Table of Contents

This is not a comprehensive summary of the course content.

MT Week 1: Introduction and Statistics Fundamentals

Random Variables, Probabilities, Kolmogorov's Axioms

A **standard convention on notations**: *capital letters* denote the random variable itself, *lower case letters* denote a particular realisation of the random variable:

$$Pr(\underbrace{X}_{\text{Random Variable}} > \underbrace{x}_{\text{Realisation}})$$

Probability Axioms / Kolmogorov's Axioms for Discrete Variables: #flashcard
For a discrete RV Y which may take one of K possible mutual exclusive outcomes $\{y_i\}_1^K$, any potential set of probability should satisfy:

- $P(Y = y_j) \geq 0 \,\forall\, j \in \{1, 2, \ldots, K\}$
- $P(Y = y_1 \text{ or } Y = y_2 \text{ or } \ldots \text{ or } . Y = y_K) = 1$
- $P(Y = y_j \text{ or } Y = y_k) = P(Y = y_j) + P(Y = y_k)$ for $j \neq k$
  This set of axioms can be generalised to incorporate continuous RVs.

Set Notations:

- Union $\cup \iff$ "or" statement
- Intersection $\cap \iff$ "and" statement
- $Y \in A \iff$ event A has happened

Sample vs Population, the Data Generating Process (DGP)

Sample, Population, and Sample Analogue:
Any object that can be calculated based on observed quantities is a **sample quantity**.
We assume there's some **data generating process (DGP)** behind an RV which generates the data we observe. Any quantity whose calculation requires the DGP is a **population quantity**.
It's common to use a hat to denote sample objects which mimic or **estimate** an unknown population object. aka **sample analogue**

DGP represents a *particular belief* about the world.

An overview of Econometrics: Econometrics can be used for 3 key aims:

- Estimate the DGP or certain characteristics about it
- Test the validity of claims of economic models
- Predict the behaviour of economic variables in situations that we do not observe using observed data (known as forecasting is TS)
  A healthy mix of economic models and statistical tools is always helpful.

Distribution Functions

Population Cumulative Distribution Function (CDF):
For the scalar random variable X, the cumulative distribution function is

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

It may also be denoted as $F_X(x)$ to emphasise this CDF is for the random variable X.
For a discrete RV, CDF will be a step function. For continuous RV, CDF will be continuous.

3 Properties for All CDFs: #flashcard

- $$\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$$
- $F(x)$ is non-decreasing in x
- $F(x)$ is right-continuous:

  $$\lim_{x \to x_0^+} F(x) = F(x_0)$$

  Any function which satisfies the above 3 properties is a potential candidate for a CDF which takes values in real numbers.

Points along the x-axis where $F(X)$ jumps up are called **atoms**, corresponding to the values where the discrete RV have a non-zero probability mess.

Empirical Distribution Functions (EDF): #flashcard
EDF is the sample analogue of CDF, which gives the proportion of observations which is less than or equal to some specific value. EDF is denoted as $\widehat{F}(x)$:

$$\widehat{F}(w) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(W_i \leq w)$$

EDF will always be a non-decreasing step function, which jumps at each of the sample values (even for continuous RVs). If the underlying RV is continuous, EDF will be more and more smooth as the sample size grows.

Density Functions

Density functions are derived from CDFs. They correspond to different things for continuous and discrete variables, and they may not exist for continuous RVs.

Density Function / Probability Mess Functions for Discrete RVs: #flashcard
The PMF for a discrete RV is defined as:

$$f(x_0) = P(X = x_0) = F(x_0) - \lim_{x \to x_0^-} F(x)$$

in other words:

$$f(x) = P(X = x) = P(X \leq x) - P(X < x)$$

It may also be written as $f_X(x)$.
The set of x for which $f(x) > 0$ is called the **support** of the distribution.

