Lecture 6

The Sampling Distribution

Section A: The Sampling Distribution of a Sample Statistic: Definition

2

Uncertainty in Sample Based Estimates

- Thus far in the course, we have summary measures for single data samples (means, proportions, incidence rates), and measures of association comparing two sample (differences in means, risk differences, relative risks, incidence rate ratios)
- We have discussed how these aforementioned sample estimates are not necessarily the population "truth", but our best estimate of some unknown truth based on an imperfect sample from a population(s)

3

Uncertainty in Sample Based Estimates

- Ultimately, it is important to recognize the potential uncertainty in a sample based estimate as it relates to the unknown truth it is estimate
- Understanding sampled based estimates vary across random samples of the same size, from the same population, will give a framework for coupling the estimate with some measure of uncertainty to make a statement about the unknown truth

4

Uncertainty in Sample Based Estimates

- This set of lectures involves defining, characterizing and estimating the theoretical sampling distribution of a sample statistic (for example, a sample mean, proportion, or incidence rate)
- Ultimately, this sampling distribution will allow for the estimation of an interval describing a plausible range of values for the unknown truth that we can only estimate, using the results from a single random sample

5

Uncertainty in Sample Based Estimates

- What is meant by uncertainty in sample based estimates, also called *sampling variability*?
- Example 1: Height distribution for one year olds in Nepal

6

Uncertainty in Sample Based Estimates

- What is meant by uncertainty in sample based estimates, also called *sampling variability*?
- Example 2: Baltimore mayoral election votes

7

Uncertainty in Sample Based Estimates

- How can the definition of *sampling variability* be formalized ?
- The sampling distribution of a sample statistic provides the answer.

8

Sampling Distribution of a Sample Statistic

- The sampling distribution of a sample statistic is a theoretical distribution, that describes all possible values of a sample statistic from random samples of the same size, taken from the same population.

9

Sampling Distribution of a Sample Statistic

- Example 1: the sampling distribution of sample mean heights of random samples of 50 Nepali children who are 12 months old

10

Sampling Distribution of a Sample Statistic

- Example 2: the sampling distribution of the sample proportion voting for Candidate A from random samples of 100 Baltimore City residents

11

Sampling Distribution of a Sample Statistic

- The sampling distribution is a theoretical entity: it cannot be observed directly, or exactly specified
- In "real-life" research only one sample from each population under study will be taken

12

Sampling Distribution of a Sample Statistic

- Lecture Sections C-D will serve to:
  - Further demonstrate and define sampling distributions by detailing the results of some computer simulations
  - Empirically show some consistent properties of sampling distributions, regardless of the sample statistics (mean, proportion, incidence rate)
  - Unveil a mathematical property that will allow for the generalization of these properties
  - Demonstrate how to estimate a sampling distribution for a sample statistics based on the results of one random sample

13

Section B: Examples, Sampling Distribution of the Sample Mean

14

Learning Objectives

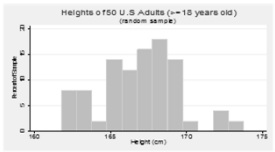
- Upon completion of this lecture section, you should be able:
  - Describe the sampling distribution of a sample mean in terms of it's composition
  - Comment on some characteristics of the sampling distributions for sample means demonstrated empirically, by simulations, including
    - General shape of the distributions
    - Center of the distributions
    - Variability of the distributions, and the relationship to the size of the samples each mean is based upon

15

Example 1: Mean Heights

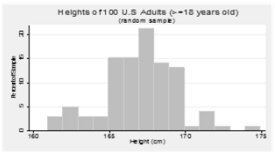
- Population #1: Height Measurements for Adults >= 18 years

Sample A: n=50



$\bar{x}$ = 166 .9 cm ; s = 2.6 cm

Sample B: n=100

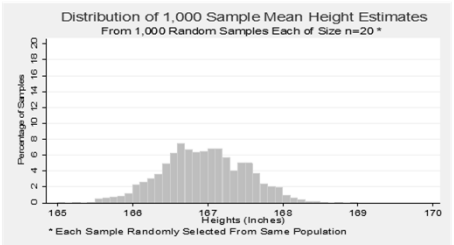


$\bar{x}$ = 167 .1 cm ; s = 2.4 cm

16

Example 1: Mean Heights

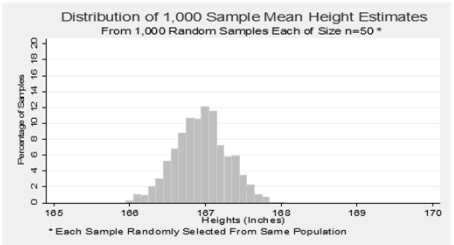
- Estimated sampling distribution: sample means from random samples of size n=20



