Sampling Distributions

Sampling Distributions

- What is a sampling distribution?
- How can information about sampling distributions be used to determine desired information related to population parameters?

Research Process

- Idea for study – Question posed
- Information collected
  - Data
    - Experiment
    - Survey
- Information – organized/analyzed
- Conclusions drawn based on analysis of data

Data

- Sample – use to make inferences about population
  - Sample mean – use to draw conclusions about population mean
    - Dependent upon sample
      - Changes based on values in sample

Data

- Use data to make inferences regarding the population
  - Problems:
    - Samples – not the population
    - Change sample then change
      - Data
      - Statistics
    - Conclusions - variable

Inference

- *A logical process of drawing conclusions from a collection of data and relationships between data and potential conclusions*
- *Assumption based on an observation*
- *A determination arrived at by reasoning; using facts to arrive at a broader conclusion*
- The reasoning involved in drawing a conclusion or making a logical judgment on the basis of circumstantial evidence and prior conclusions rather than on the basis of direct observation
- The act or process of deriving a conclusion based solely on what one already knows

Inferential Statistics

- Based on probability statements

Inferential Statistics

- **Based on probability statements *and information about the related distribution(s)***

Sampling Distribution for the Mean

- General idea for determining sampling distribution for mean $\overline{X}$:
  - Obtain sample of size n
  - Compute sample mean $\overline{X}$
  - Assuming that sampling from finite population, repeat until all possible simple random samples of size n have been obtained

Sampling Distribution for the Mean

- Amounts of money which five people have - HW

Caution about Sample Mean

Caution about Sample Mean

- Dependent upon the sample used
- If use a different sample then may get a different sample mean

To make Probability Statements for Sample Statistics …

- Need to know about the distribution of the *sample mean*

To make Probability Statements for Sample Statistics …

- Need to know about the distribution of the *sample proportion*

To make Probability Statements for Sample Statistics …

- Need to know about the distribution of the *sample statistic* of interest

Sampling Distribution for the Mean

- Distribution for all possible values of the sample mean $\overline{X}$ computed from a sample of size n from a population with mean μ and standard deviation σ.

*

Sampling Distribution for the Mean

- Distribution for all possible values of the sample mean $\overline{X}$ computed from a sample of size n from a population with mean μ and standard deviation σ.

Reminder: In real life, we may not *necessarily* have access to the subjects/units in the population. However, the variable of interest for this population does have a mean and a standard deviation.

Recall: Sampling Distribution for $\overline{X}$

- To determine sampling distribution for mean $\overline{X}$:
  - For samples of size n
  - Compute sample mean $\overline{X}$ for each sample
  - Assuming that sampling from finite population, repeat until all possible simple random samples of size n have been obtained

Nursing Home Data

http://lib.stat.cmu.edu/DASL/Datafiles/nursinghomedat.html

Datafile Name:
Nursing Home Data
Datafile Subjects:
Health , Consumer , Medical , Economics
Story Names:
Nursing Home Data
Reference:
These data are part of the data analyzed in Howard L. Smith, Niell F. Piland, and Nancy Fisher, "A Comparison of Financial Performance, Organizational Character- istics, and Management Strategy Among Rural and Urban Nursing Facilities, Journal of Rural Health, Winter 1992, pp 27-40.
Authorization:
free use
Description:

```
The data were collected by the Department of Health and Social Services of the
State of New Mexico and cover 52 of the 60 licensed nursing facilities in New Mexico
in 1988.
```

Number of cases:
52
Variable Names:

1. BED = number of beds in home
2. MCDAYS = annual medical in-patient days (hundreds)
3. TDAYS = annual total patient days (hundreds)
4. PCREV = annual total patient care revenue ($hundreds)
5. NSAL = annual nursing salaries ($hundreds)
6. FEXP = annual facilities expenditures ($hundreds)
7. RURAL = rural (1) and non-rural (0) homes

Nursing Home Data

- The distribution for the sample mean for any one of the quantitative variables would consist of all possible samples of a sample size of interest.

Nursing Home Data

- The distribution for the sample mean for any one of the quantitative variables would consist of all possible samples of a sample size of interest.
- For sample size 5, there are

$$C(52, 5) = 2{,}598{,}960$$

possible samples.

Nursing Home Data

- The distribution for the sample mean for any one of the quantitative variables would consist of all possible samples of a sample size of interest.
- For sample size 10, there are $C(52, 10) = 1.582002422 \times 10^{10}$ possible samples.

Nursing Home Data

- The distribution for the sample mean for any one of the quantitative variables would consist of all possible samples of a sample size of interest.
- For sample size 20, there are $C(52, 20) \approx 1.259946279 \times 10^{14}$ possible samples.

