Unit 1

2023年6月5日 13:03

1.1-1.4 Data Definitions

- Definitions
  - https://quizlet.com/cn/722612048/flash-cards/
- Primary vs secondary data

	Pro	Con
Primary	More reliable an specific Up to date	More time, money, effort
Secondary	Save time, effort and money	Not sure about collection process Not recent Too general

- Census vs sample

	Pro	Con
Census	More data More accuracy unbiased	Time and money consuming Unhappy experience Testing to destruction
Sample	Quicker, cheaper Repeatable	Could be biased Unrepresentative

1.5 Pearson capture-recapture formula

- Formula
  - $\frac{m}{n} = \frac{M}{N} \ or \ N = \frac{Mn}{m}$
- Assumptions
  - The population has not changed
  - The probability of being caught is equal for all individuals
  - Marks are not lost and are always recognisable
  - The sample size is large enough to be representative of the population
  - The population have been thoroughly mixed

1.6 Random sampling

- Random sampling
  - Created when each member of the population has an equal chance of being included in the sample
- Simple random sampling method
  - Each sampling unit is numbered from 1 to n
  - Generate x random number from 1 to n
  - Ignore repeats
  - Methods:
    - Random number tables
    - Random number generator on calculator or computer
    - Picking random numbers by lottery
  - Sampling units corresponding to these numbers become the sample
  - Data taken from the sample
- Simple random sampling advantages
  - Bias free
  - Easy and cheap to implement
  - Each number has a known equal chance of being selected
- Simple random sampling disadvantages
  - Not suitable when population size is large
  - Sampling frame is needed

    - Chance of being unrepresentative

1.7 Non-random sampling

- Cluster sampling
    - Used when the population can be easily put into groups e.g. towns
    - A good mixture of the population characteristics within each cluster
    - The list of clusters is the sampling frame
    - One or more cluster is picked at random and is used as the sample
- Judgment sampling
    - Use your own judgement to select a sample representative of the population
- Opportunity sampling
    - Sample taken from people who are available at time of study and meet the criteria
- Opportunity sampling advantages
    - Easy to carry out
    - No sampling frame required
    - Inexpensive
- Opportunity sampling disadvantages
    - Likely to be unrepresentative
    - Non-responses are not recorded
    - Highly dependent on individual researcher
- Quota sampling
    - Population is divided into groups according to characteristic
    - A quota group is set to try and reflect the group's proportion in the whole population
    - Interviewer set the actual sampling units (opportunity sampling with strata)
    - = stratified without sampling frame
- Quota sampling advantages
    - Allows small sample to still be representative of population
    - No sampling frame required
    - Quick, easy, inexpensive
    - Allows for easy comparison between different groups of population
- Quota sampling 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
- Systematic sampling
    - The population is ordered with a unique number each from 1 to n
    - Required elements are chosen at regular intervals i.e. take every k^th elements where $k = \frac{pop\ size\ (N)}{samp\ size\ (n)}$
    - Starting at random item between 1 and k using a random number generator
    - Select the remaining data at the chosen interval

1.8 Stratified sampling

- Stratified sampling
    - Population divided into groups / stratas
    - (Work out size of each strata)
    - Same proportion ($\frac{samp\ size\ (n)}{pop\ size\ (N)}$) sampled from each strata
    - Simple random sampling carried out in each group
    - Used when sample is large and population naturally divides into groups
- Advantages
    - Reflects population structure
    - Guarantees proportional representation of groups within population
- Disadvantages
    - Population must be clearly classified into distinct strata
    - Selection within each stratum suffers from same disadvantages as simple random sampling

1.9 Collection of data

- Types of experiments
  - Laboratory experiments
    - Carried out under controlled conditions
  - Field experiments
    - Carried out in an everyday environment
    - Can control some variables
  - Natural experiments
    - Carried out in an everyday environment
    - Researcher has no control over any variables
    - Often involves a survey
- Extraneous variable
  - Any variable that you are not interested in but that could affect the results of your experiment
- Explanatory variable = independent
- Response variable = dependent

