Course: Mathematical statistics

Week 1: Parameter, Statistic, Sampling Distribution

Lecturer: Nagulama Moses

Kumi University

March 16, 2025

Outline

Course description

- Mathematical Statistics is a branch of mathematics that focuses on collecting, organizing, analyzing, and interpreting data related to random phenomena.

- This course explores key concepts including probability distributions, sampling theory, estimation techniques, hypothesis testing, regression analysis, and statistical inference.

- It equips students with the analytical skills needed to solve real-world problems in various fields through precise and logical reasoning grounded in mathematics.

Course Goals and objective

- To introduce students to a deeper understanding of statistics.
- Enable students understand the main aspects of describing a data set and several statistical concepts.
- The objectives include; apply the sampling techniques, tests of hypothesis to real life situation, manipulate statistical data with ease in scientific research.

Intended learning outcomes

- Distinguish between parameters and statistics.
- Identify common population parameters and statistic.
- Illustrate the Central Limit Theorem (CLT) and its implications for sample mean.

What is a parameter?

- A parameter is a numerical value that describes a characteristic of a population.
- In statistics, populations refer to the entire group under study, such as all people in a country or all measurements of a certain trait.
- A parameter is generally unknown and needs to be estimated from the sample data.

Examples of Parameters

- **Population Mean (μ):** The average of all values in the population. It is calculated as:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n},$$

where n is the population size, and x_i are the individual data points in the population.

- **Population Variance (σ^2):** A measure of the spread of data points in the population. It is given by:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2,$$

where x_i is each data point in the population.

- **Population Proportion (p):** The proportion of successes in the population:

$$p = \frac{\text{Number of successes}}{n}.$$

- **Population Standard Deviation (σ):** The square root of the population variance:

$$\sigma = \sqrt{\sigma^2}.$$

Statistic

- A statistic is a numerical value that describes a characteristic of a sample. It is used to estimate the corresponding parameter of the population.

- Unlike a parameter, a statistic can be directly calculated from sample data and varies from one sample to another.

Examples of Statistic

- **Sample Mean ($\bar{x}$):** The average of the sample values. It is an estimator for the population mean (μ):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

where n is the sample size, and x_i are the individual data points in the sample.

- **Sample Variance (s^2):** A measure of the spread of sample data, calculated as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

where $\bar{x}$ is the sample mean.

- **Sample Proportion ($\hat{p}$):** The proportion of successes in the sample:

$$\hat{p} = \frac{\text{Number of successes}}{n}.$$

- **Sample Standard Deviation (s):** The square root of the sample variance:

$$s = \sqrt{s^2}.$$

Note:

- when a sample is taken then it will be constant for that sample unless when one is taken.

- Sample statistic vary from one sample to another. consequently sample statistic is a variable.

sampling method

a process of obtaining a given number of subjects from the defined population is called sampling

Reasons to sample

- Physical impossibility of checking all subjects in the population
- Cost of studying all items in the population may be quite prohibitive
- Time
- Destructive nature of the test

Methods used

(a) Simple random sampling: is a method of selecting a sample such that each element in the population has an equal chance of being included in the sample.

Simple random sampling can be done in two ways i.e.

- Lottery method which give each element each subject in the population an identification number. Need to have N pieces of paper, label them, mix them thoroughly and pick he sample until you obtain the required sample size (sampling without replacement)

- Random numbers, numbers generated by use of computers e.g. 63271, 59986, 71714, 51102, 15141, 80714. The random number contains one or more of 0,1,2,3,4,5,6,7,8,9

(b) Systematic sampling: every object in the sample is 3, 4 or 5 or include in the sample e.g. identification number 1,2,...,N, sample of size n, $\frac{N}{n} = A$.

(c) Stratified sampling: is a method obtained by dividing the population into subgroups called strata. according to various homogeneous characteristics we then randomly select members from each stratum.

(d) Cluster sampling: is a sample obtained by selecting a pre-existing group called a cluster and using all members in the cluster. sample the cluster at random and use all the sampling in the cluster.

