Mean square error

The mean square error of an estimator T of $\tau(\theta)$ is

$$MSE(T) = E_\theta[(T - \tau(\theta))^2]$$

Couple **remarks**:

- T will be superior to T^* in mean squared error in the estimation of $\tau(\theta)$ if $MSE(T \leq MSE(T^*)),\ \forall \theta$.
- Markov inequality indicates that $P(|T - \tau(\theta)| > \epsilon) \leq \frac{MSE(T)}{\epsilon^2}$
- We have $MSE(T) = VAR_\theta(T) + [b(T)]^2$

Proof.

$$\begin{aligned} E_\theta[(T - \tau(\theta))^2] &= E[(T - \mu + \mu - \tau(\theta))^2] = E[(T_\mu)^2] \\ &= E[(T-\mu)^2 + 2(\mu - \tau(\theta))(T - \mu) + (\mu - \tau(\theta))^2] \\ &= Var(T) + 0 + (\mu + \tau(\theta))^2 = Var(T) + b_T^2 \end{aligned}$$

■

Example 4.1.6 (MSE : Normal)
$X \sim N(\mu, \sigma^2)$

$$E[S'^2] = \sigma^2$$

$$E[S^2] = \frac{n-1}{n}\sigma^2$$

$$\frac{2\sigma^2}{n} = Var(S'^2) > MSE(S^2) = \frac{2\sigma^2}{n} - \frac{3\sigma^2}{n^2}$$

$\Rightarrow$ According to the MSE criterion, S^2 is better than S'^2.

4.1.5 Consistency

Consistency looks at the behavior of the sampling distribution of an estimator as $n \longrightarrow +\infty$.

Definition 4.1.6: Consistency

Let $T_n = T(X_1, \ldots, X_n)$. The estimator T_n is said to be a (weakly) consistent estimator of $\tau(\theta)$ if, for all $\theta \in \Theta$,

$$\forall \epsilon > 0 \quad \lim_{n \to +\infty} P(|T_n - \tau(\theta)| > \epsilon) = 0$$

that is, if, for all $\theta \in \Theta$, $T_n \xrightarrow{P} \tau(\theta)$.

Definition 4.1.7: Mean-square consistency

The estimator T_n is said to be a mean-square consistent estimator of $\tau(\theta)$ if, for all $\theta \in \Theta$,

$$\lim_{n \to +\infty} E_\theta[(T_n - \tau(\theta))^2] = 0$$

that is, if, for all $\theta \in \Theta$, $T_n \xrightarrow{m.s.} \tau(\theta)$.

Couple **remarks**:

- From the Markov inequality, we know that mean-square convergence implies convergence in probability, so mean-square consistency implies weak consistency.
- Since $E_\theta[(T_n - \tau(\theta))^2] = MSE(T_n) = Var(T_n) + [b(T)]^2$, a **sufficient** condition for weak sonsistency of T_n is that

$$\lim_{n \to +\infty} E[T_n] = \tau(\theta)$$
and
$$\lim_{n \to +\infty} Var[T_n] = 0$$

- Consistency is not a very restrictive property. It is useful to exclude estimators though.

4.2 Estimation Methods

4.2.1 Method of moments

The idea is to **estimate population moments by their corresponding sample moments**. Let $\theta = (\theta_1, \ldots, \theta_k)$ be a vector of unknown parameters of the population, $\mu' = E[X^r]$ will necessarily be a function of $\theta : \mu'_r = \psi_r(\theta)$. Considering the corresponding sample moments $M'_r = \sum_{i=1}^n \frac{X_i^r}{n}$ and form the system of equations $M'_r = \psi_r(\theta)$. The solution to this equation determines a methods of moments estimator of $\theta : (\phi_r(X_1, \ldots, X_n), r = 1, \ldots, k)$.

Note:

Recall the close relationship between the sample moments and the raw moemtns in the asymptotic sense.

Example 4.2.1 (A method of moment : Normal)

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$

$$\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$$
$$\begin{cases} \mu'_1 = \mu = \theta_1 \\ \sigma^2 = \mu'_2 - (\mu'_1)^2 \end{cases} \Leftrightarrow \begin{cases} \mu'_1 = \mu = \theta_1 \\ \mu'_2 = \sigma^2 + \mu^2 = \theta_1 + (\theta_2)^2 \end{cases}$$
$$\begin{cases} M'_1 = \mu \\ M'_2 = \sigma^2 - \mu^2 \end{cases} \Leftrightarrow \begin{cases} \mu = M'_1 \\ \sigma^2 = M'_2 - (M'_1)^2 = S^2 \end{cases} \Rightarrow \begin{cases} \tilde{\mu} = \bar{X} \\ \tilde{\sigma}^2 = S^2 \end{cases}$$

This is a method of moment, can also start with central moments as well.

