statistics

FOR

DATA SCIENCE

UNIT-1

feedback/corrections: vibha@pesu.pes.edu

Vibha Masti

Sampling Methods

RANDOM SAMPLING

1) Simple random sampling

- assign numbers to members of population and select randomly
- good for small population

ADVANTAGES

easier, low error, no prior information required

DISADVANTAGES

can be biased, not proportionate, hard to scale

2) Stratified random sampling

- population proportion reflected in sample
- divide population into strata/groups (gender, hair colour, age etc)

ADVANTAGES

enhanced representation, more scalable and efficient

DISADVANTAGES

classification error, time consuming, expensive

example: a student council surveys 50 students by getting random samples of 25 juniors and 25 seniors

3) Systematic Sampling

- Find the k^{th} value

ADVANTAGES

easy to select, evenly spread sample, cost effective

DISADVANTAGES

biased, no equal chance, ignored elements

example: a principal takes an alphabetised list of students and picks every fourth student from a random starting point

4) Cluster Sampling

- Population divided into non-overlapping areas (clusters)
- Each cluster microcosm of population

ADVANTAGES

convenient for geographically dispersed populations, simplified administration

DISADVANTAGES

less efficient statistically, higher sampling error, more problems

example: airline company randomly selects 5 flights and surveys everyone on them

NON-RANDOM SAMPLING

1) Convenience / Accidental

- subjects for sampling easily available
- when population not clearly defined

ADVANTAGES

easy to select, saves time and money

DISADVANTAGES

biased, sampling errors, cannot generalise

2) Judemental Sampling

- researcher chooses / is related to sample based on their judement

ADVANTAGES

minimum time

DISADVANTAGES

selection bias, sample size

3) Quota Sampling

- non-probability equivalent of stratified
- till quota is met

ADVANTAGES

minimum time

DISADVANTAGES

bias

4) Snowball Sampling

- for rare characteristic / difficulty
- from initial subject, referrals

ADVANTAGES

lowers cost

DISADVANTAGES

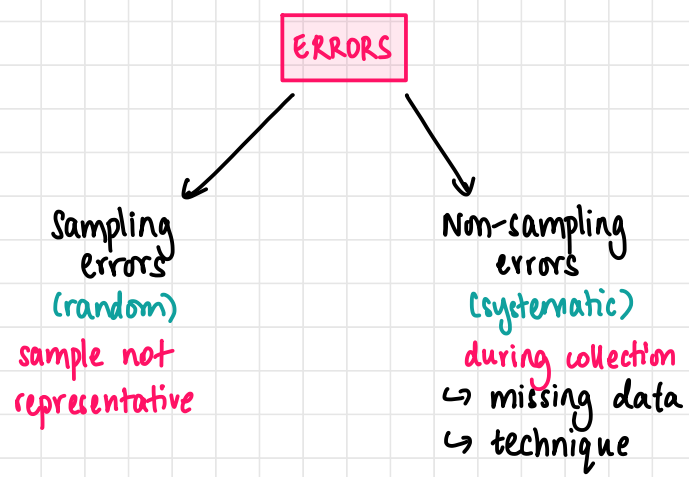
bias

selection BIAS

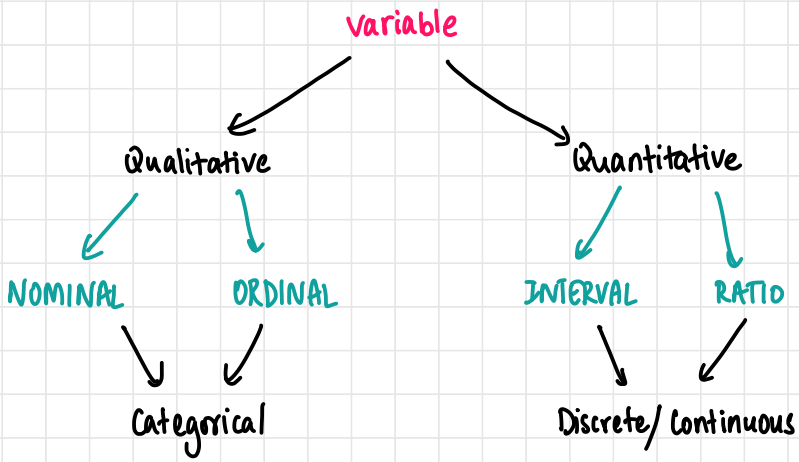
- Leave out hard to reach people
- Replace with accessible people
- Outdated sample frame

