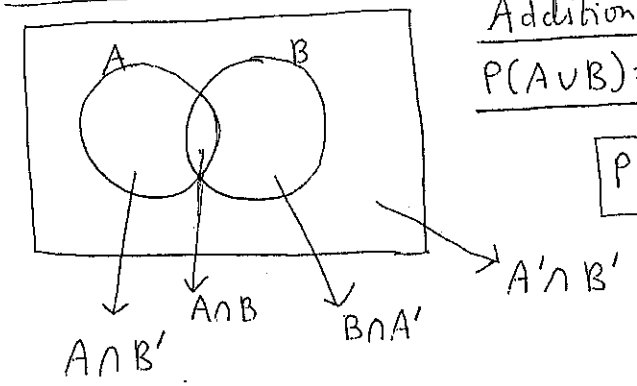
S1 Formulas

Probability



Addition Law

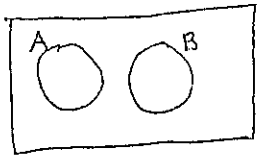
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B') = P(A) - P(A \cap B)$$

Conditional Probability

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

Mutually Exclusive
- events cannot happen at the same time.



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

Independence:

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

OR

$$P(A|B) = P(A)$$

Histograms: Frequency = $\frac{\text{Frequency}}{\text{Class Width}}$ — For continuous data

Plot F.D. against the class boundaries.

Area of bar = frequency (or, sometimes K× frequency).

Scales + Labels

P.M.C.C.

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$-1 \leq r \leq 1$

Unaffected by Coding

$$S_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

Regression - some understanding of the theory is required.

$y = a + bx$ where $b = \frac{S_{xy}}{S_{xx}}$ and $a = \bar{y} - b\bar{x}$.

Reliable predictions for x-values within range of x-data.

Line passes through pt. $(\bar{x}, \bar{y})$.

Interpret b and a in context.

Scatter diagram - scales + labels.

x - Explanatory Variable.

y - Response Variable

	Raw Data	Frequency Distribution
Mean	$\bar{x}$ or $\mu = \frac{\Sigma x}{n}$	$= \frac{\Sigma fx}{\Sigma f}$
Variance	$\sigma^2 = \frac{\Sigma x^2}{n} - \mu^2$	$= \frac{\Sigma fx^2}{\Sigma f} - \mu^2$

For grouped data, x = mid-point.

Standard Deviation = $\sqrt{\text{Variance}} = \sigma$

Coded Data: $y = a + bx \Rightarrow \begin{cases} \bar{y} = a + b\bar{x} \\ \sigma_y = b\sigma_x \end{cases}$

$$\Rightarrow \bar{x} = \frac{\bar{y} - a}{b} \text{ and } \sigma_x = \frac{\sigma_y}{b}$$

Quartiles: Q_2 = median = $\frac{n+1}{2}$

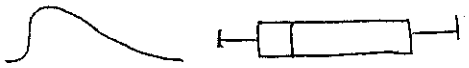
For small data sets or stem & leaf diagrams, calculate Q_1 and Q_3 directly. For larger data sets use $\frac{n+1}{4}$ or $\frac{n}{4}$.

Interpolation: Learn method for calculating quartiles, deciles and percentiles from grouped data.

Skewness

Symmetric

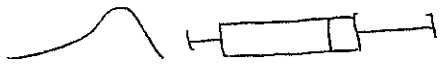
mean = median = mode



positively skewed

$Q_3 - Q_2 > Q_2 - Q_1$

mean > median > mode.



negatively skewed

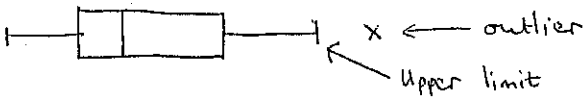
$Q_3 - Q_2 < Q_2 - Q_1$

mean < median < mode.

Comparing Box plots – Comment on median, IQR, skewness.

Outliers : $x > Q_3 + 1.5(Q_3 - Q_1)$ or $x < Q_1 - 1.5(Q_3 - Q_1)$

Formula given



Median & IQR is preferred over Mean & S.D. for skewed distributions.

Random Variables

Expected Value of X : $\mu = E(X) = \sum x\, P(X=x)$.

Variance of X : $\sigma^2 = Var(X) = E(X^2) - [E(X)]^2$

$= \sum x^2 P(X=x) - \mu^2$

$\sum P(X=x) = 1$.

$E(aX+b) = a\,E(X) + b$

$Var(aX+b) = a^2\, Var(X)$.

$$F(a) = P(X \leq a)$$

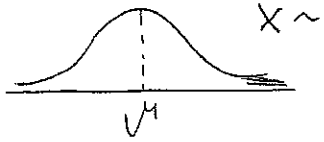
$$P(X=a) = F(a) - F(a-1)$$

Discrete Uniform Distribution

All probabilities equally likely e.g. fair die.

$E(X) = \frac{n+1}{2}$ by symmetry.

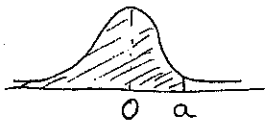
Normal Distribution



$X \sim N(\mu, \sigma^2)$

Standardise : $Z = \frac{X-\mu}{\sigma}$

Standard Normal Distribution : $Z \sim N(0, 1^2)$



$P(Z < a) = \Phi(a)$ | $P(Z > a) = 1 - \Phi(a)$

$\Phi(-a) = 1 - \Phi(a)$

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