Example 4.2.2 (A method of moment : Gamma)

$X_1, \dots, X_n \sim G(\alpha, \lambda)$

$$\begin{cases} \mu_1' = \frac{\alpha}{\lambda} \\ \mu_2' - (\mu_1')^2 = \frac{\alpha}{\lambda^2} \end{cases} \Leftrightarrow \begin{cases} \mu_1' = \frac{\alpha}{\lambda} \\ \mu_2' - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2} \end{cases} \Leftrightarrow \begin{cases} \mu_1' = \frac{\alpha}{\lambda} \\ \mu_2' = \frac{\alpha}{\lambda^2} + \frac{\alpha^2}{\lambda^2} \end{cases} \Leftrightarrow \begin{cases} \mu_1' = \frac{\alpha}{\lambda} \\ \mu_2' = \frac{\alpha+\alpha^2}{\lambda^2} \end{cases}$$

$$\Downarrow$$

$$\begin{cases} M_1' = \frac{\alpha}{\lambda} \\ M_2' = \frac{\alpha+\alpha^2}{\lambda^2} \end{cases} \Leftrightarrow \begin{cases} \alpha = M_1'\lambda \\ M_2' = \frac{M_1'\lambda + (M_1')^2\lambda^2}{\lambda^2} \end{cases} \Leftrightarrow (\cdots) \Leftrightarrow \begin{cases} \alpha = M_1'\lambda \\ \lambda = \frac{M_1'}{M_2' - (M_1')^2} \end{cases} \Leftrightarrow \begin{cases} \alpha = M_1'\lambda \\ \lambda = \frac{\bar{X}}{S^2} \end{cases} \Leftrightarrow \begin{cases} \alpha = \frac{\bar{X}^2}{S^2} \\ \lambda = \frac{\bar{X}}{S^2} \end{cases}$$

The **advantages** of using the method of momentes include 1) there is always a solution and 2) there is no need to assume much, just need the moments to exist. The **disadvantage** arises when the support depends on the parameter as moment does not take into consideration of the support of the distribution.

Recalling the properties of the sampling moments

$$E[M_r'|\theta] = \mu_r'$$
$$Var(M_r'|\theta) = \frac{[\mu_{2r}' - (\mu_r')^2]}{n} \equiv \frac{v_{rr}}{n}$$
$$Cov_\theta(M_r', M_s') = \frac{[\mu_{r+s}' - \mu_r'\mu_s']}{n} \equiv \frac{v_{rs}}{n}$$

The multivariate **central limit theorem** states that, with $M = (M_1', \dots, M_k')$ and $\mu = (\mu_1', \dots, \mu_k')$,

$$\sqrt{n}(M' - \mu) \xrightarrow{d} N_k(0, V)$$

with $V = [v_{rs}]$. The method of moments estimator of θ is a function of of M, and its asymptotic properties can be determines using the **delta method**:

$$\sqrt{n}(h(M_r') - h(\mu_r')) \xrightarrow{d} N(0, [h'(\mu_r)]^2 v_{rr})$$

4.2.2 Maximum likelihood estimator

The idea is to propose an estimate of θ that **maximizes the likelihood function**, which measures how likely it is that θ gives the true value of the parameters that generated the observed data. Formally, the MLE of θ, when it exists, is $\hat{\theta}$ such that

$$L(\hat{\theta}|x_1, \dots, x_n) \geq L(\theta|x_1, \dots, x_n) \quad \forall \theta \in \Theta$$

In most cases, it is easier and equivalent to find the value of θ which maximizes the log-likelihood function. This is done by finding the zeroes of the score function

$$\frac{d}{d\theta} \ln L(\theta|x_1, \dots, x_n) = 0$$

then verifying that the stationary point is indeed a global maximum. This is done by finding the Hessian and proving it's negative (concave).

Proof. We know that the likelihood function is always positive. And the function $\ln x$ is an increasing function. $\Rightarrow$ Value of θ that maximizes $L(\theta|\tilde{x})$ is the value of θ that maximizes $\ln L(\theta|\tilde{x})$

$$L(\theta|\tilde{x}) \propto \sum_{i=1}^{n} f(x_i|\theta) \Rightarrow \ln L(\theta|\tilde{x}) = c + \sum_{i=1}^{n} \ln f(x_i|\theta)$$

■

The MLE may not be unique (case of Normal distribution), and in many cases there will not be closed form expression for the estimate, in which case we have to resort to numerical methods.