1.10 Questionnaires and interviews

- Methods of survey

	Advantages	**Disadvantages**
Questionnaire by post / email	Cheap Convenient - no time pressure More data in - easy to send the questionnaire, large sample size Can be anonymous - more honest, answer personal questions No interviewer bias	Low response rate - refuse to do it, forget/miss it May not understand the questions May not understand the respondent's answers
Telephone	Convenient Better accuracy Clarify, ask questions Call back if missed	Miss call / not answer More time consuming / more expensive
Interview	Clarify and explain questions, ask further questions Respondents can explain answers More reliable, precise Interviewer can put people at ease when answering personal questions Hard to avoid / higher response rate - everyone interviewed answers	Time consuming / expensive May be less honest in interview and less likely to answer personal questions Interviewer bias - they may interpret answers to suit their opinions

- Pilot survey
  - A survey conducted on a small sample to test the design and methods of a much bigger main survey
  - Check for:
    - Respondents' understanding of the question
    - If the question collects the required data
    - Is the question phased correctly
    - Usefulness of the question
- Reasons for poor questions
  - Unclear
  - Leading
  - Vague and interpreted differently
  - Open question
- Characteristics of good questions
  - Short, simple and to the point
  - Should get a response from everyone
  - Language used should be easily understood
  - Should not embarrass the respondents

    - Not leading
    - Address a single issue at a time
    - Responses does not overlap
- Open question
    - Do not have any suggested responses and leave the respondent to give their own answer
    - Advantage: may reveal answer not previously considered
    - Disadvantage: there may be many different responses, difficult to analyse
    - Good for pilot survey
        - Other, please write _________
- Closed question
    - Have a set of responses for the respondent to choose from
    - Opinion scales are often used
        - Disadvantage = people may be reluctant to express strong opinions
    - Response boxes should not overlap / should be exhaustive

1.11 Problems with collected data

- Outlier / anomaly
    - Ignore if it is due to measurement or recording error
    - Don't ignore if it is not wrong
- Cleaning data
    - Identifying and correcting / removing inaccurate data values (caused by recording or other errors) or extreme values
    - Removing units or other symbols from data
    - Deciding what to do about missing data

1.12 Controlling extraneous variables

- Control group
    - Often used to test the effect of various factors in an experiment
    - Randomly selected and not subjected to any factor tested
    - The experimental group is affected
    - Two groups should be as similar as possible
        - e.g. similar age
    - Allow comparison
- Matched pair
    - Two groups of people are used to test the effects of a particular factor
    - Each individual in one group is paired with an individual in the other group who has everything in common with them except the factor being studied
    - Identical twins are very important are these
- Matched pair benefits and difficulties
    - Benefits
        - Share common features
        - Minimised difference through the matching process
        - Allow better comparison
    - Difficulties
        - Difficult to match pairs if identical twins are not available

1.13 Hypotheses

- Data planning cycle
    - Specifying and planning
        - Design the investigation e.g. write an questionnaire
        - Write an hypothesis
    - Collecting data
        - e.g. filling in questionnaires, collecting results
    - Processing and representing
        - Show results in table / graph
        - Work out averages
    - Interpreting and discussing

    - Is the hypothesis true or false?
- Characteristics of a good hypothesis
    - Need to be able to prove and test
    - Not vague, specific
    - A statement

1.14 Designing investigations

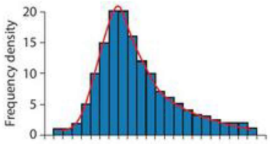
- Factors needed to consider

Factor	**Things to consider**
Time	How long to set up and carry out investigation
Cost	How much the investigation is to set up and carry out Do you need special equipment / lab / paying interviewees or participants
Ethical issues	No participant should be harmed physically or mentally Respect people's dignity and rights
Confidentiality	Will people answer sensitive questions How to keep data secure and confidential
Convenience	Is the data get locally, cheaply and in a short enough time frame
How to select your population and sample	Identify the population interested in Sampling method
How to deal with non-response	Number of response needed How many people needs to be asked to ensure this many
How to deal with unexpected results	How to investigate likely results before running a survey (pilot survey) How to deal with anomalous results

