Lezione 8 -

Random sample is a group of iid, random variables.

Sample mean, sum of r.v. over, arithmetic mean.

$$E(\bar{X}_n) = E(X_1)$$

$$Var(\bar{X}_n) = \frac{Var(X_1)}{n}$$

In general we don't know the distribution of S_n or $\bar{X}_n$, only in a few cases do we know.

Central Limit Theorem

any distribution.

Let $X_1, X_2, \dots X_n$ be a sequence of n iid r.v. with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, where $0 < \sigma^2 < \infty$, then

$$\lim_{n\to\infty} P\Big(\underbrace{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}_{\text{Standardised } \bar{X}_n} \le x\Big) = \phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \, du \text{ tho}$$

Distribution

$\sigma^2 \neq 0$, since then it would be a constant

Suppose n is large (≥ 50)

$$P\Big(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le n\Big) \approx \phi(x) \quad \text{mcal}$$

$Z \sim N(0,1)$

$$\boxed{\underbrace{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}_{Z} \underset{n,\text{ large}}{\overset{\text{approx}}{\sim}} N(0,1)} = \mathcal{L}(Z) \iff \boxed{\bar{X}_n \underset{n \text{ large}}{\overset{\text{approx}}{\sim}} N\Big(\mu, \frac{\sigma^2}{n}\Big)}$$

$\mathcal{L}$ law / Distribution of Z

The standardised version of $\bar{X}_n$, can be approximated by the standard gaussian.

$$\boxed{\iff S_n = X_1 + X_2 + \dots + X_n \underset{n \text{ large}}{\overset{\text{approx}}{\sim}} N(n\mu, n\sigma^2)}$$

We can only approximate the distribution, not the density.

$$\boxed{n \geq 50}$$

Example: Showing that it is applicable to the rest of a scientific experiment

$T =$ measurement of a variable

$T \sim N(\mu, \sigma^2)$

$T = \mu + \varepsilon$

rounding errors

$\varepsilon = X_1 + X_2 + \dots + X_n$ → all examples of error.

→ error due to atmosphere conditions

NOISE in sensor

By the CLT, $\varepsilon = X_1 + X_2 + \dots + X_n \underset{n \text{ LARGE}}{\sim} N(0, \sigma^2)$

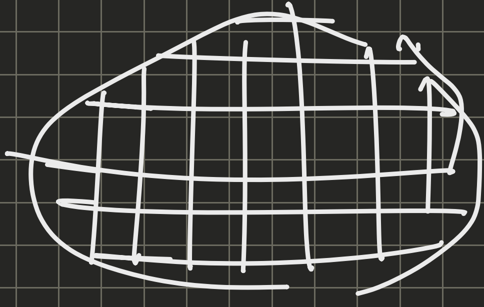
$\Rightarrow T = \mu + \varepsilon$

$\Rightarrow T \sim N(\mu, \sigma^2)$

↳ we will later be able to approximate these.

Exercise 1

Territory of the city of Despina



$n = 100$ cells each with 10000 inhabitants.

X_i = n° of underweight newborn [in year 2025] in cell i

$X_1, X_2, \dots, X_{100} \overset{iid}{\sim} Poi(\lambda = 0{,}7)$

a) Prob that in 1 cell at least 3 underweight babies?

$$P(X_1 \geq 3) = 1 - P(X_1 \leq 2) = 1 - P(X_1 = 0) - P(X_2 = 1) - P(X_1 = 2)$$

↳ same for any X_i, since we are asked for

$$= 1 - e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2!}\right] \simeq 0{,}0341$$

b) What is the approximate value of the probability that the total number of underweight babies in the whole city is at least equal to it's expectation

$$P(X_1 + X_2 + \dots + X_{100} \geq 70)$$

$$E[X_1 + X_2 + \dots + X_{100}] = 100 \cdot \lambda = 0{,}7 \cdot 100 = 70$$

$$= X_1 + X_2 + \dots + X_{100} \sim Poi(70)$$

$$X_1 + \dots + X_{100} \overset{approx}{\sim} N(n \cdot \lambda, n\lambda) = N(70, 70)$$

$100 \geq 50$ ✓

$$= \underbrace{1 - P(X_1 + \dots + X_{100} < 70)}_{\text{since this is the distribution function}} = 1 - P(X_1 + \dots + X_{100} \leq 69)$$

$$= 1 - P(X_1 + \dots + X_{100} \leq 69)$$

$$= 1 - P\left(\frac{X_1 + \dots + X_{100} - 70}{\sqrt{70}} \leq \frac{69 - 70}{\sqrt{70}}\right)$$

$$\underset{CLT}{\simeq} 1 - \Phi\left(\frac{69-70}{\sqrt{70}}\right) \simeq 1 - \Phi(-0{,}12) = \Phi(0{,}12)$$

