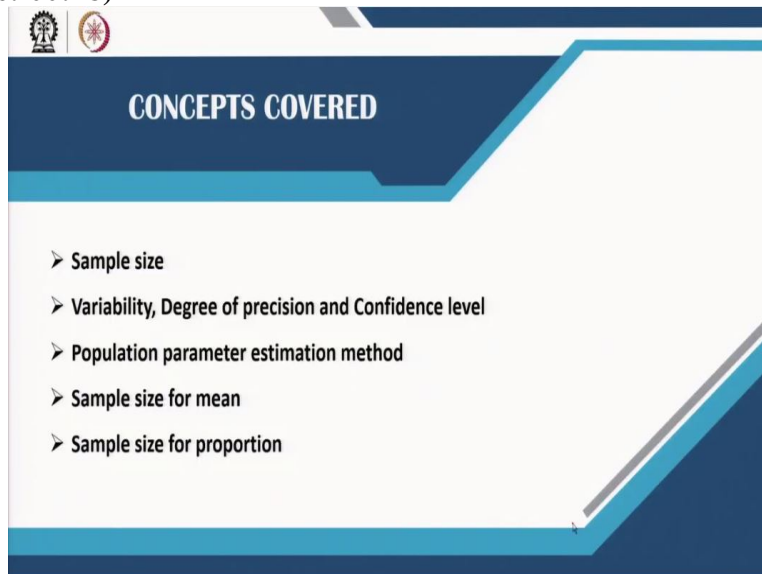


Urban Landuse and Transportation Planning
Prof. Debapratim Pandit
Department of Architecture and Regional Planning
Indian Institute of Technology - Kharagpur

Lecture-12
Sampling Theory-2

(Refer Slide Time: 00:28)



The slide features a dark blue header with the text 'CONCEPTS COVERED' in white. Below the header, a list of five topics is presented, each preceded by a right-pointing arrowhead. The slide also includes the logos of the Indian Institute of Technology (IIT) Kharagpur and the Department of Architecture and Regional Planning in the top left corner.

- Sample size
- Variability, Degree of precision and Confidence level
- Population parameter estimation method
- Sample size for mean
- Sample size for proportion

Welcome back to lecture 12, where sampling theory part 2 will be covered. In this lecture sample size, variability of population, degree of precision, confidence level, population parameter, estimation method and within that sample size for mean and sample size for proportion will be covered.

(Refer Slide Time: 00:49)

Sample size
 Sampling error reduces with the increase in sample size.
 Increase in sample size cause higher resource (i.e. time and cost) consumption.
 Equilibrium between these two extreme conditions.

Sample size determination is not only based on statistical criteria but also on financial and logistical constraints.

e.g., Comprehensive mobility plans require 1-2% of population to be surveyed for household surveys depending on city size.

Sample size can be determined based on the purpose of sample selection.

Population parameter estimation method (precision based)*
 Hypothesis testing method (power based)**

*With what precision measurement needs to be done?
 **Hypothesis testing: Two data sets are tested or compared for some statistical relationship.
 How small a difference should be detected and at what degree of certainty?

NPTEL

Background

The last lecture covered the different procedures for collecting samples or procedures for surveying. In this lecture, the different ways to calculate the sample size for a particular study will be covered. Since a sample is used instead of the total population, any parameter calculated on the sample has error. For example, the central tendency, like mean or standard deviation, of a sample will be different from that of the population. Sampling error reduces with the increase in sample size. But increased sample size will incur more time and cost and there is a need to find the optimum sample size.