Nursing Home Data

- The distribution for the sample mean for any one of the quantitative variables would consist of all possible samples of a sample size of interest.
- For sample size 30, there are $C(52, 30) \approx 2.705339196 \times 10^{14}$ possible samples.

Nursing Home Data

- $C(52, 5) = 2{,}598{,}960$
- $C(52, 10) = 1.582002422 \times 10^{10}$
- $C(52, 20) \approx 1.259946279 \times 10^{14}$
- $C(52, 30) \approx 2.705339196 \times 10^{14}$

Nursing Home Data

- $C(52, 5) = 2{,}598{,}960$
- $C(52, 10) = 1.582002422 \times 10^{10}$
- $C(52, 20) \approx 1.259946279 \times 10^{14}$
- $C(52, 30) \approx 2.705339196 \times 10^{14}$
- Question: Could we make all the possible samples????

Nursing Home Data

- $C(52, 5) = 2{,}598{,}960$
- $C(52, 10) = 1.582002422 \times 10^{10}$
- $C(52, 20) \approx 1.259946279 \times 10^{14}$
- $C(52, 30) \approx 2.705339196 \times 10^{14}$
- Question: Would we *want* to make all the possible samples????

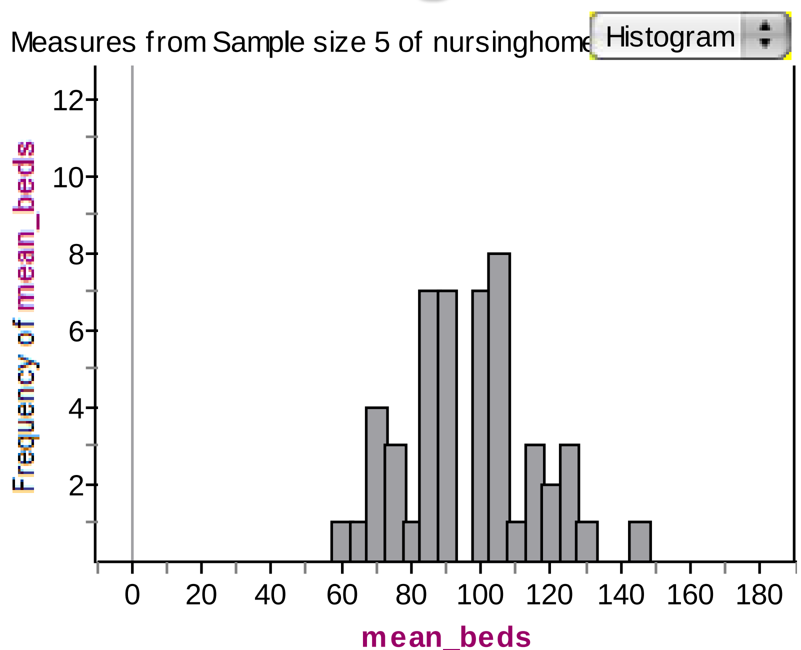
Nursing Home Data

- Keeping in mind that there are $C(52, 5) = 2{,}598{,}960$ possible samples of size 5, consider the following graphs for the distribution for the number of sample means stated and the corresponding average for this distribution of sample means.

Note: The means are provided *without* being rounded.

Nursing Home Data

- Keeping in mind that there are C(52, 5) = 2,598,960 possible samples of size 5, consider the following: 50 samples of size 5.