Bernoulli Distribution: #flashcard
$X \sim \text{Bernoulli}(\theta)$ if $P(X = 1) = \theta$ and $P(X = 0) = 1 - \theta$
Bernoulli pmf:

$$f(x) = \theta^x (1 - \theta)^{1-x}, x \in \{0, 1\}$$

Binomial Distribution: #flashcard
Binomial distribution follows the ideal of repeating Bernoulli trials n times.
Let RV X denote the number of 1's obtained in n repetitions of independent Bernoulli trials.
Binomial pmf:

$$f(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} = \frac{n!}{x!(n-x)!} \theta^x (1 - \theta)^{n-x}$$

for $x = 0, 1, 2, \ldots, n$
and $f(x) = 0$ elsewhere.

Poisson Distribution: #flashcard
Poisson distribution is used to model the number of occurrences in some time interval.
An RV $X \sim \text{Poisson}(\lambda)$ if it has pmf:

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

where λ is known as the intensity parameter: a high λ implies a higher probability of arrivals.

Density Functions of Continuous Random Variables: #flashcard
The density function of a continuous RV (aka pdf) is:

$$f(x) = \frac{dF(x)}{dx}$$

The more formal distribution is a function $f(x)$ such that:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(a)\, da$$

If no function satisfies the above relationship, then pdf doesn't exist. However, cdf will always exist.
Any function with the following properties can be a potential candidate for the pdf of a RV:

$$f(x) \geq 0 \text{ and } \int_{-\infty}^{\infty} f(x)\, dx = 1$$

Standard Normal Distribution: #flashcard
A RV $Z \sim N(0,1)$ or $Z =^d N(0,1)$ the Standard Normal Distribution (mean 0 variance 1) if it has a pdf:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right)$$

The standard normal CDF does not have a closed-form analytical solution, but can always be approximated numerically:

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(a)\, da$$

Logistic Distribution: #flashcard
Logistic cdf:

$$F(x) = \frac{1}{1+e^{-x}} \text{ for } -\infty < x < \infty$$

Logistic pdf:

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2} \text{ for } -\infty < x < \infty$$

Quantiles

Quantiles: #flashcard
Let X be a continuous random variable with a strictly increasing CDF $F(x)$ (so that the inverse F^{-1} exists). Then, the **quantile function** is given by:

$$q_\tau = Q(\tau) = F^{-1}(\tau)$$

The values of q_τ for different values of τ correspond to different **quantile** of the distribution.

- the first decile = 10^{th} percentile = $q_{0.1}$
- the first quartile = 25^{th} percentile = $q_{0.25}$
- median = $q_{0.5}$
  If the CDF has flat segments or jumps up, then the inverse function does not exist. Then, the quantiles *may not be unique*. A **formal definition of the quantile function** is:

  $$Q(\tau) = \inf\{x \in \mathbb{R} : \tau \leq F(x)\}$$

  i.e. the infimum for the set of x where $F(x) \geq \tau$.

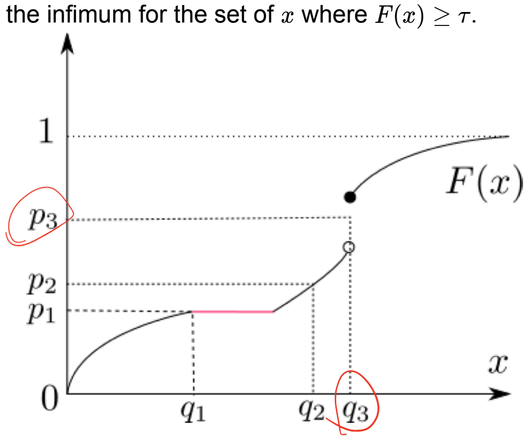


Figure: p_1, p_2 and p_3 quantiles. Source: Wikipedia.