17

Example 1: Mean Heights

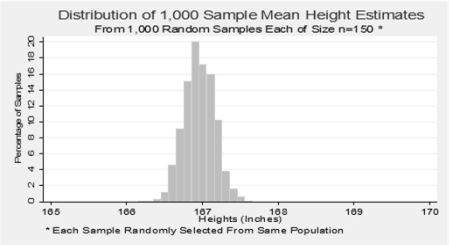
- Estimated sampling distribution: sample means from random samples of size n=50



18

Example 1: Mean Heights

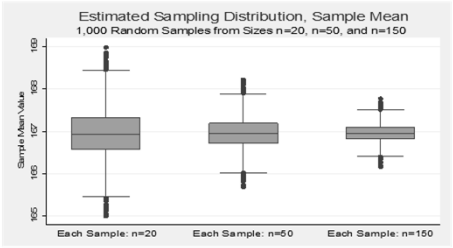
- Estimated sampling distribution: sample means from random samples of size n=150



19

Example 1: Mean Heights

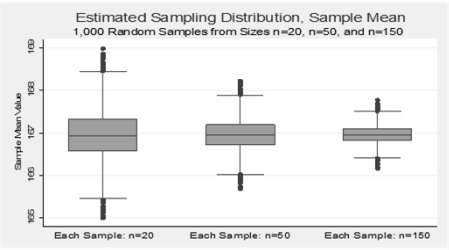
- All on one graphic



20

Example 1: Mean Heights

- What do you notice?



21

Example 1: Mean Heights

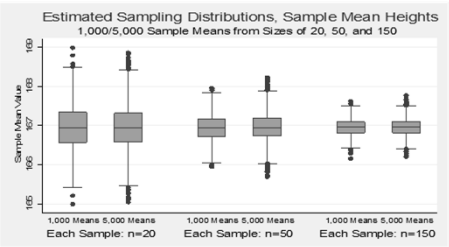
- And the truth is:
- Results from our estimated sampling distributions

Sample Size	Mean of 1,000 Sample Means	SD of 1,000 Sample Means *
n=20	167 cm	0.56 cm
n=50	167 cm	0.35 cm
n=150	167 cm	0.20 cm

22

Example 1: Mean Heights

- FYI: variation in sample means depends on size of each sample, not number of samples taken in the simulation

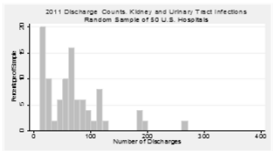


23

Example 2: Mean Discharges

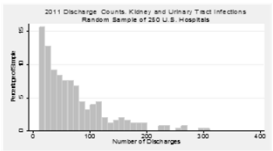
- Population #1: Hospitals in US in 2011: Discharges for Kidney and Urinary Infections[1]

Sample A: n=50



$\bar{x} = 69.1; s = 53.6$

Sample B: n=250



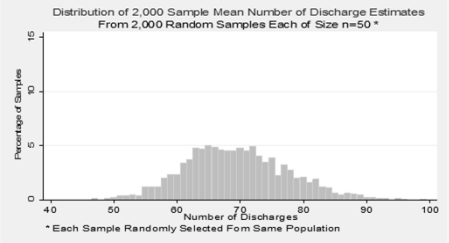
$\bar{x} = 71.7; s = 58.2$

[1] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/index.html

24

Example 2: Mean Discharges

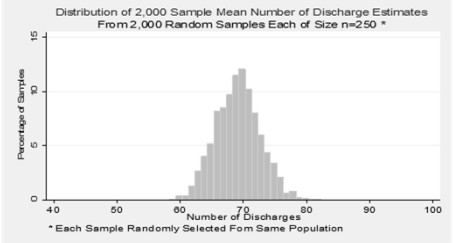
- Estimated sampling distribution: sample means from random samples of size n=50



25

Example 2: Mean Discharges

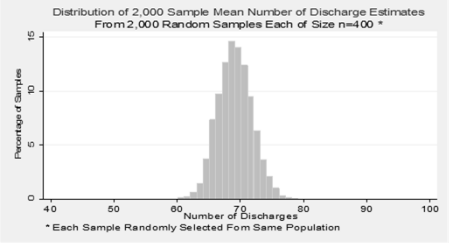
- Estimated sampling distribution: sample means from random samples of size n=250