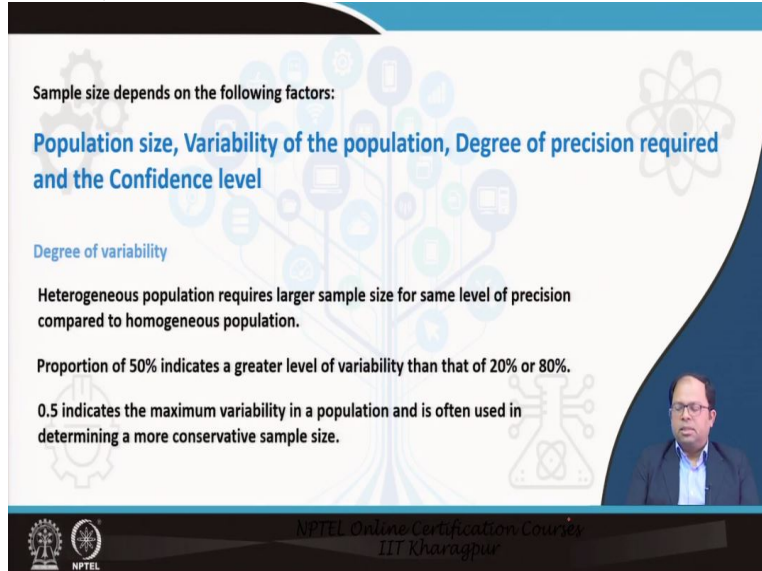
Sample size determination is not only based on statistical criteria, but also on financial and logistical constraints. Sample size can be decided based on formula, but a large number of samples may breach financial constraints. For travel diary surveys in Indian conditions, a rule of thumb is to survey 1-2% of the population in large high density urban areas and around 10% for cities with smaller population, provided it does not exceed 1000 samples.

Primarily there are two ways in which we can determine sample size. One is called population parameter estimation method and the other is hypothesis testing method.

Population parameter estimation method is based on precision. For example, how precise should be the estimate of the mean from the actual estimate of the mean of the population? Hypothesis testing method is power based. In hypothesis testing two data sets are tested or compared to find any sort of statistical relationship between them. If there is no relationship the null hypothesis gets validated. Power means how small a difference should be detected and to what degree of certainty.

But in this particular study, focus will not be on the hypothesis testing method but on the population parameter estimation method.

(Refer Slide Time: 05:57)



Sample size depends on the following factors:

Population size, Variability of the population, Degree of precision required and the Confidence level

Degree of variability

Heterogeneous population requires larger sample size for same level of precision compared to homogeneous population.

Proportion of 50% indicates a greater level of variability than that of 20% or 80%.

0.5 indicates the maximum variability in a population and is often used in determining a more conservative sample size.

NPTEL Online Certification Course
IIT Kharagpur

Sample size determination

Sample size depends on primarily four factors- population size, variability of the population, the degree of precision required and the confidence level.

A higher population may have a role in the sample size collected. But it may not always be the case. For example, the sample size will be finite even if the population is infinite.

A more heterogeneous population requires a larger number of samples for the same degree of precision. This is called the degree of variability of the population.

So, that is why whenever a group(or population) is divided into 50 50 proportion among two sub-groups it means that, maximum amount of heterogeneity is being considered. If it is like 80 or 20% that means, one sub-group is 80% another sub-group is 20%.

For a very conservative sample size selection, we take variability as a fixed value of 0.5, which actually indicates maximum variability in the population that is possible, when we are comparing 2 groups.

(Refer Slide Time: 08:24)

Degree of precision

Degree of precision or sampling error arises since a sample is being used to estimate a population parameter (mean).

Bigger samples have less sampling errors.

Error is expressed in percentage.

When the sample size is large, sample means are near to the population mean.

When sample size equals the population size, the standard error of the mean is 0.

Relationship between sampling error and sample size
(Source: https://en.wikipedia.org/wiki/Margin_of_error)

Sample size	Margin of error
2,401	4%
1,067	3%
600	4%
384	5%
96	10%

NPTEL Online Certification Courses
IIT Kharagpur

Degree of precision

Degree of precision or sampling error arises when a sample is being used to estimate a population parameter, like mean in this case. As can be seen from the diagram, for a higher degree of precision the number of samples required is high. Or it can be said that when the sample size is large, the sample mean is closer to the population mean. When the sample size equals the population size, the standard error of the mean is zero.

