Biostatistics
unit 1:

* Data collection:

[1] primary => "you" collect it for a specific use
 ↳ questionaire, survey, ...

[2] secondary => they're already collected; medical records, published research
 ↳ so might be collected for a diffrint reason.

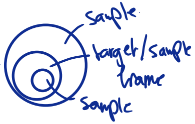
Data source ⇐ population —sampling→ sample [has to representitive of the population]
- X too small
- X too large

* Sampling Bias:
exclusion w/o scientific rations.

Types of sampling:
- representitive = probability
  - ↳ random, X Bias
  - P≠0
  - Simple Random
    - 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 => 2, 3, 5, 6, 8 "lottery method"
    - ⇑ ✓/X
    - - Time consuming
    - - expensive
    - - fast & efficient
    - - not always possible
    - - strong external validity "allows for generalizing our findings"
  - systematic
    - size / population = fraction
    - population / size = Interval
    - interval?
    - 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17
    - 6, 10, 1
  - Stratified
  - Cluster
- non-probability:
  - non-representitive (Bias)

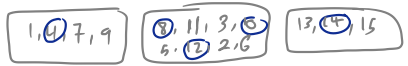
* process =>

1. Define population => 1 member is called an "element"
2. Frame => all accessible members / population elements ~>



3. Method
4. Size => cochran's formula = $\frac{Z^2 (pq)}{e^2}$
   - Z: table, depends on the CI
   - p: estimate
   - q: (1-p)
   - e: margin of error "precision"
5. excute the process

1. external validity
2. Representation. x/✓⇐

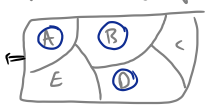


A, B, C : characteristics needed for the study & it has to be proportional to the population.

O. selected by randomly & systemically

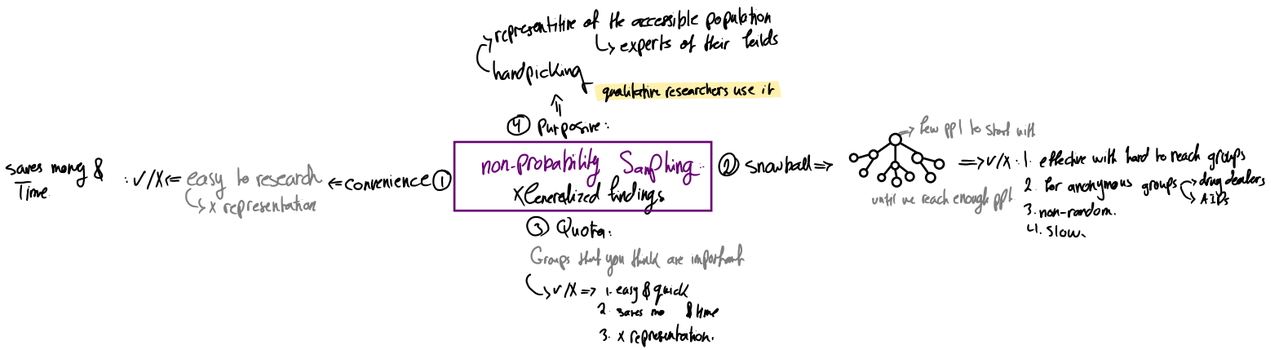
1. x sampling frame ✓/x ⇦
2. costly
3. external validity.
4. speed.
5. not always possible

