Poisson's distribution assumptions → ① $n \to \infty$
② events are independent
③ p can be very, very less
sometime p = 0 (p does not matter)

→ POISSON'S DISTRIBUTION :

IF $n \to \infty$ (very large)

$$P(X = x) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

e = 2.72 (constant)

λ = no. of objects per
- time
- length
- area
- volume

→ Scales of measurement
① nominal
② ordinal } categorical data → shown in bar / pie chart (bar → gives freq; pie → gives rel. freq.)
③ interval and ratio

① nominal : cannot apply (mathematical - quantity) sort, add, sub, multiply division operations

ex: texts, phone nos.
can be numeric as well (ex : phone no.) → cannot perform any operations on it.

② ordinal : ordered
- ex : grades (A, B, C), quality (high, low), designation
- only sort applicable
- (mathematical ops) +, −, *, / → not applicable
- exception : Likert scale
In case of reviews, we take average reviews

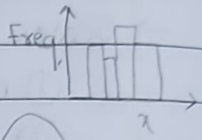
③ interval and ratio
- can apply all mathematical / arithmetic operations
- it is a scale ← ratio
- all numeric data

numerical → discrete (count)
→ continous (measure)

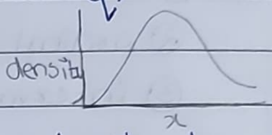
① discrete : count, variable
ex: no. of people
whole no. / integers

② continous : can be decimal also

if discrete repeats - it can be converted into categorical for better understanding



numerical data → ① histogram (freq)
② density plot

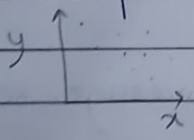


③ box plot - to understand 5 point summary

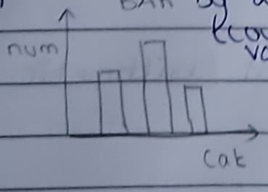
→ ① univariate → bar, histogram, pie
② bivariate

② bi-variate : a) numerical & numerical (vs)
b) num & cat (vs)
c) c & cat (vs)
d) time stamp & num (vs)

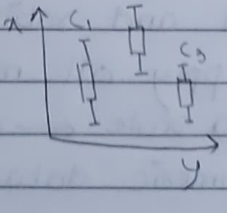
- scatterplot → covariance or correlation



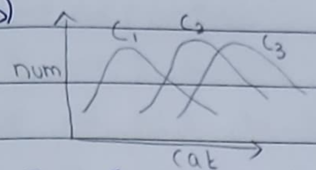
(a) BAR by aggregate (count, sum var)



(b) BOXPLOT by group

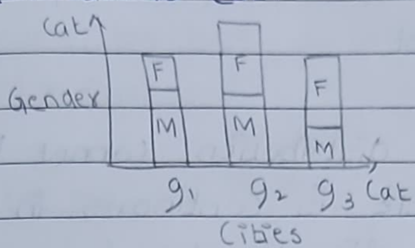


(b)

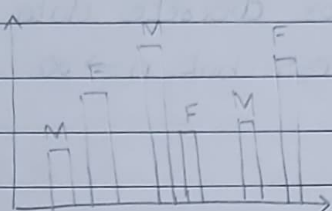


Density plot by group.

(c) cat. vs cat.

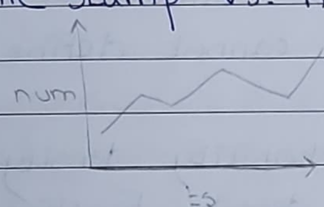


stacked bar chart.
F, M → gender



bar chart by group

d) time stamp vs. numeric.

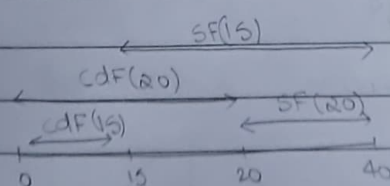


line chart.

metric : noise
seasonality
trends.

→ Data analysis ⟶ ① description - pop. data available
↳ ② inferential data - when population data is not available
- we consider sample data for analysis

→



to find prob. b/w 15 & 20

3 ways ① cdF (20) - cdF (15)
② SF (15) - SF (20)
③ 1 - [cdF 15 + SF 20]

classmate

Date

Page

→ Poisson's distribution:

$$P(X = x) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

$$= F(\lambda)$$

pmf → st.poisson.pmf

cdf → st.poisson.cdf

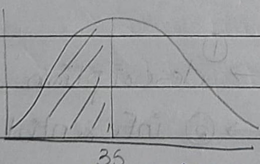
sf → st.poisson.sf

→ Cases when poisson's distribution cannot be applied:

1. when random variable is continous in nature
2. applicable only for discrete data (count)
3. when p is not less but $n \to \infty$

⇒ Continous prob. dist.