Unit 2

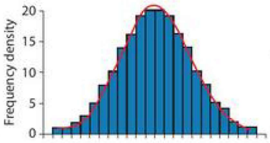
2022年10月31日 17:03

2.4 Bar charts

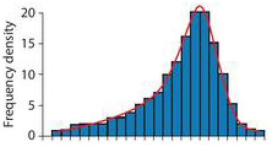
- Comparative bar charts
  - $F_1 : A_1 = F_2 : A_2 \; or \; \mathrm{F}_1 : F_2 = A_1 : A_2$
- Types of skew
  - 

    This distribution has positive skew. Most of the data values are at the lower end. Example: The age at which a person learns to write.

    The distribution is stretched out in the positive direction →.

    This distribution is symmetrical. It has no skew. Example: The lengths of leaves on a tree.

    This distribution has negative skew. Most of the data values are at the upper end. Example: The age at which a person dies.

    The distribution is stretched out in the negative direction ←.
- Interpreting skew
  - Positive skew
    - More than half of the _______ lower than the mean
    - _______ above the median has a greater spread
    - Mean > median
  - Negative skew:
    - More than half of the _______ higher than the mean
    - _______ below the median has a greater spread
    - Mean < median
  - Symmetrical
    - Equally spread out on either side of the median
    - Mean = median

Misleading diagrams

- Scales not starting at zero
- Scale not increasing uniformly
- Lines too thick to be read properly
- Axes not labelled properly
- Things at the front look larger than those at the back in 3D diagrams, angle can be distorted
- Sections separated from the main diagram can make comparisons difficult
- Dark colours tend to stand out more than light colours and may look bigger
- Differently sized bars / pictures - not sure to compare area or height
- Excluded data

Choosing the right format

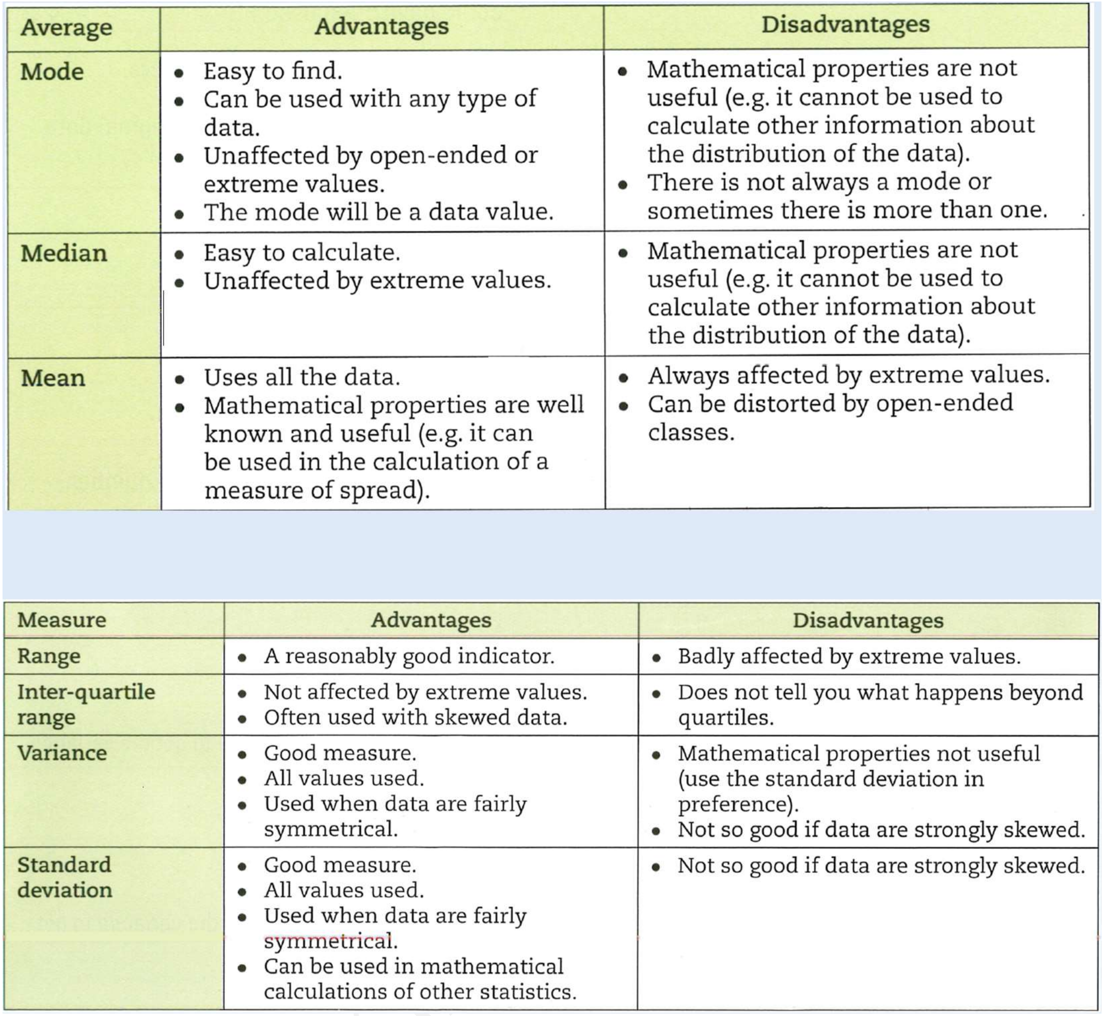
Discrete	Step frequency, frequency polygon
Continuous	Stem and leaves, cumulative frequency, histogram
Categorical	Pie, bar, pictogram, tally