(Refer Slide Time: 10:14)

Confidence level or risk level is determined using normal distribution and the Central Limit theorem.

If we draw samples of n items from the overall population (may not follow normal distribution) and develop the distribution of the sample means, it becomes increasingly normal as the sample size increases.

Normal distribution

The probability density function of the normal distribution with mean (μ) and standard deviation (σ) is given by:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty$$

Central Limit theorem

If $x_1, x_2, x_3, \dots, x_n$ is a random sample of size n drawn from any arbitrary population (with any arbitrary distribution) having mean μ and variance σ^2 , then the distribution sample mean (\bar{x}) is normally distributed with mean μ and variance σ^2/n , provided n is sufficiently large ($n > 30$).

Standard error = σ/\sqrt{n}

Sample statistic can be used to estimate population parameter.

NPTEL Online Certification Courses
IIT Kharagpur

Confidence level

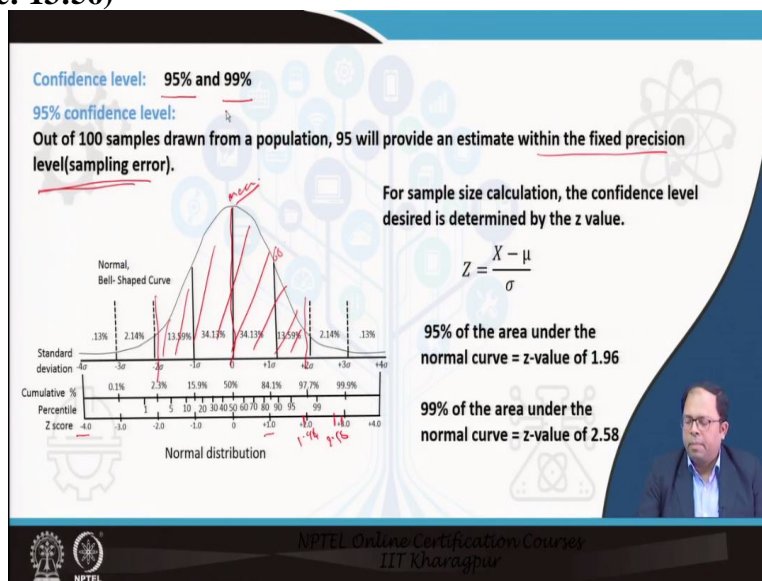
Confidence level or risk level is determined using normal distribution and the central limit theorem. According to the Central Limit theorem when n samples are drawn from a population,

which may or may not follow a normal distribution, the distribution of the sample means increasingly becomes normal as the sample size increases. With a large number of samples this could be defined as a probability density function with normal distribution with mean μ and standard deviation σ and it is given by the equation shown in the slide.

So, the central limit theorem states that if a random sample of size n drawn from any arbitrary population (with any arbitrary distribution) having mean μ and variance σ^2 square then the distribution sample mean (\bar{x}) is normally distributed with mean μ and variance σ^2/n provided that n is sufficiently large (>30).

Therefore, if the sample estimates or the sample parameters are known, one can estimate the population parameters from them. The population mean will be equal to the mean of the sample means. And the sample standard deviation will be equal to the standard deviation of the total population divided by the total number of the sample size. Using this relationship we can also determine that what should be the sample size in case we know the population parameters like μ and σ .

(Refer Slide Time: 13:56)



Confidence level

In the normal distribution curve shown in the figure around 34.13% of the observation lies on one side within one standard deviation of the mean, i.e. around 68% of the data lies within one standard deviation on both sides of the mean. Similarly, around 95% of the data lies within 2 standard deviation of the mean. So, in 95% confidence level out of 100 samples drawn from a

population 95 will provide an estimate within the fixed precision level or sampling error. So, for sample size calculation, the confidence level is determined by the Z value. Now,

$$Z = (X - \mu) / \sigma,$$

which is nothing but the Z score. The z-score is zero at the mean, 1 at a distance of 1 standard deviation (SD), 2 at a distance of 2 SD and so on.