nonresponse BIAS

- people who do not respond to survey



TYPES OF DATA



NOIR

- in increasing order of accuracy, powerfulness of measurement, precision and application

1. Nominal

- name and countable (frequency)
- data is alphabetical or numeric only
- only counting & grouping
- no other arithmatic operations
- bar graphs are appropriate
- keeping track of objects / people / names etc

2. Ordinal

- comparision between types allowed
- order matters, difference between values not
- eg: level of spiciness, satisfaction, pain scales etc

3. Interval
   - differences also meaningful
   - no absolute zero
   - eg: standardised scores, temperature

4. Ratio
   - all mathematical operations, clear absolute 0
   - eg: height, weight, real scores

Note: salary/money is typically discrete

TYPES OF STUDIES

Observational (surveys)

- do not control/interfere with sample/pop
- no treatment is given

Experiments (control + exp/treatment)

- sample/pop split into 2 groups
- treatment given to experimental group, control group given a placebo

Web Scraping

- Process of getting data from specific websites
- Not to be confused with web crawling (automatically scans through www – search engines)
- Using APIs, scrape data from certain websites
- Using Beautiful Soup to pull contents from HTML or XML pages (make sure you have permission!!!)

Process

1. Request - response
2. Parse & extract
3. Transform the data

Data Cleaning

- Missing data / discrepancies should be checked
- NaN values
- Inaccurate, incomplete, irrelevant, inconsistent
- Outlier – identify using Box plot / data summary / bar chart (categorical)
- Cleaning data is very important
- Format standardisation (eg: date, address)

Missing Data

- impute - use other values (mean/mode)
- drop - delete

Types of Statistics

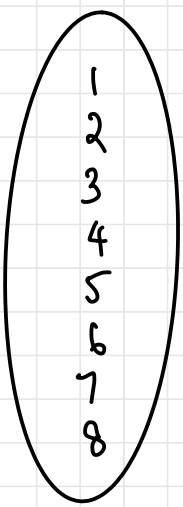
- descriptive: organise, summarise data, tables, graphs, central ten
- inferential: draw conclusions, hypothesis testing

DESCRIPTIVE STATISTICS

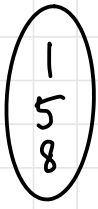
Parameter
population

Statistic
sample

Should be rep. of pop

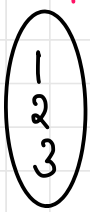


N=10



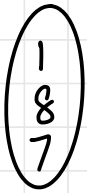
n=3

$\bar{x} = 4.67$ (mean)



n=3

$\bar{x} = 2$ (mean)



n=3

$\bar{x} = 5.33$ (mean)

statistic

parameter

$\mu = 4.5$ (mean)

Difference b/w μ and $\bar{x}$: sampling error

Central Tendency

cut off end outliers

- 3 measures
- mean, median, mode, trimmed / truncated mean

Trimmed mean

- 10% trimmed → 10% up & 10% down (80% used)

x x x x x x x x x x

Q: 5, 4, 7, 6, 8, 10, 11, 0, 7, 18 find mean, median, mode, trimmed mean (10%, 20%)

$n = 10$

mean = 7.6

median = ?

