Reminders

Binomial

- Assumptions:
  - Probability of success is constant
  - Events are independent
- Unsuitable model
  - People may collaborate to buy stuff

Poisson

- Assumptions:
  - Average/mean rate is constant
  - Events are independent
- Unsuitable model
  - The mean number of… is unlikely to be constant throughout the year, as it would likely vary across different months in a year due to seasonal fluctuations caused by factors such as the holiday seasons

Regression

- Model not suitable
  - Linear model: Scatter plot shows that y increases at a decreasing rate as x increases
  - In the long term/future prediction → will exceed maximum/minimum point, and will eventually go to negative. Not possible to increase indefinitely
- Comparing models: use r-value

Hypothesis testing

- T-test:
  - Population variance unknown (s^2)
  - N is small
  - Population X follows normal (assumption)

Sampling

Quota

To find out how often people swim in the public swimming pools, a predetermined number of males and females are stopped and surveyed on a street regarding how often they use the public swimming pools in a month. Quota Sampling is appropriate in this situation as it is easy and convenient to obtain samples in this manner since the sampling frame is not available. One disadvantage may be that the samples drawn may be biased, since they are not randomly drawn and the surveyors might select people who are more willing to participate in the survey. It is not possible to apply stratified sampling as the sampling frame is not available.

OR

Carry out a survey on the number of hours students spent on watching television every week in a primary school. Subgroups are based on age and an equal number of students is non-randomly selected from each level, for instance, the first ten students from the first class of each level (primary one, two,to primary six) is selected to participate in the survey. It is appropriate as it is easy to

obtain samples. One disadvantage is that samples drawn may be biased. Possible since the sampling frame is readily available from the school administration. The sampling units may be chosen randomly within each subgroup, with the number proportional to the relative size of the stratum.

The manager can choose to survey, say, 50 male cinema– goers and 50 female cinema–goers. He can conveniently approach anyone who walk into the cinema or queueing up to buy ticket until he has met his quota.

We can station ourselves at the exit, pick a random person within the first 100 that leaves, and then pick every 100th person after that.

Advantages

Representativeness of sample: quota sampling allows the survey to capture the responses that various groups of students. This may be preferred as certain homeroom or sports facilities may not be in as good a condition as others, and the representation of each group will ensure that the results will not be biased towards those who are often using less functional facilities or towards those who are often using the more functional facilities.

Efficiency of collecting the sample: quota sampling may be more efficient as systematic sampling in the case requires the surveyor to identify the selected respondents and to contact them, which can be time consuming

Disadvantages

Non-randomness/selection bias: quota sampling is non-random and may contain selection bias, where the surveyor chooses people who may appear friendlier or choose students in the canteen only at a selected time period. This results in certain students having no chance of being selected at all, which may affect the validity of the survey results

Non-representativeness of sample: quota sampling may result in a group being excluded entirely from the selection, which may result in t data collected being an inaccurate representation of the entire school population

Systematic

Number the club members in order from 1 to 15000 according to the name list in alphabetical order. Since k = 30, select a member randomly from the name list. Thereafter, select every 30^{th} member cycling to the start of the list if the end of the list is reached until we form a sample of 500 members

First, obtain a list of all pupils and assign sequential numbers to all of them. Next, determine the sampling interval = 950 19 50 = . Then, choose a random start by selecting a random number from 1 to 19. Finally select every 19th student subsequently until 50 pupils are obtained. Stratified sampling which could draw random samples from gender-group might be preferable as it could ensure the representation of the opinion of gender across the entire population.

Assign a number from 1 to N to each of the students, where, N represents the student population size/obtain a list of the students from the administration office in rderof their identification numbers or registration numbers. Next, determine the sampling interval size, k=1/0.02 =50. Randomly select any student from the list, say the 1^{st} student. Select every 50^{th} student thereafter (ie 51^{st}, 101^{st}), until the required sample is obtained

Disadvantage

There is a bias when the name list of the members of the fitness club have a periodic or cyclic pattern ie there is some pattern in the way the names is arranged and the pattern coincides in some way with the sampling interval of 30

Stratified

Choose 8 2-bedroom apartments and 27 4-bedroom apartments. Using the list of apartment numbers as sampling frames for the two types of apartments, use simple random sampling to select the 2 bedroom apartments and 4-bedroom apartments to be interviewed

Given the nature of the event, it will be difficult to obtain the list of all spectators, i.e. the sampling frame. Thus it is not possible to use stratified sampling.