Transformations

CDF of Transformed RVs: #flashcard
Let X be a RV with the distribution function $F_X(x) = P(X \leq x)$. Let $g(.)$ be a strictly increasing/decreasing function. Define

the new RV $Y = g(X)$. Then:

- If g is strictly increasing:

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y))$$

- If g is strictly decreasing:

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \ge g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

Probability Integral Transform: #flashcard
For $U(0,1),\ f(c) = 1, F(c) = c, Q(\tau) = \tau$
Let X be an RV. If we transform X using its CDF: $Y = F_X(X)$. Then:

$$F_Y(y) = P(Y \le y) = P(F_X(x) \le y) = P(x \le F_X^{-1}(y)) = F(F_X^{-1}(y)) = y$$

This is exactly the Uniform CDF.
Thus, *if we transform X using its CDF, we will get a uniformly distributed RV*.
Conversely, we can apply the inverse a distribution's CDF (F^{-1}) to transform a uniformly distributed RV to an RV with that distribution, which is useful for generating ransom samples.

PDF of Transformed RV: Change of Variable Formula: #flashcard
Let X be a continuous RV with pdf $f_X(x)$ and support χ. Let g be a strictly increasing/decreasing and differentiable function. Then the RV $Y = g(X)$ has support $g(\chi)$ and pdf:

$$f_Y(y) = f_X(g^{-1}(y))\ \frac{\partial g^{-1}(y)}{\partial y}$$

We can use this formula to get that, if $X \sim N(\mu, \theta^2)$, then:

$$f_Y(y) = \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$$

Moments

Not much new stuff for Mean/Var/sd

Skewness and Kurtosis: #flashcard
Skewness measures the asymmetry of the distribution. It =0 if symmetric, >0 if skewed to the right.

$$\mu_3 = \mathbb{E}\left\{\left[\frac{X - \mathbb{E}(X)}{sd(X)}\right]^3\right\}$$

but $\mu_3 = 0$ does not imply symmetry.
Kurtosis indicates the relative weight of the probability of being in the middle and the probability in the tails. i.e. how "peaked" the pdf is.

$$\mu_4 = \mathbb{E}\left\{\left[\frac{X - \mathbb{E}(X)}{sd(X)}\right]^4\right\}$$

High kurtosis $\implies$ fat tails. $N(0,1)$ has a kurtosis of 3, so we typically use another measure **excess kurtosis** $\kappa_4 = \mu_4 - 3$.
Distributions with fatter tails than normal (i.e. positive κ_4) are known as **leptokurtic/heavy-tailed distributions**.

Independence

Not much new stuff

MT Week 2: Statistics Fundamentals II

Random Vectors

Basics of Random Vectors: #flashcard
If X_1, X_2 are two scalar random variables, the 2×1 vector:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

is a **bivariate random vector**.
CDF for bivariate RVec describes the joint distribution of the 2 RVs:

$$F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$$

it needs to satisfy:

- $\lim_{x_1 \to -\infty, x_2 \to -\infty} F(x_1, x_2) = 0, \lim_{x_1 \to \infty, x_2 \to \infty} F(x_1, x_2) = 1$
- $F(x_1, x_2)$ is non-decreasing in x_1, x_2
- $F(x_1, x_2)$ s right-continuous
  PDF for continuous bivariate RVec can be obtained by:

  $$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2}$$

  due to its definition:

  $$F(x_1, x_2) = \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(a, b)\, da\, db$$

Independence and Joint CDF/PDF: #flashcard

X_{1}\statindep X_{2}\iff \begin{cases} F_{X}(x_{1},x_{2}) &=F_{X_{1}}(x_{1})F_{X_{2}}(x_{2}) \\ f_{X}(x_{1},x_{2}) &=f_{X_{1}}