0, 4, 5, 6, 7, 7, 8, 10, 11, 18

median = 7

mode = 7

10% trimmed mean ⇒ $(n)(0.10) = 10 \times 0.1 = 1$ cut off each

= 7.25

20% trimmed mean ⇒ cut off 2 each

= 7.167

Q: 30, 75, 79, 80, 80, 105, 126, 138, 149, 179, 179, 191, 223, 232, 232, 236, 240, 242, 245, 247, 254, 274, 384, 470

$n = 24$

$\text{mean} = \frac{4690}{24} = 195.42$

$\text{median} = \frac{191 + 223}{2} = 207$

mode = 80, 179, 232 (meaningless)

5% trimmed: $(24)(0.05) = 1.2 \approx 1$

drop 1

$= 190.45$

10% trimmed: $(24)(0.10) = 2.4 \approx 2$

drop 2

$= 186.55$

20% trimmed: $(24)(0.20) = 4.8 \approx 5$

drop 5

$= \frac{2717}{14} = 194.07$

Q: 39, 92, 75, 61, 45, 87, 59, 51, 87, 12,
8, 93, 74, 16, 32, 39, 87, 12, 47, 50

Find 5%, 10%, 20% trimmed mean

8, 12, 12, 16, 32, 39, 39, 45, 47, 50,
51, 59, 61, 74, 75, 87, 87, 87, 92, 93

$n = 20$

total = 300 + 766 = 1066

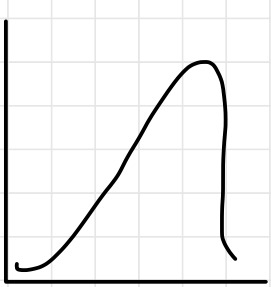
mean = 53.3

5% trimmed: $20 \times 0.05 = 1$ drop
= 53.61

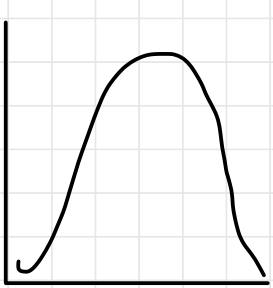
10% trimmed: 2 drop
= 53.8125

20% trimmed: 4 drop
= 54.92

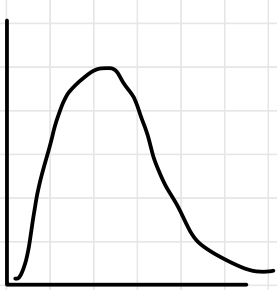
Symmetrical vs Skewed



negatively skewed



symmetric



positively skewed

Measures of Dispersion / Spread

- How data is spread
- range, variance, std. deviation

Range

- max value − min value
- misleading with outliers
- can indicate useful info for data without outliers

Variance

Population

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$$

Sample

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(n-1) used Bessel's correction*

Standard deviation

Population

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Sample

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

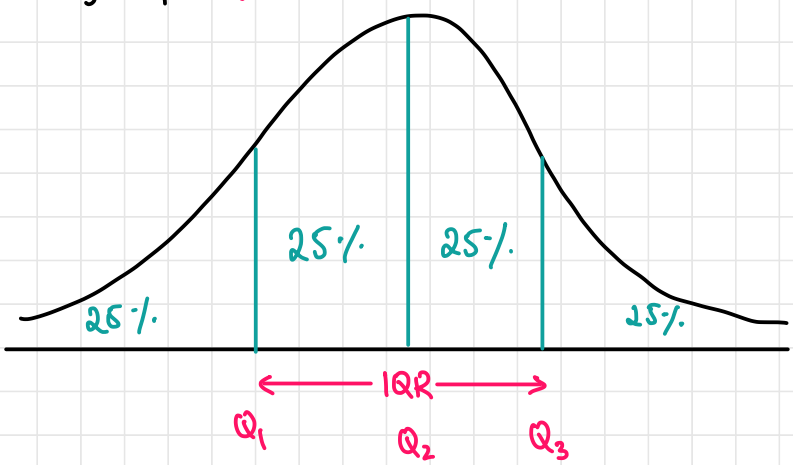
* look up

Quartiles

1. First quartile: 25^{th} percentile $= 0.25(n+1)^{th}$ term Q_1
2. Second quartile: 50^{th} percentile (median) $= 0.5(n+1)^{th}$ term Q_2
3. Third quartile: 75^{th} percentile $= 0.75(n+1)^{th}$ term Q_3

