S1

1.) DATA

→ Continuous: Can take any value in a given range
Eg. Time, Volume, Height, etc.

→ Discrete: Can only take particular values
Eg. For a Die; you can get 1, 2, 3, 4, 5, 6 only
[Singular for Dice]

⇒ STEMS & LEAF DIAGRAMS
* A way of ordering & Presenting Data

→ Plain Eg. 60, 51, 53, 42

Stems	Leaves
5	1, 2, 3 (3)
6	0 (1)

* key 5|1 = 51

* make sure the leaves digits are in order
* Always give a key

⇒ Back-to-Back

Eg. Boys: 168, 166, 172
Girls: 173, 169, 160

Boys		Girls	
(2) 8, 6	16	0, 9 (2)	* key: 6|16|0 means: Boys: 166 Girls: 160
(1) 2	17	3 (1)	

→ MEAN, MODE, MEDIAN & RANGE

* **MODE**: is that value of a variable which occurs most frequently.
  - modal class → has the highest frequency

* **MEDIAN**: is the middle value of an ordered set of Data.

$= \boxed{\frac{1}{2}(n+1)\text{th observation}}$

* **RANGE = Highest Value − Lowest Value**

* **MEAN = Average** $= \frac{\text{Sum of Observations}}{\text{Total Observations}} = \boxed{\frac{\sum x}{n} \text{ or } \frac{\sum fx}{\sum f}}$ ($\sum x$: sum of x)

Signs = $\bar{x}$; $\bar{y}$; μ

→ GROUPED FREQUENCY DISTRIBUTION

* First get class width = $\frac{\text{largest value} - \text{smallest value}}{\text{number of groups}}$

*

Class — class Width = 9.5 − 4.5 = 5

4.5 (5 — 9) 9.5 → upper class boundary = 9 + 0.5

- 4.5 ← lower class boundary = 5 − 0.5
- 5 ↓ lower class limit
- 9 ↓ upper class limit
- 7 ← midpoint = $\frac{5+9}{2}$ or $\frac{4.5+9.5}{2}$

* Cumulative Frequency = Running Total

→ HISTOGRAMS

Eg.

Class	Number of children (frequency)	Class Width	Frequency Density
(129.5–134.5) 130–134	10	5	2.0
(134.5–139.5) 135–139	22	5	4.4
(139.5–144.5) 140–144	38	5	7.6

* **Frequency Density = $\frac{\text{Frequency}}{\text{Class Width}}$**

* Upper & lower class bound can be found by taking the end of one class + beginning of next class / 2

Eg. $\frac{135+134}{2} = 134.5$

$\frac{140+139}{2} = 139.5$

* **AREA = FREQUENCY**
  FREQUENCY = Height of bar × Class Width

* **Height of Bar = $\frac{\text{FREQUENCY}}{\text{Class Width}}$** ; **$h = \frac{f}{cw}$**

* **TOTAL AREA ∝ TOTAL FREQUENCY**

↑ Propotional.

→ MEDIAN CLASS

Eg.

Visits	f	cf
0-4	32	32
5-9	71	103
10-14	20	123

* $$\text{Median} = b + \left(\frac{\frac{1}{2}n - f}{fm}\right) \times w$$

* b = lower class boundary
  f = Sum of all frequencies below b
  fm = frequency of median class
  w = class interval/width

- **Step 1**: Calculate median value
  → $\frac{1}{2}(n+1) = \frac{1}{2}(123+1) = 62$
  * This is found in class (5-9)
- **Step 2**: * Lower class boundary = 4.5 *
  * f = 32 → cf before lower class boundary
- **Step 3**: * Class width = 9.5 − 4.5 = 5
  * fm = 71 → frequency of median class

$$\text{Median} = 4.5 + \left[\left(\frac{\frac{1}{2}(123) - 32}{71}\right) \times 5\right]$$

$$= 6.58$$

→ Mean

Visits	midclass Value x	f	fx
0-4	2.~~25~~	32	~~80~~ 72
5-9	7	71	497
10-14	~~25~~ 12	20	~~500~~ 240
TOTALS		123	~~1,077~~ 809