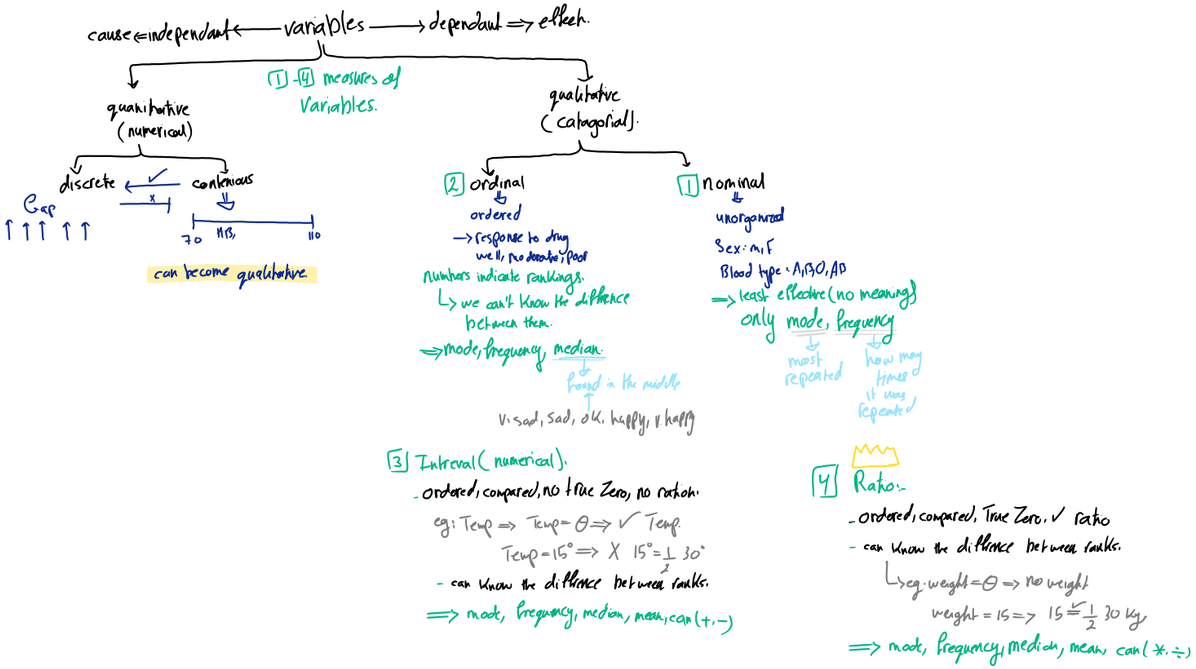
Cluster
↓
predefined groups.



Data from selected clusters
↳ all units who meet the criteria will be sampled.

usually used in interventional studies
↳ eg: immunization coverage



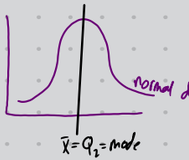


Discriptive statistics

parameter	Sample
μ	$\bar{X}$
σ^2	S^2
σ	S

parameter ← population ← Data → organized & classified into groups ⇝ organized
Sample ← Statistics ← → unorganized nor summarized ⇝ raw data

Tendency



shape distribution

[1] Mean ($\mu, \bar{X}$)

$\bar{X} = \frac{\sum X}{n}$

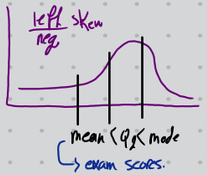
$\mu = \frac{\sum X}{N}$

always present

+, −, 0

unique

sensitive to extreme values



[2] Median

- middle value

odd ⇒ always in the middle

even ⇒ $\frac{\text{two middle digits}}{2}$

median is a more robust statistic in the case of extreme values.

frequency unaffected by extreme values

unique

[3] Mode :-

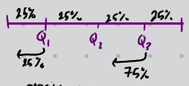
- most repeated value

not unique

could have no mode

location

[1] Quartiles



again ⇒

odd ⇒ always in the middle

even ⇒ $\frac{\text{two middle digits}}{2}$

[2] Deciles (10 parts)

[3] percintile (100 parts)

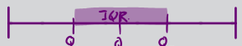
Dispersion

(spread of value).

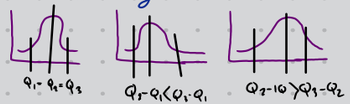
[1] Range (Max − Min).

- influenced by extreme values.

[2] IQR ($Q_3 - Q_1$)



- not influenced by outliers



[3] Standard deviation ($\sigma = \sqrt{var}$)

↑ std → ↑ variability => not a v. reliable data.

