PHT 351 Public health statistics II
BPH, Third year, sixth semester

Unit 1: Estimation

Introduction

The statistical technique of estimating unknown parameter from corresponding sample statistics is known as estimation. An estimate is a number obtain from the sample data, which we propose as the value of the parameter, and an estimator is a rule, which tells us how to come up with an estimate. In other words, any sample statistic used to estimate the population parameter is called the estimator.

For example, the sample mean $\bar{x}$, which we use for estimating the population mean, is an estimator of μ and the single numerical value of the sample mean is called an estimate of the parameter μ.

Estimation procedure can be divided into two categories: point estimation and interval estimation.

Point estimation

A point estimate is a single value that is used to estimate unknown population parameter. In point estimate we find a statistic which can be used for or to replace the true value of the population parameter for all practical purpose. A good estimator is the one, which is as close to the true value of the population parameter as possible.

Interval estimation

In interval estimation, probable range is specified within which the true value of the parameter might be expected to lie. However, this range varies with the confidence required. In this method, we first find a point estimate within which we can be reasonably confident that the true parameter will lie.

Example: If the weight of a person is recorded to be 60 kg, the measurement represents the point estimation. But if the weight of the person is shown as 60±5 kg or is recorded lying within the range 55-65 kg. This measurement gives the interval estimation.

Criteria of good estimator

1. **Unbiasedness**: The estimator is said to be unbiased if expected value of sample statistic is equal to the population parameter.
2. **Consistency**: A statistic is considered to be consistent estimator of the population parameter if as the sample size increases; the sample value of more close to the population parameter.
3. **Efficiency**: Efficiency refers to the size of the standard error of the sample statistic. The estimator with the least variance is considered as the most efficient estimator.
4. **Sufficiency**: An estimator is sufficient if it makes use of all the information in the sample. For example, mean is the sufficient estimator of the population mean because it is based on all the observation. But median and mode do not consider all the observation.

Sampling distribution of statistic

If we draw a sample of size n with simple random sampling without replacement (SRSWOR) from a given finite population of size N, then the total number of possible samples will be

$${}^{N}C_{n} = \frac{N!}{n!(N-n)!} = k \text{ (say)}$$

For each of these k values we can compute statistic $t = t(x_1, x_2, ..., x_n)$. Say the statistic is mean, standard deviation etc. which are shown below.

Sample no.	1	2	3	.	.	.	k
Statistic t	t_1	t_2	t_3	.	.	.	t_k
Mean	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	.	.	.	$\bar{x}_k$
Variance	S_1^2	S_2^2	S_3^2	.	.	.	S_k^2

The set of values the statistic so obtained for each sample, constitute the sampling distribution of mean or variance depending upon the statistic calculated. In general we call sampling distribution of statistic.

Then the mean and variance of statistic t would be,

Mean $\bar{t} = \frac{1}{n}\sum t_i$

Variance (t) = $\frac{1}{n}\sum\left(t_i - \bar{t}\right)^2 \; \forall \; i = 1, 2, ..., k$

Standard error of mean

The standard deviation of the sampling distribution of mean is called standard error of mean. Thus standard error of the sampling distribution of t is given by,

Standard error of mean, SE $(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Standard error of difference of means $(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

The above formula for SE are obtained in random sampling from an infinite population so that the sample size n is relatively very small as compared with the population size N and consequently $\frac{n}{N}$ (sampling fraction) can be neglected. Thus for a finite population of size N when a sample is drawn without replacement method we have,

Standard error of mean, SE $(\bar{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

Where $\sqrt{\frac{N-n}{N-1}}$ is called finite population correction (fpc) is applied when $\frac{n}{N} \geq 0.05$ (i.e. 5%)

Standard error of Proportion

Observed sample proportion (p) = $\sqrt{\frac{PQ}{n}}$

Difference of two proportions $(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

Thus for a finite population of size N when a sample is drawn without replacement method we have,

Observed sample proportion (p) = $\sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$

Example: A simple random sample of size 20 is drawn without replacement from a finite population of 75 units, if the number of defective units in the population is 12, find the standard error of the sample proportion.