26

Example 2: Mean Discharges

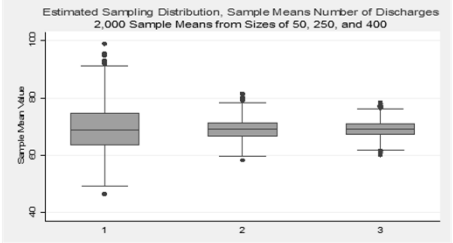
- Estimated sampling distribution: sample means from random samples of size n=400



27

Example 2: Mean Discharges

- All on one graphic



28

Example 2: Mean Discharges

- Results

29

Example 2: Mean Discharges

- And the truth is:
- Results from our estimated sampling distributions

Sample Size	Mean of 2,000 Sample Means	SD of 2,000 Sample Means
n=50	69.3	8.1
n=250	69.2	3.5
n=400	69.2	2.7

30

Summary

- Theoretical sampling distributions for sample means across random samples of the same size, from the same population, can be estimated via computer simulation
- Simulation is as useful tool for helping explore the properties of these sampling distributions. Some properties observed with the two examples in this lecture, which will be generalized shortly include:

31

Summary

- Ultimately, estimating the characteristics of a sampling distribution will be done using the results from a single random sample from a population. In lecture section D, these properties that have been demonstrated empirically via the simulations in this lecture set will be generalized.

32

Section C: Examples, Sampling Distribution of Sample Proportions and Sample Incidence Rates

33

Learning Objectives

- Upon completion of this lecture section, you should be able:
  - Describe the sampling distribution of a sample proportion, and a sample incidence rates in terms of their compositions
  - Comment on some characteristics of the sampling distributions for proportions and incidence rates, demonstrated empirically, by simulations, including
    - General shape of the distributions
    - Center of the distributions
    - Variability of the distributions, and the relationship to the size of the samples each statistic is based upon

34

Learning Objectives

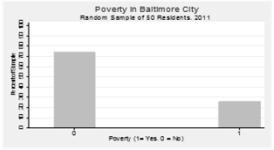
- Upon completion of this lecture section, you should be able:
  - Comment on the similarities between this lecture section's results, and the results for sampling distributions of sample means from he previous section

35

Example 1: Poverty in Baltimore

- Population #1: Baltimore City Residents

Sample A: n=50



$\hat{p} = 0.26$

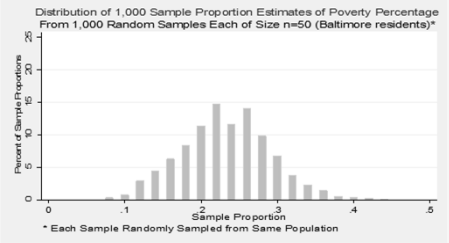
Sample B: n=100

$\hat{p} = 0.19$

36

Example 1: Baltimore City Poverty

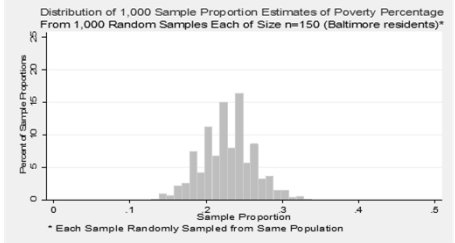
- Estimated sampling distribution: sample proportions from random samples of size n=50



37

Example 1: Baltimore City Poverty

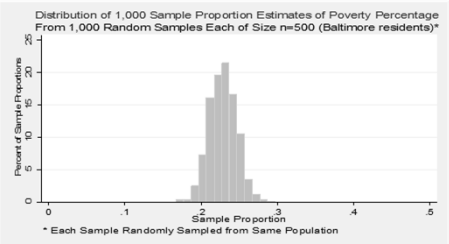
- Estimated sampling distribution: sample proportions from random samples of size n=150



38

Example 1: Baltimore City Poverty

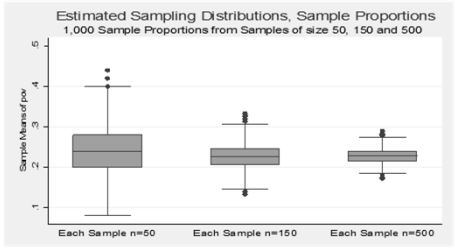
- Estimated sampling distribution: sample proportions from random samples of size n=500



39

Example 1: Baltimore City Poverty

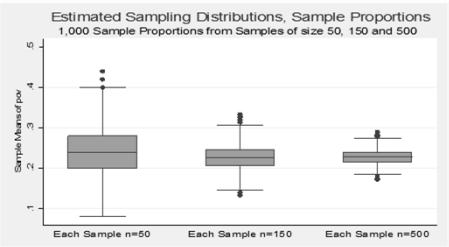
- All on one graphic



40

Example 1: Baltimore City Poverty

- What do you notice?