So, 95% of the area under the normal curve is equal to a Z score 1.96 and 99% of the area under the normal curve has a z value of 2.58. So, confidence interval states that, for a particular precision value, one can be sure that, it will lie within this particular range of observations.

So, out of 100 samples drawn from a population 95 will provide an estimate within the fixed precision level defined. This is the sampling error.

(Refer Slide Time: 17:41)

Population parameter estimation method

Parameter of interest (Proportion, Mean and Standard Deviation)
For sample size calculation we need to know some of these parameters.

- Previous studies
- If not available, a pilot study needs to be done to get the estimates for mean, standard deviation and proportion.
- Based on experience
- Rule of thumb:
Standard deviation = Range (maximum - minimum) / 4
- Strata and other information is also required.

Sample size calculation should be done using different methods such as using proportion, or mean and standard deviation and the largest sample size should be chosen.

NPTEL Online Certification Courses
IIT Kharagpur

Population parameter estimation level

In the population parameter estimation method, the term parameter is important and it means that the parameters of interest are proportion of that particular attribute in the population or the mean or the standard deviation for that particular population. Knowing some of these values one can estimate the sample size.

So if one goes for a sample survey in a town, with which he/she is not familiar with, and for which there is no sampling frame (the list of people that are there in the town) available following are the choices available. One can look for previous studies on this particular area or this particular population. If it is still not available, a small pilot study can be done and estimates for mean or standard deviation obtained. Alternately, the rule of thumb can be used, where

$$\text{Standard deviation} = \text{Range}/4$$

$$\text{Range} = \text{maximum} - \text{minimum}$$

Sample size calculation should be done using different methods and the largest sample size obtained should be used. In case this is not practical, the proportion method or the mean and standard deviation methods are the most appropriate.

(Refer Slide Time: 20:44)

Simple random sampling

Determination of Sample size for Mean

For polytomous or continuous variables.
Polytomous to dichotomous and then use a sample size based on proportion.

$$n_m = \frac{z^2 \cdot \sigma^2}{e^2}$$

Where,
 n_m is the sample size,
 z is the Z score value of the confidence level,
 e is the desired level of precision
 σ^2 is the variance of the attribute (standard deviation has to be known)

Sample size varies for each attribute due to their variances.
The highest sample size is usually adopted.

Non-response factor.

Z score for confidence level (95%) = 1.96
Standard deviation = 5
 $e = \pm 10\% = 0.1$
 $n_m = 1.96^2 \cdot 5^2 / (0.1)^2 = 9604$

NPTEL Online Certification Courses
IIT Kharagpur

Simple random sampling

In case of simple random sampling we determine the sample size by different methods. The first method is determination of sample size for mean, which is apt for polytomous (variable with multiple categories) or dichotomous (with two categories) variables. For example, if the categorization is in terms of male or female it is dichotomous. Whereas, in case of multiple age groups the categorization can be termed polytomous. Polytomous could be also converted to dichotomous by adding up the different categories into broader categories and ultimately bringing it down to two variables, where one can use a sample size based on proportion.

So, in case of determination of sample size, for mean, the formula that is used is shown in the slide, where, n_m is the sample size, Z is the Z score of the confidence level and e is the desired level of precision.

For a confidence level of 95% the z-score is 1.96. Let the standard deviation for this particular attribute be 5 and error value is +- 10% (if the actual estimate is 50, the estimate could be from 40 to 60). Based on this, the sample size comes to around 9604 samples.

The standard deviation is different for different attributes of the population. For example, in a travel diary survey, if the attribute is gender, then standard deviation will be different and the variance will be different which would result in a different sample size. The varying sample size for the different attributes is the main drawback of this method. In addition to this, the non response factor should also be considered, i.e. a lot of people do not respond to all the questions. This leads to redundant samples and provision should be made for collecting a higher number of samples than the estimate.

Compared to population parameter estimation method, the method based on proportion is more appropriate.