Interquartile range

$= Q_3 - Q_1$ IQR



Q: Find quartiles and IQR

5, 7, 12, 14, 15, 22, 25, 30, 36, 42, 53, 65

1^{st} quartile $= 0.25(12+1) = 3.25 = (3^{rd} + 4^{th})/2$
$= 13$

2^{nd} quartile $= 0.5(12+1) = 6.5$
$= (22+25)/2$
$= 23.5$

3^{rd} quartile $= 0.75(12+1) = 9.75^{th}$
$= (36+42)/2$
$= 39$

$IQR = 39 - 13 = 26$

PERCENTILE

- divides data into 100 equal parts
- to find p^{th} percentile, $\left(\frac{p}{100}\right)(n+1)$ where n is the sample size.

Q: Find 65^{th} percentile

30, 75, 79, 80, 80, 105, 126, 138, 149, 179, 179, 191, 223, 232, 232, 236, 240, 242, 245, 247, 254, 274, 384, 470

$(0.65)(25) = 16.25$

$(16^{th} + 17^{th})/2$

$= (236+240)/2 \quad = 238$

Tertile
- divide data into thirds

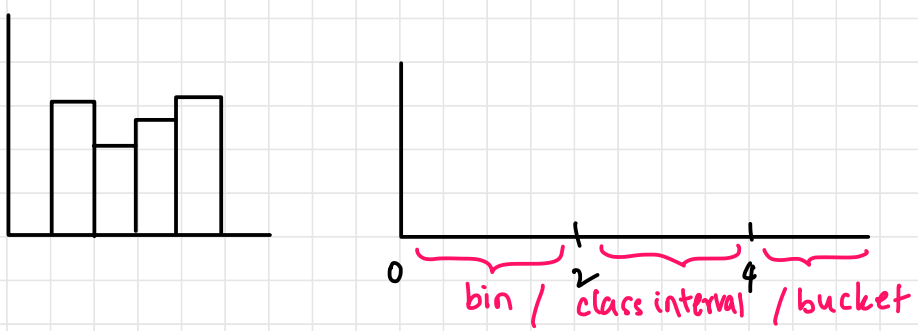
Decile
- divides data into tenths

SUMMARY STATISTICS FOR CATEGORICAL DATA

- frequencies & relative frequencies % of data in that category
- sample proportion : $\frac{\text{frequency}}{\text{sample size}}$

statquest

1. Histogram



- frequency distribution table
- does not include right end point

FREEDMAN - DIACONIS RULE

$$\text{Bin size} = \frac{2 \cdot IQR\,(x)}{\sqrt[3]{n}}$$

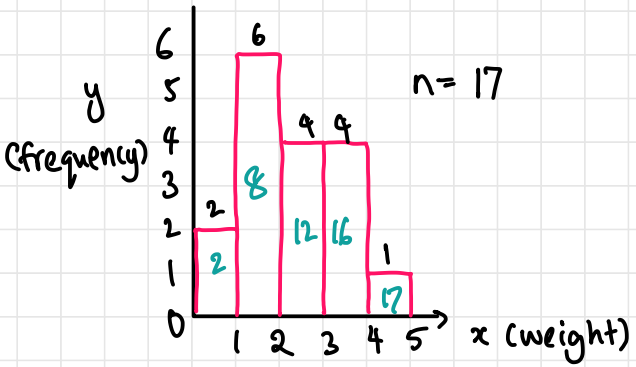
for unequal class width, height is density

$$\text{Density} = \frac{\text{relative frequency}}{\text{class width}}$$

or frequency

total density] meaningless

Mean, median, mode in histogram



n = 17

mode = [1-2]

