Statistics Review

(*) Population:

A population is the whole set of items that are of interest.

(*) Sample:

A sample is a selection of observation taken from a subset of population which is used to find out information about the population as a whole.

The size of the sample can affect the validity of any conclusion drawn.

(*) Sampling:

In random sampling, every member of the population has an equal chance of being selecting. It also removes bias from a sample

⇒ A simple random sample of size n is one where every sample of size n has an equal chance of being selected.

Types of data:

Quantitative Data:

Variables or data associated with numerical observation are called quantative data.

Qualitative Data:

Variables or data associated with non-numerial observation are called qualitative data. Example: Haire colour.

Continuous Variables:

A variable that can take any value in a given range is a continuous variable.

Discrete Variable:

A variable that can take only specific values in a given range is a discrete variable.

Measures of central tendency:

A measure of location is a single value which describes a position in a data set. If the single value describes the center of data, then it is called measure of central tendency

⇒ **Mode:** The mode or modal class is the value or class that occurs most often. (Data is qualitative)

⇒ **Median:** The median is the middle value when the data values are put in order. (Quantitative)

⇒ **Mean:** $\bar{x} = \frac{\sum x}{n}$ (quantitative)

Average

Variance: It is used to workout the spread of dataset. This makes use of the fact that each data point deviates from the mean by the amount $x - \bar{x}$.

$$\text{Var}(x) = \frac{\Sigma (x-\bar{x})^2}{n}$$

Standard Deviation:

It is the square root of variance.

$$\sigma = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n}}$$

Outliers: An outliers is an extreme value that lies outside the overall pattern of the data.

Probability Distribution:

Random variable:

A random variable is a variable whose value depends on the outcome of a random event.

- A variable can take any of a range of specific values.
- ~~A variable can take any~~
- The variable is discrete if it can only take certain numerical values.

- The range of values that a random variable can take is called sample space.

⊛⊛ A probability distribution fully describes the probability of any outcome in the sample space.

We describe probability mass fun

$P(X = x)$ — X takes a particular value x. — Random variable

(*) When all probabilities are the same, then the distribution known as a discrete uniform distribution.

$$\sum P(X = x) = 1$$

(*) Binomial Distribution:

When you are carrying out a number of trials in an experiment or survey, you can define a random variable X to represent the number of successful trials.

The sum of probabilities in a discrete probability distribution is 1.

We can model X with a binomial distribution $B(n, p)$ if:

→ there are a fixed number of trials, n

→ there are two possible outcome

→ There is a fixed probability of success, p

→ the trials are independent of each other.

→ If a random variable X has the binomial distribution $B(n, p)$ then its probability mass function is given by:

$$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r}$$

$\binom{n}{r} \rightarrow {}^{n}C_r = \frac{n!}{r!(n-r)!}$

we write → $X \sim B(n, p)$

Continuous Random Variables

It can take any one of infinitely many values. The probability that a continuous random variable takes one specific value is 0, but you can write the probability that it takes values within a given range.

$P(X = 4)$ ← X is discrete

$P(Y < 20)$ ← Y is continuous

A continuous random variable has a continuous probability distribution. Usually, it is represented using a curve on a graph.

The area under a continuous probability distribution is equal to 1.

Normal Distribution:

The continuous variables generally encountered in real life are more likely to take values grouped around a central value than to take extreme values.

The normal distribution is a continuous probability distribution that can be used to model many naturally occurring characteristics that behave this way.

For example:

- Heights of people within a given population.
- Weights of a tiger in a jungle.
- errors in scientific measurement
- Size variation in a manufactured objects.

A normal distribution has the following parameters:

- μ, the population mean
- σ^2, the population variance / Std. Deviation
- is symetrical (mean = median = mode)
- has total area under the curve equal to 1.
- has points of inflection at $\mu+\sigma$ and $\mu-\sigma$

Notation → $X \sim N(\mu, \sigma^2)$

$$X \sim N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$Z = \frac{X - \mu}{\sigma}$$

Hypothesis Testing:

A hypothesis is a statement made about the value of population parameter. P.

In order to carry out the test, we need two hypothesis.

$H_0 \longrightarrow$ The null hypothesis which we assume to be correct

$H_1 \longrightarrow$ The alternative hypothesis tells us about the parameter if one assumption is shown to be wrong.

* Hypothesis test with alternative hypotheses in the form $H_1 : P <$ ... and $H_1 : P >$ ... are called one tailed test.

(8). Hypothesis tests with an alternative hypothesis in the form $H_1: p \neq$ are called two tailed tests.

Critical values / Critical Region:

A critical region is a region of the probability distribution which, if the test statistics falls within it, would cause you to reject the null hypothesis.

