SDS – Unit 1

→ Data science → Interdisciplinary field that extracts insights from data & applies to solve problems across various domains

→ Applications → Image recog., speech recog., internet search, digital ads, price comparison websites, fraud & risk detection.

→ Data → Individual facts / items of info collected through an obsv.

Data → raw facts formatted in special way. Based on records, observations & unorganised.

Information → Data org. in such a way that they have add^n value beyond the facts themselves.

DB, Enterprise systems

→ Structured data → Data whose elems are addressable for eff. analysis. Stored in formal repo, → database. Ex: Relational data.

XML Email

→ Semi-structured data → Doesn't reside in database but has some org. prop. that make it easier to analyse. Ex: XML Data.

Analog data, GIS, AV systems

→ Unstructured data → Not org. in predefined manner or have data model. Ex: word, pdf

→ Knowledge → Body of guidelines & procedures to select/org./manipulate data to perform specific tasks on data.

→ Science → Enterprise that org. knowledge to form testable exp. & pred. of the universe.

→ 5 Vs of Big Data →
- → Volume → amt of data collected
- → variety → types of data generated
- → velocity → speed at which data is generated
- → veracity → trustworthiness of data
- → value → business value of data
- → variability → ways of using/formatting data

→ Using data →
- → Explore → find patterns
- → Predict → make informed guesses
- → Inference → quantify what you know
- bring data to model, further from prob, data given

→ Statistician → use math, eqns & stats model to ~~data~~ analyse data & make conclusions.

→ Data scientist → deliver actionable results & deploy model to the prod systems.

bring model to data; embedded in prod, need to get data

→ Stat analysis → collecting, exploring large amt of data to discover underlying patterns & trends.

→ Pop → entire collection of obj / outcomes from which info is sought.

→ Sample → subset of pop → actually obs outcomes / obj.

→ Sampling → process of selecting obs in order to make inference generalised to pop.

→ Reasons for sampling →
- → necessity → can't study whole pop.
- → practicality → more eff
- → cost eff.
- → manageability → small datasets easier
- → time saving

→ Chars of sample →
- → rep of pop.
- → app. size
- → unbiased
- → randomly selected
- → economical
- → goal-oriented

→ Tangible pop. → members are phy. obj., always finite & involves counting.

→ Conceptual pop. → not phy. obj., result of measure, not-well defined, size → large

→ Target/theoretical pop. → entire grp on which researches interested in generalising conclusions.

→ Study/accessible pop. → grp on which researches actually apply conclusions to.

→ Sampling frame → list of items/events from which pop. respondents drawn to measure. Must be rep of pop.

→ Prob sampling → every unit in pop has chance/prob of being selected → Equal prob selection (EPS) / self-weighting.

→ Non-prob sampling → not EPS. Selection based on assumptions on pop of interest, likely to prod biased sample & restricts generalisation

→ Simple random sampling →
  → EPS, selection using random no/ system/lottery.
  → Sampling frame → whole pop,
  → Random → rep of pop., used when pop ↓.
  → Small pop → lottery large → system gen no/.

Prob = $\frac{n}{N} \times 100$ n → sample size, N → pop size

→ Adv →
  → Easy to use calculate
  → EPS, rep of pop
  → ↓ sampling error
  → min knowledge of pop needed in adv

→ Disadv →
  → not useful when pop↑
  → may not rep min subgrp
  → can't be used when units → heterogeneous

→ Systematic sampling →
→ arr, pop, acc to ordering scheme & select, elem at reg, intervals. first elem → random
→ select every k^{th} element, $k = N/n$.
→ EPS. Not simple random → diff, subsets, diff prob
→ Adv, →
→ easy to select, sampling frame & easily
→ cost eff, entire pop, evenly sampled
→ ↓ chance of data contamination.
→ Disadv, →
→ might lead to bias.
→ diff, to assess precision of estimate.
→ where of elem, b/w k^{th} elem
Internally homogeneous
→ Stratified sampling → based on same shared char
→ Pop, div, into 2 grp, → strata & then sampled, using simple random.
→ Don't overlap → rep, whole pop.
→ Adeq, rep, of min, subgrp.
→ Adv →
→ ↑ rep, easy
→ ↑ stat, eff, ↑ precision
→ Disadv →
→ time consuming, ↑ cost
→ classification errors
Internally heterogeneous
→ Cluster sampling → One stage / Two stage
→ pop, div, into non-overlapping clusters & then each cluster → randomly selected.
→ Heterogeneous members in each grp.

