PL02 : Descriptive Statistics I

Video 1 : Summarizing Data for a Categorical variable.

- Categorical Data - uses labels, names or other descriptors to identify types of things.
  e.g Region, Machine, Car-make

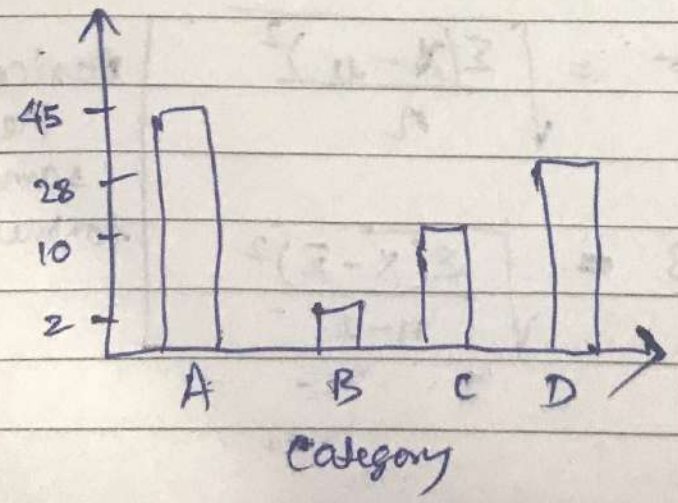
- Quantitative Data - are Numerical values that represent frequency, measurement, etc
  e.g Sales amount, Production units

- FREQUENCY DISTRIBUTION

List table of Category & Frequency

Category	Frequency
A	45
B	2
C	10
D	28

- FREQUENCY BAR CHART



- RELATIVE FREQUENCY

Relative Frequency of a class = $\frac{\text{Frequency of the class}}{n}$

e.g Relative frequency of A = $\frac{45}{85}$ = 0.52

- RELATIVE FREQUENCY BAR CHART

Same as frequency bar-chart, except relative frequency replaces frequency.

- PIE-CHART

Should only be used when there are 2 categories. It's very hard to visualize proportionally in a pie-chart.

Video 2 : Summarizing Data for a Quantitative variable

Histograms

- Buckets & Bins
  - Too few bins can create a histogram that doesn't show the shape/distribution of the underlying data, a "histoblob".
  - Too many bins create a histogram where there are too few observations in each bin and overall shape is broken up..
  - No. of bins depend on Data or decided by s/w

- HISTOGRAM

Shows the shape of the distribution of values.

Horizontal x-axis is the variable of interest
Vertical y-axis is frequency / relative frequency / percent freq

Vertical rectangle for each class or bin
Height is determined by frequency / rel freq / percent freq

No Space or gaps between bars of histogram.

- SKEW - Shapes of Histogram

(i) Left Skew - Tail is thinner on left side
(ii) Right Skew -

(iii) Symmetric - Normal Distribution, Symmetry

(iv) BIMODAL - 2 or more peaks or humps

(v) UNIFORM - No. of observations in each bin is same or almost same.

↗ Too less bin

(vi) No Pattern - Random histogram.

↗ Too many bins

- FREQUENCY Histogram / Relative Frequency Histogram

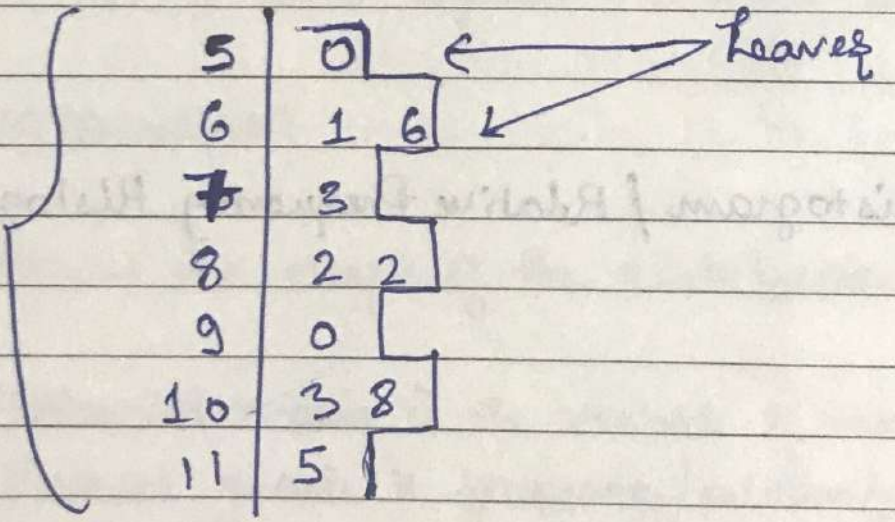
Video 3: Summarizing data for a quantitative variable using STEM & LEAF.