Measures from Sample size 5 of nursinghomedat

mean_beds	96.664

S1 = mean

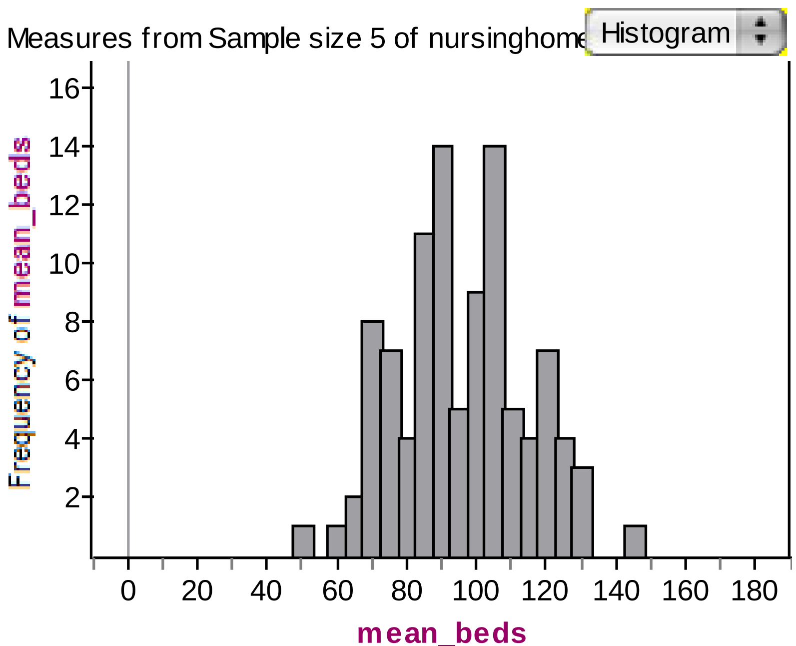
nursinghomedat

BED	93.2692

S1 = mean

Nursing Home Data

- Keeping in mind that there are C(52, 5) = 2,598,960 possible samples of size 5, consider the following: 100 samples of size 5.



Measures from Sample size 5 of nursinghomedat

mean_beds	95.798

S1 = mean

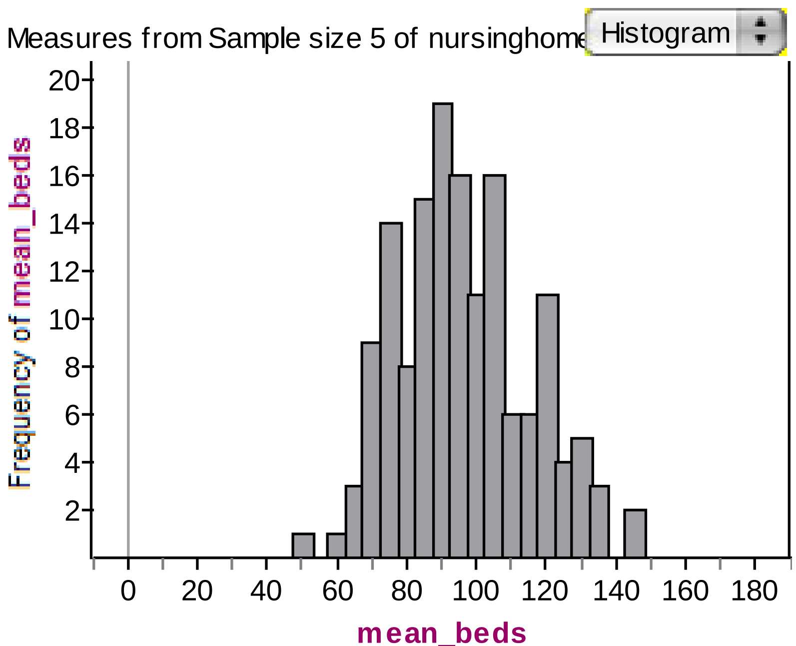
nursinghomedat

BED	93.2692

S1 = mean

Nursing Home Data

- **Keeping in mind that there are C(52, 5) = 2,598,960 possible samples of size 5, consider the following: 150 samples of size 5.**



Measures from Sample size 5 of nursinghomedat

mean_beds	96.164

S1 = mean

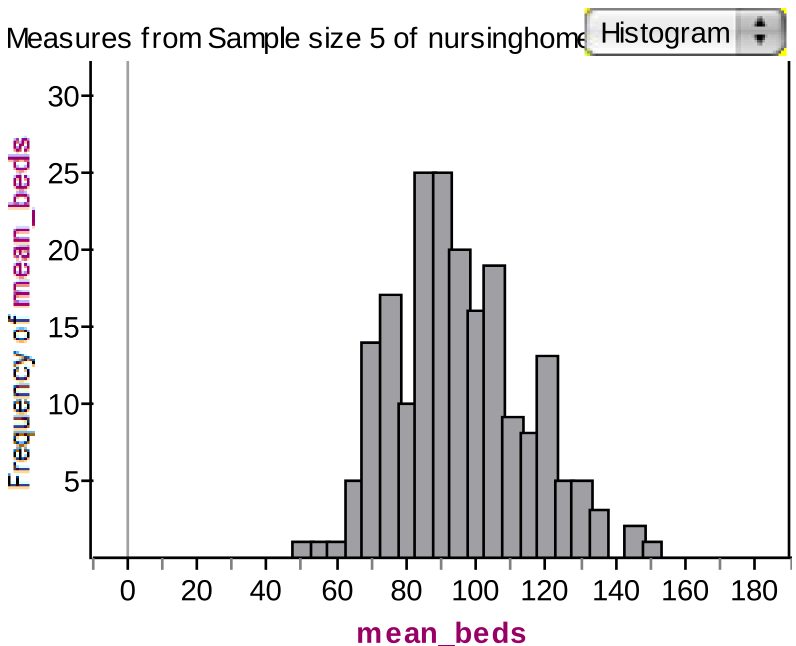
nursinghomedat

BED	93.2692

S1 = mean

Nursing Home Data

- Keeping in mind that there are C(52, 5) = 2,598,960 possible samples of size 5, consider the following: 200 samples of size 5.



