ACE 261
Fall 2002
Prof. Katchova

Lecture 7

Sampling and Sampling Distributions

Sampling and sampling distributions: Outline

- Simple random sampling
- Point estimation
- Introduction to sampling distributions
- Sampling distribution of $\bar{x}$
- Sampling distribution of $\bar{p}$
- Properties of point estimators
- Other sampling methods

2

Population parameters and sample statistics

- A population is the set of all the elements of interest.
  - Examples: all U.S. college students; all employees in a company
- A parameter is a summary measure for a population.
  μ (mean of population), σ (standard deviation of population), p (proportion of population).
- A sample is a subset of the population.
  - Examples: ACE 261 students; 15 employees of a company
- A statistic is a summary measure for a sample.
  $\bar{}$ (mean of sample), s (standard deviation of sample), $\bar{}$ (proportion of sample).

3

Statistical inference

- With proper sampling methods, the sample statistics will provide "good" estimates of the population parameters.
- The purpose of statistical inference is to obtain information about a population from information contained in a sample.

4

Simple random sampling from a finite population

- A simple random sample from a finite population of size *N* is a sample selected such that each possible sample of size *n* has the same probability of being selected.
  - Sampling with replacement – once an element has been included in the sample, it is returned to the population and can be selected again. This method is not used often.
  - Sampling without replacement – once an element has been included in the sample, it is removed from the population and cannot be selected a second time.
- In large sampling projects, computer-generated random numbers are often used to automate the sample selection process.

5

Simple random sampling from infinite population

- The population is usually considered infinite if it involves an ongoing process that makes listing or counting every element impossible.
  - Example: customers shopping at Walmart
- A simple random sample from an infinite population is a sample selected such that the following conditions are satisfied.
  - Each element selected comes from the same population.
  - Each element is selected independently.
- The random number selection procedure cannot be used for infinite populations.

6

Point estimation

- In point estimation we use the data from the sample to compute a value of a sample statistic that serves as an estimate of a population parameter.
- We refer to $\bar{x}$ as the point estimator of the population mean μ.
- s is the point estimator of the population standard deviation σ.
- $\bar{p}$ is the point estimator of the population proportion p.

7

Sampling error

- Sampling error is the absolute difference between an unbiased point estimate and the corresponding population parameter.
- Sampling error is the result of using a subset of the population (the sample), and not the entire population to develop estimates.
- The sampling errors are:
  - $|\bar{x} - \mu|$ for sample mean
  - $|s - \sigma|$ for sample standard deviation
  - $|\bar{p} - p|$ for sample proportion

8

Sampling distribution example

- A department at the U of I receives 900 applications annually from prospective students.
- The application forms contain a variety of information including the individual's SAT score and whether or not the individual desires on-campus housing.
- The director of admissions would like to know the following information:
  - the average SAT score for the applicants, and
  - the proportion of applicants that want to live on campus.
- There are 3 alternatives for obtaining the desired information:
  - Using the entire population of 900 applicants
  - Selecting a sample of 30 applicants, using computer-generated random numbers
  - Selecting a sample of 30 applicants, using a random number table (this approach is not used often)

9

Calculating population parameters

- Using the population of 900 applicants
  - SAT Scores
    - Population Mean

$$\mu = \frac{\sum x_i}{900} = 990$$

    - Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{900}} = 80$$

  - Applicants Wanting On-Campus Housing
    - Population Proportion

$$p = \frac{648}{900} = .72$$

10

Creating a random sample

- Take a sample of 30 applicants using computer-generated random numbers
  - Excel provides a function, rand(), for generating random numbers in its worksheet.
  - 900 random numbers are generated, one for each applicant in the population.
  - Then we choose the 30 applicants corresponding to the 30 smallest random numbers as our sample.
  - Each of the 900 applicants have the same probability of being included.

11

Raw data

Applicant Number	SAT Score	On-Campus Housing	Random Number
1	1008	Yes	0.25503
2	1025	No	0.58945
3	952	Yes	0.15103
4	1090	Yes	0.54278
5	1127	Yes	0.86620
6	1015	No	0.89359
7	965	Yes	0.42131
8	1161	No	0.83353

Note: Rows 9 – 900 are not shown.

