CAIE 9709 Paper 5 - Statistics

Compiled by: Dr Yu

Last updated: January 10, 2021

1 Representation of data

- Non-numerical 非数值 data are called qualitative or categorical data.
- Numerical 数值 data are called quantitative data, and are either discrete or continuous.
  - Discrete 离散型 data can take only certain values.
  - Continuous 连续型 data can take any value , possibly within a limited range.
- Data in a stem-and-leaf diagram are ordered in rows with intervals of equal width.
- In a histogram, column area $\propto$ frequency, and the vertical axis is labelled frequency density.

$$\text{Frequency density } = \frac{\text{class frequency}}{\text{class width}}$$

and

$$\text{Class frequency } = \text{class width } \times \text{ frequency density}$$

- In a cumulative frequency graph, points are plotted at **class upper boundaries**.

2 Measures of central tendency

- Measures of central tendency are the mode 众数, the mean 平均数 and the median 中位数.
- For ungrouped data, the mode is the most frequently occurring value.
- For grouped data, the modal class has the highest frequency density and the greatest height column in a histogram.
- For ungrouped data, $\bar{x} = \frac{\sum x}{n}$.
- For grouped data, $\bar{x} = \frac{\sum xf}{\sum f}$.
- The formulae for ungrouped and grouped coded data can be summarised by:
  - For ungrouped coded data:

$$\bar{x} = \frac{\sum(x-b)}{n} + b \qquad \bar{x} = \frac{1}{a}\left[\frac{\sum(ax-b)}{n} + b\right]$$

  - For grouped coded data:

$$\bar{x} = \frac{\sum(x-b)f}{\sum f} + b \qquad \bar{x} = \frac{1}{a}\left[\frac{\sum(ax-b)f}{\sum f} + b\right]$$

- For ungrouped data, the median is at the $\frac{n+1}{2}$th value.
- For grouped data, we estimate the median to be at the $\frac{n}{2}$th value on a cumulative frequency graph.

3 Measures of variation (差异性)

- Commonly used measures of variation are the range, interquartile range and standard deviation.
- A box-and-whisker diagram shows the smallest and largest values, the lower and upper quartiles and the median of a set of data.
- For ungrouped data, the median Q_2 is at the $\frac{n+1}{2}$th value.
- For grouped data with total frequency $n = \sum f$, the quartiles are at the following values.
  - Lower quartile Q_1 is at $\frac{n}{4}$ or $\frac{1}{4}\sum f$.
  - Middle quartile Q_2 is at $\frac{2n}{4}$ or $\frac{2}{4}\sum f$.
  - Upper quartile Q_3 is at $\frac{3n}{4}$ or $\frac{3}{4}\sum f$.
  - IQR $= Q_3 - Q_1$
- For ungrouped data:

$$\text{Standard deviation } = \sqrt{\text{Variance}} = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}, \quad \text{where } \bar{x} = \frac{\sum x}{n}$$

- For grouped data:

$$\text{Standard deviation } = \sqrt{\text{Variance}} = \sqrt{\frac{\sum(x-\bar{x})^2 f}{\sum f}} = \sqrt{\frac{\sum x^2 f}{\sum f} - \bar{x}^2}, \quad \text{where } \bar{x} = \frac{\sum xf}{\sum f}$$

- For datasets x and y with n_x and n_y values, respectively:

$$\text{Mean } = \frac{\sum x + \sum y}{n_x + n_y} \quad \text{and} \quad \text{Variance } = \frac{\sum x^2 + \sum y^2}{n_x + n_y} - \left(\frac{\sum x + \sum y}{n_x + n_y}\right)^2$$

- The formulae for ungrouped and grouped coded data can be summarised by:

$$\text{Var}(x) = \text{Var}(x - b)$$

$$\text{Var}(x) = \frac{1}{a^2}\text{Var}(ax - b) \quad \text{and} \quad \text{Var}(ax - b) = a^2 \times \text{Var}(x)$$

- For ungrouped coded data:

$$\frac{\sum x^2}{n} - \bar{x}^2 = \frac{\sum(x-b)^2}{n} - \left(\frac{\sum(x-b)^2}{n}\right)^2$$

and

$$\frac{\sum x^2}{n} - \bar{x}^2 = \frac{1}{a^2}\left[\frac{\sum(ax-b)^2}{n} - \left(\frac{\sum(ax-b)}{n}\right)^2\right]$$

- For grouped coded data:

$$\frac{\sum x^2 f}{\sum f} - \bar{x}^2 = \frac{\sum(x-b)^2 f}{\sum f} - \left(\frac{\sum(x-b)^2 f}{\sum f}\right)^2$$

and

$$\frac{\sum x^2 f}{\sum f} - \bar{x}^2 = \frac{1}{a^2}\left[\frac{\sum(ax-b)^2 f}{\sum f} - \left(\frac{\sum(ax-b)}{\sum f}\right)^2\right]$$

