Department of Census and Statistics

Training Division

Training Module Description

Module: Survey Sampling

Version: 1.0

Duration: 05 days

1.0 Module Description

This course covers sampling design and analysis methods useful for research and management in many fields. A well designed sampling procedure ensures that we can summarize and analyze the data with a minimum of assumptions or complications. In this course, we'll cover the basic methods of sampling and estimation and then explore selected topics and recent developments.

1.1 Objectives of the Module are;

1.1.1 To provide a good knowledge in various sampling methods and how to minimize errors in sampling.

1.1.2 Understand the concept of sampling distribution of a statistic

1.1.3 Identify the sampling distribution of a mean and the sampling distribution of a proportion

1.1.4 State and understand the Central Limit Theorem

1.1.5 Know the conditions under which you can use the normal distribution to approximate the sampling distribution of a proportion

1.1.6 Define and describe systematic sample, stratified sample, cluster sample , and Proportionate sample

1.2 Target Participants

Anyone preparing for managerial or policy making careers in governmental or private organizations with an orientation towards public service. No mathematical background beyond elementary algebra is assumed but, Participants should have a basic knowledge of statistics.

1.3 Contents (Syllabus)

1. Data Collection Methods
2. Survey Sampling Methods
3. Bias in Survey Sampling
4. Sampling Distributions
5. Difference Between Proportions
6. Difference Between Means
7. Multi Stage Designs
8. Double or Two-Phase Sampling
9. Applied Problems for Survey Sampling
10. Laboratory Session

1.4 Method of delivery

The course is delivered by means of lectures, laboratory sessions, discussion, and brainstorming sessions.

2.0 Session 1 (Day 1: 0900 – 1200hrs)

2.1 Session Description

An introduction to the Sampling is given and various methods of Methods of Data Collection

Will be discussed in detail. Lesson will be concluded after discussing advantages and disadvantages of Data Collection Methods discussed in the class.

2.2 Session Learning Outcomes

By the end of this session participants will be able to;

2.2.1 Define various Methods of Data Collection, including Census**,** Sample survey, Experiment, and Observational study

2.2.2 Analyze advantages and disadvantages of Data Collection Methods

2.3 Brief outline of the session

2.3.1 Methods of Data Collection

2.3.1.1 Census

2.3.1.2 Sample survey

2.3.1.3 Experiment

2.3.1.4 Observational study

2.3.2 Data Collection Methods: Pros and Cons

2.3.2.1 Resources

2.3.2.2 Generalizability

2.3.2.3 Causal inference

3.0 Session 2 (Day 1: 1300 – 1600hrs)

3.1 Session Description

One of the most important decisions that any researcher makes is how to obtain the type of participants needed for the study. In this discussion, participants will be given examples of each of the strategies discussed.

3.2 Session Learning Outcomes

By the end of this session participants will be able to;

3.2.1 Distinguish clearly between a parameter and a statistic

3.2.2 Identify the sampling frame

3.2.3 Learn various probability sampling strategies

3.3 Brief outline of the session

3.3.1 Population

3.3.2 Population Parameter vs. Sample Statistic

3.3.3 Sampling Frame

3.3.4 Probability vs. Non-Probability Samples

3.3.5 Probability Sampling Strategies

3.3.5.1 Simple random sampling

3.3.5.2 Systematic sampling

3.3.5.3 Stratified random sampling
3.3.5.4 Proportionate sampling
3.3.5.5 Cluster sampling

3.3.5.6 Multistage sampling

4.0 Session 3 (Day 2: 0900 – 1200hrs)

4.1 Session Description

Bias often occurs when the survey sample does not accurately represent the population. In this module participants will be introduced to various issues that leads to bias and how to minimize bias in sampling.

4.2 Session Learning Outcomes

By the end of this session participants will be able to;

4.2.1 Identify bias which occurs due to unrepresentative samples , and how to obtain a representative sample.

4.2.2 Identify bias which occur due to measurement errors

4.2.3 How to minimize sampling errors and survey bias.

4.3 Brief outline of the session

4.3.1 Bias due to unrepresentative samples

4.3.1.1 Undercoverage**.**

4.3.1.2 Nonresponse bias

4.3.1.3 Voluntary response bias

4.3.2 Bias due to measurement errors

4.3.2.1 Leading questions

4.3.2.2 Social desirability

4.3.3 Sampling errors and survey bias

5.0 Session 4 (Day 2: 1300 – 1600hrs)

5.1 Session Description

The sampling distribution of a statistic is created by repeated sampling from one proportion. In this session you will be introduced to the Central Limit Theorem, sampling Distribution of the mean, and the proportion. We discuss how these distributions are created theoretically and empirically.