* To find Mid class Value (x)

$$= \left(\frac{\text{Lower class boundary} + \text{Upper Class boundary}}{2} \right)$$

* $\text{Mean} = \frac{\sum fx}{\sum f} = \frac{809}{123} = 6.58$ (to 2 d.p.) → DO NOT FORGET TO ADD UNITS

→ Quartiles:

Q_1 = lowerquartile $= b + \left[\left(\frac{\frac{1}{4}n - f}{fm} \right) \times w \right]$ ** Similar to median formula

Q_2 = median $= b + \left[\left(\frac{\frac{1}{2}n - f}{fm} \right) \times w \right]$

Q_3 = Upper Quartile $= b + \left[\left(\frac{\frac{3}{4}n - f}{fm} \right) \times w \right]$

→ Interquartile range $= (Q_3 - Q_1)$

→ Semi-Interquartile Range (SIQR) $= \frac{1}{2}(Q_3 - Q_1)$

→ **Percentile** $= \frac{1}{100} n$

Eg. Percentile 85th

$$\therefore P_{85} = b + \left[\left(\frac{\frac{85}{100} n - f}{fm}\right) \times w\right]$$

→ **Decile** $= \frac{1}{10} n$

Eg. 3rd Decile

$$\therefore D_3 = b + \left[\left(\frac{\frac{3}{10} n - f}{fm}\right) \times w\right]$$

* Make sure you calculate the value of b, f, fm, w, etc. for EACH QUARTILE, DECILE & PERCENTILE as they are NOT the same for each of them.

* Follow the same steps mention on pg. 4.

* Remember: To identify the relevant classes, use:
  - $Q_1 = \frac{1}{4} n$
  - $Q_2 = \frac{1}{2} (n+1)$
  - $Q_3 = \frac{3}{4} n$
  - Percentile $= \frac{1}{100} n$
  - Decile $= \frac{1}{10} n$

→ Stem & leaf Diagrams

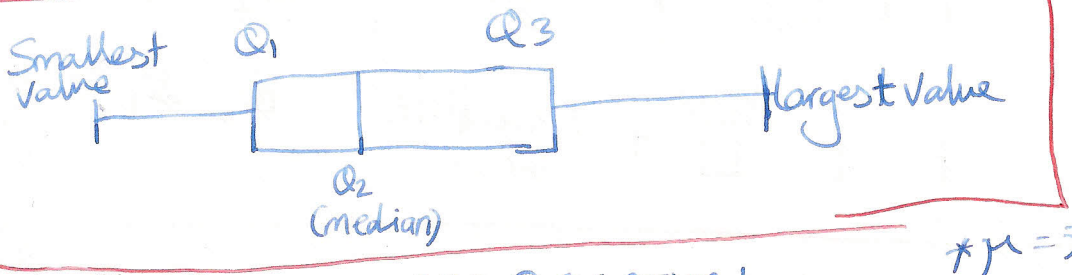
For $Q_1 = \frac{1}{4}n$

For $Q_2 = \frac{1}{2}(n+1)$

For $Q_3 = \frac{3}{4}n$

} This Tells you <u>which observation</u> to take eg. $Q_1 = \frac{51}{4} = 13$th observation $\therefore Q_1 = 32$ (taken from the Stem & leaf Diagram)

→ <u>Box Plots</u> [or Box & Whisker Diagram]



→ VARIANCE & STANDARD DEVIATION (σ^2)

* $\mu = \bar{x}$ = mean $= \frac{\Sigma x}{n} = \frac{\Sigma fx}{\Sigma f}$

* <u>VARIANCE</u> $= \sigma^2 = \frac{\Sigma (x-\mu)^2}{n} = \frac{\Sigma fx^2}{\Sigma f} - \mu^2$

* <u>STANDARD DEVIATION</u> $= \sqrt{\sigma^2} = \sqrt{\left(\frac{\Sigma fx^2}{\Sigma f} - \mu^2\right)}$