Measures from Sample size 5 of nursinghomedat

mean_beds	94.955

S1 = mean

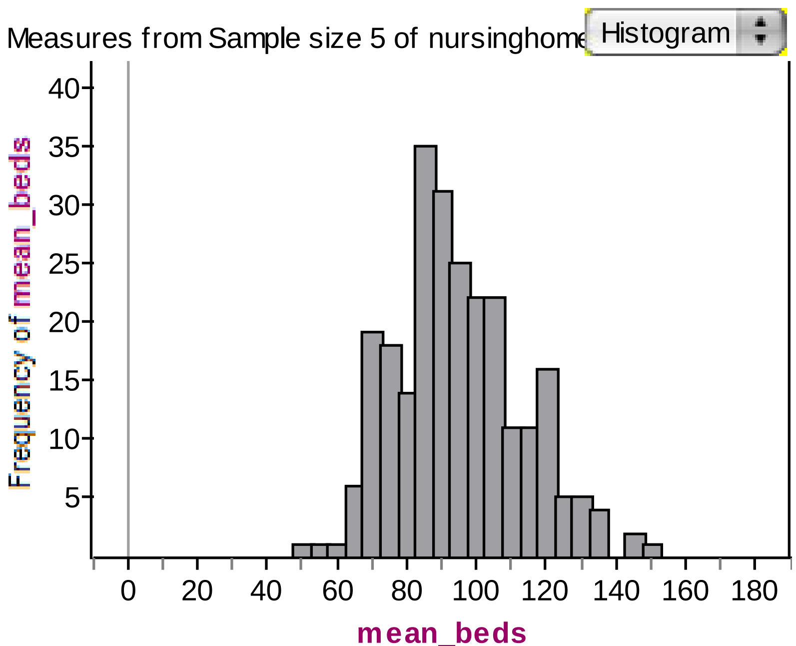
nursinghomedat

BED	93.2692

S1 = mean

Nursing Home Data

- Keeping in mind that there are C(52, 5) = 2,598,960 possible samples of size 5, consider the following: 250 samples of size 5.



Measures from Sample size 5 of nursinghomedat

mean_beds	94.5632

S1 = mean

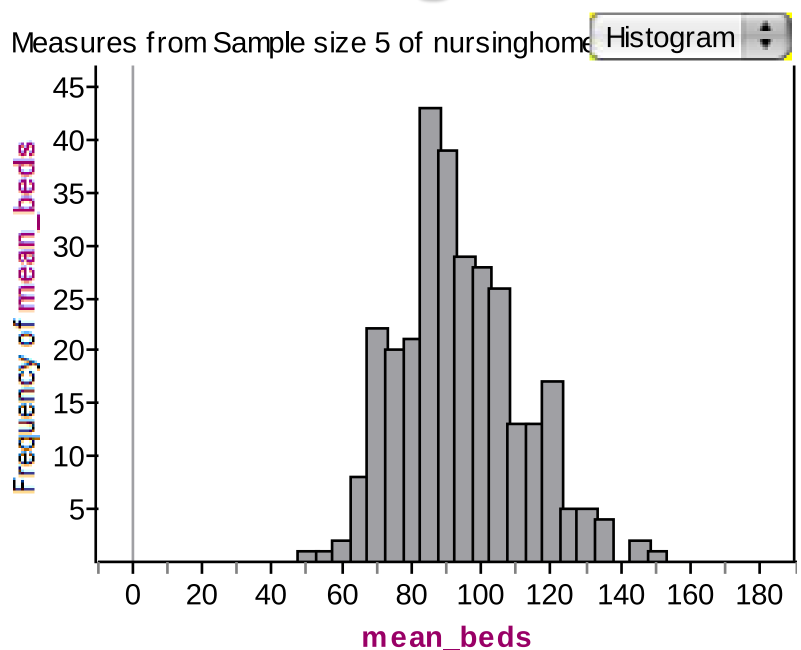
nursinghomedat

BED	93.2692

S1 = mean

Nursing Home Data

- **Keeping in mind that there are C(52, 5) = 2,598,960 possible samples of size 5, consider the following: 300 samples of size 5.**