41

Example 1: Baltimore City Poverty

- Results

42

Example 1: Baltimore City Poverty (proportion)

- And the truth is:
- Results from our estimated sampling distributions

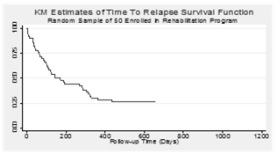
Sample Size	Mean of 1,000 Sample Proportions	SD of 1,000 Sample Proportions
n=50	0.232 (23.2%)	0.58 (5.8%)
n=150	0.229 (22.9 %)	0.34 (3.4 %)
n=500	0.229 (22.9 %)	0.18 (1.8 %)

43

Example 2: Time to Relapse (incidence rate)

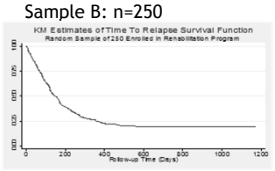
- Population #1: Substance Abusers Who Enrolled in Rehabilitation

Sample A: n=50



$\hat{IR}$ = 108 .6 events/100 person years

Sample B: n=250

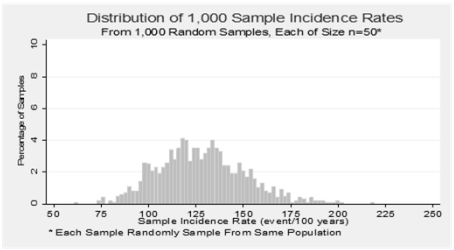


$\hat{IR}$ = 129 .1 events/100 person years

44

Example 2: Time to Relapse

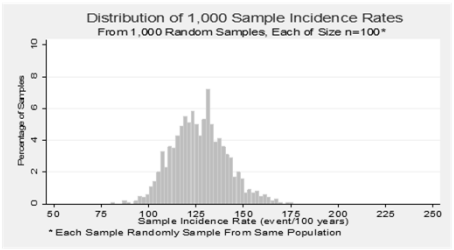
- Estimated sampling distribution: sample incidence rates from random samples of size n=50



45

Example 2: Time to Relapse

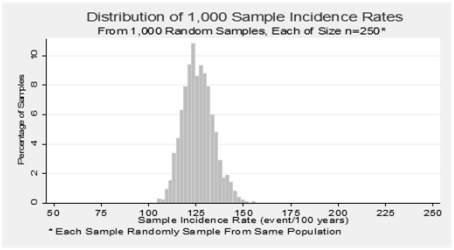
- Estimated sampling distribution: sample incidence rates from random samples of size n=100



46

Example 2: Time to Relapse

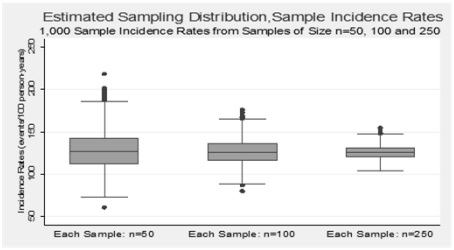
- Estimated sampling distribution: sample incidence rates from random samples of size n=250



47

Example 2: Time to Relapse

- All on one graphic



48

Example 2: Time to Relapse

- Results

49

Example 2: Time to Relapse

- And the truth is:
- Results from our estimated sampling distributions

Sample Size	Mean of 1,000 Sample Rates	SD of 1,000 Sample Rates
n=50	128.0	22.9
n=100	126.9	14.3
n=250	126.2	7.9

50

Summary

- Theoretical sampling distributions for sample proportions (and incidence rates) across random samples of the same size, from the same population, can be estimated via computer simulation
- Simulation is as useful tool for helping explore the properties of these sampling distributions. Some properties observed with the two examples in this lecture, which will be generalized shortly include:

51

Summary

- Ultimately, estimating the characteristics of a sampling distribution will be done using the results from a single random sample from a population. In lecture section D, these properties that have been demonstrated empirically via the simulations in this lecture set, and lecture set B will be generalized.

52

Section D: Estimating the Sampling Distribution From a Single Sample of Data

53

Learning Objectives

- Upon completion of this lecture section, you should be able to:
  - Explain the Central Limit Theorem (CLT), with regards to the properties of theoretical sampling distributions
  - Estimate the variability in the sampling distribution for sample means and proportions using the results from a single random sample
  - Begin to appreciate how an estimated sampling distribution can allow for the incorporation of sampling variability into the estimated sample statistic

54

Real Life Research

- In Sections B and C, we showed the results of computer simulation to illustrate some general properties of sampling distributions
- In real life research, generally, only one sample can be taken from each population under study
- How can we use the results of the single sample to estimate the "behind the scenes" theoretical sampling distribution of a sample statistic….and how can we use this to help us?