→ CODING

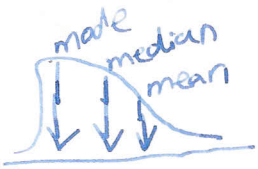
* x = Original value
* y = new 'coding' value

$$y = \frac{x-b}{a} \rightarrow \bar{y} = \frac{\bar{x}-b}{a} \rightarrow \sigma_y = \frac{\sigma_x}{a}$$

* All you are doing is simplifying longer values, in order to save time.
* to get the y values, minus any thing common eg 1 million or the smallest value, & divide by something that is divisible for all orginal values.
* Find mean & Standard Deviation of y, the replace to find of x

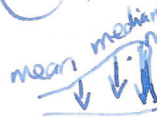
⇒ SKEWNESS & OUTLIERS

*methods



*< means 'less than'

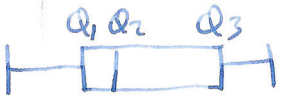
(1) Positive Skew: mode < median < mean.



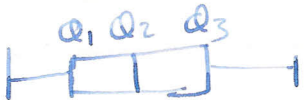
Negative Skew: mean < median < mode.

Symmetrical: mean = mode = median

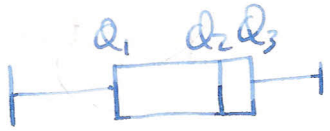
(2)



; Positive skew = $Q_2 - Q_1 < Q_3 - Q_2$



Symmetry = $Q_2 - Q_1 = Q_3 - Q_2$



Negative skew = $Q_2 - Q_1 > Q_3 - Q_2$

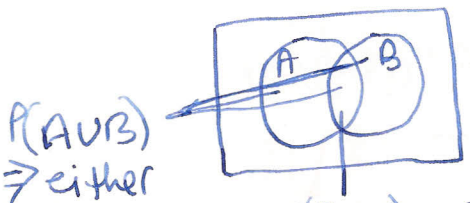
(3) Pearson's Coefficient of Skewness $= \frac{\text{mean} - \text{mode}}{\text{standard deviation}} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$

→ Outlier = is found outside the rest of the readings

2.) PROBABILITY

* $P(\text{event}) = \dfrac{\text{No. of outcomes where event happens}}{\text{Total number of possible outcomes}}$

* Sample Space = set of all possible outcomes
* Venn Diagram = shows which outcome corresponds to which events.



$P(A \cup B)$ ⇒ either A or B; Union

$P(A \cap B)$ ⇒ Both A & B; Intersection

→ ADDITION RULE

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

⇒ Complementary probability: $P(A') = 1 - P(A)$ or $P(A') + P(A) = 1$

not event A; event A

⇒ CONDITIONAL PROBABILITY

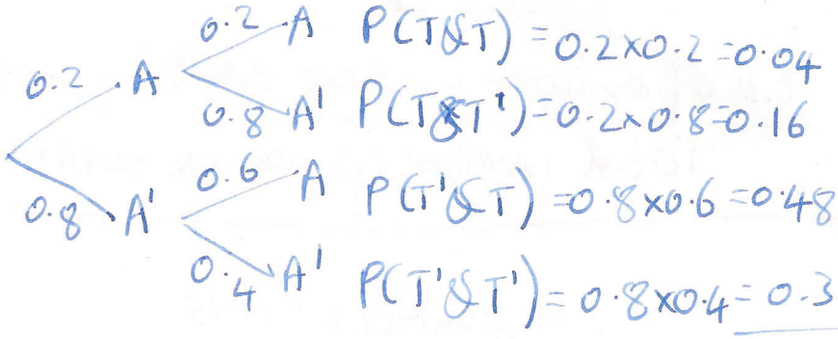
$P(m|B)$ = P(m given B has already happened)

$$P(m|B) = \frac{P(m \cap B)}{P(B)}$$

* MULTIPLICATION RULE = $P(m|B) \times P(B) = P(m \cap B)$

→ Tree diagrams: Show Probabilities of 2 or more events

Eg.



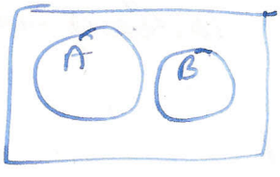
1.00 → All probabilities should add up to ONE.

