Intro to Statistics

① Descriptive statistics for exploring data	⑦ Confidence Intervals
② Producing data and sampling	⑧ Test of significance
③ Probability	⑨ Resampling
④ Normal approximation & Binomial dist	⑩ Analysis of categorial data
⑤ Sample distributions & Central limit theorem	⑪ One way Analysis of variance
⑥ Regression	⑫ Multiple Comparisons.

Descriptive statistics

→ Statistics is the science to analyze data

→ Data snooping, reproducability & multiple testing fallacy are imp in big data

→ Statistics is important in Data Science as they

1. provide skills to access if data is sufficient to answer questions
2. establishes rigorous framework for quantifying uncertainity
3. provides techniques for communicating findings.

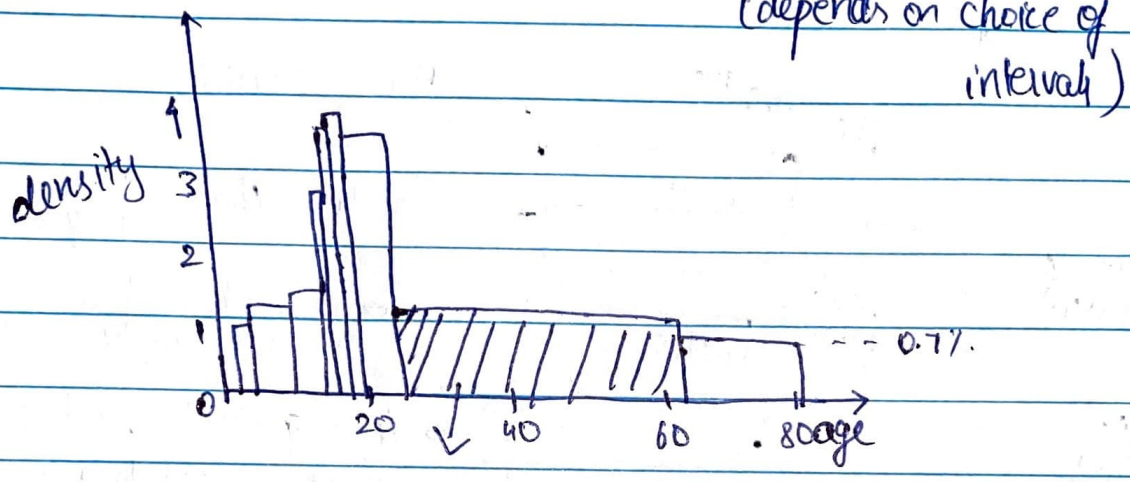
Descriptive statistics → ways to summarize data with numbers & graphs
→ Communicate information
→ Support reasoning about data

And also when data is really large, we need to summarize it first

Types of viz : Pie Chart, dot plot, bar graph, histogram

Histogram

Histogram allows bars to use different width (depends on choice of interval)



area under block is proportional to frequency

this means, total area corresponds to 100%.

→ if we are interested in figuring out what % people fall in b/w age 60-80, then we are interested in the area of that block.

→ Sometimes, we can find area without vertical scale.

Two kinds of information one can get from a histogram

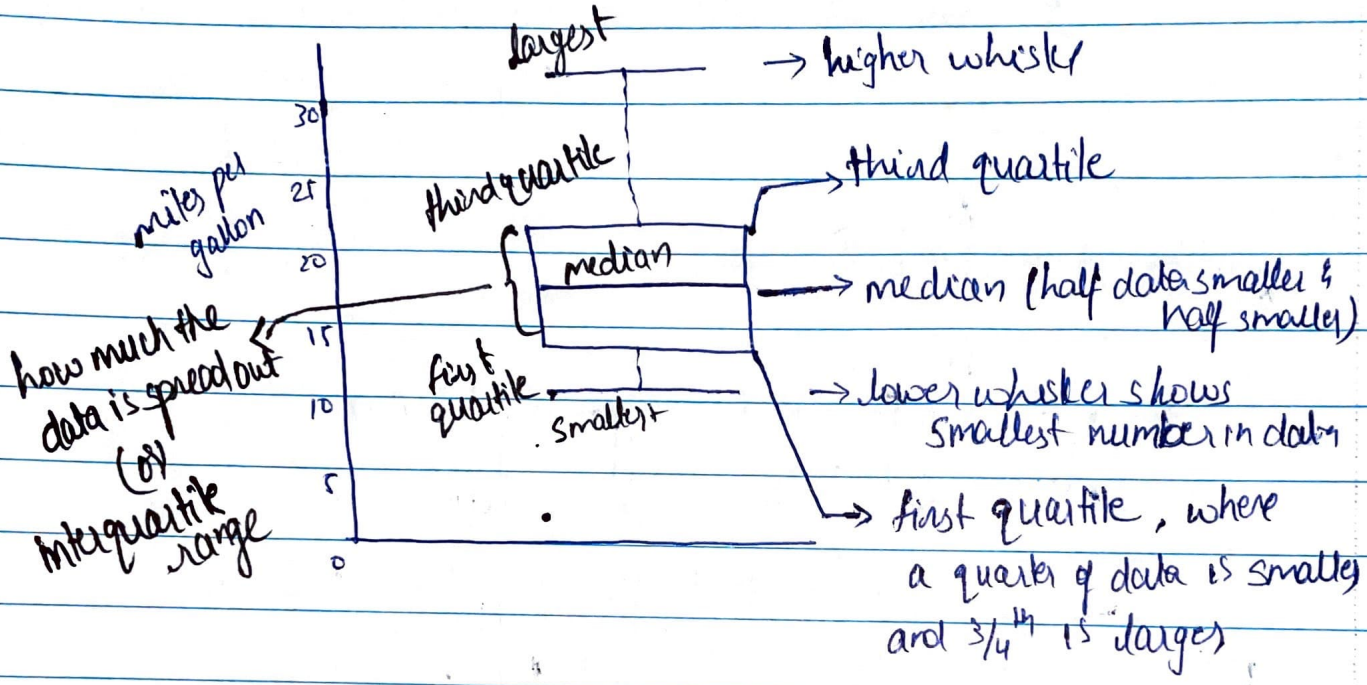
1. Density / Crowding → height of bar tells how many subjects are in one unit of horizontal scale, i.e 0.7% people fall in 60-80,

