Sampling and Estimation

Sampling

Expectation/Learning Outcome/Behavioral Objective:

-To state the distinction between a sample and a population.
-To appreciate the necessity for randomness in choosing a sample.
-To recognize that a sample mean can be regarded as a random variable and the use of Central limit theorem.

Sampling and Estimation

Estimation

Learning Outcome

Expectation/Learning Outcome/Behavioral Objective:

-To define ‘unbiased estimates’

-To calculate the unbiased estimates of the population mean and variance from a sample, using either raw or summarized data.

-To describe the meaning of confidence interval for a population mean. (The calculation of confidence interval will be taught in the respective hypothesis tests)

Sampling and Estimation

Sampling

Population

- In a statistical enquiry you often need information about a particular group. This group is known as the **population**, and it could be finite or infinite.

Sample

- A subset of a population.

Survey

- Information is collected by means of a survey. There are two types:

  (a) Census

  In a census **every** members of the population is surveyed.

  (b) Sample survey

  when a survey covers less than 100% of the population, it is known as a **sample survey**.

Bias

- The purpose of sampling is to gain information about the whole population by selecting a sample from that population. You want the sample to be representative of the population so you must give every member of the population an equal chance of being included in the sample. This should eliminate any **bias** in the selection of the sample.

- When a sampling method does over-represent or under-represent a feature of the population it is said to be biased.

- The most common approach to the task of avoiding bias is to select a **random sample**.

Sampling Methods

- Random sampling (simple, systematic, stratified)
- Non-random sampling (quota, cluster)

Random Sampling

1 Simple Random Sampling

- All possible samples of size *n* are equally likely to selected.
- Two methods of simple random sampling

  (i) drawing lots

  (ii) random number sampling

  - using random number tablets
  - calculator random number generator

Calculator random number generator

- Suppose you want to use your calculator to select a random sample of six numbers between 1 and 49.

- Press

Ran #

Shift . =

- Suppose the number you get are

0.730, 0.798, 0.369, 0.499, 0.491, 0.310,

0.135, 0.112, 0.593, 0.652, 0.015, 0.346

- 0.730, 0.798, 0.369, 0.499, 0.491, 0.310, 0.135, 0.112, 0.593, 0.652, 0.015, 0.346
- You can interpret them in various ways, for example:

  Use the first two digits to the right of the decimal point each time.

  73, 79, 36, 49, 49, 31, 13, 11, 59, 65, 01, 34

  (Ignoring repeats and numbers > 49)

  → 36, 49, 31, 13, 11, 1 OR

  Use the second and third digits to the right of the decimal point.

  (Ignoring repeats and numbers > 49)

  → 30, 10, 35, 12, 15, 46

OR

- 0.730, 0.798, 0.369, 0.499, 0.491, 0.310, 0.135, 0.112, 0.593, 0.652, 0.015, 0.346

- You can interpret them in various ways, for example:

Use all digits after the decimal point.

73, 07, 98, 36, 94, 99, 49, 13, 10,

13, 51, 12, 59, 36, 52, 01, 53, 46,

→ 7, 36, 49, 13, 10, 12

2 Systematic Sampling (for large scale sampling)

List the population in some order, and then choose every *k th* member from the list after obtaining a random starting point.

Describe how to choose a systematic sample of eight members from a list of 300.

✓ Find a suitable value of k.

$$k = \frac{N}{n} = \frac{300}{8} = 37.5 \approx 40$$

✓ Choose a random starting point. If $\boxed{Ran\ \#}$ gives 0.870, take the first member of the sample as 87 and then add 40 each time.

2 Systematic Sampling (for large scale sampling)

⇒ Sample : 27 , 67, 87, 127, 167, 207, 247, 287

Advantages : It is quick to carry out and easy to check for error.

Disadvantages : there may be a periodic cycle within the frame itself.