※ → Independent Events have No effect on Each other

$$P(A \cap B) = P(A)\,P(B)$$

Since $P(A|B) = \frac{P(A \cap B)}{P(B)}$ & $P(A|B) = P(A)$

※ → Mutually Exclusive: Events have no overlap



$\Rightarrow P(A \cap B) = 0$

Since $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, & $P(A \cap B) = 0$

$$\therefore P(A \cup B) = P(A) + P(B)$$

3.) CORRELATION

* the measure of how closely Variables are linked.

→ Product-moment Correlation Coefficient (r): measures how close to a straight line the points on a scatter graph lie.

* $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

* The formula looks complex, but is easy. All you do is plug in the correct values and thats it!

$\bar{x} = \dfrac{\Sigma x}{n}$; $\bar{y} = \dfrac{\Sigma y}{n}$

* $S_{xy} = \Sigma (x-\bar{x})(y-\bar{y}) \Rightarrow \Sigma xy - \bar{x}\bar{y} = \left[\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}\right]$

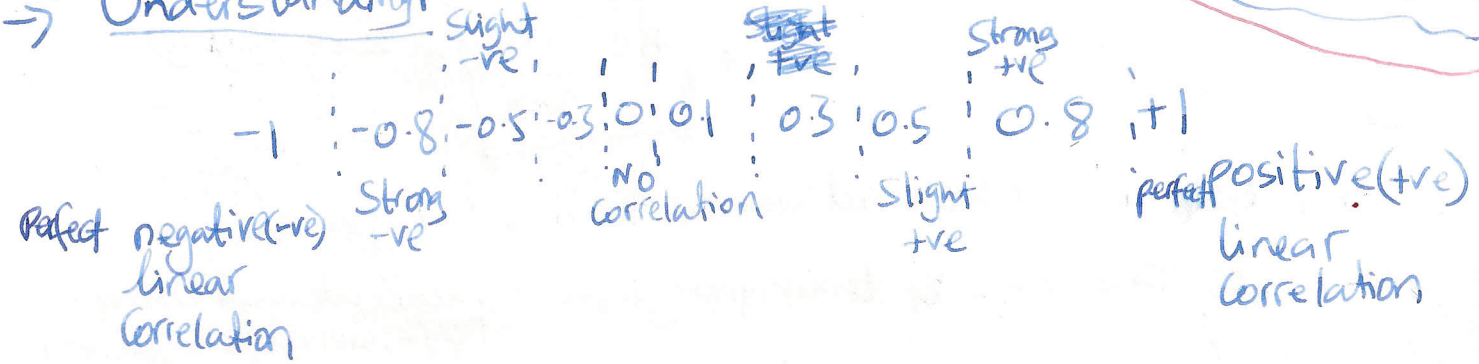
* $S_{yy} = \Sigma (y-\bar{y})^2 = \Sigma y^2 - (\bar{y})^2 \Rightarrow \left[\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}\right]$

* $S_{xx} = \Sigma (x-\bar{x})^2 = \Sigma x^2 - (\bar{x})^2 \Rightarrow \left[\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}\right]$

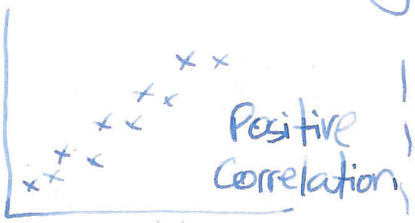
$$\therefore r = \frac{\left[\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}\right]}{\sqrt{\left[\left(\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}\right)\left(\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}\right)\right]}}$$

* r must be between +1 & −1

→ Understanding r



→ Scatter Diagrams.



Positive Correlation,

* one variable gets bigger, so does the other

* one variable gets smaller, the other becomes bigger.

negative Correlation,

no Correlation,

* The two variables are not linked.

→ Coding

$$X = \frac{x-a}{b} \quad ; \quad Y = \frac{y-c}{d}$$

* Use codes when values are massive
* Use the codes in the calculation. The PMCC(r) you get would be the same.

