Lecture 6 - Confidence interval

- **Def - Point estimation:** Determination of a single value representing a best estimate of the parameter
- **Def - Confidence interval:** Determination of a range of values where the parameter lies
- **Def - Hypothesis tests:** The process of determining if the parameter lies in a given region

Statistical model

- **Def - Statistical model:** A statistical model is characterized by a family of probability laws on the same space. Every law depends on θ . The model is denoted by

$$M = \{p(\cdot\,|\theta), \theta \in \Theta\}$$

> It's like a blueprint that helps you understand how different parts of the system are related, and how they might change over time.

4.1 Sampling and statistics

- **Random sample:** If the random variables $X_1, X_2, ..., X_n$ are independent and identically distributed (iid), then these random variables constitute a random sample of size n from the common distribution.
- **Statistic:** : Let $X_1, X_2, ..., X_n$ denote a sample on a random variable X. Let $T = T(X_1, X_2, ..., X_n)$ be a function of the sample. Then T is called a statistic.

> Example: $T(X_1, \ldots, X_n) = \overline{X}n = \frac{1}{n} \sum i = 1^n X_i$

- **Realization:** Once the sample is drawn, then t is called *the realization of* T, where $t = T(x_1, x_2, ..., x_n)$ and $x_1, x_2, ..., x_n$ is the realization of the sample.