→ Adv/ →
→ useful for geo/ diverse pop,
→ ↓ travel cost, ↓ resources, ~~[illegible]~~
→ more feasible, administration easier.
→ Disadv/ →
→ ↓ eff, ↑ cost than simple random
→ ↑ sampling error, prone to bias.

→ Convenience sampling →
→ Sampling from pop, close to hand/ convenient to researcher.
→ Adv/ → ↓ cost, ↓ time, simple, easy to imp,
→ Disadv/ → bias, not rep, of pop, sampling error

→ Judgemental Sampling →
→ Researcher chooses sample based on who they think is appr for the study. Used when limited no/ of ppl, that have expertise on area being sampled.
→ Adv/ → ↓ time, real time results, no special knowledge of stats needed
→ Disadv/ → prone to error, ↓ reliability ↑ bias, can't generalise findings

→ Quota sampling →
→ Split into mutually exclusive grps, & then judgement applied on each based on predetermined characteristics of population.
→ Adv/ → ↓ cost, fast, convenient, imp, rep, of subj grps
→ Disadv/ → impossible to det, sampling error, bias, not possible to make stat, inf,

→ Snowball sampling →
→ Survey subj, on referral from other survey subj
Existing subj, asked to nominate further subj
Used when sampling frame diff to ady
→ Ady → ↓ cost, simple, little planning, ↓ workforce
→ Disady → bias, impossible to defy sampling error rep, not guaranteed

→ sample statistic → piece of info, from fraction of pop,
→ Pop ~~statistic~~ parameter → " refers to whole pop

→ Sampling errors/random errors →
→ discrepancy b/w sample stat, & pop, param
Occurs when sample not rep, of pop,

→ Non-sampling error/ systematic error →
→ Result of errors made during data collection & processing → mainly due to sampling bias & non-response bias
↓
when sample not rep, of pop,

→ Selection bias → ~~extended~~ grp, → ↑, ↓ prob, than others
→ Non-response bias → sampling bias due to absence of certain obj/ grps,
→ Sampling variation → when two diff samples from same pop, differ from each other. items
→ Independence → knowing value of some item helps predict values of other items

- Data → Qualitative → Nominal, Ordinal
- Data → Quantitative → Interval, Ratio
- Quantitative → recorded on numeric scale.
  - Discrete
  - Continuous
- Qualitative → not recorded on numeric scale but on categories
  - Nominal → labels for categories w/o direction/order. Ex: Male, female. Can't be manipulated with math ops.
  - Ordinal → setting up inequalities, but no actual value → labels for categories with order. Ex: first, second, third.

Ordinal data → order matters but not difference b/w them
Ex: $5 < 7$ but diff b/w $3 \& 5 \neq$ diff b/w $5 \& 7$.

→ Discrete → limited no. of possible values & jumps b/w data. Ex: integer nos.
Bar graphs used to rep. discrete data.

→ Continuous → can take any value within an interval
Measured on scale or continuum.

→ Interval → data measured on scale with each pt. equidistant from each other. Measurable, ordered but no meaningful zero. Ex: Mean, Median, Mode, min, max, temp.

→ Ratio → classifies & ranks data, has true zero. Can't be -ve, manipulated by math ops.
Ex: Weight, speed, age

Page No.

→ Attribute/variable → characteristic/feature of data obj. Vary from one obj to another.

→ Prop of attributes →
  → Distinctness : =, ≠
  → Order : >, <
  → Addition : +, −
  → Multiplication : *, /

→ Nominal → distinctness
→ Ordinal → distinctness & order
→ Interval → distinctness, order & addition
→ Ratio → all 4.

→ Obs/sample → measure/survey members of sample w/o affecting them.
→ Controlled study → One grp → some treatment, Other grp → no treatment.
→ Web scraping →
  → use of program/algorithm to extract & process large amt of data from the web
  → python libraries → Beautiful Soup, Requests, Scrapy
→ Web scraper → program that goes to webpages, downloads contents, extract data out of contents & then saves data to file/db.
→ Web crawler → downloading/storing contents of large no of websites by following webpage links.