Solution: Given n = 20, N = 75

P = Proportion of defective units in the population = 12/75 = 0.16

Q = 1 – P = 1 – 0.16 = 0.84

$$SE(p) = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0.16\times 0.84}{20}}\sqrt{\frac{75-20}{75-1}} = 0.071$$

Exercise

1. The mean pulse rate of 50 people is 79.41 and a standard deviation is 0.7. Calculate standard error of the mean. **(0.099)**
2. The mean plasma potassium level for 50 adult males in a certain disease was found to be 3.35mEq/litre and the standard deviation was 0.005 mEq/litre. Calculate the standard error of mean.
3. A group of 16 subjects, a mean serum iron level was found as 148μg% with a standard deviation of 0.34μg%. Calculate the standard error of mean.
4. In a sample of 25 observations from a normal distribution with mean 98.6 and standard deviation 17.2, find the standard error of mean.
5. The Biostatistics class has a total of 60 students. Their average score in their second term exams was 70 with a standard deviation of 8. A sample of 36 these students is taken at random. Calculate the standard error of the mean for this sample. **(0.85)**
6. In a study on growth of children, one group of 100 children had a mean height of 60 cm and SD of 2.5 cm while another group of 150 children had a mean height of 62 cm and SD of 3 cm. Calculate standard error of difference of the mean. **(0.35)**
7. Case fatality rate of liver failure is 70 percent. There are 74 cases during a year. What is the standard error of sample 74? **(0.053)**
8. In a survey of 300 children in the age group 0 – 5 years, showed 15% prevalence rate of protein calorie malnutrition. Calculate the standard error of proportion of malnutrition. **(0.021)**
9. Occurrence of pyorrhea among 100 persons who did not brush their teeth was found to be 10, while in another sample of 100 persons who brushed their teeth, the occurrence was found to be 2. Calculate the standard error of proportion.
10. A sample of 20 persons drawn from a certain village showed the number of attacks of cold per person is: 3, 4, 0, 7, 5, 1, 6, 5, 8, 8, 2, 6, 0, 2, 4, 6, 1, 5, 2, 6. Calculate the standard error of mean.
11. Blood serum cholesterol level of 10 subjects is: 260, 277, 250, 240, 255, 245, 278, 288, 263, 290. Calculate the standard error of mean.

Relationship between sample size and standard error

- As the standard error decreases, the precision (accuracy) with which the sample mean can be used to estimate population mean increases.
- In other word, as the standard error decreases, the value of any sample mean will probably be closer to the value of the population mean.
- As S.E. of a sample mean varies inversely with the square root of sample size *n* therefore, If we increase sample size then only S.E. of a sample mean decreases and the value of sample mean will be probably be closer to the value of the population mean.

Illustration

If population standard deviation σ =100.

Case I: Consider sample of size *n* = 10

S.E. of sample mean = $\frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{10}} = 31.63$

Case II: Consider sample of size *n* = 100

S.E. of sample mean = $\frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{100}} = 10$

From this as we have increase *n* from 10 to 100 the valve of standard error decreases.

Central Limit Theorem

Central Limit Theorem relates the shape of the population distribution and the shape of the sampling distribution of the mean. It state,

If $\bar{x}$ be the mean of a random sample of size *n* drawn from a population having mean μ and standard deviation σ, then the sampling distribution of sample mean $\bar{x}$ is approximately a normal distribution with mean μ and S.D. = S.E. of $\bar{x}$, provided the sample size *n* is sufficiently large ($n \geq 30$)

It assures us that whatever the shape of the population distribution, the sampling distribution of the mean approaches normal as the sample size increases.

The significance of the central limit theorem is that it permits us to use sample statistics to make inference about population *n* parameters without knowing anything about the shape of the frequency distribution of that population other than what we can get from the sample.

Example: The distribution of annual income of a hospital staff has a mean salary (μ) = Rs 34420 and a standard deviation (σ) of Rs 17076. If you draw a random sample of 30 hospital staff, what is the probability that their average earning is (i) more than Rs 40000 and (ii) between Rs 40000 and Rs 42000?

Solution:

Given Mean (μ) = 34420, standard deviation (σ) = 17076, sample size (n) = 30

For sampling distribution, standard normal variate is given by

$$Z = \frac{\bar{X} - \mu}{SE(\sigma_{\bar{x}})} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - 34420}{\frac{17076}{\sqrt{30}}}$$

(i) $P(\bar{X} > 40000) = ?$

When $\bar{X} = 40000, Z = 1.78$

$P(\bar{X} > 40000) = P(Z > 1.78) = 0.0375$ from normal table

(ii) $P(40000 \le \bar{X} \le 42000) = ?$

$P(40000 \le \bar{X} \le 42000) = P(1.78 \le Z \le 2.42) = 0.029$ from normal table

Determination of sample size

Determining the appropriate sample size for an investigation whether it is laboratory investigation or any other clinical trial is an essential step in the statistical design and is usually a difficult one. An adequate sample size should be drawn on a study so that it yields reliable results. Here we will develop a formula for determining the sample size for estimating the population parameters viz. mean and proportion.

Sample size for estimating population mean

For the determination of sample size the three factors must be known.

1. The desired confidence level, which determines the value of Z, the critical value from the standard normal distribution.
2. The acceptable sampling error E (i.e. the difference between sample mean and population mean.
3. The population standard deviation σ.