- Show proportion - pie charts / composite bar charts
- Exact value - stem and leaves

Unit 3

2023年6月12日 13:45

3.11 Deciding upon the most appropriate average/ measure of spread

- 

Average	Advantages	Disadvantages
Mode	• Easy to find. • Can be used with any type of data. • Unaffected by open-ended or extreme values. • The mode will be a data value.	• Mathematical properties are not useful (e.g. it cannot be used to calculate other information about the distribution of the data). • There is not always a mode or sometimes there is more than one.
Median	• Easy to calculate. • Unaffected by extreme values.	• Mathematical properties are not useful (e.g. it cannot be used to calculate other information about the distribution of the data).
Mean	• Uses all the data. • Mathematical properties are well known and useful (e.g. it can be used in the calculation of a measure of spread).	• Always affected by extreme values. • Can be distorted by open-ended classes.

Measure	Advantages	Disadvantages
Range	• A reasonably good indicator.	• Badly affected by extreme values.
Inter-quartile range	• Not affected by extreme values. • Often used with skewed data.	• Does not tell you what happens beyond quartiles.
Variance	• Good measure. • All values used. • Used when data are fairly symmetrical.	• Mathematical properties not useful (use the standard deviation in preference). • Not so good if data are strongly skewed.
Standard deviation	• Good measure. • All values used. • Used when data are fairly symmetrical. • Can be used in mathematical calculations of other statistics.	• Not so good if data are strongly skewed.

3.12-3.13 Comparing distributions and making estimates

- A full comparison needs a **measure of central tendency** Median / Mean

  a **measure of dispersion (spread)** (Range), IQR, s.d.

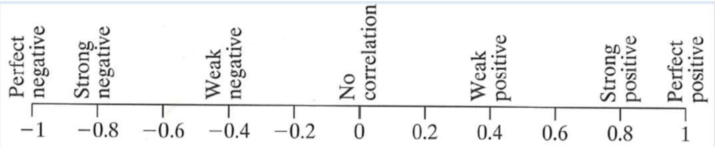
  and if possible a **comparison of skewness** +ve, -ve, symmetrical