Other sampling Methods include

- Purposive sampling: Where the researcher deliberately selects specific individuals or data points based on predefined criteria.
- Sequence sampling: Used when data is collected in a sequential manner, meaning each sample influences the decision to continue or stop sampling.
- Double sampling: Where an initial sample is drawn and analyzed, followed by a second sample to refine estimates or make better decisions.
- multi-stage sampling: Where the selection process occurs in multiple stages, instead of selecting individuals directly from the entire population.

Key Differences Between Parameter and Statistic

population parameter	sample statistic
population mean μ	$\overline{x}$ sample mean
population variance σ^2	s^2 sample variance
population proportion P	$\hat{\theta} = \frac{x}{n}$ sample proportion
$\mu_1 - \mu_2$	$\hat{\theta} = \overline{x_1} - \overline{x_2}$
$P_1 - P_2$	$\hat{\theta} = \frac{x_1}{n_1} - \frac{x_2}{n_2} = \hat{P}_1 - \hat{P}_2$

Sampling Distribution

- Sampling distribution is the probability distribution of a sample statistic.
- We need to observe that if a samples are randomly selected with replacement the sample means will be different from the population mean μ.
- Sampling error is the difference between the sample statistic and the corresponding population parameter i.e. $(\bar{x} - \mu)$

Example

A professor of maths is supervising a total of 4 PhD students the number of hours per week the 4 students spend in developing their research proposal are $2, 6, 4, 8$.

(a) Calculate the mean number of hours a student spend per week and the standard deviation for the number of hours a student spend in developing a proposal.

(b) if random samples of size 2 are taken from the population with replacement

(i) compute the mean number of hours for each sample

(ii) Determine the probability distribution for the sample mean $\overline{x}$

(iii) calculate the mean of the sample means

solution

if μ is population mean and σ^2 population variance

$$\mu = \frac{2+6+4+8}{4} = 5$$

$$\sigma = \sqrt{\frac{(2-5)^2+(6-5)^2+(4-5)^2+(8-5)^2}{4}}$$

$$\sigma = 2.226$$

sample number	sample	sample mean($\overline{x}$)
1	2,2	2
2	2,6	4
3	2,8	5
4	6,2	4
5	6,6	6
6	6,4	5
7	6,8	7
8	4,2	3
9	4,6	5
10	4,4	4
11	4,8	6
12	8,2	5
13	8,6	7
14	8,4	6
15	8,8	8
16	2,4	3

probability distribution

$\bar{x} = x$	2	3	4	5	6	7	8
$P(\bar{x} = x)$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Denote the mean of the sample means by

$$\mu_{\bar{x}} = E(\bar{x}) = 2.\frac{1}{16} + 3.\frac{2}{16} + 4.\frac{3}{16} + 5.\frac{4}{16} + 6.\frac{3}{16} + 7.\frac{2}{16} + 8.\frac{1}{16}$$

$$= \frac{80}{16} = 5$$

Theorem

- If all possible samples of size n are taken with replacement from the same population with mean μ, standard deviation σ then the sampling distribution of the sample mean is approximately normally distributed with mean of the sample means $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ i.e. $\approx N(\mu_{\bar{x}}, \sigma^2_{\bar{x}}) = N(\mu, \sigma^2)$

Proof

- Denote i^{th} observation in a sample x_i.
- Different samples will give different observations of x_i and therefore x_i is an observation on a random variable x_i.
- Suppose the random variable x_1 has μ and σ^2 the same will be true for all other random variables $x_2, x_3, ..., x_n$ resulting from similar observations each with mean μ and variance σ^2 (Hogg,2012).

$$\mu_{\bar{x}} = E(\bar{x}) = \frac{E(x_1 + x_2 + x_3 + ... + x_n)}{n}$$

$$= \frac{1}{n} E(x_1 + x_2 + x_3 + ... + x_n)$$

$$= \frac{1}{n} E(x_1) + E(x_2) + ... + E(x_n) = \frac{n\mu}{n} = \mu$$

$$\sigma_{\bar{x}}^2 = var(\frac{x_1 + x_2 + ... + x_n}{n})$$

$$= \frac{1}{n^2} Var(x_1 + x_2 + ... + x_n)$$

$$= \frac{1}{n^2} var(x_1) + var(x_2) + ... + var(x_n)$$

$$= \frac{1}{n^2}\sigma^2 + \sigma^2 + ... + \sigma^2$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Standard deviation

$$\sigma = \frac{\sigma}{\sqrt{n}}$$

Standard error. the standard deviation of the sampling distribution of a statistic is called a standard error.