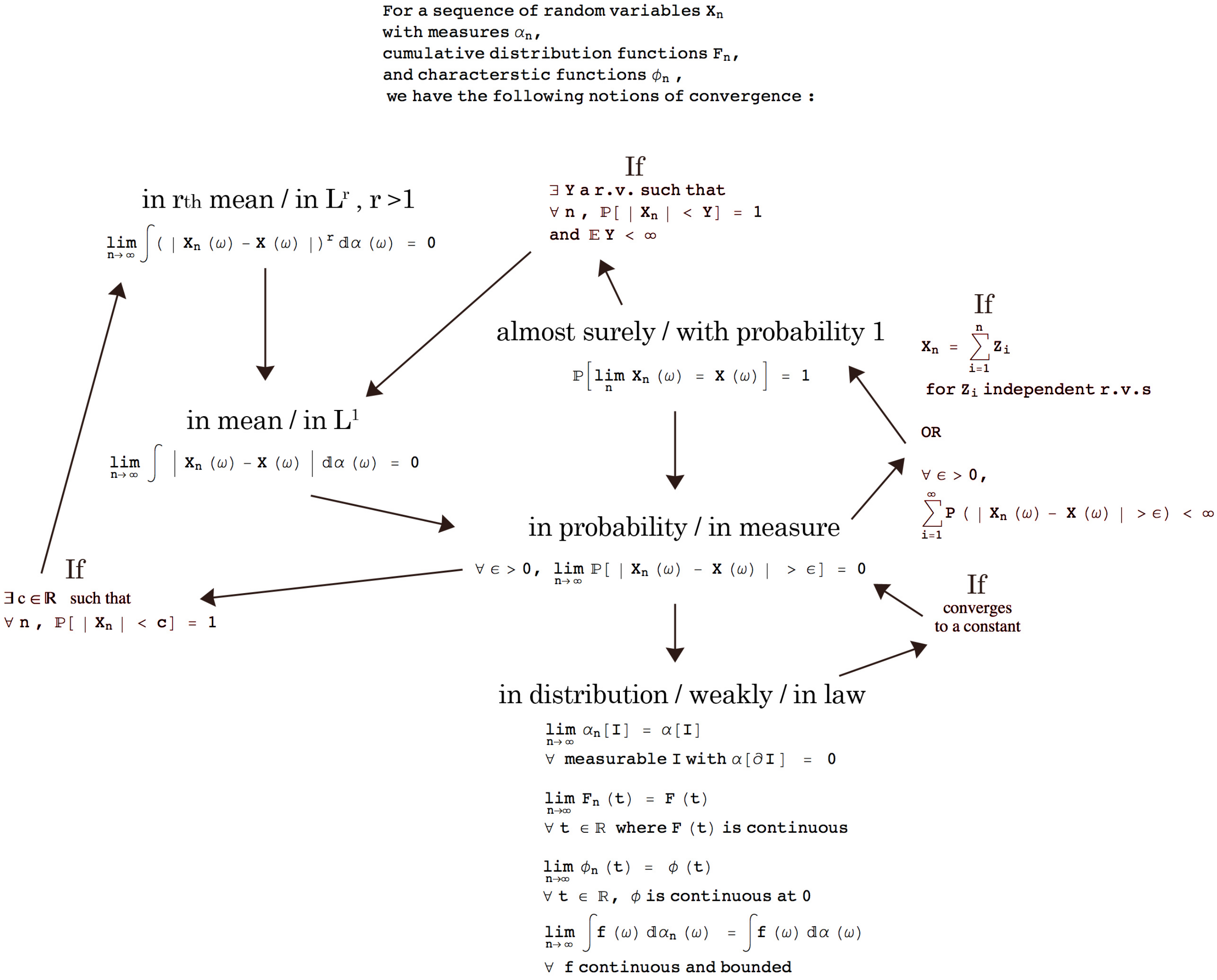
Estimator

- **Def - estimator:** An estimator $\hat{\theta}(X_1, ..., X_n)$ is a statistic which aims at estimating a quantity θ

> e.g. parameter, variance

- **Def - estimate:** The realization $\hat{\theta}(x_1, ..., x_n)$
- **Def - Bias:** $b_\theta(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$
- **Def - Mean squared error:** $\mathbb{E}((\hat{\theta} - \theta)^2) = b_\theta(\hat{\theta}) + Var(\hat{\theta})$

- A sequence of estimators $\hat{\theta}_n$ is **consistent** if $\hat{\theta}_n \to \theta$ as $n \to \infty$
  1. (Most used) Strong consistency if almost sure convergence
  2. Mean Square Consistency if L^2 convergence
  3. Consistency in probability if convergence in probability
- **Def - asymptotically unbiased:** $\mathbb{E}(\hat{\theta})_n \xrightarrow{n\to\infty} \theta$
- An estimator $\hat{\theta}_n$ of θ is said to be **asymptotically Normal** if $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in law when $n \to \infty$ to a centered normal distribution


For a sequence of random variables X_n
with measures α_n,
cumulative distribution functions F_n,
and characterstic functions ϕ_n ,
we have the following notions of convergence :
in rth mean / in L^r , r >1
$\lim_{n\to\infty} \int (|X_n(\omega) - X(\omega)|)^r d\alpha(\omega) = 0$
If
$\exists$ Y a r.v. such that
$\forall$ n , $\mathbb{P}[|X_n| < Y] = 1$
and $\mathbb{E}$ Y $< \infty$
almost surely / with probability 1
$\mathbb{P}\left[\lim_{n} X_n(\omega) = X(\omega)\right] = 1$
If
$X_n = \sum_{i=1}^{n} Z_i$
for Z_i independent r.v.s
OR
$\forall \epsilon > 0$,
$\sum_{i=1}^{\infty} P(|X_n(\omega) - X(\omega)| > \epsilon) < \infty$
in mean / in L^1
$\lim_{n\to\infty} \int |X_n(\omega) - X(\omega)| d\alpha(\omega) = 0$
in probability / in measure
$\forall \epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}[|X_n(\omega) - X(\omega)| > \epsilon] = 0$
If
$\exists c \in \mathbb{R}$ such that
$\forall$ n , $\mathbb{P}[|X_n| < c] = 1$
If
converges
to a constant
in distribution / weakly / in law
$\lim_{n\to\infty} \alpha_n[I] = \alpha[I]$
$\forall$ measurable I with $\alpha[\partial I] = 0$
$\lim_{n\to\infty} F_n(t) = F(t)$
$\forall$ t $\in \mathbb{R}$ where F (t) is continuous
$\lim_{n\to\infty} \phi_n(t) = \phi(t)$
$\forall$ t $\in \mathbb{R}$, ϕ is continuous at 0
$\lim_{n\to\infty} \int f(\omega) d\alpha_n(\omega) = \int f(\omega) d\alpha(\omega)$
$\forall$ f continuous and bounded