→ LINEAR REGRESSION

* This is the 'Statistics' way of find equation of the best fit line.

→ Regression line (lineaf best fit) ⇒ $\boxed{y = a + bx}$

* $\boxed{a = \bar{y} - b\bar{x}}$ or $\frac{(\Sigma y)}{n} - b\frac{(\Sigma x)}{n}$

* $\boxed{b = \frac{S_{xy}}{S_{xx}}}$ or $\frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}$

* Residual = Observed y-value − Estimated y-value
  - the value taken from graph − Value calculat using regression line equation.

4.) PROBABILITY DISTRIBUTION

* X (upper case) → name of random variable
  x (lower case) → particular value
* Random Variable → no fixed value
* Discrete Random Variable – only has a certain number of Random Variables.

Eg. Rolling a die with 4 faces.

x:	1	2	3	4
$P(X=x)$:	$\frac{k}{1}$	$\frac{k}{2}$	$\frac{k}{3}$	$\frac{k}{4}$

(possible results; probabilities)

$P(X=x)$: ↑ Probability function

$$\sum_{\text{all } x} P(X=x) = 1$$ → all probabilities add up to 1.

$$\therefore \frac{k}{1} + \frac{k}{2} + \frac{k}{3} + \frac{k}{4} = 1 \Rightarrow k = \frac{12}{25}$$

→ mode = most likely value (with) = highest probability

Eg. 2

x:	1	2	3	4
$P(X=x)$:	$\frac{12}{25}$	$\frac{6}{25}$	$\frac{4}{25}$	$\frac{3}{25}$

(a) $P(X=1) = \frac{12}{25}$

(b) $P(X=0) = 0$

(c) $P(X<3) = \frac{6+12}{25} = \frac{18}{25}$
↳ $P(2)+P(1)$

(d) $P(X \leq 3) = \frac{12+6+4}{25} = \frac{22}{25}$
↳ $P(2)+P(1)+P(3)$

(e) $P(X \geq 1) = P(X \leq 4)$

(f) $P(2 \leq X < 4) = \frac{6+4}{25} = \frac{10}{25}$
↳ $P(x=2)+P(x=3)$

(g) mode = 1 (It has the highest probability)

→ Distribution function = Cumulative Distribution Function

$$\boxed{F(x_0) = P(X \leq x_0)}$$

$\therefore F(3) = P(X \leq 3)$

Eg.

$$P(X=x) = \begin{cases} \frac{kx}{6} & \text{for } x = 1,2,3 \\ \frac{k(7-x)}{6} & \text{for } x = 4,5,6 \\ 0 & \text{otherwise} \end{cases}$$

* Re-write this in the table format. It's always useful.

x :	1	2	3	4	5	6
$P(X=x)$:	$\frac{k}{6}$	$\frac{2k}{6}$	$\frac{3k}{6}$	$\frac{3k}{6}$	$\frac{2k}{6}$	$\frac{k}{6}$

$$\left(\underbrace{\frac{k}{6} + \frac{2k}{6} + \frac{3k}{6}}_{\frac{kx}{6}} + \underbrace{\frac{3k}{6} + \frac{2k}{6} + \frac{k}{6}}_{\frac{k(7-x)}{6}}\right) = 1 \Rightarrow \therefore k = \frac{1}{2}$$

* $F(3) = P(X \leq 3) \Rightarrow P(x=1) + P(x=2) + P(x=3) = \frac{6}{12} = \frac{1}{2}$

* You can add a row for Cumulative distribution:

→

x :	1	2	3	4	5	6
$F(x)$:	$\frac{k}{6}$	$\frac{3k}{6}$	$\frac{6k}{6}$	$\frac{9k}{6}$	$\frac{11k}{6}$	$\frac{12k}{6}$ ← This should be 1

* It's just like Cumulative frequency.