Advantages

Stratified sampling ensures that households occupying two-bedroom apartments and households occupying four-bedroom apartments are proportionately represented in the sample

Hypothesis testing

Since p-value 0.0382 0.05 = < , we reject H0 and conclude that there is sufficient evidence, at 5% level of significance, that the population mean time for a

student to complete the project exceeds 30 hours.

It is not necessary to assume that the population follows a normal distribution for the test to be valid because since n =150 is large, by Central Limit Theorem, X follows normal distribution bv

The meaning of at the 5% significance level is that there is a probability of 0.05 that it was wrongly concluded that the company had understated the mean amount of loans borrowed by its clients.

There is a probability of k% that the test will conclude that the mean mass of garderner wholemeal bread is not 400g, when it is actually 400g

Binomial

Two assumptions are: • the colour of each car observed must be independent of the colour of any other car observed. • the probability that any one car observed is red is the same throughout the sample.

The eye colour fo a person is independent of that of another person

Deals closed may not be independent of one another as customers may collaborate to buy cars as a group for better bargaining power.

Normal

Part (iii) includes the cases in part (ii) as well as other cases. For example, 6 7 < < C and T > 56 so that C T+ > 62 is included in part (iii) but not in part (ii). Hence, the answer for (iii) would be greater than the answer in (ii)

One possible reason could be that the mean number of grand pianos sold every week would not be constant from one week to another because of seasonal fluctuations such as sales, holidays, the economic climate etc.

This means that as the company buys more and more components from supplier A, it is more likely that a randomly chosen component that is faulty was supplied by A.

We assume that the thicknesses of all textbooks are independent of one another.

Poisson

The mean number of cars joining the immigration checkpoint queue for any subinterval of the same length of time within 1 hour is constant.
OR cars join the immigration queue independently of one another, throughout the entire hour

Sales of all refrigerators are independent of one another

Average rateof the refrigerators being sold is constant

Unsuitability

The mean number of cars joining the immigration checkpoint queue every hour may not be constant due to peak periods as there may be more cars heading to or returning from work

2 brands of refrigerator are in the same price range and they can be competing in terms of sales

The average re of refrigerators sold is unlikely to be a constant due to sale, festive seasons, economic conditions etc

The mean number of guests checking into the hotel per hour is unlikely to be constant throughout the year. The number of guests checking into the hotel is likely to vary across different months in a year due to seasonal fluctuations caused y factors such as the holiday seasons

Regression and correlation

Even though $r \approx -0.912$ indicates a high negative linear correlation between the variables, x and t, the linear model may not be suitable as the scatter diagram shows that the points follow a curved nature. Also, it is not possible to have $x = -11.7$ as concentration of blood cannot be negative.

Using the GC, the product moment correlation coefficient, $r \approx 0.970$ (3s.f.). Since value of r is close to 1 which suggests a strong positive linear correlation, a linear model seems appropriate.

With the point P removed, the values of t increases as x increases, but by decreasing amounts. Hence it is consistent with a model of the form

The value of $x = 8.0$ is outside the data range of x. Despite the high value of r, such extrapolation made the estimate unreliable.

The scatter diagram in (i) shows F increasing at an increasing rate as v increases, so a quadratic model is a better one.

Model B is the most appropriate since x increases at an increasing rate as T increases, which is the same as the shape shown in the scatter diagram

Reliability

Since x=40 is within the range of values of x, [11,97] and the product moment correlation coefficient, -0.994 has an absolute value that is close to 1, suggesting a strong linear correlation between the variables y and 1/x, therefore the estimate is reliable

Unsuitability

A linear model is not appropriate because the scatter diagram indicates that as increases, s is increasing at a decreasing rate which is not a linear relationship. Furthermore, a linear model will mean that the product's monthly sales in Singapore will increase indefinitely with the increase of the number of promoters. This is not realistic in the context of the question as the product's monthly sales will likely slow down and perhaps decrease due to market saturation.

A linear model may not be appropriate since there is a certain limit to how fast a person can complete the distance. This is also evident in the scatter diagram, which shows a slight reduction in the rate of decrease of the record time.

A linear model predicts the average diameter will keep increasing indefinitely without any limit. Therefore a linear model isnot appropriate.

A quadratic model predicts that the average diameter will eventually attain a maximum value an thereafter decrease as the age increases, till it eventually takes on negative values. This is not possible, therefore a quadratic model is not appropriate.

A quadratic model would show the record time increasing again in the future, which is impossible. Hence a quadratic model would not be appropriate