GUIDE to Answering Qualitative Questions for Statistics (assumptions, explain, comment etc)

(some edits updated 20 Oct 2023)

This guide hopefully helps you to answer qualitative questions more confidently.

SCAN this QR code to access to the soft copy.

1. Binomial Distribution

Typical Context:

- Question talks about some characteristic of the random variable being studied, eg, faulty object/ rotten fruit.
- X is the random variable for the number of (faulty objects/rotten fruits), out of n (objects)
- There is a probability of success, p, refers to the probability of that characteristic happening.
- State clearly in mathematical notations: **X ~ B (n, p)**

QUESTION: State, in the context of the question, **two assumptions** needed to model X by a binomial distribution.

PHRASING SUGGESTION:

1. The event of a (*object/fruit*) is (*faulty/rotten*) is independent of the event of any other (*object/fruit*) being (*faulty/rotten*).
(this actually conveys the idea that the EVENT of the faulty thing happening to one object does not affect anyther object being faulty.)

2. The probability that each(*object/fruit*) is (*faulty/rotten*) is constant throughout for all (*objects/fruits*).
(this phrasing on probability constant should be cited especially for those questions that mention something like "On average, 8% of mugs are faulty." This is because if given as an average value, we need to assume that the probability of selecting something of the characteristic needs to be constant for it to model by binomial.)

Note the key words: EVENT INDEPENDENT; PROBABILITY CONSTANT.

Never use the phrasing "probability of…is independent of…" because only events are independent AND NOT the probability value that is independent.

2. Normal Distribution

QUESTION: Explain whether the normal distribution is a suitable model.

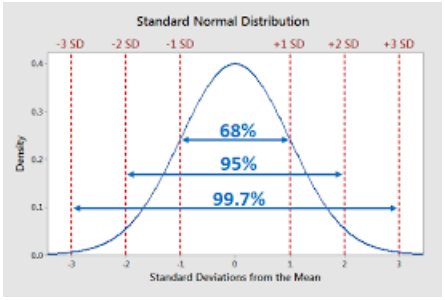
***(Context**: It was found that the time a teenager spends per day on Twister social media is a random variable with a mean of 4 minutes and a standard deviation of 2 minutes.)*

PHRASING SUGGESTION:

Let *T* be the time spent by a teenager on Twister.

Suppose $T \sim \mathrm{N}\left(4, 2^2\right)$,

Then according to the empirical rule, **99.7%** of the values of *T* will lie within 3 standard deviations. This means $4 \pm 3(2) = (-2, 10)$ which contains a significant range of negative values which does not make sense with time. Hence, the normal distribution is not a suitable model.



(pls note question mentions data context has mean of 4 minutes and a standard deviation of 2 minutes DOES NOT MEAN that it is normal distribution automatically.)

3. Hypothesis Testing (Significance testing)

- Denote *X* as the random variable for the life span of the front tyres.
- Denote μ to be the population mean life span of the front tyres.
- Denote H_0: μ =20000
- Denote H_1: μ > 20000

(A) Explain 5% Level of Significance

5% level of significance means that there is a 0.05 probability that the test will indicate to reject the claim (Ho) that the speed of PMD riders is 10km/h or less when in fact the claim is true.

(recall defn of LOS: Thus, the **significance level** $\alpha \times 100\%$ is the percentage (probability) of rejecting H_0 (*write the stem phrase of problem context*) given that H_0 is true.)

(B) Explain the p-value in context (eg if p-value is 0.03)

The *p*-value represents the smallest level of significance for which the null hypothesis that the mean speed of PMD riders is 10km/h or less would be rejected. (*cite the context which is about mean speed of PMD riders*)

(Recall defn of P-value: The ***p*-value** (observed significance level) is the smallest value of the significance level α for which the null hypothesis would be rejected.)

(C) Conclusion statement

To test $H_0 : \mu = 70$ against $H_1 : \mu \neq 70$

If Conclusion is Do Not Reject Ho, write:

Do Not Reject Ho. There is **insufficient** evidence at the 5% level of significance to conclude that the mean mass of the cereal packets has changed (*H1 stem phrase*).

If Conclusion is Reject Ho, write:

Reject Ho. There is **sufficient** evidence at the 10% level of significance to conclude that the mean mass of the cereal packets has changed (*H1 stem phrase*).