(Refer Slide Time: 25:31)

Determination of sample size from proportion

Cochran's formula for calculating sample size when the population is infinite

$$n_s = \frac{(z^2 \cdot p \cdot q)}{e^2}$$

Where,
 n_s is sample size,
 z is the Z score value of the confidence level,
 p is the proportion of the attribute in the population,
 $q = 1 - p$
 e is the level of precision.

Z score for confidence level (90%) = 1.645
 $p = 0.5$ (50 percent variability)
 $q = 0.5$
 $e = \pm 5\% = 0.05$
 $n_s = 1.645^2 \cdot 0.5 \cdot 0.5 / (0.05)^2 = 270$

Z score for (99% c.l.) = 2.58
 $p = 0.8$
 $q = 0.2$
 $e = \pm 5\% = 0.05$
 $n_s = 2.58^2 \cdot 0.8 \cdot 0.2 / (0.05)^2 = 426$

NPTEL Online Certification Courses
 IIT Kharagpur

Determination of sample size from proportion

For determining the sample size from proportion, one popular formula is Cochran's formula for calculating sample size. When the population is infinite the formula is shown in the slide.

$$n_s = (z^2 \cdot p \cdot q) / e^2$$

So here n_s is sample size, z is the Z score value of the confidence level and p is the proportion of the attribute in the population and $q = 1 - p$.

Compared to the earlier formula, pq is taking care of the variability. p is the proportion of that attribute in the population and q is the $1 - p$.

For a confidence interval of 90%, the Z score is around 1.645. For a maximum variability of 50% p is taken as 0.5, and then q becomes 0.5. Error is taken $\pm 5\%$. Putting all the values, the sample size comes to around 270 samples.

Increasing this error to 10% will further reduce the sample size. The sample size also reduces with variability. Increasing the confidence level from 90 to 95% or 99%, on the other hand will increase the sample size. So, here using Z score of 2.58 (for 99% confidence interval) and reduced variability with $p = 0.8$ and $q = 0.2$ we get a sample size of 426. This also shows that the impact of confidence level is pretty high compared to variability.

(Refer Slide Time: 28:05)

Cochran's formula (population size is finite)

$$n = \frac{n_s}{1 + \frac{(n_s - 1)}{N}}$$

Finite population correction factor

$N =$ population
 $n_s =$ sample size determined using Cochran's formula for infinite population.

If sample size estimated is more than 5 percent of total population then only this correction factor is used which reduces the sample size.

Let us assume population=50000
 So, $4160/50000 * 100 = 8.3\text{percent}$.

Z score for (99% c.i.) = 2.58
 $p = 0.5$
 $q = 0.5$
 $e = \pm 2\% = 0.02$

Cochran's formula for infinite population
 $n_s = 2.58^2 \cdot 0.5 \cdot 0.5 / (0.02)^2 = 4160$

Correction factor

$$n = 4160 / (1 + (4160 - 1) / 50000)$$

$$n = 4160 / 1.08$$

$$n = 3851 \text{ (new sample size)}$$

NPTEL Online Certification Courses
 IIT Kharagpur

Cochran's formula for finite population

If the sample size for a finite population is estimated to be more than 5% of the total population, there is a need to incorporate a correction factor which can reduce the sample size. Incorporating the correction factor the formula is shown in the slide.

$$n = n_s / (1 + (n_s - 1) / N)$$

Where, N is equal to the total population and n_s is equal to the sample size determined using Cochran formula for Infinite population. So, for a total population of around 50,000 and based on the previous calculation (Z score is 2.58, variability is 0.5, e is $\pm 2\%$) the samples that need to be collected is 4160.

So, this population is approximately 8.3% of the total population. So, the correction factor is utilized, reducing the sample to 3851, which is less than the sample for an infinite population.