[4] variation ($\frac{\sum (X-\bar{X})^2}{n-1}$ / $\frac{\sum (X-\mu)^2}{N}$)

✱ Standard error(e) :-

$\frac{S}{\sqrt{n}}$ → ↑ e ⇝ ↓ error.

data classification

[1] Spacial

Geographical

[2] Temporal

over a period of time

" chronological "

[3] Qualitative

Simple

2 classes

Black | White

A	5
B	8
C	7
Total	20

A	5	0.25	25
B	8	0.4	40
C	7	0.35	35
Total	20		

Frequency

$(\frac{X}{T})$ Relative frequency

$(\frac{X}{T} * 100)$ percent

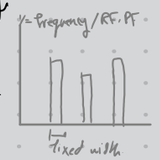
Manifold

different classes

B → dark grey | light grey

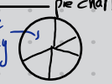
W → soft white | brown

Bar graph



pie chart

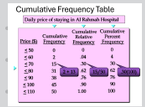
relative frequency



[4] Quantitative

frequency

- cumulative
- Histograme ogive
- Dot plot

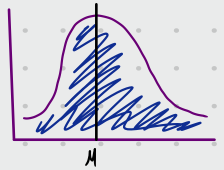


Approximate Class Width = (109 − 50)/6 = 9.83 ≈ 10

Daily price ($)	Frequency
50-60	2
60-70	13
70-80	16
80-90	7
90-100	7
100-110	5
Total	50

$\frac{\text{largest} - \text{smallest}}{T}$

normal distribution ⇒ $X \sim N(\underset{\text{mean}}{\mu}, \underset{\text{var}}{\sigma^2})$ ——— $Z = \frac{X-\mu}{\sigma}$ ———→ Standard normal ⇒ $Z \sim N(0,1)$

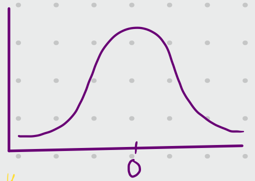


AUC = prob = 1

- + answ → above the mean
- − answ → below the mean

to say the Zscore = 2.3 ~> above the mean by 2.3 std.

* the numbers are on the opposite in the table



- $p(Z<K) = \square$
- $p(Z>K) = 1 - p(Z<K)$ or $p(Z>-K)$
- $p(a<Z<b) = p(Z<b) - p(Z<a)$
- $p(Z=K) = 0$

$\sigma_① < \sigma_②$

Central limit thereorm:-

sample is distributed normally

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ → std = $\frac{\sigma}{\sqrt{n}}$

→ ① from a normally distributed population.

② $n \geq 30$ ~> because of skewed distribution.

Inferential statistics
- → relation between variables
- → confidence interval
- → Test of hypothesis

Research hypothesis:

* qualitative has no hypothesis
- predict between dependant vs independant variable
- contain the population

~> difference ———→ significant ⇒ we reject the null hypothesis

H_0: null ~> hypothesized parameter "fail to reject"

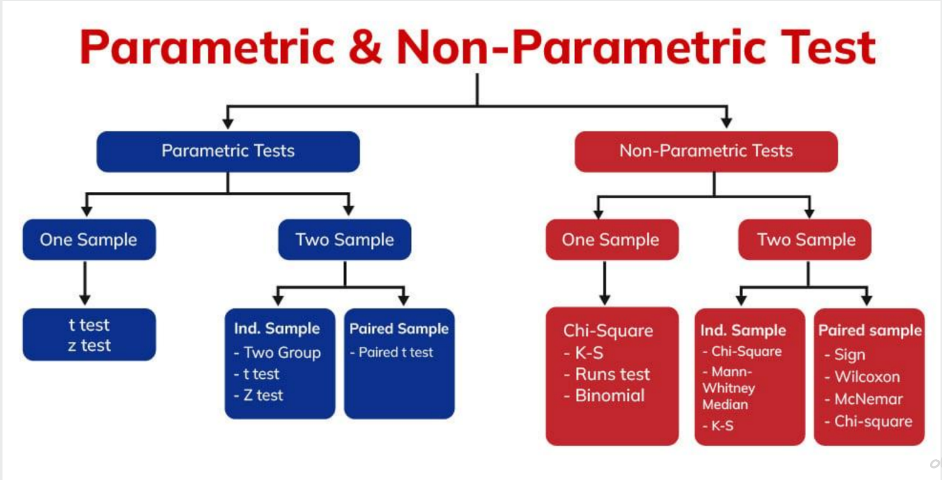
H_1: alternative ~> "accepted" if H_0 is rejected

