A Level Mathematics - Statistics

Xingzhi Lu (2129570@concordcollege.org.uk)

2024

1 Statistical sampling

1.1 Populations and samples

1.1.1 Definitions

Population: The whole set of items that are of interest

Census: Observes or measures every member of a population

Sample: A selection of observations taken from a subset of the population which is used

Sampling unit: Individual units of a population

Sampling frame: A list of all people or item that can potentially be involved in the sample

1.1.2 Census

Advantages

- Gives a completely accurate result

Disadvantages

- Time consuming and expensive
- Cannot be used when the testing process destroys the item
- Hard to process large quantity of data

1.1.3 Sample

Advantages

- Less time consuming and expensive than a census
- Fewer people have to respond
- Less data to process than in a census

Disadvantages

- The data may be less accurate
- The sample may not be large enough to give information about small sub-groups of the population

1.2 Random sampling methods

1.2.1 Simple random sampling

Method

1. Each sampling unit is numbered from 1 to n
2. Generate x random number between 1 to n using random number generator / lottery picks / random number tables (or draw out x names with lottery pick)
3. Sampling units corresponding to these numbers become the sample
4. Data taken from the sample

Advantages

- No bias
- Easy and cheap to implement for small populations and small samples
- Each sampling unit has a known and equal chance of selection

Disadvantages

- Not suitable when the population size or the sample size is large
- A sampling frame is needed

1.2.2 Systematic sampling

Method

1. The population is ordered with a unique number each from 1 to n
2. Required elements are chosen at regular intervals i.e. take every kth elements where $k = \dfrac{\text{Population size}}{\text{Sample size}}$
3. Starting at random item between 1 and k using a random number generator
4. Select the remaining data at the chosen interval

Advantages

- Simple and quick to use
- Suitable for large samples and large populations

Disadvantages

- A sampling frame is needed
- It can introduce bias if the sampling frame is not random

1.2.3 Stratified sampling

Method

1. Population divided into groups / stratas
2. Same proportion ($\dfrac{\text{Sample size}}{\text{Population size}}$) sampled from each strata (Work out size of each strata)
3. Simple random sampling carried out in each group
4. Used when sample is large and population naturally divides into groups

Advantages

- Sample accurately reflects the population structure
- Guaranteees proportional representation of groups within a population

Disadvantages

- Population must be clearly classified into distinct strata
- Selection within each stratum suffers from the same disadvantages as simple random sampling

1.3 Non-random sampling methods

1.3.1 Opportunity sampling

Method

1. Sample taken from people who are available at time of study and meet the criteria

Advantages

- Easy to carry out
- No sampling frame required
- Inexpensive

Disadvantages

- Likely to be unrepresentative
- Non-responses are not recorded
- Highly dependent on individual researcher

1.3.2 Quota sampling

Method

1. Population divided into groups according to a given characteristic
2. A quota group is set to try and reflect the group's proportion in the whole population
3. An interviewer or researcher selects a sample that reflects the characteristics of the whole population (opportunity sampling)

Advantages

- Allows a small sample to still be representative of the population
- No sampling frame required
- Quick, easy, inexpensive
- Allows for easy comparison between different groups of population

Disadvantages

- Non-random sampling can introduce bias
- Population must be divided into groups, which can be costly or inaccurate
- Increasing scope of study increases number of groups, adding time or expense
- Non-responses are not recorded

2 Data presentation and interpretation

2.1 Types of data

Quantitative: Associated with numerical observations

Qualitative: Associated with non-numerical observations

Continuous: Can take any value in a given range

Sampling unit: Can only take specific values in a given range

2.2 PMCC

$a = \dfrac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma(x_i - \overline{x})^2}\sqrt{\Sigma(y_i - \overline{y})^2}}$
$b = y - a\overline{x}$

2.3 Interpreting distributions

Measuring central tendency: Mean / median / mode

Measuring variation: Variance / standard deviation / range / interpercentile ranges

2.3.1 Standard deviation / variance

$S_{xx} = \Sigma(x - \overline{x})^2 = \Sigma x^2 - \dfrac{(\Sigma x)^2}{n}$
$Var(x) = E(x^2) - (E(x))^2 = E((x - E(x))^2)$
$\sigma = \sqrt{\dfrac{S_{xx}}{n}} = \sqrt{\dfrac{\Sigma x^2}{n} - (\dfrac{\Sigma x}{n})^2}$

2.3.2 Sample variance

$S^2 = \dfrac{\Sigma(x_i - \overline{x})^2}{n-1} = \dfrac{1}{n-1}(\Sigma x_i^2 - \overline{x}^2 n) = \dfrac{1}{n-1}(\Sigma x_i^2 - \dfrac{(\Sigma x)^2}{n})$