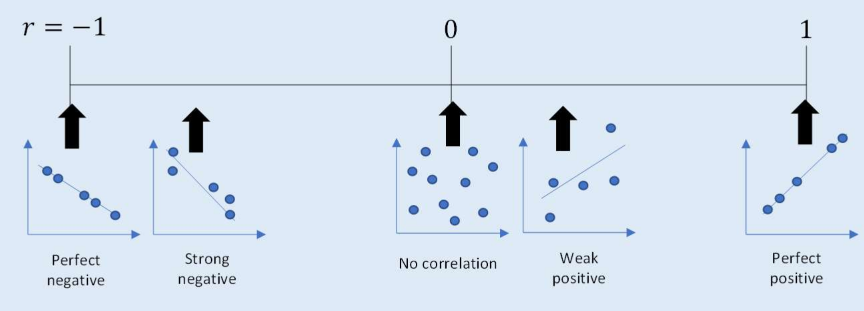
  Also, try to make a statement **in the context of the question**
- Describing relations
  - Gradient: increasing more per unit / less per unit / faster / slower

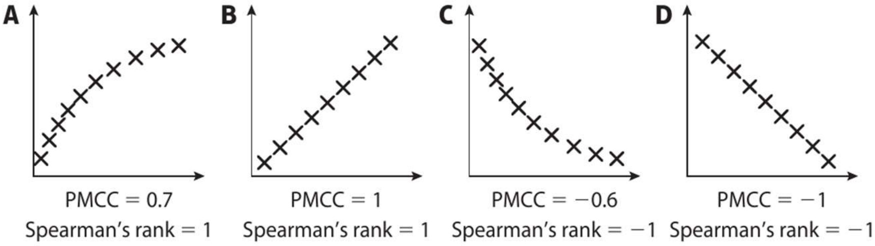
Unit 4

2024年2月27日 14:29

4.7-4.9 Correlation coefficients

- Spearman's rank correlation coefficient / r_s
  - $r_s = 1 - \dfrac{6\sum d^2}{n(n^2-1)}$
  - Values tied = given mean of rankings
- 

- Pearson's product moment correlation coefficient
  - A measure of linear correlation
  - 

- 

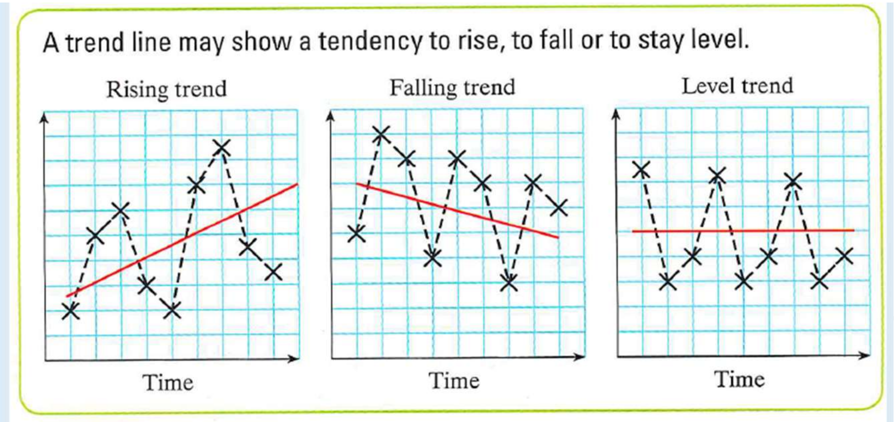
Unit 5

2023年9月19日 14:53

<u>5.1 Line graphs and time series</u>

- Time series
  - A set of observations (data values) taken over a period of tine

<u>5.2-5.3 Trend lines and variations</u>

- Types of trends
  - 

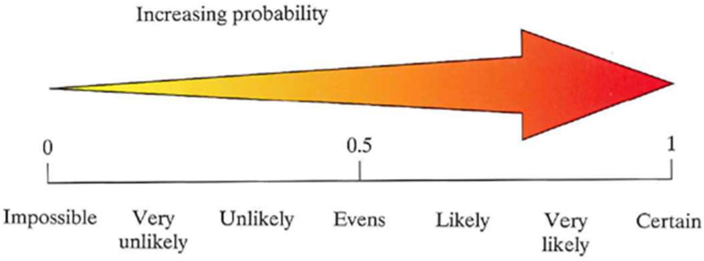
    A trend line may show a tendency to rise, to fall or to stay level.