12

Randomly selected sample

Applicant Number	SAT Score	On-Campus Housing	Random Number
12	1107	No	0.00027
773	1043	Yes	0.00192
408	991	Yes	0.00303
58	1008	No	0.00481
116	1127	Yes	0.00538
185	982	Yes	0.00583
510	1163	Yes	0.00649
394	1008	No	0.00667

Note: Rows 9 – 900 are not shown.

13

Calculating sample statistics

- Point estimates
  - $\bar{x}$ as point estimator of μ

$$\bar{x} = \frac{\sum x_i}{30} = \frac{29{,}910}{30} = 997$$

  - s as point estimator of σ

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{29}} = \sqrt{\frac{163{,}996}{29}} = 75.2$$

  - $\bar{p}$ as point estimator of p

$$\bar{p} = 20/30 = .68$$

14

Point estimators

- If we do not use the population, we can select a sample, calculate different sample statistics and use these as estimates for population parameters.
- But there is a problem!
  - Unfortunately, there can be many randomly selected samples from the same population.
  - Every sample will produce different estimates of the population parameters.

15

Sampling distribution of $\bar{x}$

- The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the sample mean $\bar{x}$.
- Expected Value of $\bar{x}$

$$E(\bar{x}) = \mu$$

where:

μ = the population mean

16

Sampling distribution of $\bar{x}$

- Standard Deviation of $\bar{x}$
- Finite population

$$\sigma_{\bar{x}} = \left(\frac{\sigma}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}}$$

Infinite population

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

  - A finite population is treated as being infinite if n/N < .05.
  - $\sqrt{\frac{N-n}{N-1}}$ is the finite correction factor.
  - $\sigma_{\bar{x}}$ is referred to as the standard error of the mean.

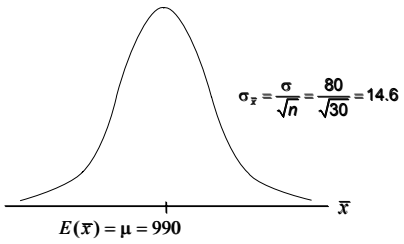
17

Sampling distribution of $\bar{x}$

- The central limit theorem enables us to conclude that the sampling distribution of $\bar{x}$ can be approximated by a normal probability distribution if we use a large ($n \geq 30$) simple random sample.
- When the simple random sample is small ($n < 30$), the sampling distribution of $\bar{x}$ can be considered normal only if we assume the population has a normal probability distribution.

18

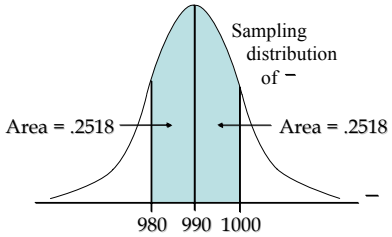
Sampling distribution of $\bar{x}$



Sampling distribution of $\bar{x}$

- What is the probability that a simple random sample of 30 applicants will provide an estimate of the population mean SAT score that is within plus or minus 10 of the actual population mean μ ?
- In other words, what is the probability that $\bar{x}$ will be between 980 and 1000?

Sampling distribution of $\bar{x}$



- Using the standard normal probability table with z = 10/14.6= .68, we have area = (.2518)(2) = .5036

Sampling distribution of $\bar{p}$

- The sampling distribution of $\bar{p}$ is the probability distribution of all possible values of the sample proportion $\bar{p}$
- Expected value of $\bar{p}$

$$E(\bar{p}) = p$$

where:

p = the population proportion

Sampling distribution of $\bar{p}$

- Standard deviation of $\bar{p}$
- Finite Population

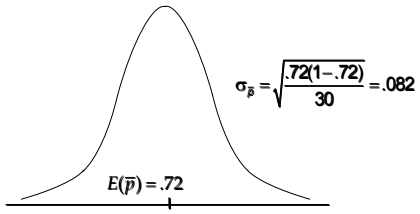
$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Infinite Population

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

– $\sigma_{\bar{p}}$ is referred to as the standard error of the proportion.