55

Some Common Characteristics

- Some common themes emerged in sections B and C. Regardless of the type of data we were summarizing (continuous, binary, time-to-event) with the appropriate sample statistics (mean, proportion, incidence rate), the resulting sampling distribution:
  - Was generally symmetric, i.e. "approximately" normal regardless of the size of the sample each sample statistics was based upon
  - Was centered (i.e. on average) at the true value of the population level quantity being estimated by the statistic
  - Had variability that systematically decreased the larger the sample each estimate was based upon

56

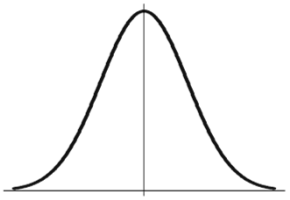
Some Common Characteristics

- There is a mathematical theorem that generalizes these properties called the Central Limit Theorem (CLT).
- Basically, the CLT states that the (theoretical) sampling distribution of a sample statistic will:
  - Be approximately normal
  - Have average of the true, population level value being estimated
  - *Have variability that is a function of the variation of individual values in the population (sd) and the size of the sample the statistic is based upon:* This variability in sample statistics across multiple samples of the same size is called **Standard Error**

57

Some Common Characteristics

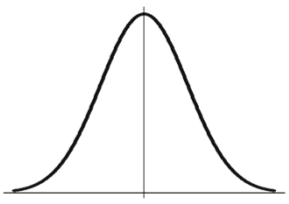
- The CLT : if we were to take multiple random samples of the same sizes from the same population, and look at the distribution of the sample estimates across these multiple samples, it would be:



58

Some Common Characteristics

- Example: sample means based on samples of size n



59

Example 1 : Blood Pressure Data, 113 Men

- Example 1: Systolic blood pressure (SBP) measurements from a random sample of 113 adult men taken from a clinical population

(Estimate of μ): $\bar{x} = 123.6$ mmHg

(Estimate of σ): $s = 12.9$ mmHg

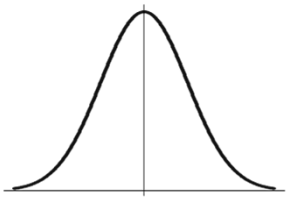
60

Example 2: Length of Stay Data

- Example 2: Length of stay claims at Heritage Health with an inpatient stay of at least one day in 2011 (12,928 claims)

(Estimate of μ): $\bar{x} = 4.3$ days

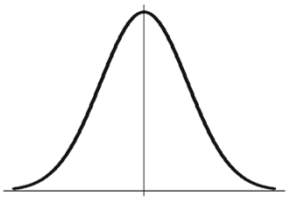
(Estimate of σ): $s = 4.9$ days



61

Some Common Characteristics

- Example: sample proportions based on samples of size n



62

Example 3: Maternal/Infant HIV Transmission

- Results

Results. From April 1991 through December 20, 1993, the cutoff date for the first interim analysis of efficacy, 477 pregnant women were enrolled; during the study period, 409 gave birth to 415 live-born infants. HIV-infection status was known for 363 births (180 in the zido-

$$\hat{p} = \frac{53}{363} \approx 0.15\ (15\%)$$

Of the 363 births whose HIV status was assessed (up to 18 months after birth) , 53 infants were HIV infected.

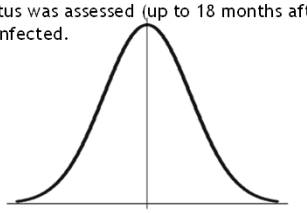
63

Example 3: Maternal/Infant HIV Transmission

- Results

Of the 363 births whose HIV status was assessed (up to 18 months after birth) , 53 infants were HIV infected.

$$\hat{p} = \frac{53}{363} \approx 0.15\ (15\%)$$



64

So How Will This Help Us?

65

Summary

- In "real life" research, only one sample will be taken from each population being studied
- The sampling distribution for the sample summary measure of interest (mean, proportion, or incidence rate) can be estimated coupling the results of the CLT with information from the single sample from a population

66

Summary

- Ultimately, this process will enable the creation of interval that gives a range of possibilities for the unknown, population level, value of the quantity being estimated
- Estimated Standard Error of a Sample Mean estimated from n observations:
- Estimated Standard Error of a Sample Proportion estimated from n observations:

67