- Trend line
  - Show the general trend of the data
- Seasonal variation
  - Variation in a time series following a regular time period
- Assumptions for extrapolating
  - Overall trend continues
  - Seasonal trend continues

Unit 6

2023年10月20日　　14:28

6.1-6.2 The Meaning of Probability and Experimental Probability

- Probability
  - A numerical measure of the chance of an event happening
  - 0 = impossible
  - 1 = certain
  - 

- Trial
  - The act of testing / doing something
- Outcome
  - Result of a trial
- Event
  - A set of one or more successful outcomes
  - A subset of the sample space
- Expected frequency
  - The number of times you expect the event to happen
- Theoretical probability
  - Uses mathematics rather than an experiment to determine the chance of something happening.
- Experimental probability
  - Aka estimated probability
  - The probability of an event happening based on an experiment or observation.
  - $\text{expermental probability} = \dfrac{\text{number of trials with successful outcome}}{\text{total number of trials}}$
  - * Larger sample size / higher number of trials = estimate for probability gets closer to the true value

6.3 Risk

- Risk
  - The probability of an negative event happening
  - $\text{risk of an event}=\dfrac{\text{number of trials in which event happens}}{\text{number of trials}}$
- Absolute risk
  - The probability of an event happening
- Relative risk
  - How many times more likely an event is to happen for one group compared to another group
  - $\text{relative risk for a group}=\dfrac{\text{risk for those in group}}{\text{risk for those not in group}}$

6.4-6.5 Sample space and Venn diagrams

- Sample space
  - A list of all possible outcomes
- Sample space diagram
  - Using a table to represent the sample space
  - Can be used to find the probability of a set of events

6.6-6.7 Mutually Exclusive, Exhaustive events and the General Addition Law

- Mutually exclusive
    - Cannot happen at the same time
    - $P(A \cap B) = 0$
    - $P(A \cup B) = P(A) + P(B)$
- Exhaustive
    - A set of event is exhaustive if the set contains all possible outcomes
- General addition law
    - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Unit 7

2023年11月21日 23:25

7.1 Simple index number

- Index number = $\frac{\text{current price}}{\text{base year price}} \times 100$

7.2 RPI, CPI, GDP and weighted index mean

- RPI
  - Retailed Price Index
  - Shows the rate of change of prices in everyday life
    - e.g. mortgage payments, food, heating and petrol
  - The UK government uses the RPI to set the interest rate for student loans
- CPI
  - Consumer Price Index
  - Measures the rate of price changes in everyday life but does not include mortgage payments
  - State benefits and pensions in the UK are updates each year in line with the CPI
- GDP
  - Gross Domestic Product
  - The value of goods and services a country produces within a stated time period (usually one year)
- Recession
  - An economy is in recession when its GDP falls in two or more successive quarters
- Weighted index number = $\frac{\sum \text{index number} \times \text{weight}}{\sum \text{weight}}$

7.3 Chain base index number

- Chain base index numbers
  - Compare prices from each year / month / week with the previous year / month / week etc.
  - The chain base index number for each year is that year's price as a percentage of the previous year's price, written without the percentage sign

7.4 Crude and standardised rates of change

- Crude rate = $\frac{\text{number of deaths, births, etc.}}{\text{total population}} \times 1000$
- Standard population in an age group = $\frac{\text{number of in age group}}{\text{total population}} \times 1000$
- Age specific crude rate = $\frac{\text{number of deaths, births, etc. in that age group}}{\text{total population in that age group}} \times 1000$
- Age specific standardised rate = $\frac{\text{age specific crude rate}}{1000} \times$ age specific standard population
- Standardised rate = $\sum$ age specific standardised rate

$$= \sum \frac{\text{age specific crude rate}}{1000} \times \text{standard population of that age group}$$