3 Stratified Sampling

- The population is split into distinguishable layers or strata (eg age , occupation)
- separate random samples are then taken from each stratum and put together to form the sample.

Competent carriers employs 320 drivers, 80 administrative staff and 40 mechanics. A committee to represent all the employees is to be formed. The committee is to have 11 members and the selection is t be made so that there is as close a representation as possible without bias towards any individuals or groups. Explain how this could be done.

3 Stratified Sampling

- The population is split into distinguishable layers or strata (eg age , occupation)
- separate random samples are then taken from each stratum and put together to form the sample.

Competent carriers employs 320 drivers, 80 administrative staff and 40 mechanics. A committee to represent all the employees is to be formed. The committee is to have 11 members and the selection is t be made so that there is as close a representation as possible without bias towards any individuals or groups. Explain how this could be done.

Non-Random Sampling

1 Cluster Sampling

A sample obtained by sampling some f, but not all of, the possible subdivisions within a population. These subdivisions, called clusters, often occur naturally within the population.

Advantage : no need to have a complete sampling frame f the whole population.

Disadvantage : non-random.

A town has 7500 primary school children in 250 classes, each with an average class size of 30. If you want to select a sample of 90 children.

- Use the classes as clusters and take a sample consisting of three classes (90 children)

2 Quota Sampling

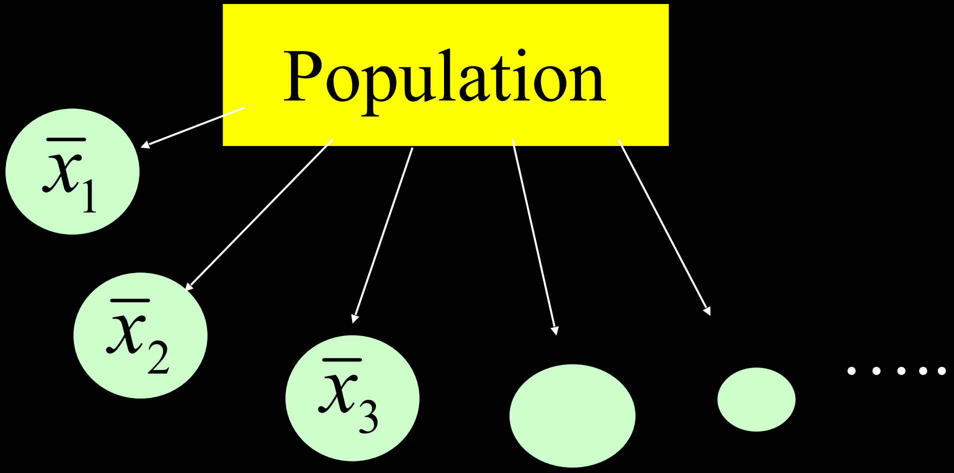
The population is divided into groups in term of age, sex, income level and so on. Then the interviewer is told how many people to interview within each specified group, but is given no specific instructions about how to locate them and fulfill the quota.

Disadvantage : non-random, there is possibility of bias in the selection process.

The distribution of the sample mean

- Take a random sample of *n* observations from a population.
- From a finite population, sampling should be with replacement to ensure that the observation are independent.
- Repeat the procedure until you have taken all possible samples of sizes *n*, calculating the sample mean of each one.
- Form a distribution of all the sample means.

Sampling distribution of means



$$\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$$

The mean and variance of the sampling distribution of means

Consider a population X in which

$$E(X) = \mu \quad and\, Var(X) = \sigma^2$$

Take *n* independent observations $x_1\,, x_2\,, x_3\,, \ldots, x_n$ from *X*.

The sample mean, $\bar{X}$

$$E(\bar{X}) = \mu \quad and \quad var(\bar{X}) = \frac{\sigma^2}{n}$$

$\frac{\sigma}{\sqrt{n}}$ = standard error of the mean

(a) The distribution of $\bar{X}$ when the population of X is normal (any ample size)

$If\ \ X \sim N(\mu, \sigma^2)$ then

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Example 1:

A shipment of steel bars will be accepted if the mean breaking strength of a random sample of ten steel bars is greater than 250 pounds per square inch. In the past, the breaking strength of such bars has had a mean of 235 and a variance of 400.