- when random variable is continous
- since data is continous, cannot define probability at a given point
- But you can define probability density using pdf → probility density function $\left(P(x)/x\right)$
- since continous, no difference between $P(X<35)$ and $P(X \leq 35)$ → so we always use $P(X \leq 35)$



35

$P(X > 35)$ and $P(X \geq 35)$ are same

Formula for bell curve:

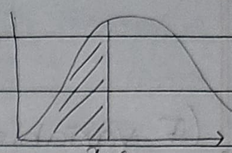
$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}}$$

↑

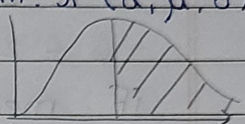
pdf

$e \rightarrow$ exponent
$\pi \rightarrow 3.14$
$\sigma \rightarrow$ std. deviation
$\mu \rightarrow$ mean

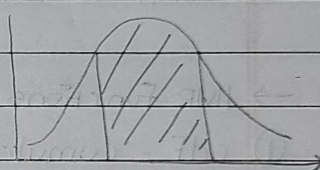
$\Rightarrow$ normal distribution :



- never touches x-axis.
- $P(x \le a) = \int_{-\infty}^{a} F(x).dx \rightarrow$ st. norm. cdf (a, μ, σ)
- $P(x \ge a) = \int_{a}^{+\infty} F(x).dx \longrightarrow$ st. norm. sf (a, μ, σ)



- $P(a \le x \le b) = \int_{a}^{b} F(x).dx$



a b

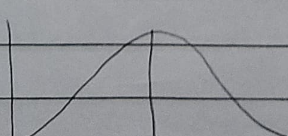
3 ways : ① st. norm. cdf (b, μ, σ)
 $-$ st. norm. cdf (a, μ, σ)
② st. norm. sf (a, μ, σ) $-$ st. norm. sf (b, μ, σ)
③ $1 -$ [st. norm. cdf (a, μ, σ) + st. norm. sf (b, μ, σ)]

Z :
$\rightarrow$ $F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

- $Z = \frac{x-\mu}{\sigma}$
- no units, units cancelled
- if normal distribution followed. $\rightarrow -4 \le z \le +4$
- standard normal distribution.

$Z = \frac{x-\mu}{\sigma}$

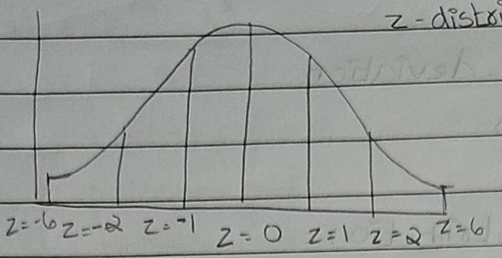
~~For~~ $Z = x$ when $\boxed{\mu = 0, \sigma = 1}$



mean = 0

$Z=0$

z/t1



(if variable follow normal dist.)

$-1 \le z \le +1 \Rightarrow 68.2\%$ (area shaded)

$-2 \le z \le +2 \Rightarrow 95.4\%$

$-3 \le z \le +3 \Rightarrow 99.72\%$ — area not shaded $3.14 \times 10^{-6} = 0.00000314$

$-6 \le z \le +6 \Rightarrow 99.999969$

if mean = median = 0 } to determine if it is normal dist.
k = 0

$1 - 0.00000314$ = area of graph

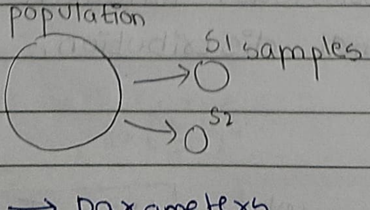
→ IMP Functions to use:

1. cdf - cumulative probability - ascending (lowest to given pt)
2. SF - cum. prob. - descending (given pt to highest)
3. isf - inverse survival function (inverse of SF)
4. ppf - point percentile function (inverse of cdf)

⇒ Inferential Statistics:

Inferential statistics →
1. sampling
2. estimation
3. hypothesis testing

1. Sampling

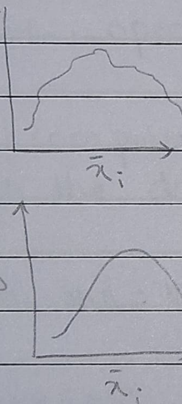


population → parameters
sample → statistic

		parameter	statistic
numerical	mean	μ	$\bar{x}$
	variance	σ^2	s^2
	SD	σ	s
categorical	proportion	P	$\hat{p}$

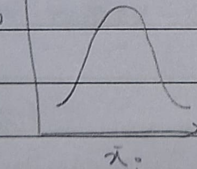
μ and $\bar{x}$ are not same
for all above parameter $\neq$ statistic

→ suppose, 100 samples
n = sample size
= 15 ~~object~~ observations



$\bar{x}_i$

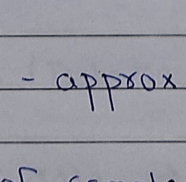
n = 30



$\bar{x}_i$

(after $n = 30$
normal dist. observed)
$n \geq 30$ approx. normal dist.

n = 50



$\bar{x}_i$

CLT (Central Limit Theorem)
aka law of large numbers

CLT :

i) $n \geq 30$ - approx. normal. dist.
ii) $\bar{x} \simeq \mu$
iii) S.D. of sample mean $\simeq \frac{\sigma}{\sqrt{n}}$ → std. error

~~iv)~~

Sampling

Probabilistic / Stochastic

① Scientific

a) simple random sampling

b) Stratified random sampling

c) Cluster sampling

d) systematic sampling

non - Probabilistic (not used)

① non - scientific

a) comfortable and convinient

b) quota sampling

c) judgemental sampling
biased

d) snowball sampling

ESA
=
DIFF.