

S1

Representation of Sample Data

Stem and leaf diagram

Each row represents a stem, indicated by the number to the left of the vertical line. The digits to the right are the leaves associated with the stem

Grouped frequency distribution

For the class 5-9:

4.5 is the lower class boundary

5 is the lower class limit

9 is the upper class limit

9.5 is the upper class boundary

5 is the class width ($9.5 - 4.5$)

7 is the class mid-point

Histogram

Area \propto Frequency

Total area \propto Total Frequency

Methods for summarising sample data (location)

Mode is the most frequent value

Median is the middle value

Quartiles are 25% 50% and 75% of the way through an ordered sample

Mean, or μ is defined by $(\sum x) / n$

Methods for summarising data (dispersion)

Variance

$$\sigma^2 = \frac{\sum(x - \mu)^2}{n}$$

Unbiased estimator of the population:

$$s^2 = \frac{\sum(x - \mu)^2}{n - 1}$$

Skew

Positive: $Q_2 - Q_1 < Q_3 - Q_2$

Negative: $Q_2 - Q_1 > Q_3 - Q_2$

Symmetry: $Q_2 - Q_1 = Q_3 - Q_2$

Probability

P(event A or event B) = $P(A \cup B)$

P(event A and event B) = $P(A \cap B)$

P(not event A) = $P(A^c)$

Complementary Probability

$$P(A^c) = 1 - P(A)$$

Addition Rule

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

Conditional Probability

$$P(A \text{ given } B) = P(A | B) = P(A \cap B) / P(B)$$

Multiplication Rule

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

Independent Events

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

Mutually exclusive

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

Correlation

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

Where:

$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \cdot \sum y_i}{n}$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

r is a measure of linear correlation.

$r = 1 \Rightarrow$ perfect positive linear correlation

$r = -1 \Rightarrow$ perfect negative linear correlation

$r = 0 \Rightarrow$ no linear correlation

Regression

Explanatory or independent variable

A variable set independently of another variable

Response or dependant variable

A variable whose values are dependant on another, independent, variable.

Linear regression of line y on x

$$y = ax + b$$

Where:

$$b = S_{xy} / S_{xx}$$

$$a = y_{\mu} - bx_{\mu}$$

y_{μ} means the arithmetic mean of y , and x_{μ} means the arithmetic mean of x

Discrete random variables

Discrete random variable X

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

$$\mu = E(X) = \sum xP(X = x)$$

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

Properties of expected values and variance

$$E(aX + b) = aE(X) + b$$

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

Cumulative distribution function F(x)

$$0 \leq F(x) \leq 1$$

$$F(x_0) = P(X \leq x_0) = \sum P(X = x)$$

The normal distribution

$$\text{Mean} = \mu$$

$$\text{Variance} = \sigma^2$$

Given that $X \sim N(\mu, \sigma^2)$, then:

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