→ Web scraping process →
  → Req → resp
  → Parse and extract
  → Transform data
→ Python libraries →
→ Requests / urllib2 → download webpages
→ BeautifulSoup / LXML → Parsing HTML
→ Pandas / Matplotlib / ~~Excel~~ Numpy / Scipy → handling / transforming data.
→ Pandas / csv → read csv files

→ Data Quality →
  → Dirty data → incomplete, noisy/error, inconsistent, intentional,
  → Imp, data quality → better decision making, more confidence, ↓ risk, ↑ results.

→ Data cleaning → detecting/correcting corrupt/inaccurate data from dataset.
  → Detect, resolve, treat

→ Outliers → data distinctly diff, from other pts,

→ Stats → science of data → collecting, classifying, analysing numerical info, & providing conclusions for it.
→ Descriptive stat, → org, summ, display of data → final result table, graph, etc
→ Inferential stat, → using sample to draw conclusions about pop → final result prob, scores
  ↓ estimating parameters
  ↓ taking sample stat & saying smthg abt pop param
  ↓ hypothesis test
  ↓ sample data used to answer research qs,

→ Handling missing data →
→ Drop obs → only when we sure missing data is not informative.
→ Drop feature →
→ Impute missing → replace with avg / median / weighted mean

→ Measures of central tendency →
→ Mean ($\bar{x}$) = $\frac{\Sigma x_i}{n}$, $\frac{\Sigma x_i f_i}{\Sigma f_i}$, avg value
→ Median → middle value
→ Mode → most common value

→ Trimmed mean → arrange samples in order & trimming equal no, of them from each end.
Used for large, extremely skewed distributions
sample size n, p% trimmed mean then no, of data pts, to be trimmed → np/100

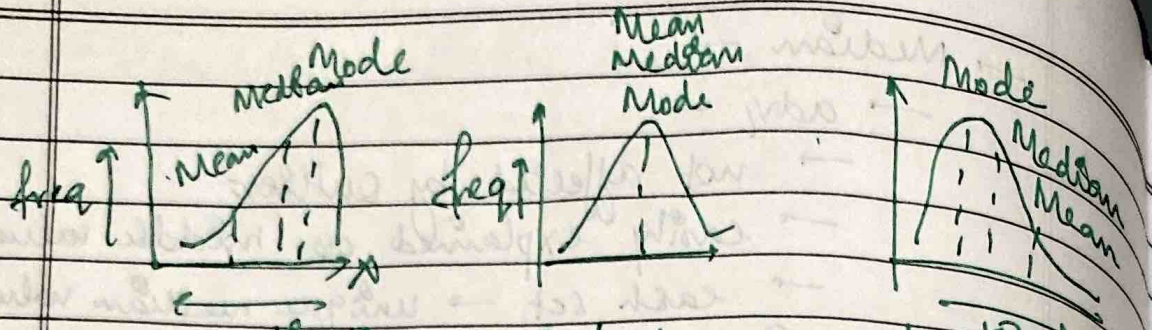
→ Mean →
→ adv, →
→ accounts for all available info, can be combined with other means to find overall mean.
→ easy way to rep, entire data with single no,
→ each data set → unique mean value
→ disadv, →
→ very sensitive
→ easily affected by outliers
→ only used with interval/ratio data.

→ Median →
  → adv, →
    → not affected by outliers
    → easily explained as middle value.
    → each set → unique median value.
  → disadv, →
    → value perceived as it is.
    → cannot be utilised for algebraic treatment

→ Empirical formula: Mode = 3Median − 2Mean

→ Mode →
  → adv, →
    → quick & easy to compute
    → unaffected by outliers
    → used on any level of measurement
  → disadv, →
    → terminal statistic
    → irrelevant if no repeating values
    → many values of have same count → irrelevant

→ Skewness → measure of assymetry of distribution about its mean.
→ Symmetric distribution → left & right sides equally balanced about mean. Mean, median & mode same.
→ Skewed distribution → " not balanced. Mean & median further toward skew than mode.
↑ distance b/w median & mean → ↑ skewness



left skewed / -ve skewed (mode > median > mean)

Perfectly symmetric

right skewed / +ve skewed (mode < median < mean)

Nominal → mode
Ordinal → median
Interval/ratio (not skewed) → mean
" (skewed) → median

→ Measures of dispersion → (Absolute, Relative) → how homogeneous/heterogeneous data is.
→ Absolute → exp. ran in terms of avg of deviation (the mean/SD, etc.)
→ Relative → distribution of two/more datasets.
Ex: coeff of mean, coeff of SD, etc.