Expectation and Variance of a Random Vector: #flashcard

- For the 2×1 random vector $X = (X_1, X_2)'$, we have

$$\boxed{\mathbb{E}(X)} = \mathbb{E}_X(X) = \mathbb{E}_X \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}_X(X_1) \\ \mathbb{E}_X(X_2) \end{pmatrix} = \begin{pmatrix} \mathbb{E}_{X_1}(X_1) \\ \mathbb{E}_{X_2}(X_2) \end{pmatrix} = \boxed{\begin{pmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \end{pmatrix}}.$$

$$= \begin{bmatrix} \mathbb{E}_{X_1}[X_1] \\ \mathbb{E}_{X_2}[X_2] \end{bmatrix}$$

- Note that

$$\boxed{\underline{\mathbb{E}_X(X_1)} = \int \int x_1 f_X(x_1, x_2) dx_1 dx_2 = \int x_1 f_{X_1}(x_1) dx_1 = \underline{\mathbb{E}_{X_1}(X_1)}}$$

which follows from the relation $f_{X_1}(x_1) = \int f_X(x_1, x_2) dx_2$, i.e. marginalising the joint density with respect to X_2.

- The variance of the (bivariate) random vector is given by

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}\left\{[X-\mathbb{E}(X)]\,[X-\mathbb{E}(X)]'\right\} \\
&= \mathbb{E}\left\{\begin{bmatrix} X_1-\mathbb{E}(X_1) \\ X_2-\mathbb{E}(X_2)\end{bmatrix}[X_1-\mathbb{E}(X_1), X_2-\mathbb{E}(X_2)]\right\} \\
&= \mathbb{E}\begin{pmatrix} [X_1-\mathbb{E}(X_1)]^2 & [X_1-\mathbb{E}(X_1)][X_2-\mathbb{E}(X_2)] \\ [X_1-\mathbb{E}(X_1)][X_2-\mathbb{E}(X_2)] & [X_2-\mathbb{E}(X_2)]^2 \end{pmatrix} \\
&= \begin{pmatrix} \mathbb{E}\left\{[X_1-\mathbb{E}(X_1)]^2\right\} & \mathbb{E}\left\{[X_1-\mathbb{E}(X_1)][X_2-\mathbb{E}(X_2)]\right\} \\ \mathbb{E}\left\{[X_1-\mathbb{E}(X_1)][X_2-\mathbb{E}(X_2)]\right\} & \mathbb{E}\left\{[X_2-\mathbb{E}(X_2)]^2\right\} \end{pmatrix} \\
&= \boxed{\begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) \\ \mathrm{Cov}(X_1,X_2) & \mathrm{Var}(X_2) \end{pmatrix}}.
\end{aligned}$$

- This is called the variance (or also the covariance) matrix. This matrix is, by definition, symmetric.

- These definitions extend straightforwardly to column vectors with more than two rows.

This Var-Cov matrix is positive semi-definite.

Linear Transformations of a Random Vector: #flashcard

- Let X be a $p \times 1$ random vector, let A be a non-random $q \times 1$ vector, and let B be a non-random $q \times p$ matrix.
  Then

$$\underset{(q\times 1)}{Y} = \underset{q\times 1)}{A} + \underset{(q\times p)(p\times 1)}{BX}$$

  is a $q \times 1$ random vector.

- The expected value of this transformation is given by

$$\mathbb{E}(Y) = \mathbb{E}(A+BX) = A + B\mathbb{E}(X).$$

- The variance of this transformation is given by

$[B]_{q\times p}$ $\left[\begin{smallmatrix}\text{VarCov} \\ \text{Matrix}\end{smallmatrix}\right]_{p\times p}$ $[B']_{p\times q}$

$$\mathrm{Var}(Y) = \mathrm{Var}(A+BX) = \mathrm{Var}(BX) = \underline{B\mathrm{Var}(X)B'}.$$

$= \mathbb{E}[(BX - \mathbb{E}[BX])(BX - \mathbb{E}[BX])']$

$= \mathbb{E}[B(X - \mathbb{E}[X])(X - \mathbb{E}[X])'B'] = B\,\mathbb{E}[(X-\mathbb{E}[X])(X-\mathbb{E}[X])']B'$

Families of Distributions

Chi-Squared Distributions: #flashcard

χ^2 with DOF = 1 is just the $N(0,1)$ squared

- Let $Z \sim N(0,1)$ and define $Y = Z^2$. Then

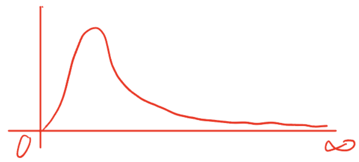
$$\boxed{Y \sim \chi_1^2.}$$

- In other words, Y has the **chi-squared** distribution with one degree of freedom.