(i) What is the probability, assuming the breaking strengths are normally distributed, that one randomly selected steel bar will have a breaking strength in the range of 245 to255? 0.1498

(ii) What is the probability that the shipment will be accepted? 0.0089

Example 2:

A liquid drug is marketed in phials containing a nominal 1.5 ml but the amounts can vary slightly. The volume in each phials may be modelled by a normal distribution with mean 1.55 ml and standard deviation σ ml. The phials are sold in packs of 5 randomly chosen phials. It is required that in less than 0.5% of packs will the total volume of the drug be less than 7.5 ml. Find the greatest possible value of σ.

The Central Limit Theorem

(b) The distribution of $\bar{X}$ when X is not normally distributed

When samples are taken from a population that is **not normally distributed**, the sampling distribution Takes on the characteristics normal shape as the sample size increases.

→For large *n* the distribution of the sample mean is **approximately normal**.

This result is known as Central Limit Theorem.

For sample taken from a non-normal population with mean μ and variance σ^2, by the Central Limit theorem, $\bar{X}$ is approximately normal and

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

provided that the sample size, *n* is large ($n \geq 30$).

Example 3:

A manufacturer of light bulbs says that its light bulbs have mean lifetime of 700 hours and a standard deviation of 120 hours. You purchased 144 of these bulbs with the idea that you would purchase more if the mean lifetime of your sample was longer than 680 hours. What is the probability that you will not buy again from this manufacturer? 0.0228

Estimation

Unbiased Estimates of Population Parameters

Suppose that you do not know the value of a particular parameter of a distribution, for example the mean or variance.

A random sample from the distribution is taken and use it in some way to make an **estimate** of the value of your unknown parameter.

This estimate is **unbiased** if the average (or expectation) of a large number of values taken in the same way is the true of the parameter.

Point Estimates

- The best unbiased estimate of μ, the population mean, is μ? where

$$\mu ? = \bar{x} = \frac{\sum x}{n} \qquad , \bar{x} = \text{mean of the sample}$$

Point Estimates

- The best unbiased estimate of σ^2, the population variance, is $\sigma\boxed{?}$ where

$$\sigma\boxed{?} = \frac{n}{n-1}s^2 \qquad , s^2 = \text{variance of the sample}$$

$$\sigma\boxed{?} = \frac{n}{n-1}\left[\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2\right]$$

$$\sigma\boxed{?} = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]$$

$$\sigma\boxed{?} = \frac{1}{n-1}\sum\left(x-\bar{x}\right)^2$$

Note: Calculator in SD mode $\sigma\boxed{?} = x_{\sigma_{n-1}}$

Point Estimates

If the sample values are given as **grouped data**,

- an unbiased estimate of μ, the population mean is

$$\mu\boxed{?} = \bar{x} = \frac{\sum fx}{\sum f}$$

- an unbiased estimate of σ², the population variance is

$$\sigma\boxed{?}^{2} = \frac{1}{\left(\sum f\right)-1}\left[\sum x^{2} f - \frac{\left(\sum xf\right)^{2}}{\sum f}\right]$$

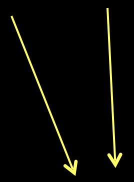
Example 4:

Thirty oranges are chosen at random from a large box of oranges. Their masses, x grams, are summarized by $\sum x = 3033$ and $\sum x^2 = 306676$. Find, to 4 significant figures, unbiased estimates for the mean and variance of the mass of an orange in the box. The oranges are packed in bags of 10 in a shop and the shopkeeper told customers that most bags weight more than a kilogram. Show that the shopkeeper's statement is correct indicating any necessary assumption made in your calculation.