We assume that test statistics can be modelled by Binomial distribution

↑ For Simplicity

(9)

How to find the accepted hypothesis:
H0, Null Hypothesis will be accepted if the data lie in Accepted region.
H1, Alternative Hypothesis will accepted if the data lie in Critical Region.

For example:

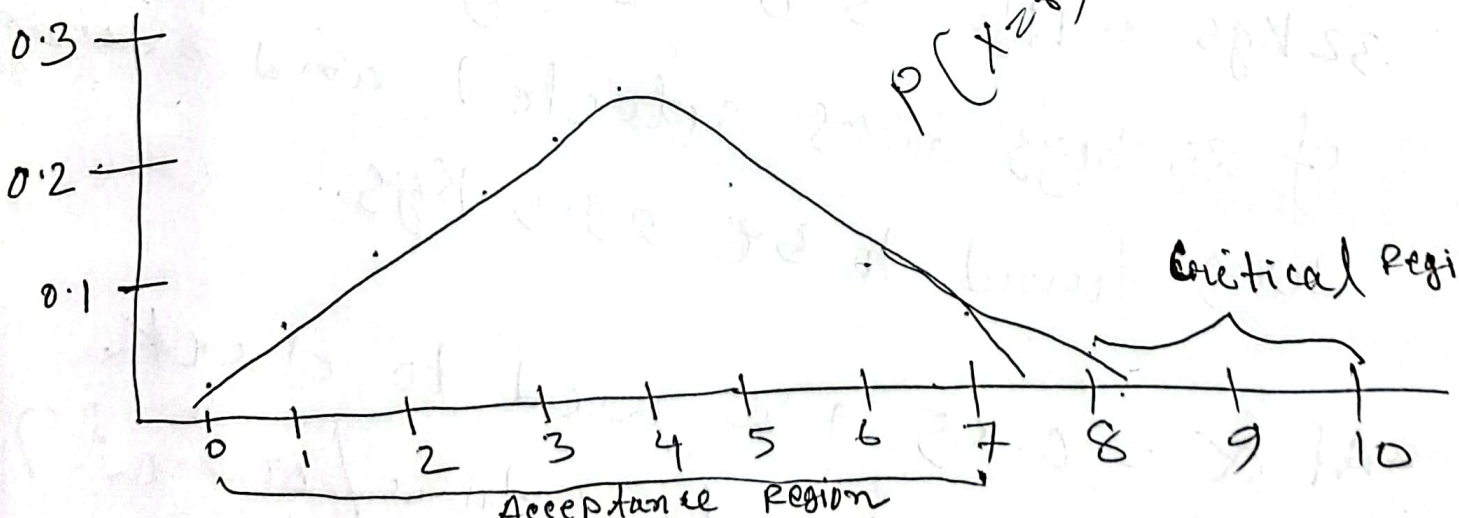
A test statistic is modelled as $B(10, p)$ and hypothesis test at the 5% significance level uses

$H_0: p = 0.4$, $H_1: p > 0.4$.

Assuming H_0 to be true, X has the following distribution:

$X \sim B(10, 0.4)$

$P(X=r) = \binom{n}{r} p^r (1-p)^{n-r}$



$P(X=10) = 0.0001$, $P(X=9) = 0.0016$ and

$P(X=8) = 0.0106$. $P(X=7) = 0.0425$

$P(X=6) = 0.1114$

Z test:

$$Z = \frac{\bar{x} - m}{\sigma / \sqrt{n}}$$

$\bar{x}$ → Sample mean
m → population mean
σ → Standard dev
n → number of Samples.

(2)

Example:

A complain was registered stating that the boys in the municipal schools were ~~carfed~~ underfed.

Average weight of the boys 10, is 32 Kgs with S.D = 9 Kg. A sample of 25 boys was selected and average was found to be 29.5 Kgs.

At $\alpha = 0.05$ we need to check whether the complaint true / or not?

Using z test:

$$z = \frac{29.5 - 32}{9/\sqrt{25}}$$

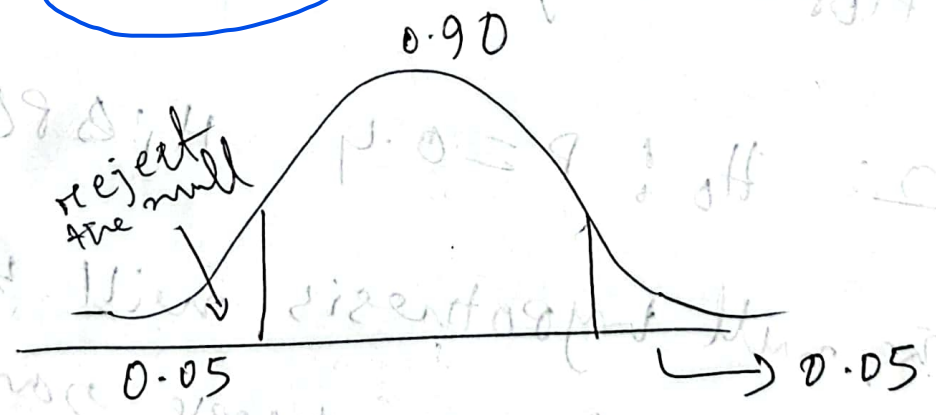
$$z = -1.39$$

H_0: $\mu = 32$

H_1: $\mu < 32$ ← the complain is true

Now, use z value to find out the p value.