- The density function for χ_1^2 is given by

$$\boxed{f_Y(y) = \left(\frac{1}{\sqrt{2\pi}}\right) y^{-1/2} \exp(-y/2) \qquad \text{for } y > 0.}$$

- Unlike the normal or t distributions, the chi-squared distribution is skewed to right, and its support consists of $(0, \infty)$.



- If $Z_1, Z_2, ..., Z_n$ are iid standard normal random variables, then

$$\boxed{X = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2}$$

Summation of n $N(0,1)$ squared
→ Chi-square with DOF = n

(**chi-squared** with n degrees of freedom), with density function

$$\boxed{f_X(x) = \left(\frac{1}{c_n}\right) x^{\frac{n}{2}-1} \exp(-x/2) \quad \text{for } x > 0}$$

for some normalisation constant c_n which ensures that $\int_0^\infty f_X(x)dx = 1$.

Closed under Convolution: #flashcard
f

- If $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$, and X and Y are **independent**, then

$$\boxed{X + Y \sim \mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).}$$

- Hence, the class of normal distributions is said to be closed under convolution.

i.e. we can add them up

- In fact, linear combinations of independent normals are always normal.

- If $X \sim \chi_m^2$ and $Y \sim \chi_n^2$, and X and Y are **independent**, then $X + Y \sim \chi_{m+n}^2$.

- Hence, the class of chi-squared distributions is said to be closed under convolution.

- **Not all distributions share this property.**

t and F Distributions: #flashcard
f

- The t and F distributions also frequently come up in hypothesis testing (though generally in so called "exact" inference, which is concerned with cases where theoretical results hold independent of the sample size).

- The **t distribution** (William Sealy Gosset, aka A.Student) is directly related to the normal and chi-squared distributions:

- Suppose $Z \sim \mathrm{N}(0,1)$ and $W \sim \chi^2_n$ are two **independent** random variables. Then

$$\boxed{T = \frac{Z}{\sqrt{W/n}} \sim t_n;}$$

  that is, T has a (Student) t distribution with n degrees of freedom.

- The **F distribution** (Ronald Fisher; George Snedecor) is obtained by a particular combination of two chi-squared random variables:

- Suppose $W_1 \sim \chi^2_m$ and $W_2 \sim \chi^2_p$ are two **independent** random variables. Then

$$\boxed{W = \frac{(W_1/m)}{(W_2/p)} \sim F_{m,p};}$$

  that is, W has an F distribution with (m,p) degrees of freedom.

- Graphically, the F distribution density looks very similar to that of the chi-squared distribution: both are skewed to right, and both have supports $(0,\infty)$.

Asymptotic Theory

Convergence of Random Sequences

Convergence in Probability: #flashcard
X_n converges in probability to c ($\text{plim}_{n\to\infty} X_n = c$ or $X_n \to^p c$) if and only if:

$$\forall\, \epsilon, \delta > 0, \exists n^* \in \mathbb{N}\; s.t.\; n > n^* \implies P(|X_n - c| < \epsilon) > 1 - \delta$$

equivalently:

$$\forall\, \epsilon > 0, \lim_{n\to\infty} P(|X_n - c| < \epsilon) = 1$$

Intuitively, this means we can let the probability of X being arbitrarily close to c with a probability arbitrarily close to 1 by choosing a large enough n.
This is weaker than both almost sure convergence and mean square convergence.
Note that probability limits need not to be non-random. It's possible for one RV to converge to another RV.