5.2 Session Learning Outcomes

By the end of this session participants will be able to;

5.2.1 Understand the Central Limit Theorem

5.2.2 Understand the sampling distribution of the Mean

5.2.3 Understand the sampling distribution of the Proportion

5.3 Brief outline of the session

5.3.1 Variability of a Sampling Distribution
5.3.2 Central Limit Theorem
5.3.3 Sampling Distribution of the Mean
5.3.4 Sampling Distribution of the Proportion

6.0 Session 5 (Day 3: 0900 – 1200hrs)

6.1 Session Description

Many statistical applications involve comparisons between two independent sample proportions, and two independent sample means. In this session participants are introduced to the theory of deference between proportions and how to solve problems. And also the will learn the theory and problems of deference between means.

6.2 Session Learning Outcomes

By the end of this session participants will be able to;

6.2.1 Understand the Difference Between Proportions
6.2.2 Understand the Difference Between Means

6.3 Brief outline of the session

6.3.1 Difference Between Proportions: Theory
6.3.2 Difference Between Proportions: Problem
6.3.3 Difference Between Means : Theory
6.3.4 Difference Between Means : Problem

7.0 Session 6 (Day 3: 1300 – 1600hrs)

7.1 Session Description

In Section1, we introduce multi-stage design and give a few practical examples. We then provide the notations for two stage design. The unbiased estimators for two stage design with simple random sampling at each stage is discussed. Then we discuss the ratio estimator for two stage design with simple random sampling at each stage.

In Section 2, we discuss the two stage design with primary units selected with probability proportional to size and secondary units selected with simple random sampling. The Hansen-Hurwitz estimator is computed for this situation. We also show how to compute the estimated variance for the H-H estimator.

7.2 Session Learning Outcomes

By the end of this session participants will be able to;

7.2.1 know why and when to use multi-stage sampling

7.2.2 compute unbiased estimator and its estimated variance for the two stage design when srs is used at each stage

7.2.3 compute ratio estimator and its estimated variance for the two stage design when srs is used at each stage

7.2.4 compute the Hansen-Hurwitz estimator and its estimated variance when primary units are selected with probability proportional to size and secondary units selected with srs

7.3 Brief outline of the session

7.3.1 Two stages with S.R.S. at each stage

7.3.1.1 multi-stage design
7.3.1.2 two stages with simple random sampling at each stage
7.3.1.3 unbiased estimator
7.3.1.4 ratio estimator

7.3.2 Two Stages with Primary Units Selected by Probability Proportional to Size and Secondary Units Selected with S.R.S.

8.0 Session 7 (Day 4: 0900 – 1200hrs)

8.1 Session Description

In Section 1, we introduce double sampling and discuss application of double sampling for ratio estimation. We then provide the formula for the variance of the ratio estimator while double sampling is used. An example is given to illustrate how to conduct the double sampling and how to compute the ratio estimator as well as the estimated variance of the estimator. The allocation in double sampling is then discussed.

In Sections 2, double sampling for stratification is discussed. An example is then used to illustrate the use of double sampling for stratification and the computation of the estimate as well as the estimated variance of the estimate.

8.2 Session Learning Outcomes

By the end of this session participants will be able to;

8.2.1 know how to use double sampling to collect information for ratio estimation

8.2.2 compute the optimal allocation in double sampling for ratio estimation

8.2.3 know how to use double sampling to collect information for stratification

8.2.4 compute the estimate when double sampling is used to collect information for stratification

8.2.5 compute the estimate variance of the estimate when double sampling is used to collect information for stratification

8.3 Brief outline of the session

8.3.1 Double Sampling for Ratio Estimation

8.3.1.1 double sampling
8.3.1.2 ratio estimation with double sampling
8.3.1.3 allocation in double sampling for ratio estimation

8.3.2 Double Sampling for Stratification

8.3.2.1 double sampling for stratification

8.3.2.2 example

9.0 Session 8 (Day 4: 1300 – 1600hrs)

9.1 Session Description

In Section 1, we discuss non-sampling error, which tend to persist even if sample size gets larger and larger. We then discuss one common non-sampling error: nonresponse. We then discuss how to use double sampling to adjust for non-response in the form of call backs. Then an example is given to illustrate how to compute the estimate and the estimated variance of the estimate. We then provide the formula for selecting the optimal number for call back. In section 1, we also discuss how to design surveys to reduce non-response.