* Discrete Uniform Distribution:

μ = Mean = $\left(\frac{a+b}{2}\right)$; Variance = $\left[\frac{(b-a+1)^2 - 1}{12}\right]$

$= \left(\frac{(n+1)(n-1)}{12}\right)$

also same as Expected Value

* EXPECTED VALUES, MEAN & VARIANCE

Mean Expected Value = $E(X) = \sum x P(X=x)$

Variance = $E(X^2) - \left(E(X)\right)^2 = \sum x^2 P(X=x) - \left[\sum x P(X=x)\right]^2$

Eg.

x:	1	2	3	4
P(X=x):	12/25	6/25	4/25	3/25

* Expected Value = $E(X) = \sum x P(X=x) =$

$\left(1 \times \frac{12}{25} + 2 \times \frac{6}{25} + 3 \times \frac{4}{25} + 4 \times \frac{3}{25}\right)$

$= 48/25 = 1.92$

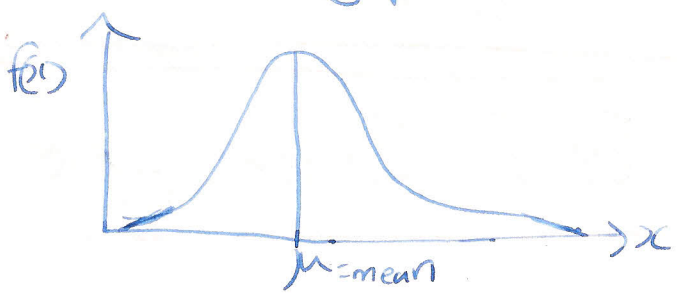
* Variance = $\sum x^2 P(X=x) - \left(\sum x P(X=x)\right)^2 = E(X^2) - \left(E(X)\right)^2$
→ Var(x)
⇒ σ^2 $= \frac{24}{5} - \frac{2304}{625} = \frac{696}{625} = 1.1136$

* Expected Value & Variance formulas for functions

* $E(aX) = aE(x)$
* $E(aX+b) = aE(x)+b$
* $Var(aX) = a^2 Var(x)$
* $Var(aX+b) = a^2 Var(x)$

⇒ THE NORMAL DISTRIBUTION

* Continuous Distribution ⇒ Area = Probability
  ↳ has no gaps

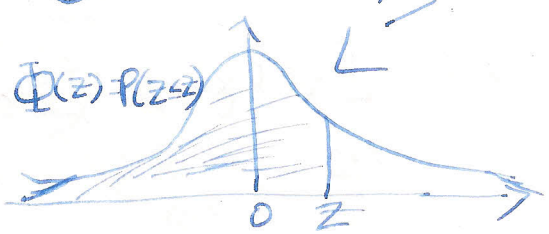


* We 'Normalise' a variable to use the Z-tables.

$X \sim N(\mu, \sigma^2)$ ⇒ X is Normally distributed with mean μ & Variance σ^2

To 'normalise' to Z

* $X \sim N(\mu, \sigma^2) \rightarrow \frac{X-\mu}{\sigma} = Z, \quad Z \sim N(0,1)$



Eg. $X \sim N(5, 16)$

(a) $P(X < 7)$

* first 'Normalise'

$$N(0,1^2) \sim \frac{X-\mu}{\sigma} = Z$$

$$\therefore P(X<7) = P\left(Z < \frac{7-5}{4}\right) = P(Z < 0.5) = 0.6915$$

* On the table, look for 0.50 under 'z' and write the value opposite it, under 'Φz'

(b) $P(X>9)$

$$\boxed{P(Z>a) = 1 - \Phi(a)}$$

$$\therefore P(X>9) = P\left(Z > \frac{9-5}{4}\right) = P(Z>1) = 1 - P(Z<1) = 1 - 0.8413 = 0.1587$$

* Remember $\Phi(1) = P(Z<1)$

↳ cumulative distribution function.

(c) $P(5<X<11) = P\left(\frac{5-5}{4} < Z < \frac{11-5}{4}\right) = P(0<Z<1.5)$

$$= P(Z<1.5) - P(Z<0)$$
$$= 0.9332 - 0.5$$
$$= 0.4332$$

$$\boxed{P(a<Z<b) = \Phi(b) - \Phi(a) = P(Z<b) - P(Z<a)}$$

* $\boxed{\Phi(-2) = 1 - \Phi(2)}$