- STEM & LEAF Display - shows

  - the Rank order of the data.
  - the shape of the distribution of data.
  - Modal qualities of the data.

  - The leaf is always the last single digit.
  - The stem is formed from all digits left of the last digit.

- { 50, 61, 66, 73, 82, 82, 90, 103, 108, 115 }



- Stem & leaf is a sideways histogram
- It gives data in a Ordinal fashion & dataset can be reconstructed also
- It can show Modal Properties

Video 4 : Summarizing data using Tables

(Describing 2 or more variable with table)

- CROSSTABULATIONS (CROSSTABS)

  - frequency table, Histograms, Steam & Leaf Displays relates to Describing a single variable.

  - Crosstab is a table summary for two variables.

  - Variables can be Categorical or Quantitative.
  - Quantitative variables are often placed into bins or classes as in histogram.

  - Used in advanced statistics : chi-square, ANOVA

  - Size of crosstab is no. of categories of one variable multiplied by no. of categories of second variable.

  - Frequencies, Agg func, Percent,

region	appliance a	b	c	d
east				
west				
north				
south				

Video 5: Mean, Median and Mode

- Measuring the "CENTER" OF data.
- Mean - technically called Arithmetic mean.
  It is the average of all observations.
- Median - The middle observation of the data.
  after sorting from smallest to largest
  if there are ODD no. of observations.
  - If the no. of observations is an even no,
    the median is the mean of the two middle values
- Mode - The observation that occur most often
  In data. A dataset can have one mode,
  multiple modes, or none.
- Mean = $\bar{x} = \frac{\Sigma x}{n}$
- Median - Sort Smallest to Largest
  - If Odd length, $\frac{n}{2}$
  - If even length, $\frac{p + (p+1)}{2}$ $\left(p = \frac{n}{2}\right)$
- Mode - Observation that occuress most.

- Mean can be influenced by Extreme Observations.

- TRIMMED MEAN

  (i) Sort smallest to larget
  (ii) Remove the no. of observation from both ends.

  - Trimmed means are better for a single variable (Univariate).

- Mean, Median and Mode provide information about the "center of data".
- Mean can easily be influenced by extreme values
- Large difference b/w mean & median could be a warning sign.
- A trimmed mean could also provide insights.

Video 6: Mean of Grouped Frequencies

- GROUPED FREQUENCIES

  - In the absence of the Original Data, with only a Frequency Count Table or Histogram.

    we could still find the "CENTER" of data with a very close approximation.

  - The higher the count in each bin and narrower the bin range, the better the approximation will be -

Frequency Table

lower limit	Upper Limit	Frequency
A	B	10
B+1	C	7
C+1	D	15
D+1	E	21

(i) Find mid point of each bin interval

$$\text{midpoint} = \frac{\text{upper band} + \text{lower bound}}{2}$$

(ii) To apply "Weighting Process"

Multiply each mid-point by its frequency.

(iii) Sum all the values from step 2.

(iv) Divide the sum from step 2 by overall no. of obs.

to get Mean Frequency

Check ACCURACY $= \left(1 - \frac{\text{Original} - mf}{mf}\right) \times 100$

Video 7: Standard Deviation of Grouped frequency

$$S_x = \sqrt{\frac{\Sigma (m - \bar{x}')^2 f}{n-1}} \qquad s^2 = (S_x)^2$$

- Grouped calculations are often close to actual data values.

Video 8: Percentiles, Quartiles, Quintiles & Deciles

- These all are variations of the same thing (percentiles)
- – They do not have to be actual value in the dataset.

  – Observations are always sorted from Smallest to Largest.

- Percentiles represent the No. of Values out of the total that are at or below that percentile.

- These measure of location are RELATIVE.

- Location formula – $L_p = \frac{P}{100}(n+1)$

  where L is a location in the sorted data.
  p is the percentile you are looking for
  n is the no. of observations in the data.