The central limit theorem (CLT)

If a random samples of size $n \geq 30$ are taken from a population i.e finite or infinite with a mean μ and variance σ^2, then the sampling distribution of the sample mean $\bar{x}$ is approximately normally distributed with

$$\mu_{\bar{x}} = \mu, \sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

hence the quantity

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is the value of the standard normal random variable $N(0, 1)$

Note

1. when the original variable (population) is normally distributed then the distribution of the sample means will be normally distributed regardless of the sample size.
2. when the distribution of the population might not be normal provided that sample size $n > 30$ then the distribution of the sample mean will be approximately normal.
3. if the sample size is large, the central limit theorem can be used to answer qtns about the sample means in the same way the normal distribution can be used to answer questions about the individual variables.

Example

The mean score in a mathematics test in a class is 75 with a standard deviation of 8. a student is chosen at random from the class. what is the probability that he scored a mark greater than 65%. (assume the scores are normally distributed)

$$\mu = 75, \sigma = 8$$

let X be the students score

$$x = N(75, 64), P(x > 65)$$

$$z = \frac{x - \mu}{\sigma} = \frac{65 - 75}{8} = -1.25$$

$$P(x > 65) = P(z > -1.25) = 0.5 + \phi(1.25)$$

$$= 0.5 + 0.3944 = 0.8944$$

Example

Uganda broadcasting corporation reported that children below 5 years watch a TV of an average of 25hrs a week. Assume that a number of hours children watch TV is normally distributed with $\sigma = 3$ hours. a random sample of 20 children below 5 years is selected. find the probability that the mean number of hours they watch TV is greater than 26.3 hours.

solution

$$\mu = 25, n = 20, \sigma = 3, N(25, 9)$$

$\bar{x}$ be the mean number of hours

$$P(\bar{x} > 26.3)$$

sampling distribution of sample mean

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{26.3 - 25}{\frac{3}{\sqrt{20}}} = 1.938$$

$$P(z > 1.938) = 1 - P(z < 1.938) = 0.0268$$

Example

Ball bearings provided by certain factory is found to have a mean mass of 5.02g with a standard deviation of 0.3g. find the probability that a random sample of 100 ball bearing chosen from this rot will have a mean mass

(i) between 4.96 to 5.00g

(ii) more than 5.0g

(iii) At most 5.20g

solution

$$\mu = 5.02, \sigma = 0.3, n = 100$$

let $\bar{x}$ be the mean mass of the random sample

$$P(4.96 < \bar{x} < 5.00), n = 100(n > 30)$$

by CLT the distribution of $\bar{x}$ is $\approx N(5.02, \frac{0.09}{100})$ consequently $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

$$P(4.96 < \bar{x} < 5.00)$$

$$z_1 = \frac{4.96 - 5.02}{\frac{0.3}{\sqrt{100}}} = -1.33$$

$$z_2 = \frac{5.00 - 5.02}{\frac{0.3}{\sqrt{100}}} = -0.66$$

$$P(4.96 < \bar{x} < 5.00) = 0.4082 - 0.2454 = 0.2286$$

$$P(\bar{x} > 5.10), z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{5.10 - 5.02}{\frac{0.3}{\sqrt{100}}}$$

$$P(\bar{x} > 5.10) = 0.0032$$

$$P(\bar{x} \leq 5.20), z = \frac{5.20 - 5.02}{\frac{0.3}{\sqrt{100}}}, z = 6$$

$$P(\bar{x} \leq 5.20) = P(z \leq 6) = 1$$

References

- Hogg,R;Mckean,J;Craig,A(2012).Introduction to mathematical statistics, 7th edition, pearson Prentice Hall, 2012.

- Hastings K.J,(1997) Probability and statistics, Addison Wesley reading,massachusetts.

Thank You!

Next Lecture: Sampling distribution