To develop a formula for determining the sample size, we have,

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

From above on simplification we get,

$$n = \left(\frac{Z\sigma}{E}\right)^2$$

where, n = required sample size

Z = significant value of Z at desired confidence level

E = $|X - \mu|$ = maximum allowable error

σ = standard deviation of population

Sample size for estimating population proportion

For the determination of sample size the three factors must be known.

1. The desired confidence level, which determines the value of Z, the critical value from the standard normal distribution.
2. The acceptable sampling error E (i.e. the difference between sample mean and population mean.
3. The population proportion of success P.

To develop a formula for determining the sample size, we have,

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

From above on simplification, we get,

$$\therefore n = \frac{Z^2 PQ}{E^2} = PQ\left(\frac{Z}{E}\right)^2$$

Where, n = required sample size
Z = significant value of Z at desired confidence level
P = population proportion of success
Q = 1 – P
E = $|p - P|$ = maximum allowable error permitted in the estimate.

Remark:
If no previous sample study has been taken so that we do not know the value of P, then we assume P to be equal to 0.5.

Exercise

1. The difference in cholesterol levels of obese patients treated long time back with some control and experimental diets was 20.5. The prior information available showed a standard deviation of 49.0 with a confidence of probability of 0.99, if we want to estimate the size of the sample, what is the sample size, we need for the study? (38)
2. A laboratory attendant would like to determine how fast students can go over a standard procedure in the calculator. Assuming that it is known from previous studies that $\sigma = 20$ seconds, he would like to determine how large a sample he should take to be 95% confident that his estimate will be within 15 seconds of the true mean. (7)
3. Suppose we want to estimate the mean weight of Nepalese men and we want to be 95% confident that our estimate is within ±2 kg of the actual mean. Since our previous study allows us to estimate that the standard deviation of men's weight is $\sigma = 18.4$ kg. Determine the appropriate sample size. (326)
4. Suppose it is necessary to know how many Kwashiorkor cases would be required to be at least 95 percent confident that the error in estimating the true proportion of Kwashiorkor children treated successfully by means of the sample proportion will not exceed 0.15. (43)
5. A team of medico research experts feels confident that a new drug they have developed will cure about 80% of the patients. How large should the sample size be for the team to be 98% certain that the sample proportion of cure within ±2% of the proportion of all cases that the drug will cure? (2172)
6. A public health survey is to be done in an urban area of Kathmandu to estimate the proportion of children ages 0 to 14 having adequate polio immunization. The final estimate should be within 0.05 of the true proportion with probability 0.98. What is the minimum sample size required?

Factors influencing sample size

- Margin of error
  A margin of error will get narrower as the sample size increases. The margin of error selected depends on the precision needed to make population estimates from a sample
- Confidence level
  As the confidence level increases, so too does the sample size. Researchers will chose a higher confidence level in order to reduce the chance of making a wrong conclusion about the population from the sample estimate.
- Proportion (or percentage)/ standard deviation

Confidence interval

Confidence interval defines a range of values within which our population parametric value is expected to lie. It is also known as confidence limit. The limit within which the null hypothesis should lie with specified probabilities is called confidence limits or fiducial limits. 95% confidence is the degree of confidence most commonly used. In general, a 95% confidence interval estimate is interpreted as follows: if all possible samples of the same size n are taken and their sample means are computed, there is 95% chance that the true population mean will lie somewhere within the interval around their sample means and only 5% of them do not. However wider confidence interval such as 99% is also used. The closer a point lies to the middle of the confidence interval, the more likely it is representative of the population. The width of the confidence interval depends on the standard error of the statistic and the degree of confidence we have chosen. If sample values lies between the confidence limits, the null hypothesis is accepted; if it does not, the hypothesis is rejected at the specified level of significance. The calculation of confidence interval is performed with reference to a probability distribution e.g. the normal distribution, t distribution and χ^2 distribution. Out of these the most frequently used methods for calculating confidence interval involves the use of normal distribution.

Confidence interval for estimating population mean

For 95% confidence limit, from area under normal probability curve we have,

$P(-1.96 \leq Z \leq 1.96) = 0.95$

or $P\left(-1.96 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95$

or $P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$

The values $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ are called 95% confidence limit or fiducial limit for mean of the population corresponding for a given sample.

Similarly for 99% confidence limit for estimating population mean $= \bar{x} \pm 2.58\frac{\sigma}{\sqrt{n}}$

$\therefore$ Confidence interval for estimating population mean $= \bar{x} \pm Z_\alpha S.E(\bar{x})$

$$= \bar{x} \pm Z_\alpha \frac{\sigma}{\sqrt{n}}$$

Where $\bar{x}$ = sample mean

Z_α = value of Z at α% level of significance

σ = population standard deviation

n = sample size

This shows that if the data are assumed to be normally distributed the sample mean and standard error of mean are used to calculate confidence intervals for the mean.

