Statistics

→ Statistics = The science of learning from data
→ It gives us methods & tools to:
- collect data
- Summarize data
- Analyze data
- interpret result
- make decisions under uncertainty

Statistic VS Statistics

→ Statistic → A single number or summary from data
Ex→ avg exam score, minimum Temp, % of retired people in a survey.

- Statistics → Academic disciple
→ Develops methods to analyze, interpret, and apply statistics
→ works with experts to make meaningful conclusions.

why statistics matters :

(1) summarizing data :→ makes huge data sets understandable.
Ex→ average, graphs, percentiles.

(2) Avoiding misleading info→ Provides structure to check if claims are valid.
Ex→ opinion polls with margin error

(3) Decission making under uncertainty
Ex→ Should a patient undergo preventive treatment if at risk?

(4) understanding variation → not just average, but spread of data.
Ex→ weather, election, demand for a product

(5) Prediction / forecasting
→ weather, elections, demand for a Product

6. Measuring difficult concepts
   - Easy : age, height
   - Hard : mood, Political ideology.
   - Statistics helps design rigorous measurement methods

7. Efficient data collection
   - Balance btw too much data (costly) vs too little data (inaccurate)

Where Data come from

-> Why Does Data source matter?
   -> knowing how data is generated is crucial- it influence which statistical methods are valid.
   -> Befor analyzing -> ask
      - where did the data come from
      - was it designed or organic?
      - can it be assumed i.i.d?

Two main types of data

1) Organic | Process data

- Data generated naturally by ongoing processes.
- usually large-sale (Big data), collected automatically
- Example:
  - -> Financial transaction, stock markets
  - -> netflix viewing history
  - -> web browsing activity
  - -> sports performance metrics
  - -> sensor data (Temp, pallution)

Features:
- -> massive in size -> requires computational resources
- -> often used to find patters & trends
- -> harder to clean & prepare for analysis

(2.) Designed data collection

- Data gathered from purposeful studies to answer specific questions.
- Example
  - → surveys (opinions from a sample of people)
  - → interviews with sampled individuals
  - → Extracting and coding specific tweets for research.

Features
- → smaller in size
- → Easier to work with
- → Designed with a clear research objective
- → more controlled & rigorous than organic data

Big data

→ Refers mainly to organic process data

→ challanges!
- Storing, cleaning, Processing.
- Required strongcs + statistical methods to make usable datasets

i.i.d. data (independend & identically Distributed)

→independent → each observation doesn't affect others

→identically distributed → all observations come from the same statistical distribution

Ex→ Final exam score in a large stats class - each students gredes independently, score may follow a normal distribution

→ Why i.i.d. is important

→ many statistical method assume i.i.d.

→ Allows estimation of:

→ mean, variance, percentile

→ Precision of these estimates

→ Non-i.i.d. cases (when assumptions break)

→ dependence : students cheat → scores not independent

→ Difference distributions: males vs females may have different score distributions.

→ Group effects: Students in different discussion sections may perform differently

in such cases ⊗

→ must use different statistical methods to handle dependencies.

→ Always ask:

1. did data come from organic process or designed collection?

2. can we assumed i.i.d?

→ Choice of statistical procedure depends on these answers.

Variable type

→ Why variable matters?

→ The type of variable determines:

→ How we summarize it (mean, frequency)

→ what analysis methods make sense

~~NHaNes~~

NHANES DATA

ID	BMI	Race	Age	Adult
62161	23.3	3	22	1
62163	17.3	5	14	0
62164	23.2	3	44	1
62165	27.2	4	16	0
62202	29.7	1	36	1
...	...	...	...	...

Race

1. Mexican American
2. Other Hispanic
3. Non-Hispanic white
4. Non-Hispanic Black
5. Other

Adult

0 - Age is less than 18

1 - Age is greater than equal to 18.

Two main types of variable

(1.) Quantitative (Numerical):

→ Represent measurable quantities

→ Arithmetic (like average, sum) makes sense

Subtypes:

→ continuous → Can take any value within a range

Ex→ BMI, height, weight, time to run a mile, age

→ Discrete → only specific countable numbers

→ Ex→ no of children in a household

numbers of pets

2. categorical (qualitative)

→ classifies items into groups
→ Arithmetic doesn't make sense
(can average race codes meaningfully)

Subtypes:

→ ordinal → has a natural order/ranking
ex → class standing (freshman → sophomore
→ Junior → Senior)

→ nomial → no natural order
ex → race, marital status, eye color etc