4 Probability

- Probabilities are assigned on a scale from 0 (impossible) to 1 (certain).
- When one object is randomly selected from n objects, Pr(selecting any particular object) $= \frac{1}{n}$.
- $$\Pr(\text{event}) = \frac{\text{Number of favourable equally likely outcomes}}{\text{Total number of equally likely outcomes}}$$
- $\Pr(A) + \Pr(A') = 1$.
- In n trials, event A is expected to occur $n \times \Pr(A)$ times.
- $A \cup B$ means 'A or B' and $A \cap B$ means 'A and B'.
- Mutually exclusive events have no common favourable outcomes.
  For mutually exclusive events A and B,
  $$\Pr(A \text{ or } B) = \Pr(A \cup B) = \Pr(A) + \Pr(B)$$
- Non-mutually exclusive events have at least one common favourable outcome.
  For any two events A and B,
  $$\Pr(A \text{ or } B) = \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$
- Independent events can occur without being affected by the occurrence of each other.
  Events A and B are **independent** if and only if
  $$\Pr(A \text{ and } B) = \Pr(A \cap B) = \Pr(A) \times \Pr(B) \quad \text{and} \quad \Pr(A|B) = \Pr(A|B')$$
- For any two events A and B,
  $$\Pr(A \cap B) = \Pr(A) \times \Pr(B|A) \quad \text{and} \quad \Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

5 Permutations and combinations

- We define 0! and
  $$n! = n(n-1)(n-2)\cdots \times 3 \times 2 \times 1,$$
  for any integer $n > 0$.
- A key word that points to a permutation is **arranged**.
  A permutation is a way of selecting and arranging objects in a particular order.
- Key words that point to a combination are **chosen** and **selected**. A combination is a way of selecting objects in no particular order.
- From n distinct objects, there are:
  - ${}^nP_n = n!$ permutations of all n objects.
  - ${}^nP_r = \frac{n!}{(n-r)!}$ permutations of all r objects.
  - $\frac{n!}{p! \times q! \times r! \times \cdots}$ permutations in which there are $p, q, r, \ldots$ of each type.
  - ${}^nC_r = \frac{n!}{r!(n-r)!}$ combinations of r objects.

6 Probability distributions

- A discrete random variable can take only certain values and those values occur in a certain random manner.
- A probability distribution for a discrete random variable is a display of all its possible values and their corresponding probabilities.
- For the discrete random variable X:
  - $\sum p = 1$.
  - $\mathrm{E}(X) = \sum xp$.
  - $\mathrm{Var}(X) = \sum x^2 p - \{\mathrm{E}(X)\}^2$.

7 The binomial and geometric distributions

- A binomial distribution can be used to model the number of successes in a series of n repeated independent trials where the probability of success on each trial, p, is constant.
  - If $X \sim B(n, p)$, then
$$\Pr(X = r) = \binom{n}{r} p^r (1-p)^{n-r}.$$
  - $\mathrm{E}(X) = \mu = np$
  - $\mathrm{Var}(X) = \sigma^2 = np(1-p) = npq$, where $q = 1 - p$.
- A geometric distribution can be used to model the number of trials up to and including the first success in a series of repeated independent trials where the probability of success on each trial, p, is constant.
  - If $X \sim \mathrm{Geo}(p)$, then
$$\Pr(X = r) = p(1-p)^{r-1}, \quad \text{for } r = 1, 2, 3, \ldots$$
  - $\mathrm{E}(X) = \mu = \frac{1}{p}$.
  - $$\Pr(X \leq r) = 1 - (1-p)^r \quad \text{and} \quad \Pr(X > r) = (1-p)^r$$
  - The mode of all geometric distributions is 1.

8 The normal distribution

- A continuous random variable can take any value, possibly within a range, and those values occur by chance in a certain random manner.
- The probability distribution of a continuous random variable is represented by a function called a probability density function or PDF.
- A normally distributed random variable X is described by its mean and variance as $X \sim N(\mu, \sigma^2)$.
- The standard normal random variable is $Z \sim N(0, 1)$.
- When $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$ has a standard normal distribution, and the standardised value $z = \frac{X-\mu}{\sigma}$ tells us how many standard deviations x is from the mean.

- $X \sim B(n, p)$ can be approximated by $N(\mu, \sigma^2)$, where $\mu = np$ and $\sigma^2 = npq$, provided that n is large enough to ensure that $np > 5$ and $nq > 5$.
- Continuity corrections must be made when a discrete distribution is approximated by a continuous distribution.