Almost Sure Convergence: #flashcard
X_n converges almost surely to c ($X_n \to^{a.s.} c$) if and only if:

$$P\left(\lim_{n\to\infty} |X_n - c| = 0\right) = 1$$

Intuitively, this means there exist some states of the world where $X_n \not\to c$, but they will never realise in the limit.
Almost sure convergence implies convergence in probability:

$$X_n \to^{a.s.} c \implies X_n \to^p c$$

Convergence in Mean Square: #flashcard
X_n converges in mean square to c ($X_n \rightarrow^{m.s.} c$) if and only if:

$$\lim_{n \to \infty} \mathbb{E}\left[(X_n - c)^2\right] = 0$$

Intuitively, this means, in the limit the expectation of mean square deviation from c is 0.
Convergence in mean square implies convergence in probability:

$$X_n \rightarrow^{m.s.} c \implies X_n \rightarrow^{p} c$$

Weak Law of Large Numbers: #flashcard
Let $X_1, \ldots, X_n$ be an *iid sequence* with mean $\mathbb{E}\left[X_i\right] = \mu$ and variance $Var(X_i) = \sigma^2$. Then:

$$\frac{1}{n} \sum_{i=1}^{n} X_i \rightarrow^{p} \mathbb{E}\left[X_i\right]$$

Proof:

- In particular, remember that for an iid sequence $X_1, \ldots, X_n$, defining $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$ we have

$$\boxed{\mathbb{E}(\overline{X}) = \mu \qquad \text{and} \qquad \operatorname{Var}(\overline{X}) = \frac{\sigma^2}{n}.}$$

- It follows that

$$\boxed{\mathbb{E}[(\overline{X} - \mu)^2] = \operatorname{Var}(\overline{X}) = \frac{\sigma^2}{n} \to 0, \qquad \text{as } n \to \infty.}$$

$\lim_{n \to \infty} \mathbb{E}[(\overline{X} - \mu)^2] = 0$

- This establishes m.s. convergence of $\overline{X}$ to c, implying $\overline{X} \xrightarrow{p} c$.

Law of Large Numbers

Kolmogorov's LLN: #flashcard
Let $X_1, \ldots, X_n$ be an *iid sequence* with mean $\mathbb{E}\left[X_i\right] = \mu$ (no variance requirement). Then:

$$\frac{1}{n} \sum_{i=1}^{n} X_i \rightarrow^{a.s.} \mathbb{E}\left[X_i\right]$$

Continuous Mapping / Slutsky's Theorem: #flashcard
If $X_c \rightarrow^{p} c$ and the function $g(.)$ is continuous at c, then:

$$g(X_n) \rightarrow^{p} g(c)$$

equivalently:

$$\text{plim}_{n \to \infty} g(X_n) = g(\text{plim}_{n \to \infty} X_n)$$

The only requirement on g is to be continuous at the limit. It can, for instance, be non-linear.
Also note that the expectation operator does not share this property (Jensen's Inequality).

- **First:** If $X_n \xrightarrow{d} X$ and the function g is continuous, then $g(X_n) \xrightarrow{d} g(X)$.
- This results is also referred to as a version of the **continuous mapping theorem.** *Remember we have already seen a version of this for convergence in probability.*
- **Second:** If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, and the function $g(\cdot,\cdot)$ is continuous at the point (X, c), then $g(X_n, Y_n) \xrightarrow{d} g(X, c)$. In particular:

$$\boxed{\begin{aligned} g(a,b) = a+b \quad &\Rightarrow \quad X_n + Y_n \xrightarrow{d} X + c; \\ g(a,b) = ab \quad &\Rightarrow \quad X_n Y_n \xrightarrow{d} Xc; \\ g(a,b) = a/b \quad &\Rightarrow \quad X_n / Y_n \xrightarrow{d} X/c. \end{aligned}}$$