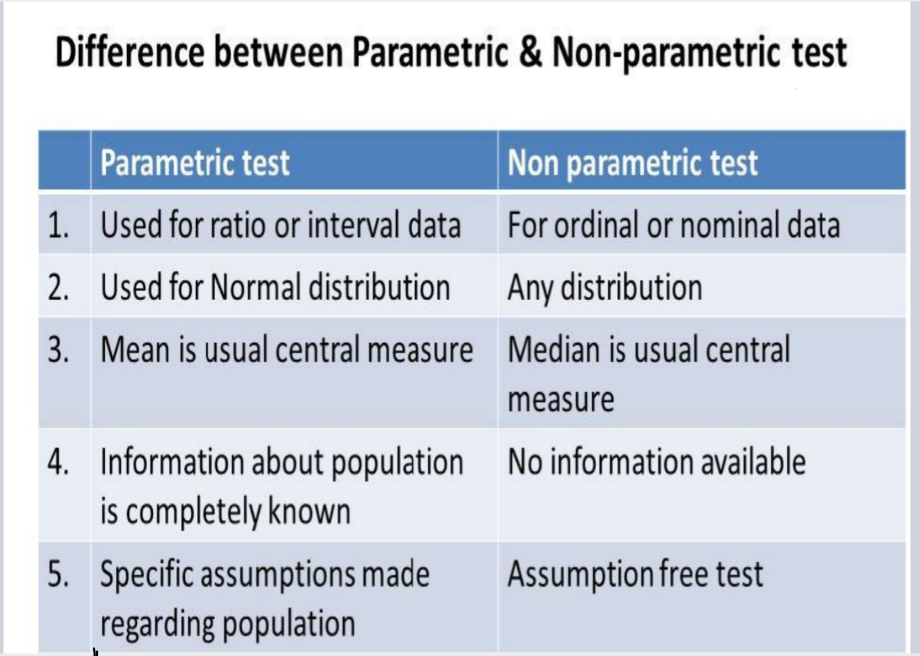
	H_0 true	H_0 false
Rej H_0	Type I (α) error	power of test ($1-\beta$)
Acc H_0		Type II (β) error

type I, α: p(rej H_0 when it's true)

type 2, β: p(acc H_0 when it's false) ⇒ more "wrong"

$1-\beta$ = the power of test — rej H_0 when it's false





Difference between Parametric & Non-parametric test

	Parametric test	Non parametric test
1.	Used for ratio or interval data	For ordinal or nominal data
2.	Used for Normal distribution	Any distribution
3.	Mean is usual central measure	Median is usual central measure
4.	Information about population is completely known	No information available
5.	Specific assumptions made regarding population	Assumption free test

→ T test
- → one sample
- → 2 samples with unknown variances.

→ ANOVA ~> 3≤ samples
- → data as an interval/ratio (quantitative).

Chi-square:

H_0: variables are independant (no association)

H_1: variables are dependant (✓ association).

observed values ⇒

expected value = $\frac{\text{ROW} * \text{col}}{\text{Total}}$

→ $\chi^2 = \sum \frac{(O-E)^2}{E}$ (O → observed, E → expected)

distance O & E ↑, χ^2 ↑ ⇒ rej H_0

O & E ↓, χ^2 ↓ ⇒ acc H_0

* p value ≤ 0.05 (α, significance) ⇒ we reject H_0 ⇒ ✓ association

p value > 0.05 ⇒ we fail to reject H_0 ⇒ ✗ association