S1 Revision Notes

Representation and summary of Data – Location

Key Words: Continuous, discrete, mean, median, mode, cumulative frequency tables, class width/midpoint, coding

- *Mean:* $\mu = \frac{\sum x}{n}$ *or* $\frac{\sum fx}{\sum f}$
- *For large data: use coding to work out mean of coded data, then equate to coding for original mean*
- *Linear interpolation (for some x): using your bivariate date, sketch x against f(x). Plot some a where* $a < x$ *and some b where* $b > x$ *and consider the gradient of* $\frac{f(b)-f(a)}{b-a} = \frac{f(x)-f(a)}{x-a}$
  - *Used to calculate values in cumulative frequency tables*
- *Combined mean of set A (size: n, mean: x) and of set B (size: m, mean: y) is given by* $\frac{nx+my}{n+m}$

Representation and summary of Data – Measures of dispersion

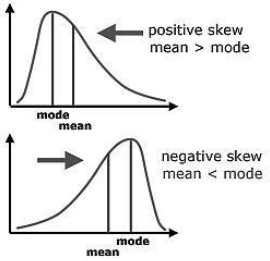
Key Words: range, lower/upper quartiles, interquartile range, percentiles, deviation, variance, standard deviation

- *Variance* $\sigma^2 = \frac{\sum x^2}{n} - \mu^2 = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$
- *Standard deviation: how much the members of a group differ from the mean value for the group*
- *For large data: use coding to work out standard deviation of coded data*
  - *Adding or subtracting DOESNT affect SD, multiplying or dividing does! (Same to the SD!)*
- *Percentiles splits the data up into xth Percentile* $\frac{n}{100}$ *(might need to calculate percentile range)*

Representation of Data

Key Words: stem and leaf diagram, outliers, box plot, histograms, positive/negative skew,

- *An outlier (typically) is any value:*
  - *Greater than the upper quartile + 1.5 x interquartile range*
  - *Less than the lower quartile – 1.5 x interquartile range*
- *Box plot: (outlier), (lowest value), (lower quartile), (median), (upper quartile), (highest value), (outlier)*
- *Histograms (area is proportional to frequency):*
  - $\therefore frequency\ density \times class\ width = frequency$
  - *Frequency density always on the y-axis*
- *Skew is calculated as follows*
  - $Q_2 - Q_1 < Q_3 - Q_2 = +ve$
  - $Q_2 - Q_1 > Q_3 - Q_2 = -ve$
  - $Mode < mean = +ve$
  - $Mode > mean = -ve$
- *Using* $x = \frac{3(mean-median)}{standard\ deviation}$ *if*
  - $x = 1 = perfect\ positive\ skew$
  - $x = -1 = perfect\ negative\ skew$



Probability

Key Words: Venn diagrams, Mutually Exclusive, Exhaustive, Independent, Complementary, Conditional Probability

- *Mutually Exclusive: events that cannot happen at the same time e.g. passing and failing an exam*
  - $P(A\ \mathbf{or}\ B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$ *if NOT ME*
  - $P(A\ \mathbf{or}\ B) = P(A \cup B) = P(A) + P(B)$ *IF ME* $(P(A \cap B) = 0)$
- *Exhaustive: events where all possible outcomes are included e.g. throwing a head or a tail on a fair coin*
  - $P(A\ or\ B) = 1$
- *Independent: one event has no effect on another event occurring e.g. throwing a 1 or a 2 on a fair dice*
  - $P(A\ \mathbf{and}\ B) = P(A \cap B) = P(A) \times P(B)$ *if independent events*
- *Complementary:* $P(A') = 1 - P(A)$
- *Conditional Probability:* $P(A\ \mathbf{given}\ B) = P(A|B) = \frac{P(A \cap B)}{P(B)}$
- *Multiplication rule:* $P(A\ \mathbf{and}\ B) = P(A \cap B) = P(A) \times P(B|A) = P(B)\ x\ P(A|B)$
- *Addition rule:* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Correlation

Key Words: scatter diagram, PMCC "r", coding

- $r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$ $s_{xy} = \sum xy - \frac{\sum x \sum y}{n}$ $s_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$ $s_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$
- $-1 \leq r \leq 1$ *where:*
  - *-1 is perfect negative* ***linear*** *correlation, 0 is no linear correlation and 1 is perfect positive*
- *Use coding to deal with easier values! The PMCC is not affected by any coding*

Regression

Key Words: independent (explanatory) variable, dependent (response) variable, residuals, regression line, interpolation, extrapolation

- *A distance, e, from a point (on scatter diagram) to the of best fit is called a residual*
  - *Outliers have relatively large residuals*
  - *The line of best fit aims to minimise* $\sum e^2$ *- called the regression line of y on x*
  - *Equation of the regression line:* $y = a + bx, where\ b = \frac{S_{xy}}{Sxx}\ and\ a = \bar{y} + b\bar{x}$
- *Interpolation: estimate values* ***within*** *the range of data*
- *Extrapolation: estimate values* ***outside*** *the range of data*

Discrete Random Variables

Key Words: random variable, discrete random variable, cumulative frequency distribution

- *For discrete random variable:* $\sum P(X = x) = 1$
- *In a cumulative frequency distribution* $F(x) = P(X \leq x)$
- $E(X) = \sum xP(X = x)$
- $Var(X) = E(X^2) - \left(E(X)\right)^2$
- $E(aX + b) = aE(X) + b$
- $Var(aX + b) = a^2Var(X)$
- *For a Discrete uniform distribution over the values 1,2,3,..,n:*
  - $E(X) = \frac{n+1}{2}$
  - $Var(aX + b) = \frac{(n+1)(n-1)}{12}$

Normal Distribution

Key Words: random variable, mean, standard deviation

- *If* $X \sim N(\mu, \sigma^2)$
  - $Z \sim N(0,1^2)$ *where* $Z \sim \frac{X-\mu}{\sigma}$ *(called standardising)*
- $\Phi(z) = P(Z < z)$
- $P(X \leq Z \leq Y) = \Phi(Y) - \Phi(X)$