Interval Estimates

Another way of using a sample value to estimate an unknown population parameter is to construct an interval, known as a confidence interval. This is an interval that has probability of including the parameter.

The interval = (a , b)



Confidence limits

Interval Estimates

Calculating a confidence interval

(a) Confidence interval for μ, the population mean

- Of a normal distribution
- With known variance σ^2
- Using any size sample, *n* large or small

- Of a non-normal distribution
- With known variance σ^2
- Using large sample, $n \geq 30$.

Interval Estimates

(a) Confidence interval for μ, the population mean

In general, a 100(1− α)% confidence interval for the population mean for a sample of size *n* taken from a normal population with variance σ^2 is given by

$$\left(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right)$$

Where $\bar{x}$ is the sample mean and the value of z is such that

$$\Phi(z) = 1 - \frac{1}{2}\alpha$$

Example 5:

The volume of milk in litre cartons filled by a machine has a normal distribution with mean μ litres and standard deviation 0.05 litres. A random sample of 25 cartons was selected and the contents, x litres, measured. The results are summarized by $\sum x = 25.11$. Calculate

(a) A symmetric 98% confidence interval for μ,

(b) The width of a symmetric 90% confidence interval for μ based on the volume of milk in a random sample of 50 cartons.

Example 6:

The heights of men in a particular district are distributed with mean μ cm and the standard deviation σ cm.

On the basis of the results obtained from a random sample of 100 men from the district, the 95% confidence interval for μ was calculated and found to be

(177.22 cm, 179.18 cm).

Calculate

(a) The value of the sample mean,

(b) The value of σ,

(c) A symmetric 90% confidence interval for μ.

Interval Estimates

Calculating a confidence interval

(b) Confidence interval for μ, the population mean

- Of a normal/non-normal distribution
- With unknown variance σ^2
- Using a large sample, *n*.

Interval Estimates

(b) Confidence interval for μ, the population mean

Given a large sample ($n \geq 30$) from any population, a 100(1− α)% confidence interval for the population mean is given by

$\bar{x}$ = sample mean

$$\left(\bar{x} - z\frac{\sigma\boxed{?}}{\sqrt{n}}, \bar{x} + z\frac{\sigma\boxed{?}}{\sqrt{n}}\right)$$

The value of z is such that

$$\Phi(z) = 1 - \frac{1}{2}\alpha$$

where $\sigma\boxed{?}^{2} = \frac{n}{n-1}s^{2}$ *or* $\sigma\boxed{?}^{2} = \frac{1}{n-1}\left[\sum x^{2} - \frac{\left(\sum x\right)^{2}}{n}\right]$

Example 7:

A random sample of 40 precision resistors is checked as a part of a control process. The mean resistance of the resistors should be 200 ohms. The results of the measurements on this sample are summarized by

$$\sum (x - 200) = -0.488 \,, \sum (x - 200)^2 = 1.4776$$

where x ohms is the resistance of a resistor.

(i) Calculate a 95% confidence interval for the population mean resistance of the resistors. (199.93, 200.05)

(ii) Does the confidence interval support the hypothesis that the population mean resistance is 200 ohms? yes

Interval Estimates

Calculating a confidence interval

(c) Confidence interval for μ, the population mean

- Population is normal distribution
- With unknown variance σ^2
- Sample size n is small. ($n < 30$)

Interval Estimates

(c) Confidence interval for μ, the population mean

100(1− α)% confidence interval for the population mean is given by

$$\left(\bar{x} - t\frac{\sigma}{\sqrt{n}}, \bar{x} + t\frac{\sigma}{\sqrt{n}}\right)$$

$\bar{x}$ = sample mean

t is the value from a t (n - 1) distribution

where $\sigma^2 = \dfrac{n}{n-1}s^2$ *or* $\sigma^2 = \dfrac{1}{n-1}\left[\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}\right]$

Example 8:

The time, *t* minutes, taken by 18 children in an infant reception class to complete a jig-saw puzzle were measured. This results are summarized by

$$\sum x = 75.6 \, , \sum x^2 = 338.1.$$

(i) Stating your assumptions, calculate a symmetric 95% confidence interval for the population mean time for children to complete the puzzle. Assume Normal population and random sample (3.65, 4.75)

(ii) The manufacturers of the puzzle indicate a mean completion time of 5 minutes. What conclusion might be made about the children in the class? Class appears better than average, suggesting that they are not a random sample.