2. Percentage → area = width × height.

⇒ $\underset{(60-80)}{20} \times 0.7\%/\text{year} \Rightarrow 14\%$ subjects fall in range

Boxplot (Box-and-Whisker) :- (five number summary)

Boxplot depects 5 key points of data



→ Boxplot shows less data than histogram but takes less space and is well suited to compare several datasets.

Scatterplot :-

Used to depict data that comes in pairs.

→ Usually used to visualize relation b/w 2 points/variables.

⇒ Purpose of statistical analysis is to compare observed data to a reference So, it is very useful to provide context.

Numerical summary measures :-

For summarizing data with one number, use mean or median

$$\text{Mean} = \text{Average} = \frac{\sum n_i x_i}{\sum n_i} = \frac{\text{sum of all terms}}{\text{total terms}}$$

Median : Number where half data is larger & half is smaller.

$$\text{median} = \begin{cases} \left(\frac{n+1}{2}\right)^{th} \text{ observation; } n = \text{odd} \\ \left(\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}\right)^{th} \text{ observation; } n = \text{even} \end{cases}$$

Mode : Most frequent number

$$\text{mode} = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right), \text{ where}$$

$l \rightarrow$ lower boundary of modal class
$h \rightarrow$ size of modal class
$f_m \rightarrow$ frequency corresponding to modal class
$f_1 \rightarrow$ frequency preceeding modal class
$f_2 \rightarrow$ frequency proceeding modal class.

→ If histogram is symmetric then mean = median.
→ when histogram is skewed right, mean > median. Better use median.

Percentile :-

→ If top 10% reports household income of $135,000 then that means 90% have income < $135,000, then that point is called 90th percentile

→ top 25% has $85,000, then that point is 75th percentile or third quartile

→ 50th percentile = median.

⇒ interquartile range = (third quartile − first quartile)

A more commonly used measure of spread is standard deviation

$\bar{x}$ = average of x_i $\forall i \in 1,n$

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

⇒ the difference of each number from average, square it & sum it to average and find root of that average.

→ If data is increased by x%, mean, median, S, interquat % ↑ x%

by value x, mean ↑x, median ↑x, others remain same.

⇒ Both mean($\bar{x}$) & standard deviation(s) are used to express data, while mean is for giving us a measure of center & second is to give measure of spread.

⇒ If data is skewed, use median

Q

(1) 1, 2, 3, 4, 5

$$\text{mean} = \frac{1+2+3+4+5}{5} = 3$$

$$S = \sqrt{\frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5}} = \sqrt{\frac{4+1+0+1+4}{5}} = \sqrt{2}$$

$$= 1.4142$$

5% increase in data,

~~1.05 + 2.05 + 3.05 + 4.05 + 5.05~~

$$\frac{1.05+2.1+3.15+4.2+5.25}{5} = 3.15$$

$$S = \sqrt{\frac{(1.05-3.15)^2+(2.1-3.15)^2+(3.15-3.15)^2+(4.2-3.15)^2+(5.25-3.15)^2}{5}}$$

$$= \sqrt{\frac{11.025}{5}} = \sqrt{2.205} = 1.4849$$

Change = 7.07%