In general, the confidence interval for any statistic is given by,

Confidence interval = $t \pm Z_\alpha$ S.E(t)

Remarks:

1. If σ^2 is unknown then for large sample its estimate provided by the sample variance s^2 is used to obtain the confidence limit for μ.

   i.e. Confidence interval = $\bar{x} \pm Z_\alpha \frac{s}{\sqrt{n}}$

2. The above relation for estimating the limits for μ can be used only for infinite population. In case of simple random sampling without replacement from a finite population of size N, then

   Confidence interval = $\bar{x} \pm Z_\alpha \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

The following table gives common confidence level and their Z values

Confidence level $(1-\alpha)$	50%	68.26%	90%	95%	96%	98%	99%	99.73%
Z_α	0.6745	1	1.645	1.96	2.05	2.33	2.58	3

Note: when no reference to the confidence level is given then always take Z_α= 3.

Confidence interval for difference of means

If $\bar{x}_1$ and $\bar{x}_2$ are the sample means of two large independent random sample of sizes n_1 and n_2 drawn from two infinite population with mean μ_1 and μ_2 and standard deviation σ_1 and σ_2 respectively. The $(1-\alpha)$% confidence interval for estimating the difference of mean $(\mu_1 - \mu_2)$ is given by,

Confidence limit = $\left(\bar{x}_1 - \bar{x}_2\right) \pm Z_\alpha$ S.E$\left(\bar{x}_1 - \bar{x}_2\right)$

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

If σ_1^2 and σ_2^2 are unknown, then their estimates provided by corresponding sample variance s_1^2 and s_2^2 i.e. for large sample, $\hat{\sigma}_1^2 = s_1^2$ and $\hat{\sigma}_2^2 = s_2^2$

The confidence interval for estimating the difference of mean $(\mu_1 - \mu_2)$ is given by,

Confidence limit = $\left(\bar{x}_1 - \bar{x}_2\right) \pm Z_\alpha \, S.E\left(\bar{x}_1 - \bar{x}_2\right)$

$$= \left(\bar{x}_1 - \bar{x}_2\right) \pm Z_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Confidence interval for estimating population proportion

Confidence interval for estimating population proportion = $p \pm Z_\alpha S.E(p)$

$$= p \pm Z_\alpha \sqrt{\frac{pq}{n}}$$

The confidence interval for estimating the difference of two proportion ($p_1 - p_2$) is given by,

Confidence limit = $\left(p_1 - p_2\right) \pm Z_\alpha \, S.E\left(p_1 - p_2\right)$

$$= \left(p_1 - p_2\right) \pm Z_\alpha \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Exercise

1. In maintaining the quality control on the production of optical lenses, the production manger wants to estimate the mean thickness of the lenses produced. A random sample of 100 lenses revealed a mean of 0.50 millimeter thickness. The population standard deviation is known to be 0.15 millimeter. Calculate a 99 percent confidence interval for the mean thickness of the population of optical lenses produced. (0.4613, 0.5387)
2. We wish to estimate the average number of heart beats per minute for a certain population at 98% confidence level. The average number of heart beats per minute for a sample of 49 subjects was found to be 90. Assume that 49 patients constitute a random sample and that population is normally distributed with a standard deviation of 10. (86.6681, 93.3319)
3. 100 transdermal patches have been removed from a batch and the stability of the active agent determined at 25^{o}C. The observed mean and standard deviation of degradation rate constant for the therapeutic agent was 0.09 ± 0.01per day. Calculate the 80% confidence interval of the estimates of the mean. (0.089, 0.091)
4. A new therapeutic agent has been developed to promote diuresis. In a clinical trial, the diuretic effects of this new agent and a commercial agent were assessed. In this the urine was collected over a 12 hour period after administration of a single tablet. The mean ± standard deviation volume of urine collected in the group of 65 patients who received the new therapeutic agent was 48.8 ± 9.1L, where as in the control group of 95 patients (who received the commercially available preparation) the volume was 37.9 ± 4.6L. Calculate the 95% confidence intervals for the difference in the urine volume induced by the two therapeutic agents. (10.9 ± 2.40L)
5. In a random sample of 400 items from a large consignment, 20 items were found to be defective. Find 99% confidence limits for the percentage of defectives in the consignment. (2.19, 7.81)
6. In a marketing survey for the introduction of a new product in a town, a sample of 400 persons was drawn. When they were approached for sale, 280 of them purchased the product. Find the 95% confidence limits for the percentage of persons who would buy the product in the town. (65.5% , 74.5%)