Example 4.2.3 (MLE : Poisson)

$X_1 \ldots, X_n \sim Po(\lambda),\ \lambda > 0$

$$L(\lambda|\tilde{x}) \propto e^{-n\lambda}\lambda^{\sum x_i} \Rightarrow \ln L = c - n\lambda + \sum x_i \ln \lambda$$
$$S(\lambda|\tilde{x}) = \frac{\partial \ln L}{\partial \lambda} = -n + \frac{\sum x_i}{\lambda}$$
$$S(\lambda|\tilde{x}) = 0 \Leftrightarrow -n + \frac{\sum x_i}{\lambda} = 0 \Leftrightarrow \lambda = \bar{x}$$
$$\frac{\partial^2 \ln L}{\partial \lambda^2} = \frac{-\sum x_i}{\lambda^2} < 0 \Rightarrow \text{maximum} \Rightarrow \hat{\lambda} = \bar{x}$$

Example 4.2.4 (MLE : Gamma)

$X_1 \ldots, X_n \sim G(\alpha, \lambda),\ \lambda$ is known

$$L(\alpha|\tilde{x}) \propto \prod_{i=1}^{n} \frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda} \propto \frac{\lambda^{n\alpha}}{(\Gamma(\alpha))^n} \left(\prod_{i=1}^{n} x_i\right)^\alpha$$
$$\ln L(\alpha|\tilde{x}) = c(\lambda, \tilde{x}) + n\alpha \ln \lambda - n \ln \Gamma(\alpha) + \alpha \sum \ln x_i$$
$$S(\alpha|\tilde{x}) = n \ln \lambda - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum \ln x_i$$
$$S(\alpha|\tilde{x}) = 0 \Leftrightarrow n \ln \lambda - n\psi(\alpha) + \sum \ln x_i \Rightarrow \text{needs to be solved numerically}$$

Properties of the MLE:

- **Invariance**: If $\hat{\theta}$ is the MLE of θ, and τ is a one-to-one function of θ, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$. This can be generalized to situations where $\tau(\cdot)$ is not one-to-one as well, as long as it's reasonable.
- If T is **sufficient** and there is a MLE of θ, then this estimator is a function of T.

  Proof. If sufficient then

  $$L(\theta|\tilde{x}) \propto f(\tilde{x}|\theta) = g(T(\tilde{x}); \theta)h(\tilde{x})$$
  $$\ln L(\theta|\tilde{x}) = c(\tilde{x}) + \ln g(T(\tilde{x}); \theta)$$
  $$\frac{\partial \ln L}{\partial \theta} = \frac{\partial g(T(\tilde{x}); \theta)}{\partial \theta} = 0$$

  The solution to this equation will be a function of $T(\tilde{x})$ ■
- The **most efficient** estimator of θ, if it exists, is also the MLE of θ.

  Proof. If there is a most efficient estimator, then

  $$S(\theta|\tilde{x}) = a(\theta)[T(\tilde{x}) - \theta]$$
  $$S(\theta|\tilde{x}) = 0 \Leftrightarrow \theta = T(\tilde{x})$$

  The MLE is the most efficient estimator of θ. ■

  Note:

  If I've found the most efficient estimator, no need to find the MLE becuase they're the same. But finding the MLE first then saying it's the most efficient estimator is wrong.
- Under regularity conditions, the MLE of θ satisfies

  $$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, [I_X(\theta)]^{-1})$$

  It is possible to replace $I_{X_1,\ldots,X_n}(\theta)$ either by

- The Fisher information evaluated at MLE: $nI_X(\hat{\theta})$
- The observed Fisher information: $H(\theta) = -\frac{\partial^2}{\partial\theta^2} \ln L(\hat{\theta}|X_1, \ldots, X_n)$

- Above results combined show that **MLE is BAN**(Best Asymptotically Normal) and is consistent and asymptotically most efficient.

Proof. Proof of consistency,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, [I_X(\theta)]^{-1})$$
$$\Downarrow a_n(T_n - \theta) \xrightarrow{d} T \Rightarrow T_n \xrightarrow{p} \theta$$
$$\hat{\theta} \xrightarrow{p} \theta \Rightarrow \hat{\theta} \text{ is consistent}$$

Asymptotic variance of $\hat{\theta} = \frac{1}{nI_X(\theta)}$ = CRLB under regularity conditions.

$$\ln L(\theta|\tilde{x}) = \sum \ln f(x_i|\theta)$$
$$S(\theta|\tilde{x}) = \sum S(\theta|X_i)$$
$$S(\theta|\tilde{x}) = \underbrace{S(\hat{\theta}|\tilde{x})}_{=0} + (\hat{\theta} - \theta)\frac{\partial S}{\partial\theta} + R \Rightarrow \text{1st order Taylor Expansion}$$
$$S(\theta|\tilde{x}) = (\hat{\theta} - \theta)\frac{\partial S}{\partial\theta}$$
$$\downarrow \text{ sum of iid r.v.}$$
$$E[S(\theta|\tilde{x})] = 0$$
$$E\left[-\frac{\partial S}{\partial\theta}\right] = I_{\tilde{x}}(\theta)$$

■