In Sections 2, the technique of interpenetrating subsample is discussed. An example is used to show this technique which takes into consideration the interviewer effect.

In Section 3, we discuss the case we do not know whether an element belongs to the subpopulation until after it has been sampled and how to estimate the mean and total of this subpopulation. As example is then given to illustrate the method.

9.2 Session Learning Outcomes

By the end of this session participants will be able to;

9.2.1 understand the difference between sampling error and non-sampling error

9.2.2 use double sampling to adjust for non-response by call backs

9.2.3 find the optimal allocation for the number of callbacks

9.2.4 use interpenetrating subsample technique to take care of interviewer effect

9.2.5 estimate the mean and total over subpopulation

9.3 Brief outline of the session

9.3.1 Double Sampling for Non-response

9.3.1.1 non-sampling error and double sampling

9.3.1.2 selecting the number of call backs

9.3.1.3 variance of the mixture distribution

9.3.1.4 designing surveys to reduce non-response

9.3.2 Interpenetrating Subsample

9.3.2.1 interpenetrating subsamples

9.3.2.2 example

9.3.3 Estimation of means and totals over subpopulation

9.3.3.1 estimation of means and totals over subpopulation

9.3.3.2 example

10.0 Session 9 (Day 5: 0900 – 1200hrs)

10.1 Session Description

In Section 1, we discuss non-sampling error, which tend to persist even if sample size gets larger and larger. We then discuss one common non-sampling error: nonresponse. We then discuss how to use double sampling to adjust for non-response in the form of call backs. Then an example is given to illustrate how to compute the estimate and the estimated variance of the estimate. We then provide the formula for selecting the optimal number for call back. We also discuss how to design surveys to reduce non-response.

In Sections 2, the technique of interpenetrating subsample is discussed. An example is used to show this technique which takes into consideration the interviewer effect.

In Section 3, we discuss the case we do not know whether an element belongs to the subpopulation until after it has been sampled and how to estimate the mean and total of this subpopulation. As example is then given to illustrate the method.

10.2 Session Learning Outcomes

By the end of this session participants will be able to;

10.1.1 understand the difference between sampling error and non-sampling error
10.1.2 use double sampling to adjust for non-response by call backs
10.1.3 find the optimal allocation for the number of callbacks
10.1.4 use interpenetrating subsample technique to take care of interviewer effect
10.1.5 estimate the mean and total over subpopulation

10.2 Brief outline of the session

10.2.1 Double Sampling for Non-response
- 10.2.1.1 non-sampling error and double sampling
- 10.2.1.2 selecting the number of call backs
- 10.2.1.3 variance of the mixture distribution
- 10.2.1.4 designing surveys to reduce non-response

10.2.2 Interpenetrating Subsample
- 10.2.2.1 interpenetrating subsamples
- 10.2.2.2 example

10.2.3 Estimation of means and totals over subpopulation
- 10.2.3.1 estimation of means and totals over subpopulation
- 10.2.3.2 example

11.0 Session 10 (Day 5: 1300 – 1600hrs)

11.1 Session Description

In this session participants are expected to gain a good practical experience of the theories learned in this course. Demonstration is based on MS Excel but participants are free use any advanced statistical package like SPSS Statistics.

11.2 Session Learning Outcomes

By the end of this session participants will be able to;

11.2.1 gain a good experience to put theory into practice

11.3 Brief outline of the session

11.3.1 Adding-in the data analysis tool pack to excel
11.3.2 Creating Random Numbers in Excel 2007
11.3.3 Random Number Generator
11.3.4 Sampling

Material for further reading and useful web links: *Sampling* by Steven Thompson, 2nd edition.

TESTS

(Note: Participants are not supposed to see the reply first.)

12.0 Model MCQ Paper

12.1 Which of the following statements are true?

I. A sample survey is an example of an experimental study.
II. An observational study requires fewer resources than an experiment.
III. The best method for investigating causal relationships is an observational study.

(A) I only
(B) II only
(C) III only
(D) All of the above.
(E) None of the above.