4. Correlation and Regression

When drawing scatter diagram

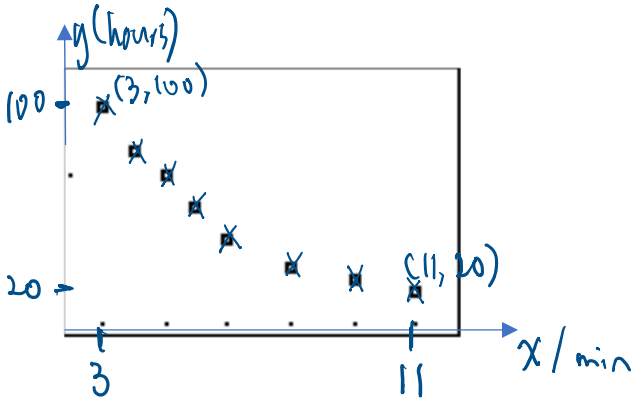
[Use crosses (x), and must indicate : (i) axes labes and units, (ii) the minimum and maximum value on each axis, (iii) label the end-points coordinates too.
The points (if for example x values are in equal increments must show equal spacing.

*The **relative positioning** of the points must be proportionate. Use the GC axes markings as a guide.*

Sometimes question requires to add onto the diagram a regression line, this line will help you to position your points more proportionately]

Answering questions on after drawing scatter diagram, comment on whether linear model is suitable (more for H2 Maths, ignore for H1 Maths) *(no r value is calculated yet at this point).*

[Use GC obtain scatter diagram, observe the diagram; mention data points has a curvi-linear trend and mention something like " y increases as x increases but by decreasing amounts"]



The scatter diagram shows that as *x* increases, the rate of decrease in *y* becomes smaller and *y* appears to approach a value.. Thus a linear model is not suitable as a linear model requires the rate of decrease to be constant.

Comment on value of correlation coefficient (in context)

[cite 3 things: close to 1, indicate strong positive linear correlation betweeen the 2 variables, name the variables in context and not in terms of symbols]

The value of r (0.968) **is close to 1**, indicating a **strong positive linear correlation between** the 2 variables x and y. *(cite in words the names of the variables eg height and weight) , e.g.,*

From GC, r = 0.964
The value of r is positive and close to 1, indicating a strong and positive linear correlation between the gross domestic product per capita and the plastic waste per capita per day.

Comment when estimate is reliable

[cite 3 key-points: within data range (cite it), interpolation, r close to 1]

The estimate is reliable since ***x* = 1.25** is within the data range **(0.5 to 2.0**), hence it is an interpolation. In addition, $|r|$ is close to 1 which indicates a strong positive (or negative) linear correlation between *x* and *y*.

Comment when estimate is NOT reliable

[cite 2 key-points: not within data range (cite it), extrapolation,

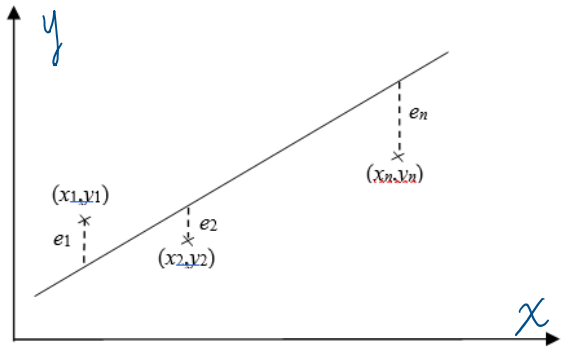
Or [x may be within data range but |r| not close to 1]

The estimate is NOT reliable since ***x* = 2.50** is not within the data range **(0.5 to 2.0**), hence it is an extrapolation.

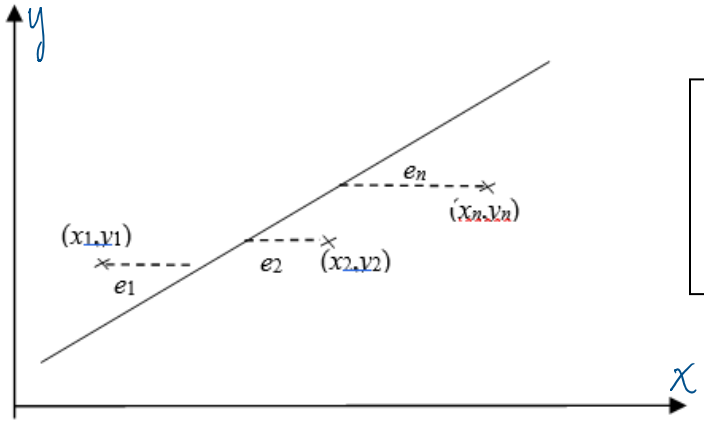
OR The estimate is not reliable as r= 0.5 is not close to 1, although x value is within the data range (ref BMQ3)

Since *x* is outside the given range of data, the linear relation may no longer hold, therefore the estimate is unreliable.