Location of a quartile ($n=12$) $L_{25} = \frac{25}{100}(12+1) = 3.25$

Location of a median $L_{50} = \frac{50}{100}(12+1) = 6.5$

Location of 3rd quartile $L_{75} = \frac{75}{100}(12+1) = 9.75$

- Sort from smallest to largest.

25th Percentile 1st Quartile (Q1) | 50th Percentile Median (Q2) | 75th Percentile 3rd Quartile (Q3)

1	2	3	4	5	6	7	8	9	10	11
29500	54K	54K	65600	70400	73600	78800	80400	91200	94.7K	992K

500K

$L_{25} = 3.25$ $L_{50} = 6.5$ $L_{75} = 9.75$

$Q1 = 54000 + .25(65600 - 54000) = 56900$ ← 1st Quartile

$Q2 = 73600 + .5(78800 - 73600) = 76200$ ← 2nd Quartile.

$Q3 = 91200 + .75(94700 - 91200) = 93825$ ← 3rd Quartile

- None of the values Q1, Q2, Q3 are actual values in dat.
- Interpretation: Approximately one quarter or 25% have salaries at or below 56900
  Approximately 50% of workers have, have salaries at or below 76200.
  - so on.

∴ Percentile value = lower location value + location decimal × (upper − lower)

Percentile of a value = $\frac{x + 0.5y}{n}$

Percentile of 70400 = $\frac{4 + 0.5(1)}{12} = .38$ (Rounded)

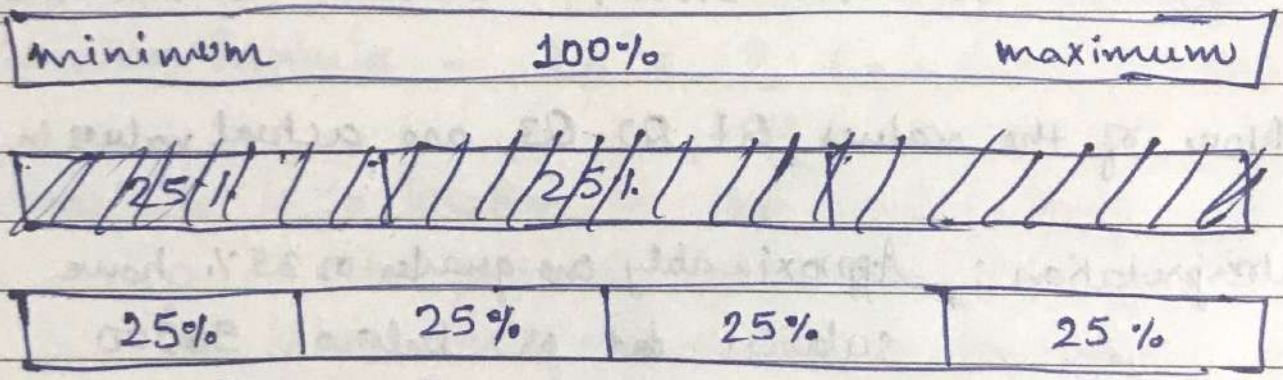
x is no. of observations up to, but not including the value, n the value
y is the count of observations equal to the value
n is the no. of observations in the data.

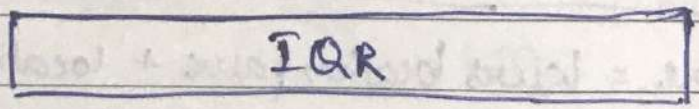
Video 9: IQR and Box Plots

- A simple way to visualize shape of our data
  - We can tell if our data is pulled (skewed) in one direction
  - An easy way to identify outliers.
  - There are 2 main types of Box plot
  - Box plots also called BOX & Whisker plot.