Measures from Sample size 5 of nursinghomedat

mean_beds	93.7127

S1 = mean

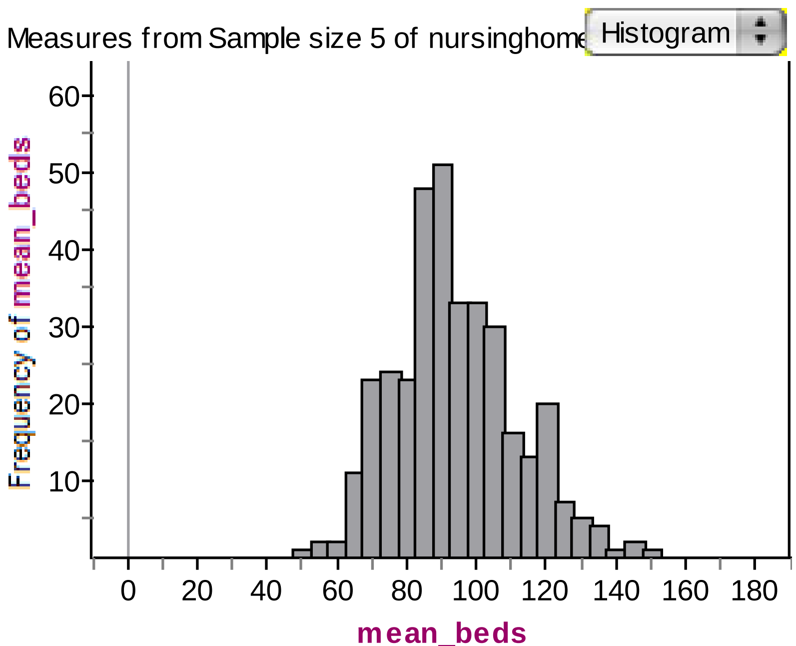
nursinghomedat

BED	93.2692

S1 = mean

Nursing Home Data

- Keeping in mind that there are C(52, 5) = 2,598,960 possible samples of size 5, consider the following: 350 samples of size 5.



Measures from Sample size 5 of nursinghomedat

mean_beds	93.672

S1 = mean

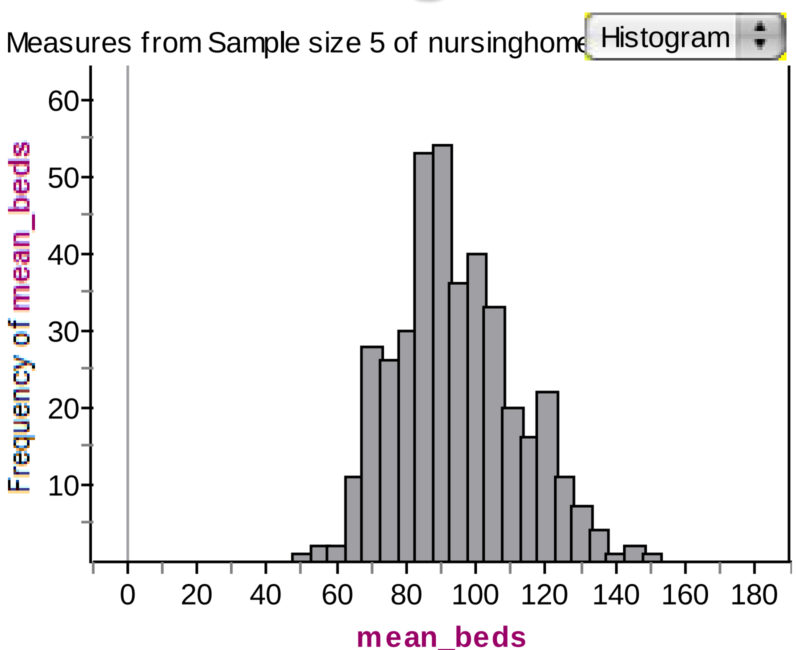
nursinghomedat

BED	93.2692

S1 = mean

Nursing Home Data

- **Keeping in mind that there are C(52, 5) = 2,598,960 possible samples of size 5, consider the following: 400 samples of size 5.**