Least Square Regression Lines of *y* on *x* VS *x* on *y* (the meanings and explanations)



For regression line of *y* on *x*, the values of *x* are considered as accurate and the sum of squares of deviation in the *y*-direction is minimized.



For regression line of *x* on *y*, the values of *y* are considered as accurate and the sum of squares of deviation in the *x*-direction is minimized.

5. General common problems (could be in probability, binomial or normal distribution scenarios)

QUESTION: Explain why the answer to part (iii) is greater than the answer to part (ii). (see below)

> (ii) Find the probability of the event that both a randomly chosen chicken has a selling price exceeding $7 and a randomly chosen turkey has a selling price exceeding $55.
> (iii) Find the probability that the total selling price of a randomly chosen chicken and a randomly chosen turkey is more than $62.
> ***[Question from N2007 H2/P2/Q8]***

Solving gives (ii) answer is 0.160 and (iii) answer is 0.392

PHRASING SUGGESTION:

It is because the **event** in (ii) is a **subset** of the **event** in (iii).
For example, the selling price of a chicken and turkey could be $5 and $58 respectively in (iii) but this event is not possible in (ii), since (ii)'s event requires chicken selling price as exceeding $7.
[*Special Note: Recognise that (iii)'s event encompasses more cases ($10 chickens with $53 Turkey to make $63 > 62*) *which is not possible for (ii) as turkey needs exceeds $55 in (ii).]*

QUESTION: Explain "Random Sample" in context.
(Read this together with the Binomial Assumptions and see the difference in phrasing.)

PHRASING SUGGESTION:

A random sample means that every bottle of hand sanitiser in the population (produced by the factory) **has an equal chance of being selected (or chosen)** to form the sample.
In addition, the selection of one bottle of hand sanitiser is **independent** the selection of any other bottle of hand sanitiser.
(i.e., each bottle from the population has an equal chance of being selected, and the bottles are independently selected from one another to form the sample.).

Note that a **reason** to do **a random sample** (considered as a method to select a sample from a population to do some investigation, test etc) **is to avoid bias**.

QUESTION: Explain whether there is a need to make any assumption about the population distribution of X (eg, speed of PMD riders).

PHRASING SUGGESTION:

There is no need for any assumptions to be made about the population distribution of the speed of PMD riders (X) since n = (*quote the value*) is large,

by the Central Limit Theorem, the **sample mean speed** of PMD riders ($\overline{X}$) will follow a normal distribution approximately.

6. Finding s-squared (given the various forms of data presented)

MF26 formulae sheet quoted as below:

Sampling and testing

Unbiased estimate of population variance:

$$s^2 = \frac{n}{n-1}\left(\frac{\Sigma(x-\bar{x})^2}{n}\right) = \frac{1}{n-1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)$$

(Need able to use the MF 26 effectively.)

Q1 sample data presented as $n = 200$, $\sum(x-2) = -41.8$, $\sum(x-2)^2 = 140.38$

Q2 sample data presented as $n = 80$, $\sum x = 6120$, $\sum x^2 = 476955$,

Q3 sample data presented as "A random sample of 50 watermelons produced is weighed and its mean mass and variance are recorded as 1.78 kg and 0.12 kg^2 respectively". (n =50)

Q1 $s^2 = \frac{1}{n-1}\left(\sum(x-2)^2 - \frac{(\sum(x-2))^2}{n}\right)$

$= \frac{1}{199}\left(140.38 - \frac{(-41.8)^2}{200}\right)$

(application of 2nd formula in MF26; note the transformed data by -2, does not affect the spread)

Q2 $s^2 = \frac{1}{80-1}\left[476955 - \frac{6120^2}{80}\right] = \frac{8775}{79} = 111.0759494 \approx 111$

(direct application of 2nd formula in MF26)

Q3 $s^2 = \frac{n}{n-1}\sigma_n^2 = \frac{n}{n-1}$(variance of the sample of n data points)

$s^2 = \frac{n}{n-1}$(sample variance) *(indirect application of 1st formula in MF26)*

$= \frac{50}{49}(0.12) = \frac{6}{49}$

DOO#WKH#EHVW$$$#

Compiled by LKS (20 Oct 2023)