12.2 An auto analyst is conducting a satisfaction survey, sampling from a list of 10,000 new car buyers. The list includes 2,500 Ford buyers, 2,500 GM buyers, 2,500 Honda buyers, and 2,500 Toyota buyers. The analyst selects a sample of 400 car buyers, by randomly sampling 100 buyers of each brand.

Is this an example of a simple random sample?

(A)Yes, because each buyer in the sample was randomly sampled.
(B)Yes, because each buyer in the sample had an equal chance of being sampled.
(C)Yes, because car buyers of every brand were equally represented in the sample.
(D) No, because every possible 400-buyer sample did not have an equal chance of being chosen.
(E) No, because the population consisted of purchasers of four different brands of car.

12.3 Which of the following statements are true?

I. Random sampling is a good way to reduce response bias.
II. To guard against bias from undercoverage, use a convenience sample.
III. Increasing the sample size tends to reduce survey bias.
IV. To guard against nonresponse bias, use a mail-in survey.

(A)I only
(B) II only
(C) III only
(D) IV only
(E) None of the above.

12.4 Which of the following statements are true?

I. Random sampling is a good way to reduce response bias.
II. To guard against bias from under coverage, use a convenience sample.
III. Increasing the sample size tends to reduce survey bias.
IV. To guard against non-response bias, use a mail-in survey.

(A) I only
(B) II only
(C) III only
(D) IV only
(E) None of the above.

12.5 . Which of the following is an example of random sampling techniques?

(a) Taking the name of every person in a telephone book

(b) Generating a list of numbers by picking numbers out of a hat and matching these numbers to names in the telephone book

(c) Taking every tenth or twentieth name from a list of everybody in the telephone book

12.6 . If the people whose views you need are, for example, all under 50 years old, both men and women, and all have children under 11, then the interviewers will be asked to find and interview people of the same type. When they have finished interviewing you will have [a sample] – a set of [respondents] – all of whom are under 50, half of whom are women, and all of whom have children under 11. It will be a small simmering-down, [a cross-section], of all the people you're interested in.

What kind of sampling does this example use?

(a) Random sampling

(b) Systematic sampling

(c) Quota sampling

12.7. When every member of the accessible population has an equal chance of being selected to participate in the study, the researcher is using

A. Sample.
B. accessible population
C. Target population.
D. World.

12.8. When every member of the accessible population has an equal chance of being selected to participate in the study, the researcher is using

A. Simple random sampling.
B. Stratified random sampling.
C. Convenience sampling.
D. Purposive sampling.

12.9. If a researcher selected five schools at random and then interviewed each of the teachers in those five schools, the researcher used

A. Simple random sampling.
B. Stratified random sampling.
C. Cluster random sampling.
D. Two-stage random sampling.

12.10. Which of the following is an example of a random sampling method?

A. systematic sampling
B. convenience sampling
C. purposive sampling
D. cluster random

12.11. The best sample is one that is

A. A systematic sample.
B. Convenient.
C. Representative of the population.
D. Purposefully selected.

Questions 12.12 to 12.14 refer to the following research situation:
A researcher who wanted to determine the benefits of using a new beginning algebra study technique obtained permission from a school district to select 50 high school students. The researcher selected 50 beginning algebra students at random. The researcher selected 25 of these 50 students to participate in the new study program.

12.12. The researcher gave a training session on traditional study techniques to the other 25 students and asked them to use these methods.
The most likely target population in this study is

A. Algebra students in the district.
B. All students in the district.
C. All algebra students.
D. The 25 students who learned the new study techniques.

12.13. The method of sampling used in the study is

A. Simple random sampling.
B. Stratified random sampling.
C. Cluster sampling.
D. Convenience sampling.

12.14. The greatest threat to external validity in this study is

A. The division of the sample into two groups of 25.
B. The use of only 50 students in the sample.
C. The use of students from only one district.
D. The use of only two different study techniques.

12.15. The purpose of stratified random sampling is to make certain that

A. Every member of the population has an equal chance of being selected for the sample.
B. The sample proportionately represents individuals from different categories of the population.
C. The participants chosen for the study are the ones most likely to react to the treatment.
D. The sample is more representative of the target population than the accessible population.

12.16. Population generalizability refers to

A. Conclusions researchers make about a random sample.
B. Conclusions researchers make about information uncovered in research study.
C. The degree to which a sample represents the population of interest.
D. The degree to which results of a study can be extended to other settings or conditions.