Law of large numbers and central limit theorem

- **Theo - Strong LLN:** $X_1, X_2, \ldots$ sequence of iid. integrable in (L^1) r.v with $\mu = \mathbb{E}[X_1]$ Then,

$$\bar{X_n} \xrightarrow[n\to\infty]{\text{a.s.}} \mu$$

- **Theo - Central limit theorem (CTL):** $X_1, ..., X_n$ sequence of i.i.d integrable r.v with $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = Var(X_1) < +\infty$ Then,

$$W_n = \frac{\bar{X_n} - \mu}{\mu/\sqrt{n}} \xrightarrow[n\to\infty]{distrib.} \mathcal{N}(0,1)$$

Method of moments (MoM)

Consists in finding:

- a function $g(\theta) \in \mathbb{R}^d$ invertible, with g^{-1} continuous
- a function $\psi(x) \in \mathbb{R}^d$ such that $\mathbb{E}(|\psi(X_1)|) < \infty$
- and have g such that $g(\theta) = E(\psi(X_1))$ for all $\theta \in \Theta$

Methods of moment estimator:

$$\hat{\theta} = g^{-1}(\frac{1}{n}\sum_{i=1}^{n}\psi(X_i))$$

Maximum likelihood estimator

- **Def - Likelihood of θ:** $L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$

> quantify how well a set of observations, $x_1, x_2, ..., x_n$, fit a given distribution described by the PDF, f_θ.

- **Def - Maximum likelihood estimator:** $\hat{\theta} = argmax L(\theta)$

4.2 Confidence intervals

- Let $X_1, ..., X_n$ be a sample of a r.v. X having pdf $f(x;\theta)$, $\theta \in \Theta$
- Let $0 < \alpha < 1$ be specified.
- Let L and U be two statistics.
- We say that the interval (L, U) is a $(1-\alpha)100\%$ confidence interval for θ if

$$P_\theta(\theta \in (L, U)) = 1 - \alpha$$

- $1 - \alpha$ is called the confidence coefficient of this interval.

> For example, if we have a sample of size n from a population with an unknown parameter, θ, and we want to construct a 95% confidence interval for θ, we would set $\alpha = 0.05$. This means that the probability of the true value of θ lying within the confidence interval is 0.95, or 95%.

Quantile reminder *(skip)*

The quantile reminder is calculated as the difference between the sample quantile and the theoretical quantile, divided by the standard deviation of the sample quantile. For example, let Q_α be the α-quantile of a given distribution, and let $\hat{Q}_\alpha$ be the sample quantile, corresponding to the α-quantile of the empirical distribution of the data. The quantile reminder is then given by:

$r_\alpha = \frac{\hat{Q}\alpha - Q\alpha}{s_\alpha}$

where s_α is the standard deviation of the sample quantile.

Exact confidence interval: Gaussian model

With n i.i.d samples, $X_i \sim N(0,1)$ and estimators
$\bar{X_n} = \frac{1}{n} \sum i = 1^n X_i$
and
$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X_n})^2$

- Assuming μ known,

$$Z_n = \sqrt{n} \frac{\bar{X_n} - \mu}{\sigma} \sim N(0,1)$$

- When μ unknown,

$$W_n = \sqrt{n} \frac{\bar{X_n} - \mu}{\sqrt{\hat{\sigma}_n^2}} \sim T_{n-1}$$

4.2.1 Confidence intervals for difference in means

Confidence Intervals for Difference in Means

> A confidence interval for the difference in means is a range of values that is likely to contain the true difference between the means of two populations, with a certain level of confidence. It is used to estimate the difference between the means of two populations when the means are unknown and the data are collected from random samples from each population.

- Using the CLT,

$$Z = \frac{\hat{\Delta} - \Delta}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} - \frac{\hat{\sigma}_2^2}{n_2}}}$$

where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the sample variance estimates are the sample variance estimates

> Δ represents the difference between the means of two populations, and $\hat{\Delta}$ is the sample estimate of this difference

- **Difference estimator:** $\hat{\Delta} = \bar{X_n} - \bar{Y_n} = \hat{p}_1 - \hat{p}_2$
- **Variance:** $Var(\hat{\Delta}) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
- $(1-\alpha)100\%$ **confidence interval:**

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} - \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$