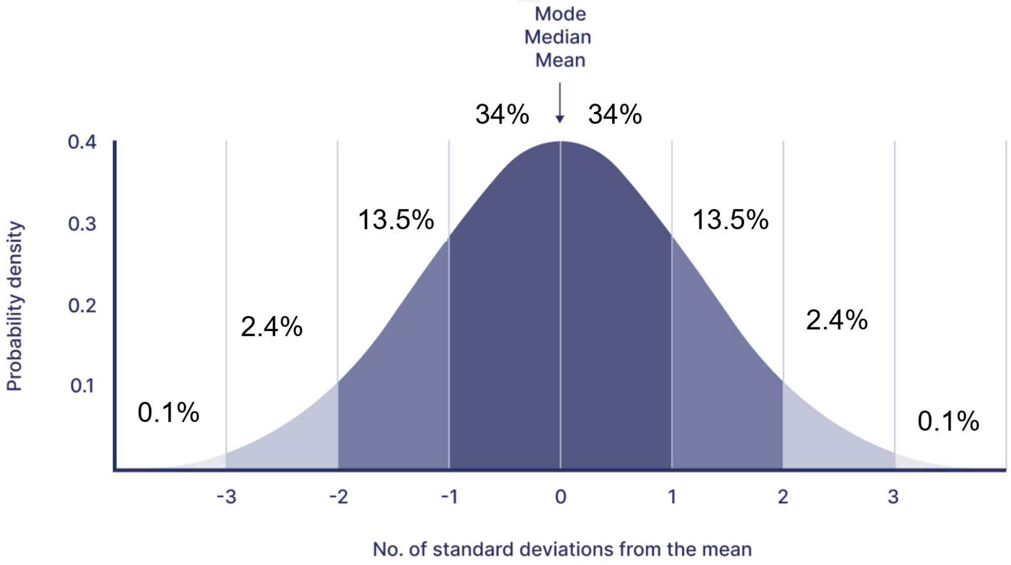
Unit 8

2024年1月23日 14:56

8.1 Binomial distribution

- Probability distribution
  - A list of all the possible outcomes of an experiment, together with their probabilities
- Binomial distribution conditions
  - There must be a fixed number of trials
  - There are only two possible outcomes for each trial: success and failure
  - The probability of success (p) and failure (q) are fixed
  - The probability of success in any trial is independent of the outcomes of other trials
- X ~ B(n, p)
- Mean value / expected value = np

8.2 Normal distribution

- Conditions of normal distribution
  - The data is continuous
  - The distribution is symmetrical and bell-shaped
  - The mode, median and mean are approximately equal
- Properties of normal distribution
  - Symmetrical about the mean (skew is close to 0)
  - Mode = median = mean
  - 68% of observations lie within ± 1 standard deviation of the mean
  - 95% of observations lie within ± 2 standard deviation of the mean
  - 99.8% (virtually / nearly all) of observations lie within ± 3 standard deviation of the mean
  - 

- Sketching normal distribution diagram
  - Peak = mean
  - ★ Higher standard deviation = more spread out, lower peak
- X ~ N(mean, <u>variance</u>) / X ~ N(mean, <u>(standard deviation)^2</u>)

8.3 Standardised scores

- Calculation
  - $z = \dfrac{score - mean}{s.d.}$
  - Score > mean = positive, score < mean = negative

8.4 Quality assurance and control charts

- Quality assurance

  - Involves checking samples to ensure that the product of a manufacturing process meets the required standards
  - Mean of sample = more closely distributed than individual samples
- Control charts
  - A time series used for quality assurance
  - Assume normally distributed:
    - 95% between two warning limits ($\mu \pm 2\sigma$)
    - Inside warning limits = process in control, acceptable
    - 99.8% between two action limits ($\mu \pm 3\sigma$)
    - Between warning + action: another sample is taken to check that nothing has gone wrong, production not stopped if still in warning limit, outside warning limit = stop production and reset machines
    - Outside action limits = process gone wrong, production stopped + machines are reset
- Quality control charts for the ranges of samples
  - Action and warning limits given in exam
  - May not have a lower warning / action limit