$= \frac{1+2}{2} = 1.5$

median

pos = (0.5)(n+1) = 9

9 lies in [2-3]

$$\text{median} = \text{lower} + \frac{(\text{pos} - \text{cf}) \times \text{C.W}}{\text{freq}}$$

$$= 2 + \frac{(9-8)}{4} \times 1 = 2.25$$

Quartiles

Q_1 = pos = $(0.25)(n+1) = 4.5$ lies in [1-2]

$$Q_1 = \text{low} + \frac{(\text{pos} - \text{cf})}{\text{freq}} \times \text{cw}$$

$$= 1 + \frac{(4.5-2)}{6} \times 1 = 1.42$$

Q_3 = pos = $(0.75)(18) = 13.5$ [3-4]

$$Q_3 = 3 + \frac{(13.5-12)}{4} \times 1 = 3.375$$

IQR = 1.96

$$\text{mean} = \frac{2 \times \left(\frac{0+1}{2}\right) + 6 \times \left(\frac{1+2}{2}\right) + 4\left(\frac{2+3}{2}\right) + 4\left(\frac{3+4}{2}\right) + \left(\frac{4+5}{2}\right) \times 1}{17}$$

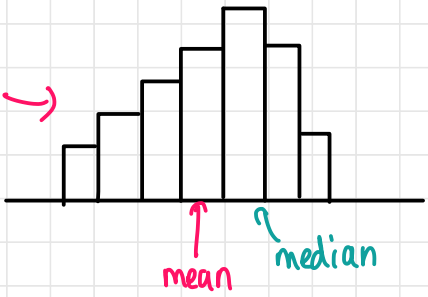
$$= 2.265$$

density :- $\frac{\text{rel. freq}}{\text{class wid}}$

Histogram Distributions

1) Symmetric
   - normal, Gaussian
   - mean = median = mode

2) Left-skewed
   - mean < median
   - negatively skewed



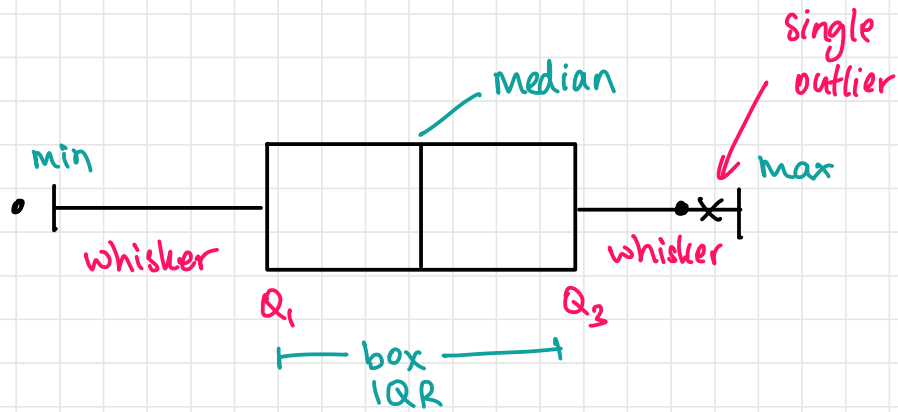
3) Right-skewed
   - mean > median
   - distribution of wealth