Measures from Sample size 5 of nursinghomedat

mean_beds	94.038

S1 = mean

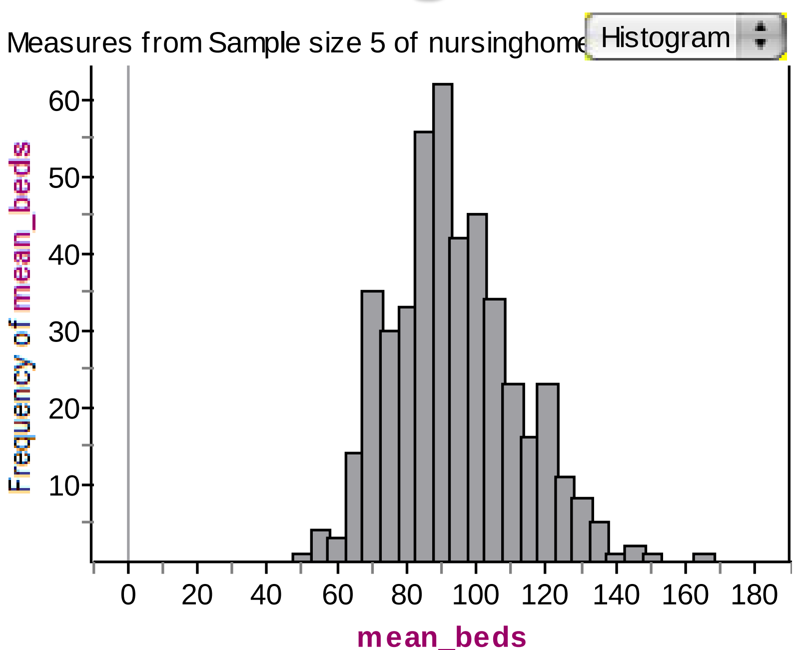
nursinghomedat

BED	93.2692

S1 = mean

Nursing Home Data

- Keeping in mind that there are C(52, 5) = 2,598,960 possible samples of size 5, consider the following: 450 samples of size 5.



Measures from Sample size 5 of nursinghomedat

mean_beds	93.4316

S1 = mean □ □

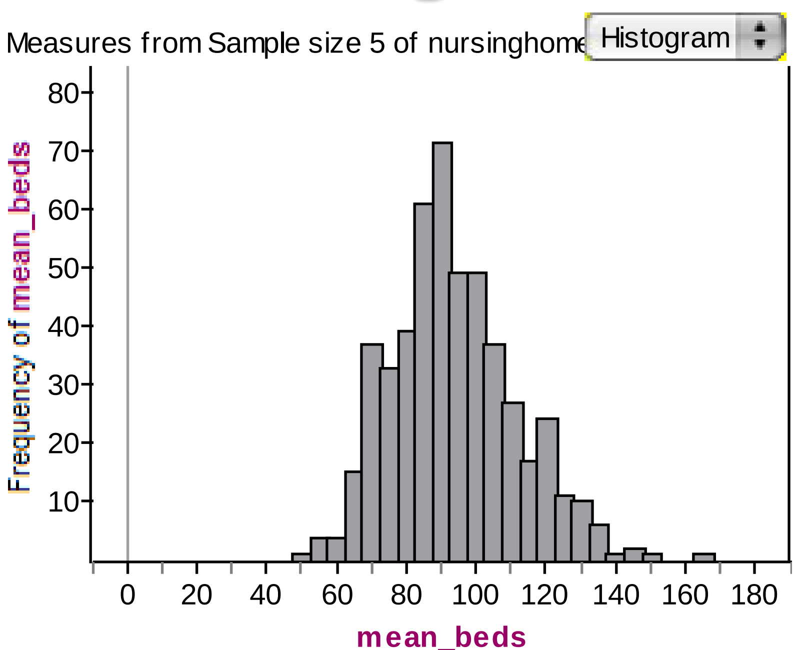
nursinghomedat

BED	93.2692

S1 = mean □ □

Nursing Home Data

- Keeping in mind that there are C(52, 5) = 2,598,960 possible samples of size 5, consider the following: 500 samples of size 5.



Measures from Sample size 5 of nursinghomedat

mean_beds	93.38

S1 = mean □ □

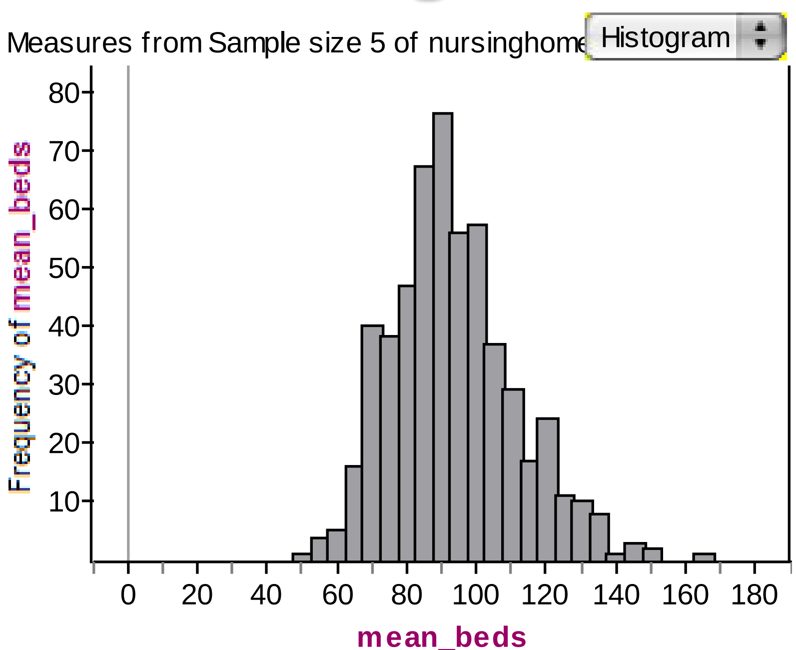
nursinghomedat

BED	93.2692

S1 = mean □ □

Nursing Home Data

- Keeping in mind that there are $C(52, 5) = 2{,}598{,}960$ possible samples of size 5, consider the following: 550 samples of size 5.