2.4 Transform to linear regression

2.4.1 Exponential

- $y = ab^x \rightarrow \ln y = x \ln b + \ln a$
- x-axis $= x$, y-axis $= \ln y$, gradient $= \ln b$, y-intercept $= \ln a$

2.4.2 Power

- $y = ax^b \rightarrow \ln y = b \ln x + \ln a$
- x-axis $= \ln x$, y-axis $= \ln y$, gradient $= b$, y-intercept $= \ln a$

2.4.3 Logarithmic

- $y = a \ln x \rightarrow$ kept the same
- x-axis $= \ln x$, y-axis $= y$, gradient $= a$

3 Statistical distributions

3.1 Binomial distribution

3.1.1 Notation

$X \sim B(n, p)$

3.1.2 Probability calculation

$\mathrm{P}(x) = \binom{n}{x} p^x q^{n-x}$

3.1.3 Assumptions

- There are a fixed number of trials, n
- There are two possible outcomes only (success and failure)
- There is a fixed probability of success, p
- The trials are independent of each other

3.1.4 Approximation of binomial distribution

If n is large ($n \geq 35$) and p is close to 0.5, then $X \sim N(np, np(1-p))$
When estimating probability ($n \in \mathbf{N}$):

- $\mathrm{P}(X > n) \approx \mathrm{P}(X > [n + 0.5])$
- $\mathrm{P}(X \geq n) \approx \mathrm{P}(X \geq [n - 0.5])$
- $\mathrm{P}(X = n) \approx \mathrm{P}([n - 0.5] < X < [n + 0.5])$
- $\mathrm{P}(X < n) \approx \mathrm{P}(X < [n - 0.5])$
- $\mathrm{P}(X \leq n) \approx \mathrm{P}(X \leq [n + 0.5])$

3.2 Normal distribution

3.2.1 Notation

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

3.2.2 Conditions

- Mean = median = mode
- Continuous variable
- Symmetrical distribution

3.2.3 Shape of distribution

- Symmetrical shape (mean = median = mode)
- Bell-shaped curve with asymptotes at each end
- Total area under curve = 1
- Has points of inflection at $\mu + \sigma$ and $\mu - \sigma$
- Approximately 68% of data lies within 1 s.d. from mean, 95% within 2 s.d., 99.7% (nearly all) within 3 s.d.

4 Hypothesis testing

4.1 Definitions

Null hypothesis H_0: $\theta = \theta_0$

Alternative hypothesis H_1: $\theta \neq \theta_0$ / $\theta > \theta_0$ (right tail) / $\theta < \theta_0$ (left tail)

Significance level: Probability of rejecting H_0 when assuming H_0 is true

4.2 Test on proportion / probability of success if binomial: $B(n, p)$

4.2.1 By critical value

If stats test $t^* >$ cv: reject H_0
Else if stats test $t^* <$ cv: accept H_0

4.2.2 By p value

If $P(t \geq t^*) < \alpha$: reject H_0
Else if $P(t \geq t^*) > \alpha$: accept H_0

4.2.3 Test population mean with unknown s.d.

$n > 35$: CLM, see below
$n < 35$: t-test, see below

4.3 Probability calculation

If n is large enough ($n \geq 35$), then sample mean $\overline{x}$ is normally distributed: $\overline{x} \sim N(M_{\overline{x}}, \sigma^2_{\overline{x}})$ (Central limit theorem)

- $M_{\overline{x}} = M$
- $\sigma^2_{\overline{x}} = \dfrac{1}{n}\sigma^2 \to \sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$

If $n \geq 35$, sample proportion $\hat{p}$ with an attribute is normally distributed: $\hat{p} \sim N(M_{\hat{p}}, \sigma^2_{\hat{p}})$

- $M_{\hat{p}} = p$ (mean of $\hat{p}$ is population proportion p)
- $\sigma^2_{\hat{p}} = \dfrac{p(1-p)}{n}$

t-test: if sample size $n < 35$, $\overline{x} \sim t(M, (\dfrac{S}{\sqrt{n}})^2)$, degree of freedom $= n - 1$

- $S = \sqrt{\dfrac{\Sigma(x_i - \overline{x})^2}{n-1}}$
- t-test stats $= \dfrac{\overline{x} - M}{\dfrac{S}{\sqrt{n}}}$
- If t-test stats $>$ critical value (check on data sheet) then reject H_0

4.4 Hypothesis test for zero correlation

- H_0: $\rho = 0$
- H_1: $\rho \neq 0$
- Check the data sheet for cv
- Sample $r >$ cv $=$ reject H_0, else: not reject H_0