- Percentiles and Quartiles visualized

Sort Smallest to largest

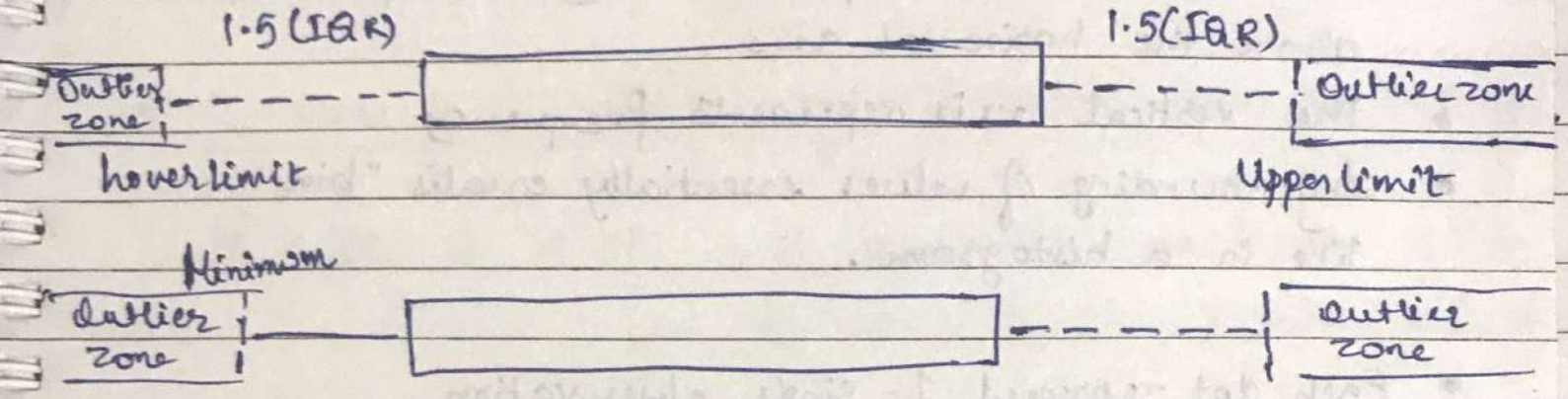




→ IQR is simply the middle 50% of the data.

→ IQR = Q3 − Q1

- OUTLIER ZONE



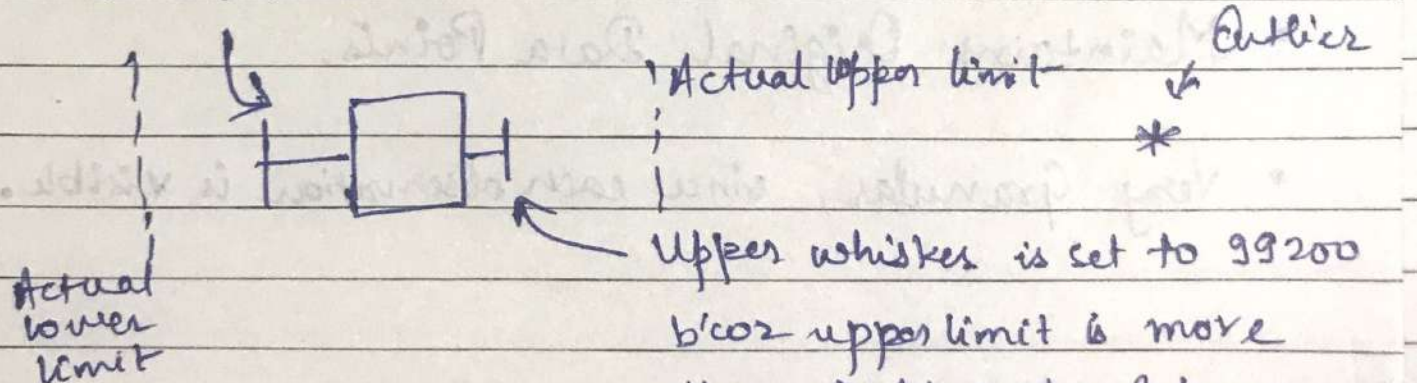
IQR = Q3 − Q1 = 36925

lower limit = Q1 − (1.5 × IQR) = 1512.5

Upper limit = Q3 + (1.5 × IQR) = 149212.5

Since 1512.5 is lower than lowest value, 29500.

∴ lower whisker



Video 10 : Dot Plots

- Each observation is represented by a dot placed above the horizontal axis.
- The vertical axis represents frequency.
- Any rounding of values essentially creates "bins" like in a histogram.

- Each dot represent 1- single observation.

- Visualize shape of the data.
- Tell if data is pulled (skewed) in one direction.
- An easy way to identify outliers.
- Makes comparing characteristics of data between categories very easy using color or shape.

- Maintains Original Data Points.

- Very Granular, since each observation is visible.