12.17. The degree to which results of a study can be extended to other settings or conditions describes

A. Population generalizability.
B. Conclusions researchers make about a random sample.
C. Conclusions researchers make about information uncovered in research study.
D. Ecological generalizability.

12.18. Which of the following characteristics suggests using a qualitative sampling technique?

A. The intense level of interaction between the research and participants
B. The desired depth of information required to understand the phenomenon of interest
C. The long timeframe for conducting the research
D. All of these

12.19. Which type of sampling strategy is exemplified by selecting two types of

Individuals—those who are extremely happy and those who are extremely sad.

A. Snowball
B. Intensity
C. Homogeneous
D. Purposive

12.20. Which of the following characteristics clearly differentiates probability and purposive sampling?

A. Sample sizes are typically the same.
B. Probability sampling starts with a defined population and selects a sample from it, while non-probability sampling starts with a sample and defines the population relative to the characteristic of that sample.
C. Probability sampling makes it difficult to generalize from the sample to populations, while non-probability sampling makes it easy to do so.
D. More than one technique can be used to select a sample with either approach.

12.21. One of these questions showed that 25% of the population favored a 7-day waiting period between application for purchase of a handgun and the resulting sale, while the other question showed that 70% of the population favored the waiting period. Which produced which result and why?

(A) The: first question probably showed 70% and the second question 25% because of the lack of randomization in the choice of pro-gun and anti-gun subjects as evidenced by the wording of the questions.

(B) The first question probably showed 25% and the second question 70% because of a placebo effect due to the wording of the questions.

(C) The first question probably showed 70% and the second question 25% because of the lack of a control group.

(D) The first question probably showed 25% and the second question 70% because of response bias due to the wording of the question.

(E) The first question probably showed 70% and the second question 25% because of response bias due to the wording of the question.

12.22. Ann Landers, who wrote a daily advice column appearing in newspapers across the country, once asked her readers, "If you had it to do over again, would you have children?" Of the more than 10,000 readers who responded, 70% said no. What does this show?

(A) The survey is meaningless because of voluntary response bias.

(B) No meaningful conclusion is possible without knowing something more about the characteristics of her readers.

(C) The survey would have been more meaningful if she had picked a random sample of the 10,000 readers who responded.

(D) The survey would have been more meaningful if she had used a control group.

(E) This was a legitimate sample, randomly drawn from her readers and of sufficient size to allow

the conclusion that most of her readers who are parents would have second thoughts about having children.

12.23. To find the average occupancy size of student-rented apartments, a researcher picks a simple random sample of 100 such apartments. Even after one follow-up visit, the interviewer is unable to make contact with anyone in 27 of these apartments. Concerned about no response bias, the researcher chooses another simple random sample and instructs the interviewer to continue this procedure until contact is made with someone in a total of 100 apartments. The average occupancy size in the final 100-apartment sample is 2.78. Is this estimate probably too high or too low?

(A) Too low, because of undercoverage bias.

(B) Too low, because convenience samples overestimate average results.

(C) Too high, because of undercoverage bias.

(D) Too high, because convenience samples overestimate average results.

(E) Too high, because voluntary response samples overestimate average results.

12.24. Consider the following three events:

I. Although 18% of the student body are minorities, in a random sample of 20 students, 5 are minorities.

II. In a survey about sexual habits, an embarrassed student deliberately gives the wrong answers.

III. A surveyor mistakenly records answers to one question in the wrong space.

Which of the following correctly characterizes the above?

(A) I, sampling error; Il, response bias; ill, human mistake

(B) I, sampling error; n, non-response bias; ill, hidden error

(C) I, hidden bias; n, voluntary sample bias; ill, sampling error

(D) I, under-coverage error; n, voluntary error; ill, unintentional error

(E) I, small sample error; n, deliberate error; ill, mistaken error

12.25. A researcher plans a study to examine the depth of belief in God among the adult population. He obtains a simple random sample of 100 adults as they leave church one Sunday morning. All but one of them agrees to participate in the survey. Which of the following are true statements?

I. Proper use of chance as evidenced by the simple random sample makes this a well designed Survey.

II The high response rate makes this a well-designed survey.

Ill. Selection bias makes this a poorly designed survey.

(A) I only

(B) II only

(C) III only

(D) I and II

(E) None of these statements is true.