Unimodal vs **Polymodal**
single mode — multiple modes

- helps find outlier

2. Box Plot

- box-and-whisker plot



Outliers

lower whisker

$= Q_1 - (1.5)(IQR)$

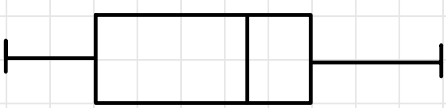
upper whisker

$= Q_3 + (1.5)(IQR)$

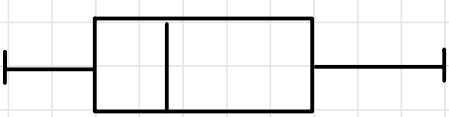
outliers outside this range

1) Symmetric
   - normal, Gaussian
   - mean = median = mode

2) Left-skewed
   - mean < median
   - negative



3) Right-skewed
   - mean > median
   - distribution of wealth
   - positive

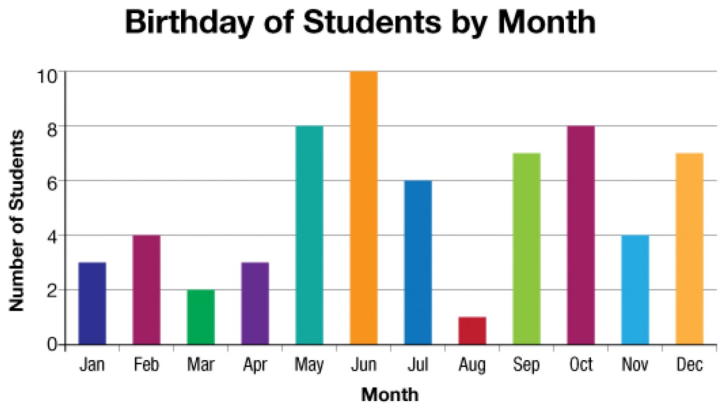


Unimodal vs Polymodal
single mode multiple modes

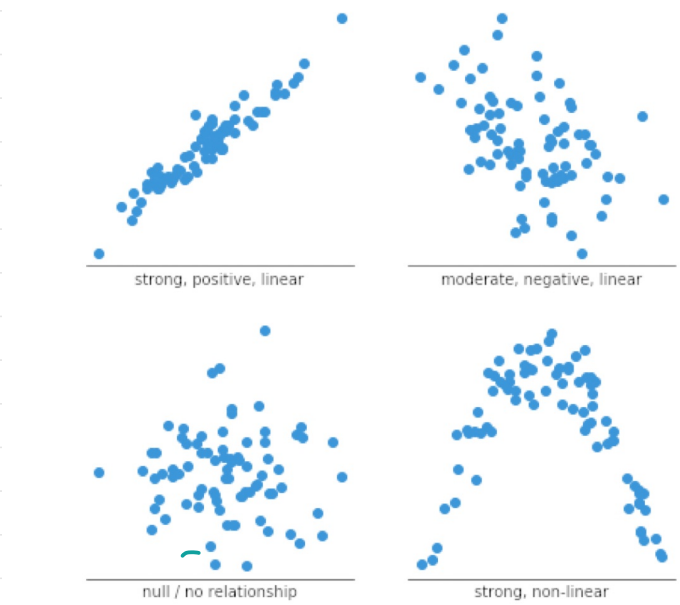
- helps find outlier

Scatter Plots, Bar Charts, Heatmaps

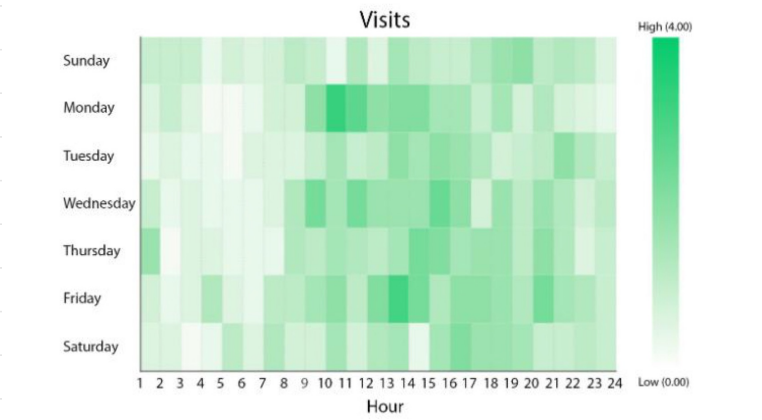
Bar Chart



Scatter Plot



Heatmap


Visits
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Hour
High (4.00)
Low (0.00)