Look up the z table → p value is $p = 0.0823$ > 0.05



3

Example

An election candidate believes she has the support of 40% of the residents in a particular town. A Researcher wants to test, at the 5% significance level, whether the candidates is over-estimating her support. The researcher asks 20 people whether they support the candidate or not. 3 people say that they do.

→ over-estimated her support

Ans: $H_0 : P = 0.4$ $H_1 : P < 0.4$

The null hypothesis will be rejected if 3 or fewer people say they support the candidate is less than 5% given that $P = 0.4$.

$P(X \leq 3) = 0.016$ (From the Binomial distribution table)

If less than significance level is rejected, if higher, it will be accepted

∴ H_0 is rejected. Since $0.016 < 0.05$

Example 02: (9)

The standard treatment for a particular disease has $\frac{2}{5}$ (0.4) probability of success. A certain doctor has undertaken research in this area and has produced a new drug which has been successful with 11 out of 20 patients. The doctors claims that the new drug represents an improvement on the standard treatment.

Test, at 5% significance level, the claim made by the doctor.

⇒ $H_0 : P = 0.4$ (It means that there is no improvement)

$H_1 : P > 0.4$ (It means that there is improvement)

(b) work out the critical region:

$$P(X \geq 13) = 1 - P(X \leq 12)$$
$$= 1 - 0.979$$
$$= 0.02 < 0.05$$

Look 95% cumulative success rate of probability distribution

Therefore, it is less than the significance level. we need to reject the hypothesis if the value is 13 or more.

Since, 11 doesnot lie in the critical region we accept H_0 hypothesis.

We can conclude that there is no significant ~~div~~ evidence of improvement.

* Here 1 to 12 is acceptance region. And 13 to 20 is critical region. Since 11 is in acceptance region that means null hypothesis(h0) is accepted.
* That means doctor claim is not true. Because it is not improved.

Example 03 (5)

A polling organisation claims that the support for a particular candidate is 35%. It is revealed that the candidate will pledge to support local charities if needed. The pooling organisation think that the level of support will go up as a result. It taken a new poll of 50 votes.

a. Describe the test statistics and state suitable null and alternative hypothesis.

b. Using 5% level of significance, find the critical region for a test to check the belief.

c. In new poll, 28 people are found to support the candidate. Comment on this observation in the light of critical region.

a. H_0 : $P = 0.35$ (Level of support did not increase)

H_1 : $P > 0.35$ (Level of support has increased)

b. 5% Significance level; we can get $P(X \leq 23) = 0.96$. Then

$$P(X \geq 24) = 1 - 0.96$$
$$= 0.04 < 0.05$$

Our critical region starts from 24. It means that for the value from 24 we need to reject the null hypothesis.

From 1 to 23 is accepted region. From 24 to 50 is critical region and the 28 lie into the critical region that's why alternative is acceptable and null is rejected.

c. 28 people supported the candidate

In that case, 28 is lies in the critical region. Thus, we reject the null hypothesis. It indicates that level of support has risen in P.

Two-tailed Test:

A one tailed test is used to test when it is claimed that the probabili has either gone up or gone down. A two tailed test is used when it is thought that the probability has changed in either direction.

6 * For a two-tailed test, halve the significance level at the end you are testing.

$X \sim B(n, p)$, then the expected outcome is np. If the observed value, x is lower than this then consider $P(X \leq x)$. If the observed value is higher than the expected value then consider $P(X \geq x)$.

Ans:

The proportion of people eating vegetarian meals at Enrico's is $\frac{1}{3}$.

Hypothesis:

$H_0 : p \geq \frac{1}{3}$ → Habit not change, $H_1 : p \neq \frac{1}{3}$ → Habit changed

H_0 is true $X \sim B(10, \frac{1}{3})$

$$P(X \leq 1) = P(X=0) + P(X=1)$$

$$= \left(\frac{2}{3}\right)^{10} + 10\left(\frac{2}{3}\right)^{9}\left(\frac{1}{3}\right)$$

$$= 0.01734 + 0.08670$$

$$= 0.104 > 0.025$$

Hypothesis accepted.

The expected value would be $10 \times \frac{1}{3} = 3.33$

The observed value is 1, less than the expected value. So consider $P(X \leq 1)$