- This result is referred to as **(Generalised) Slutsky's Theorem.**

Asymptotic Distribution

Convergence in Distribution: #flashcard
Let $\{X_n\}$ be a sequence of RV and Y be an RV. X_n converges in distribution to Y ($X_n \to^d Y$) iff:

$$\forall\, x \text{ where } P(X \leq x) \text{ is continuous}: \lim_{n\to\infty} P(X_n \leq x) = P(Y \leq x)$$

Equivalently:

$$\forall\, x \text{ where } F_X(x) \text{ is continuous}: \lim_{n\to\infty} F_X(x) = F_Y(x)$$

Convergence in probability implies convergence in distribution, but convergence in distribution does not imply convergence in probability! (Convergence in distribution is weaker)

$$\begin{cases} X_n \to^p Y \implies X_n \to^d Y \\ X_n \to^d Y \not\Longrightarrow X_n \to^p Y \end{cases}$$

Basic/Lindebery-Levy Central Limit Theorem: #flashcard
Suppose $\{X_n\}$ is an iid sequence with $\mathbb{E}[X_i] = \mu$ and $Var(X_i) = \sigma^2$. Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then:

$$\sqrt{n}\left(\frac{\overline{X} - \mu}{\sigma}\right) \to^d N(0,1)$$

Equivalently:

$$\sqrt{n}(\overline{X} - \mu) \to^d N(0, \sigma^2)$$

Note that $\{X_n\}$ can have any distribution, and the iid requirement can be relaxed with additional moment conditions.
Finite-sample Statement:In finite sample,

$$\sqrt{n}\left(\frac{\overline{X} - \mu}{\sigma}\right) \sim^a N(0,1)$$

where $\sim^a$ denotes "approximately distributed with". This holds $\forall\, n \in \mathbb{N}$, but the approximation becomes unreliable with small n.

Confidence Interval and Hypothesis Testing

Hypothesis Testing: #flashcard
f

- Type I error: rejecting a true H_0.
- Type II error: failing to reject a false H_0.
- Power of a test: $\mathrm{P}\left(\text{reject } H_0 | H_0 \text{ false}\right)$.
- Level of significance: P (reject $H_0|H_0$ true); the probability of making a Type I error.
- Even when H_0 is true, there is a very small yet non-zero probability that $(\widehat{\theta}-r)/\sigma$ will be VERY large. Seeing this very large value, we may reject H_0, which is a Type I error.
- The principle: choose a Type I error probability (level of significance) that you are comfortable with and do the test. Typically people choose 1%, 5% and (less typically) 10%.
- The corresponding cut-off value c_α is called the critical value.

Suppose:

$$\frac{\hat{\theta}-\theta}{\Sigma} \sim N(0,1) \tag{2}$$

$$\boxed{\begin{aligned} &\text{Case 1}: H_0: \theta \geq r \quad \text{vs} \quad H_1: \theta < r, \\ &\text{Case 2}: H_0: \theta \leq r \quad \text{vs} \quad H_1: \theta > r, \\ &\text{Case 3}: H_0: \theta = r \quad \text{vs} \quad H_1: \theta \neq r. \end{aligned}}$$

- Our test statistic is
$$\boxed{\frac{\widehat{\theta}-\theta}{\Sigma} \sim N\left(0,1\right).}$$
- Let α be the level of significance. Then, reject H_0 in favour of H_1 at α level of significance if
$$\boxed{\text{Case 1}: \frac{\widehat{\theta}-r}{\Sigma} < -c_\alpha, \qquad \text{Case 2}: \frac{\widehat{\theta}-r}{\Sigma} > c_\alpha, \qquad \text{Case 3}: \left|\frac{\widehat{\theta}-r}{\Sigma}\right| > c_{\alpha/2}.}$$
- Otherwise we fail to reject H_0 (we never say ACCEPT).