$= 0{,}54776$

Approximation of the Poisson distribution function with the Gaussian distribution function

$$X \sim Poi(\lambda) \Rightarrow X \overset{approx}{\sim} N(\lambda, \lambda)$$

if $\boxed{\lambda \geq 20}$ this is valid

This is because we can see:

$$X \overset{d}{=} X_1 + \dots + X_n \;, \rightarrow X_1, \dots, X_n \overset{iid}{\sim} Poi\left(\frac{\lambda}{n}\right)$$

e.g $n=100 \geq 50$ ✓ $\quad \simeq N(\lambda, \lambda)$

This allows us to use the CLT.

Exercise 2

The energy efficiency of some type of solar panel

$$X \sim f(x,\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0<x<1 \\ 0 & \text{otherwise} \end{cases} \qquad 0<x<1 \quad \text{a.s.}$$

almost surely

$\theta = 0{,}75$

a) probability that the energy efficiency of a panel is larger than 60%

$$P(X > 60\%) = \int_{0,6}^{1} \theta x^{\theta-1} dx = x^{\theta}\Big|_{0,6}^{1} = 1-(0{,}6)^{0,75} \simeq 0{,}3183$$

b) $n=51$ panels:

? probability that at most 25 of the selected panels exceed 60% of the energy efficiency.

S = n° of panels which exceed threshold [60%]

$S \sim Bin(n=51, p=0{,}3183) \rightarrow P(S \leq 25)$ (*)

$> 60\% \rightarrow$ success

↳ n° of successes out of $\overset{n=}{51}$ Bernoulli trials.

$$Y \sim Bin(n,p) \qquad Y \overset{d}{=} V_1 + V_2 + \dots + V_n \;, V_i \overset{iid}{\sim} Be(p)$$

$$\overset{CLT}{\simeq} N(np, np(1-p))$$

↳ $E[V_i]$

We could also apply Poi as a limit of the Binomial distribution

We can use CLT if:

$n \geq 50$

$np > 5$

$n(1-p) > 5$

In this case:

$n = 51 \geq 50$

np and $n(1-p) > 5$ ✓

$$S \overset{approx}{\sim} N(np = 16{,}23333, np(1-p) = 11{,}06624)$$

$$\circledast \; P(S \leq 25) = P\left(\frac{S - np}{\sqrt{np(1-p)}} \leq \frac{25 - 16{,}23333}{\sqrt{11{,}06624}}\right) \overset{CLT}{\simeq} \phi(\quad) \simeq \phi(2{,}64) = 0{,}9959$$

Exercise 3

$$X_1, X_2, \dots, X_{51} \overset{iid}{\sim} U(0, 3.2)$$

$$\bar{X}_{51} = \text{sample mean} = \frac{X_1 + X_2 + \dots + X_{51}}{51}$$

$$E[\bar{X}_{51}] = E[X_i] = \frac{b+a}{2} = 1{,}6$$

$$Var[\bar{X}_{51}] = \frac{Var[X_i]}{51} = \frac{\frac{(b-a)^2}{12}}{51} = \frac{\frac{(3.2-0)^2}{12}}{51} = 0{,}0167$$

b) Compute and approximate value for the probability that the sample mean $\bar{X}_{51}$ is larger than 1.98

since this is the distribution function

$$P(\bar{X}_{51} > 1,98) = 1 - P(\bar{X}_{51} \leq 1,98)$$

$$\bar{X}_{51} \overset{\text{approx}}{\sim} N(1.6,\ 0,0167)$$

$n = 51 > 50$

$$\circledast \quad 1 - P\left(\frac{\bar{X}_{51} - E[\bar{X}_{51}]}{\sqrt{Var(\bar{X}_{51})}} \leq \frac{1,98 - 1,6}{\sqrt{0,0167}}\right) \overset{CLT}{\simeq} 1 - \phi(2,94) = 1 - 0,9984 = 0,0016$$

Populations and Samples

↳ Statistic is to learn information about a population, so we take a sample and measure the characteristic of interest of said sample.

We randomly select n subject of the population, for each we take the interested values and each measurement is a realization of the variable x_i, which is the actual value in their body at any moment.

We want to estimate the density based on measured values

Next lesson we start statistics.