Confidence Intervals for a difference of population means, $\mu_1 - \mu_2$

(a) **The population variances σ_1^2 and σ_2^2 are known**

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Example 9:

Data were collected for the mass (in grams) of tomatoes of a particular species grown under glass and outdoors. The results can be summarized as follows.

Under glass : $\sum x = 9568.5, \quad \sum x^2 = 918141.55, \quad n_x = 100.$

outdoor : $\sum y = 9012.1, \quad \sum y^2 = 814725.65, \quad n_y = 100.$

Assuming that the data can be treated as independent random samples, calculate a 95% confidence interval for the difference in population masses for tomatoes grown under glass and outdoors. (4.15, 6.97)

Confidence Intervals for a difference of population means, μ_1 - μ_2

(b) The populations have a common variance, σ^2 , which is known.

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm z_{\frac{\alpha}{2}} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example 10:

The same physical fitness test was given to a group of 100 scouts and to a group of 144 guides. The maximum score was 30. The guides obtained a mean score of 26.81 and the scouts obtained a mean score of 27.53. Assuming that the fitness scores are normally distributed with a common population standard deviation of 3.48, Find a 90% confidence interval for the difference between population means for the scouts and guides.

Confidence Intervals for a difference of population means, $\mu_1 - \mu_2$

(c) The populations have a common variance, σ^2 , which is unknown. For large samples.

$$\sigma\boxed{?}^{2} = \frac{n_1 s_1^{\;2} + n_2 s_2^{\;2}}{n_1 + n_2 - 2}, \quad (s_1^{\;2} \text{ and } s_2^{\;2} \text{ are sample variances})$$

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm z_{\frac{\alpha}{2}}\,\sigma\boxed{?}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example 11:

Two Statistics teachers, Mr. Chalk and Mr. Talk, conducted an experiment, measuring the distances for several shots in a golf game.

Denoting the distance Mr. Chalk hits the ball by *x* metres, the following results were obtained:

$$n_1 = 40, \sum x = 4080, \sum \left(x - \bar{x}\right)^2 = 1132$$

Denoting the distance Mr. Talk hits the ball by *x* metres, the following results were obtained:

$$n_2 = 35, \sum y = 3325, \sum \left(y - \bar{y}\right)^2 = 1197$$

Example 11:

Assuming that the populations have a common variance, Find a 99% confidence interval for the difference between population means for their distances for several shots in the golf game.

Confidence Intervals for a difference of population means, $\mu_1 - \mu_2$

(d) The populations have a common variance, σ^2 , which is unknown.
For small samples.

$$\sigma\boxed{?}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}, \quad (s_1^2 \text{ and } s_2^2 \text{ are sample variances})$$

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{n_1+n_2-2}\,\sigma\boxed{?}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example 12:

A Mathematics teacher at a large school wishes to find a confidence interval for the difference in performance at a mental arithmetic test between students in Year 9 and students in Year 11. He takes random samples of size 15 from each year and obtains the following results for the scores, measured out of 100.

Year	sample mean	unbiased estimate of population variance
9	62.1	4.58^2
11	66.4	5.24^2

Example 12:

		unbiased estimate of
Year	sample mean	population variance
9	62.1	4.58^2
11	66.4	5.24^2

Find a 98% confidence interval for the difference between population means for the two years. You may assume that the scores are distributed normally and that the variances for the two years are equal. (-0.133 , 8.73)