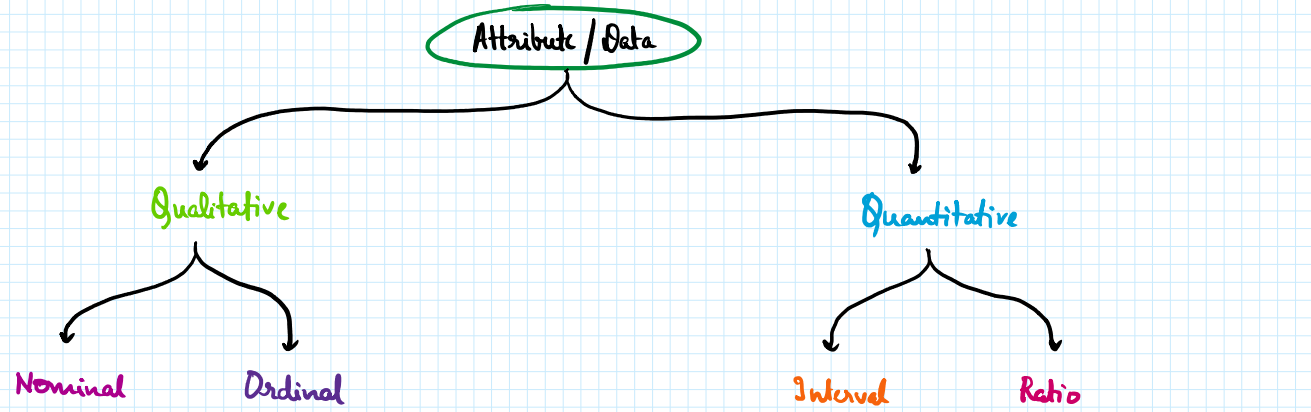
2. More Basics, Graphical Representations

08 August 2024 08:56

Errors in sampling
- Sampling / random error : sample not representative
- Non-sampling error : mistakes in data collection



ATTRIBUTE / VARIABLE / FEATURE / DIMENSION
- Represents characteristic / feature of object
- Can vary for each observation

TYPES OF STUDIES

Observational studies: Do not affect members of sample

Controlled studies: Assign elements to groups and apply certain treatments to certain groups
- Control group: Isolates effect of independent variable to be studied
- Experimental group: Change in value of independent variable to see effect on dependent variable

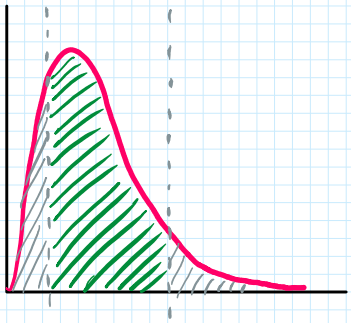
p% TRIMMED MEAN

Gets rid of outliers

$\frac{p}{100}(n)$ = No of observations to be removed before taking mean

NOTE: Weighted mean

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$



$$\text{Mean} - \text{mode} = 3(\text{mean} - \text{median})$$

Skewed & Symmetric Distribution

→ measure of asymmetry about mean

Symmetric distribution ⟶ mean ≈ median ≈ mode

Skewed distribution ⟶ median, mode lies towards skew

mode < median < mean ⟶ positively skewed

mean < median < mode ⟶ negatively skewed

Quantiles

Cut off points that divide distributions into equal groups

$\left(\frac{p}{100}\right)(n+1)$ = position of quantile (in terms of percentile)

Eg: Quartile, decile etc.