→ Range → highest value − lowest value
→ adv →
→ easy to calculate
→ simplest of measure of dispersion
→ indy of change of origin
→ disadv →
→ affected by fluctuations
→ dep. on change of scale
→ affected by outliers
→ not reliable

→ Percentile → comp, b/w particular value & rest of the values in dataset.

percentile rank $R = (P/100) * (N+1)$

→ Quartile → divide list of nos into quarters.

25th percentile ← Q_1 → middle no, b/w smallest no, & median

50th ← Q_2 → median

75th ← Q_3 → middle no, b/w median & largest no,

$Q_1 = 0.25(n+1)$

$Q_2 = 0.5(n+1)$

$Q_3 = 0.75(n+1)$

→ Interquartile Range $= Q_3 - Q_1$

↓

distance b/w 25th & 75th percentile

→ variance → measure of spread of recorded values

↓ → small → closer to mean

large → far from mean

Sample variance $s^2 = \Sigma (x - \bar{x})^2 / (n-1)$

$s^2 = \frac{1}{n-1}\left(\Sigma x^2 - n\bar{x}^2\right)$.

Population variance $\sigma^2 = \Sigma (x - \bar{\mu})^2 / N$

→ Standard Deviation → deviation of elem of dataset from mean value of distribution

→ Histogram →

mode → $\frac{\text{lower} + \text{upper}}{2}$ (highest bin height)

Q_1, Q_3, median → $\text{lower} + \left[\frac{(\text{pos} - \text{cf})}{\text{freq}}\right] \times$ class width

Q_1 → pos $= 0.25(n+1)$, median → pos $= 0.5(n+1)$, Q_3 → pos $= 0.75(n+1)$

mean → $\frac{\Sigma f \left(\frac{\text{lower} + \text{upper}}{2}\right)}{N}$

→ Data visualisation → translating data to visual context.

→ Histograms → used for continuous data. area prop. freq, width = class interval

learn histogram construction

$$\text{Bin Size} = 2\left[\frac{IQR(x)}{3\sqrt{n}}\right]$$

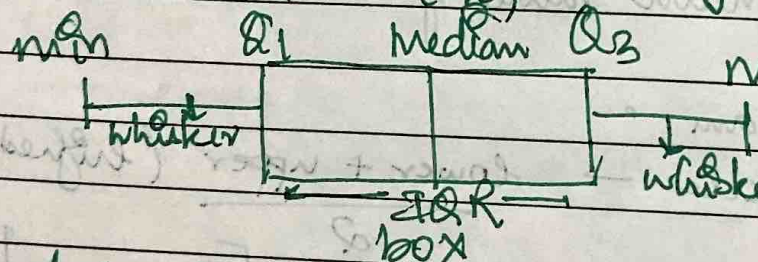
→ Scatterplot → shows whether two var correlated. Explanatory var (ind.) → x-axis & response var. (dep.) → y-axis.
More data cluster along one line → more the variables are correlated.

Pearson coefficient $r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left(n\Sigma x^2 - (\Sigma x)^2\right)\left(n\Sigma y^2 - (\Sigma y)^2\right)}}$

x → value of data point on x-axis
y → " " y-axis
n → no. of data points

r > 0 → +ve correlation
r < 0 → -ve "
r = 0 → no "
r = +1/-1 → perfect correlation

→ Boxplot → way of summarising set of data on interval scale.



Useful when distribution skewed & helps find Outliers.

min = Q1 - 1.5 (IQR)
max = Q3 + 1.5 (IQR)

+ve skew → Q2 closer to Q1 than Q3

-ve skew → Q2 closer to Q3 than Q1

→ Barchart → summarises categorical data.
x-axis → categories, y-axis → counts, perc, etc.

→ Heatmap → shows magnitude of phenomenon as color in 2D. Show relation b/w two variables plotted on each of the axes.
Heatmaps used in search engine optimisation
↓
process of improving quality & quantity of web traffic to website/webpage from search engines

→ Correlograms → variant of heatmap that replaces variables on the axes with numeric variables.

→ Dot plot/line plots →
→ x-axis → range of values, dots → obs.
→ Dot-height → rep. freq of obs. values
→ Useful when n → small & many repeated values present.