Measures from Sample size 5 of nursinghomedat

mean_beds	93.1735

S1 = mean

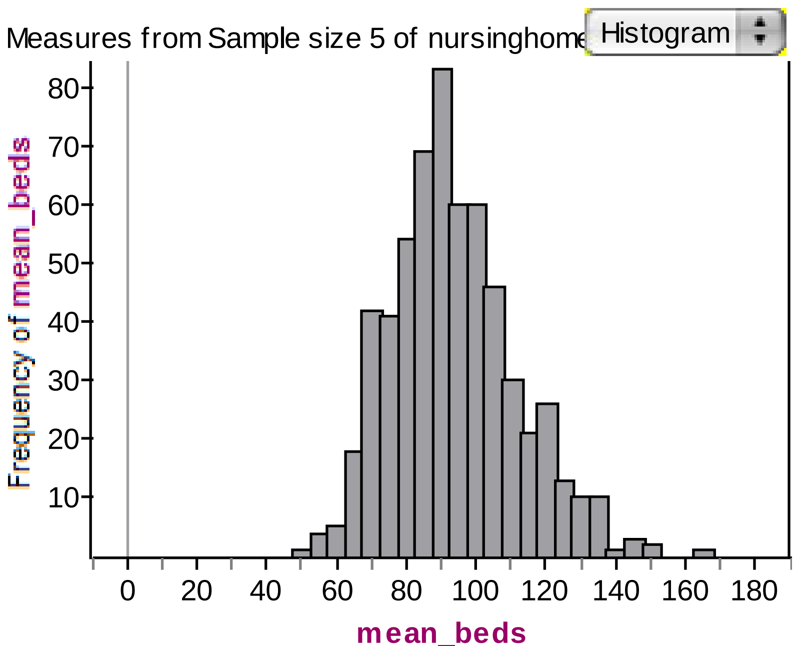
nursinghomedat

BED	93.2692

S1 = mean

Nursing Home Data

- Keeping in mind that there are C(52, 5) = 2,598,960 possible samples of size 5, consider the following: 600 samples of size 5.



Measures from Sample size 5 of nursinghomedat

mean_beds	93.413

S1 = mean

nursinghomedat

BED	93.2692

S1 = mean

Nursing Home Data

mean for data: 93.2692

Sample Size	Mean
50	96.664
100	95.798
150	96.164
200	94.955
250	94.5632
300	93.7127
350	93.672
400	94.038
450	93.4316
500	93.38
550	93.1735
600	93.413

Nursing Home Data

mean for data: 93.2692

Sample Size	Mean
50	96.664
100	95.798
150	96.164
200	94.955
250	94.5632
300	93.7127
350	93.672
400	94.038
450	93.4316
500	93.38
550	93.1735
600	93.413

Examining these means, we see the *Law of Large Numbers* in action …

Nursing Home Data

mean for data: 93.2692

Sample Size	Mean
50	96.664
100	95.798
150	96.164
200	94.955
250	94.5632
300	93.7127
350	93.672
400	94.038
450	93.4316
500	93.38
550	93.1735
600	93.413

… as additional *observations* are added, the difference between the population mean and the sample mean changes …

Nursing Home Data

mean for data: 93.2692

Sample Size	Mean
50	96.664
100	95.798
150	96.164
200	94.955
250	94.5632
300	93.7127
350	93.672
400	94.038
450	93.4316
500	93.38
550	93.1735
600	93.413

… as additional *samples* are added, the difference between the population mean and the sample mean changes …

Nursing Home Data

mean for data: 93.2692

Sample Size	Mean
50	96.664
100	95.798
150	96.164
200	94.955
250	94.5632
300	93.7127
350	93.672
400	94.038
450	93.4316
500	93.38
550	93.1735
600	93.413

… as additional *samples* are added, the difference between the population mean and the sample mean approaches zero.

Law of Large Numbers

- ***As the sample size increases,* the sample mean, $\overline{X}$, and the population mean, μ, *become closer in value.***

Law of Large Numbers

- As additional observations are added to the sample, the *difference* between the sample mean, $\overline{X}$, and the population mean, μ, approaches zero.

Sampling Distribution

- As the sample size n increases, the standard deviation for the distribution of the $\overline{X}$ decreases.