Sampling distribution of $\bar{p}$



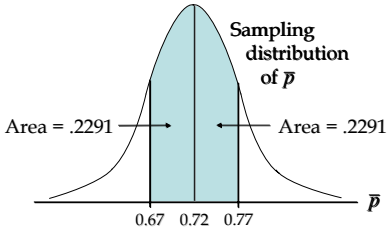
- The normal probability distribution is an acceptable approximation since np = 30(.72) = 21.6 $\geq$ 5 and $n(1 - p)$ = 30(.28) = 8.4 $\geq$ 5.

Sampling distribution of $\bar{p}$

- What is the probability that a simple random sample of 30 applicants will provide an estimate of the population proportion of applicants desiring on-campus housing that is within plus or minus .05 of the actual population proportion?
- In other words, what is the probability that $\bar{p}$ will be between .67 and .77?

25

Sampling distribution of $\bar{p}$



- For z = .05/.082 = .61, the area = (.2291)(2) = .4582. The probability is .4582 that the sample proportion will be within +/-.05 of the actual population proportion.

26

Properties of Point Estimators

- Before using a sample statistic as a point estimator, statisticians check to see whether the sample statistic has the following properties associated with good point estimators.
  - Unbiasedness
  - Efficiency
  - Consistency

27

Properties of Point Estimators: Unbiasedness

- If the expected value of the sample statistic is equal to the population parameter being estimated, the sample statistic is said to be an unbiased estimator of the population parameter.

28

Properties of Point Estimators: Efficiency

- Given the choice of two unbiased estimators of the same population parameter, we would prefer to use the point estimator with the smaller standard deviation, since it tends to provide estimates closer to the population parameter.
- The point estimator with the smaller standard deviation is said to have greater relative efficiency than the other.

29

Properties of Point Estimators: Consistency

- A point estimator is consistent if the values of the point estimator tend to become closer to the population parameter as the sample size becomes larger.

30

Other Sampling Methods

- Stratified random sampling
- Cluster sampling
- Systematic sampling
- Convenience sampling
- Judgment sampling

31

Stratified Random Sampling

- Example: The basis for forming the strata might be department, location, age, industry type, etc.
- USDA's farm data is based on a stratified random sampling.
- The population is first divided into groups of elements called strata. The elements within each stratum are as much alike as possible (i.e. homogeneous group).
- A simple random sample is taken from each stratum. Formulas are available for combining the stratum sample results into one population parameter estimate.
- Advantage: Allows for a smaller total sample size.

32

Cluster Sampling

- Example: A primary application is area sampling, where clusters are city blocks or other well-defined areas.
- The population is first divided into separate groups of elements called clusters. Each cluster is a representative small-scale version of the population (i.e. heterogeneous group).
- All elements within each sampled (chosen) cluster form the sample.
- Advantage: The close proximity of elements can be cost effective (i.e. many sample observations can be obtained in a short time).
- Disadvantage: This method generally requires a larger total sample size than simple or stratified random sampling.

33

Systematic Sampling

- Example: Selecting every 100^{th} listing in a telephone book after the first randomly selected listing.
- This method has the properties of a simple random sample, especially if the list of the population elements is a random ordering.
- Advantage: The sample usually will be easier to identify than it would be if simple random sampling were used.

34

Convenience Sampling

- Example: A professor conducting research might use student volunteers to constitute a sample.
- The sample is identified primarily by convenience.
- It is a nonprobability sampling technique. Items are included in the sample without known probabilities of being selected.
- Advantage: Sample selection and data collection are relatively easy.
- Disadvantage: It is impossible to determine how representative of the population the sample is.

35

Judgment Sampling

- Example: A reporter might sample three or four senators, judging them as reflecting the general opinion of the senate.
- The person most knowledgeable on the subject of the study selects elements of the population that he or she feels are most representative of the population.
- It is a nonprobability sampling technique.
- Advantage: It is a relatively easy way of selecting a sample.
- Disadvantage: The quality of the sample results depends on the judgment of the person selecting the sample.

36