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

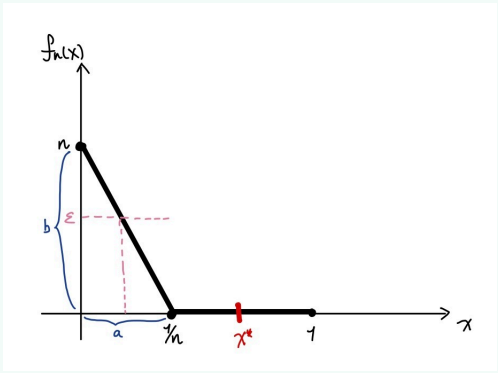
Notions of **convergence of sequences**: let $f_n, f : [0, 1] \to \mathbb{R}$

- 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

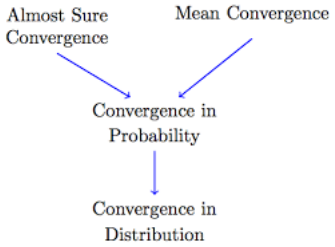
$$F_n(x) \to F_n$$

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,

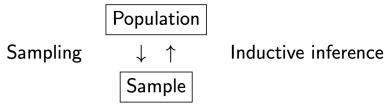
$$X_1, \ldots, X_n \overset{iid}{\sim} B(1, \theta)$$

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

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

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 momnets**, $\mu'_k = E[X^k]$ 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

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

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

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

■