Mean and Standard Deviation for a Sampling Distribution

- Suppose a simple random sample of size n is taken from a population that has a mean μ and standard deviation σ.

Mean and Standard Deviation for a Sampling Distribution

- Suppose a simple random sample of size n is taken from a population that has a mean μ and standard deviation σ.
- The mean for the distribution of $\overline{X}$ is the same as the population mean

Mean and Standard Deviation for a Sampling Distribution

- Suppose a simple random sample of size n is taken from a population that has a mean μ and standard deviation σ.
- The mean for the distribution of $\overline{X}$ is the same as the population mean
  - That is, $\mu_{\overline{x}} = \mu$.

Mean and Standard Deviation for a Sampling Distribution

- Suppose a simple random sample of size n is taken from a population that has a mean μ and standard deviation σ.
- The standard deviation of the sampling distribution of $\overline{X}$ is $\sigma_{\overline{x}}$ and $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$.

Mean and Standard Deviation for a Sampling Distribution

- Suppose a simple random sample of size n is taken from a population that has a mean μ and standard deviation σ.
- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ is referred to as the *standard error of the mean.*

Sampling Distribution from Population that is Approximately Normal

- If the population is normally distributed then the distribution of sample means is normally distributed.

Central Limit Theorem

- The sampling distribution of $\overline{X}$ becomes approximately normal as the sample size n increases.

Central Limit Theorem

- The sampling distribution of $\overline{X}$ becomes approximately normal as the sample size n increases.
- How large must the sample be???

Central Limit Theorem

- The sampling distribution of $\overline{X}$ becomes approximately normal as the sample size n increases.
- How large must the sample be???
  - The sample size must be at least 30.

Central Limit Theorem

- The sampling distribution of $\overline{X}$ becomes approximately normal as the sample size n increases.
- How large must the sample be???
  - For sample size n, $n \geq 30$.

Population with mean μ and standard deviation σ	Distribution of sample means	Mean of sample means	Standard deviation of sample means
Not normal with n < 30	Not normal	$\mu_{\bar{x}} = \mu$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Not normal with n ≥30	Approximately normal	$\mu_{\bar{x}} = \mu$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Normal	Normal for any sample size	$\mu_{\bar{x}} = \mu$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Population with mean μ and standard deviation σ	Distribution of sample means	Mean of sample means	Standard deviation of sample means
Not normal with n < 30	Not normal		
Not normal with n ≥30	Approximately normal	$\mu_{\overline{x}} = \mu$	$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$
Normal	Normal for any sample size		

Sample Proportion

- Consider a sample of size n obtained from a population for which each element/individual either has or does not have a particular characteristic. The sample proportion, denoted p, is given by

$$p = \frac{x}{n}$$

where x is the number of elements/individuals in the sample that have the characteristic. **

Sample Proportion

- The sample proportion, $\hat{p}$, is a statistic that estimates the population proportion, p.

$$\hat{p} = \frac{x}{n}$$

$$\hat{p} = \frac{\text{number in sample with charateristic}}{\text{number in sample}}$$

Sample Proportion

- The sample proportion, $\hat{p}$, is a statistic that estimates the population proportion, p.

$$\hat{p} = \frac{x}{n}$$

$$\hat{p} = \frac{\text{number in sample with charateristic}}{\text{number in sample}}$$

- Sample proportion, $\hat{p}$, estimates the population proportion, p.

Sampling Distribution of $\hat{p}$

- Having the sample size no more than 5% of the population size (that is, $n \leq 0.05N$) is necessary so that the result obtained from one individual is *independent* of the result obtained from the rest of the individuals.

***

Sampling Distribution of $\hat{p}$

- For a simple random sample of size n for which the sample size is less than 5% of the population size, that is, $n \leq 0.05N$,
  - the sampling distribution of $\hat{p}$ is approximately normal provided

$$np(1 - p) \geq 10.$$

Sampling Distribution of $\hat{p}$

- For a simple random sample of size n for which the sample size is less than 5% of the population size, that is, $n \leq 0.05N$,
  - the mean of the sampling distribution of $\hat{p}$ is $\mu_{\hat{p}} = p$.

Sampling Distribution of $\hat{p}$

- For a simple random sample of size n for which the sample size is less than 5% of the population size, that is, $n \leq 0.05N$,
  - the standard deviation of the sampling distribution of $\hat{p}$ is

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

Sampling Distribution of $\hat{p}$

- It is important to note that the mean of the sampling distribution for $\hat{p}$ is $\mu_{\hat{p}} = p$ <u>and</u> the standard deviation of the sampling distribution of $\hat{p}$ is

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

if $np(1 - p) \geq 10$ <u>or</u> $np(1 - p) \ngeq 10$