Note: Variance, SD

$$V(X) = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \text{(sample)}$$

$$= E(X^2) - [E(X)]^2$$

$$SD(X) = \sqrt{V(X)}$$

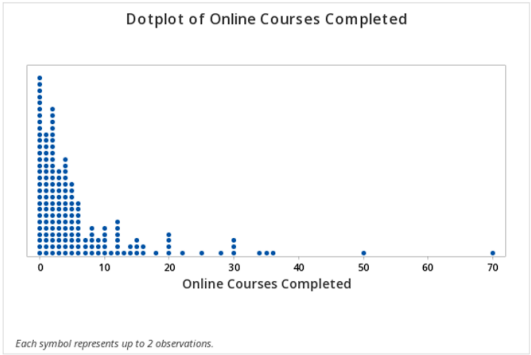
Diagrammatical Representation of Data

Stem and Leaf Plot

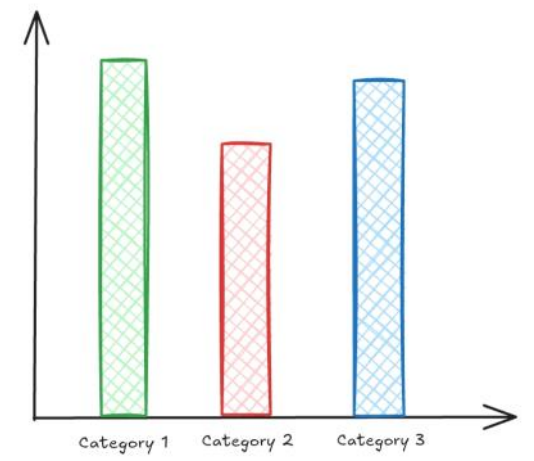
Dataset: 23, 23, 45, 65, 62, 27

Stem	Leaf
2	3, 3, 7
4	5
6	2, 5

Dotplots



Bar Chart



Frequency Distribution Table

Number of bins = $k = \sqrt{n}$ [OR] $2\sqrt[3]{n}$ [OR] $\log_2 n$

Width of bins = $h = \left(\frac{\max - \min}{k}\right) = \frac{2(IQR)}{\sqrt[3]{n}}$

$1 - < 3 \rightleftharpoons [1, 3)$

Relative freq. $= \frac{f_i}{\Sigma f_i}$

Density $= \frac{\text{Relative freq.}}{\text{Class width}}$

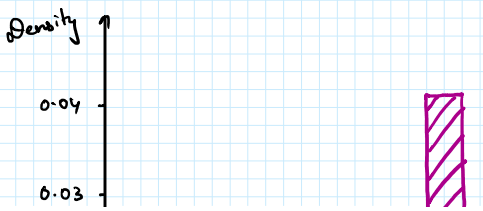
Histogram

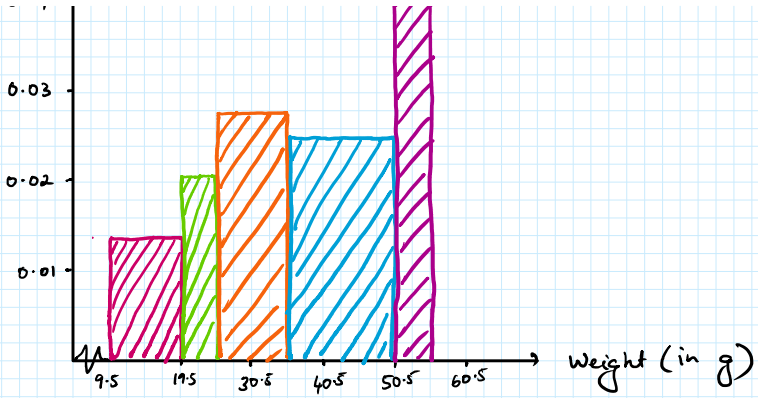
For a given frequency distribution:

- Bin widths equal $\longrightarrow$ Plot bars of height freq/ rel. freq./ density
- Bin widths unequal $\longrightarrow$ Plot bars of height density

Example: Plot a histogram for the following data where 48 objects were measured in grams

Class	Frequency	Relative Frequency	Density
10-19	6	0.125	0.01388
20-24	4	0.083	0.02075
25-34	12	0.25	0.02778
35-50	18	0.375	0.025
51-55	8	0.167	0.04175





Example: Histogram for rainfall in Los Angeles from 1965-2006.

0.2, 3.7, 1.2, 13.7, 1.5, 0.2, 1.7, 0.6, 0.1, 8.9, 1.9, 5.5, 0.5, 3.1, 3.1, 8.9, 8.0, 12.7, 4.1, 0.3, 2.6, 1.5, 8.0, 4.6, 0.7, 0.7, 6.6, 4.9, 0.1, 4.4, 3.2, 11.0, 7.9, 0.0, 1.3, 2.4, 0.1, 2.6, 4.9, 3.5, 6.1, 0.1