(Refer Slide Time: 30:13)

Yamane's formula (based on population)

Considering, 95% confidence level and p (proportion of the attribute in the population) = 0.5.

$$n_t = \frac{N}{1 + N(e^2)}$$

N = population
 e is the level of precision.
 n_t = sample size

$N = 50000$
 $e = \pm 2\% = 0.02$
 $n_t = 50000 / (1 + 50000(0.02^2)) = 2381$

$N = 50000$
 $e = \pm 5\% = 0.05$
 $n_t = 50000 / (1 + 50000(0.05^2)) = 397$

[10%]

NPTEL Online Certification Course
 IIT Kharagpur

Yamane's formula

Yamane proposed a formula where the population can be used directly instead of the variability in the population. So, here we can say that, if we consider a confidence interval of 95% and the proportion of attributes of the population is considered fixed at 0.5%, then the sample size is given by the formula shown in the slide.

$$n_t = N / (1 + N * e^2)$$

Where, N is the population and e is the level of precision and n_t is the sample size.

For $N = 50,000$, with an e value of $\pm 2\%$, the total sample size required is 2381. Whereas, for $e = \pm 5\%$, the sample size is 397. It can be seen that the increase in the precision increases sample size by a huge order. In certain studies the precision level required may not be that high, and the sample size turns out to be low.

(Refer Slide Time: 31:44)

Sample size requirement

Stratified random samples (Variances of subpopulations, strata, or clusters along with variability in population) *deft.*

As per analysis:

- Mean, Frequencies (any size)
- Multiple regression, ANOVA etc. : 200-500 samples *350*

As per distribution of data:

- Normal distribution: 30 to 200
- Skewed distribution

NPTEL Online Certification Courses
IIT Kharagpur

Conclusion

The methods discussed here are valid for sample size collection in simple random sampling, whereas, for stratified random sampling, one needs to consider not only the variances of the population, but also the variances of each sub population, the different strata, clusters, etc.

In this case some new factors get introduced into the equation, which is called a deft factor which takes care of the sample subgroup variances and so on.

Different people have designed different formulas for sample size, and there are no hard and fast rules for determining sample size. It has to be dependent on the analysis that one is doing. Thus, for estimating only mean and frequency of a particular variable from the sample, the sample size can be much smaller. Whereas, for a detailed analysis like regression analysis, one requires approximately 200 to 500 samples. For Logit regression a standard thumb rule is to have around 350 samples.


For a normally distributed data, a lower sample size (around 30 to 200) is required. But, if the distribution is skewed, probably more samples are required. So, these are the different aspects that also influences sample size.

(Refer Slide Time: 34:20)

References:

REFERENCES

- Bryman, A. (2003). Quantity and Quality in Social Research. In Quantity and Quality in Social Research. <https://doi.org/10.4324/9780203410028>
- Bryman, A. (2012). Social research methods Bryman. OXFORD University Press. <https://doi.org/10.1017/CBO9781107415324.004>
- ICF International. 2012. Demographic and Health Survey Sampling and Household Listing Manual. MEASURE DHS, Calverton, Maryland, U.S.A.: ICF International
- Designing Household Survey Samples: Practical Guidelines, ST/ESA/STAT/SER.F/98, Department of Economic and Social Affairs, Statistics Division, United Nations, New York, 2005.
- Glenn D. Israel, Determining sample size, PEOD6, University of Florida, IFAS extension, November 1992.



(Refer Slide Time: 34:29)

CONCLUSION

Various methods and rule of thumbs are followed for sample size determination.

In addition to population characteristics, sample size depends on the desired degree of precision and confidence level required by the surveyor.

Sample size determination based on proportion is more robust compared to sample size determined using population variance.

Sample size also depends on the type of analysis and the distribution of the data.

To conclude, it can be said that various methods and rule of thumbs are followed for sample size determination. In addition to population characteristics, sample size depends on desired degree of precision and confidence level required by the surveyor. So, for a more precise result and higher confidence level more would be the sample size.

Sample size determination based on proportion is more robust compared to sample size determination using population variance. And finally, sample size also depends on the type of analysis and the distribution of that data. Thank you.