$Q_3 = \frac{3}{4}(42+1) = 32.25^{th}$ value

$Q_1 = \frac{1}{4}(43) = 10.75$

$Q_3 = \frac{5.5+6.1}{2} = 5.8$

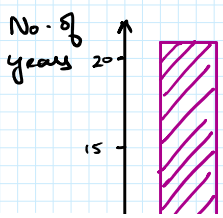
$Q_1 = \frac{0.6+0.7}{2} = 0.65$

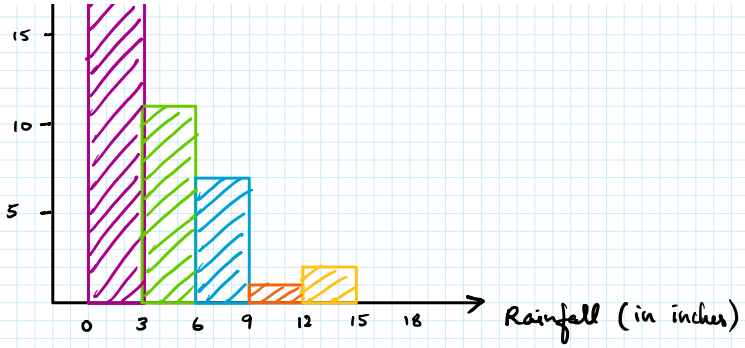
$IQR = Q_3 - Q_1 = 5.15$

$h = \frac{2(IQR)}{\sqrt[3]{n}} = 2.83 \approx 2.9 \approx 3$ (bin width)

$k = \frac{range}{h} = \frac{13.7-0}{3} = 4.xx \approx 5$

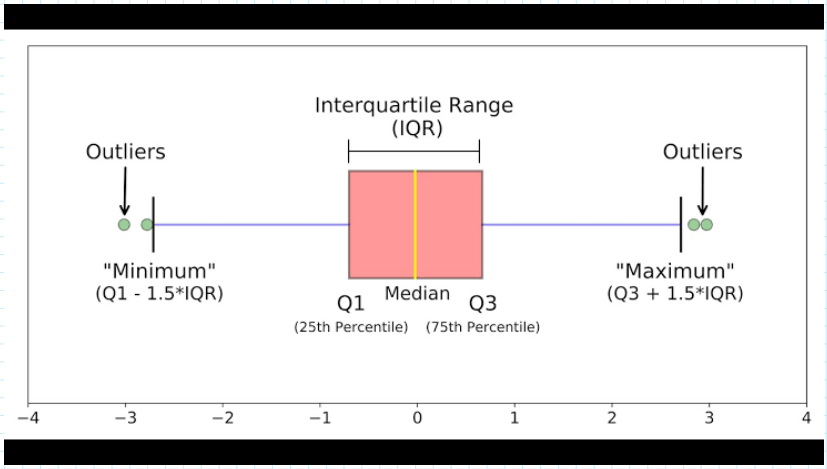
Class	Frequency	Relative Frequency	Density
0-3	21	0.5	
3-6	11	0.262	
6-9	7	0.167	
9-12	1	0.024	
12-15	2	0.048	





Box Plot / Box and Whisker Plot

- 5 number summary plot : minimum, Q_1, Q_2, Q_3, maximum
- Represents shape of distribution, central value and variability



Note: Outliers

Data value considered an outlier if:

Data value $< Q_1 - 1.5(IQR)$

Data value $> Q_3 + 1.5(IQR)$

Example: Draw a box plot for the given data

10, 11, 12, 25, 25, 27, 31, 33, 34, 34, 35, 36, 43, 50, 59

$n = 15$

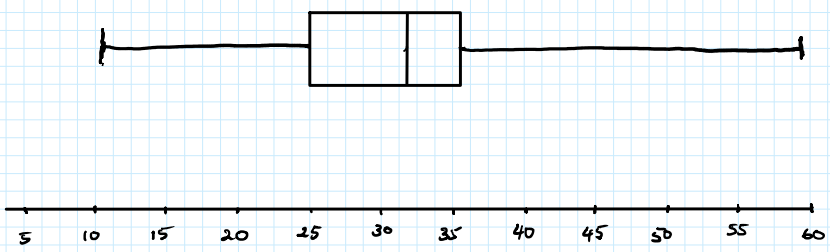
$Q_1 = 0.25(16) = 4^{th}$ position $= 25$

$Q_2 = 0.5(16) = 8^{th}$ position $= 33$

$Q_3 = 0.75(16) = 12^{th}$ position $= 36$

Minimum $= 11$

Maximum $= 59$



Example: 6, 7, 13, 17, 20, 25, 39, 41, 43, 49, 51, 62

Bar Charts	Histograms
• Qualitative data	• Quantitative data
• Gaps b/w bars	• No gaps
• Bars can be reordered	• Cannot be reordered
• Width of bars are the same	• Width need not be the same