Confidence Intervals

Confidence Intervals

- What is a confidence interval?
- How are confidence intervals used?
- How are confidence intervals interpreted?

Research Process

Research Process

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

Data

Data

- In general, from a sample

Data

- In general, from a sample
  - Use sample data (and statistics) to make inferences about *population*

Data

- In general, from a sample
  - Use sample data (and statistics) to make inferences about *population*
    - Use *sample mean* to draw conclusions about *population mean*
    - Use *sample proportion* to draw conclusions about *population proportion*

Data

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

Inference

- *Assumption based on an observation*
- A conclusion obtained on the basis of evidence and reasoning
- The act or process of deriving a conclusion based solely on what one already knows

Statistical Inference

- The theory, methods, and practice of forming judgments about the parameters of a population and the reliability of statistical relationships, typically on the basis of random sampling

Statistical Inference

- **The theory, methods, and practice of forming judgments about the parameters of a population and the reliability of statistical relationships, typically on the basis of random sampling**
  - ***A logical process of drawing conclusions from a collection of data and relationships between data and potential conclusions***

Inferential Statistics

- Based on probability statements

Inferential Statistics

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

Reminder about Samples

Reminder about Samples

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

To make Probability Statements for Sample Statistics …

- Need to know about the distribution of the
  - *Sample mean*
  - *Sample Proportion*

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:

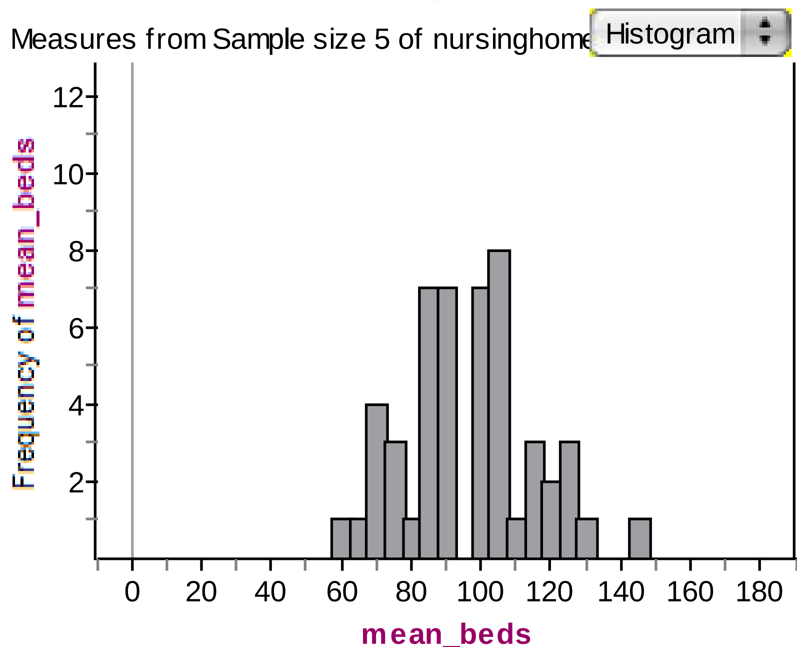
1. BED = number of beds in home
2. MCDAYS = annual medical in-patient days (hundreds)
3. TDAYS = annual total patient days (hundreds)
4. PCREV = annual total patient care revenue ($hundreds)
5. NSAL = annual nursing salaries ($hundreds)
6. FEXP = annual facilities expenditures ($hundreds)
7. RURAL =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

- 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

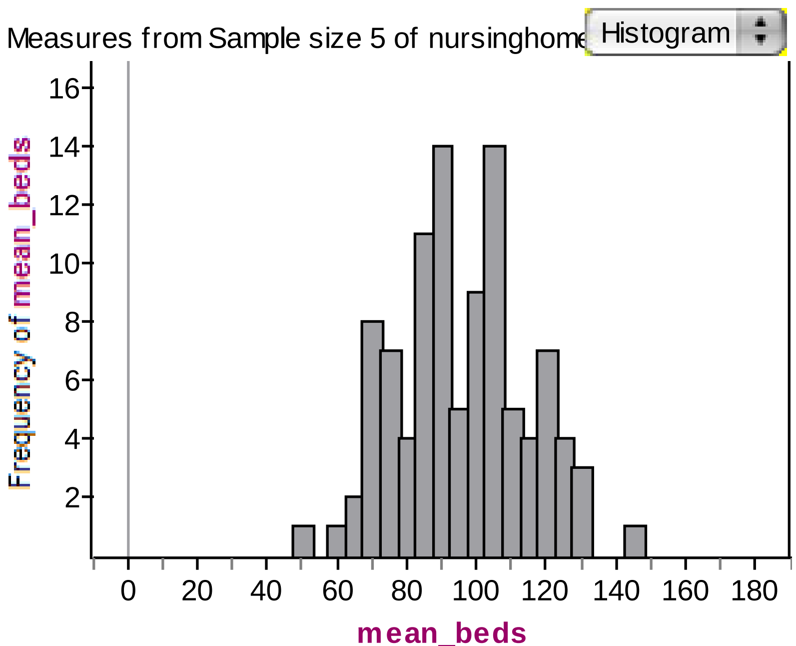
nursinghomedat

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

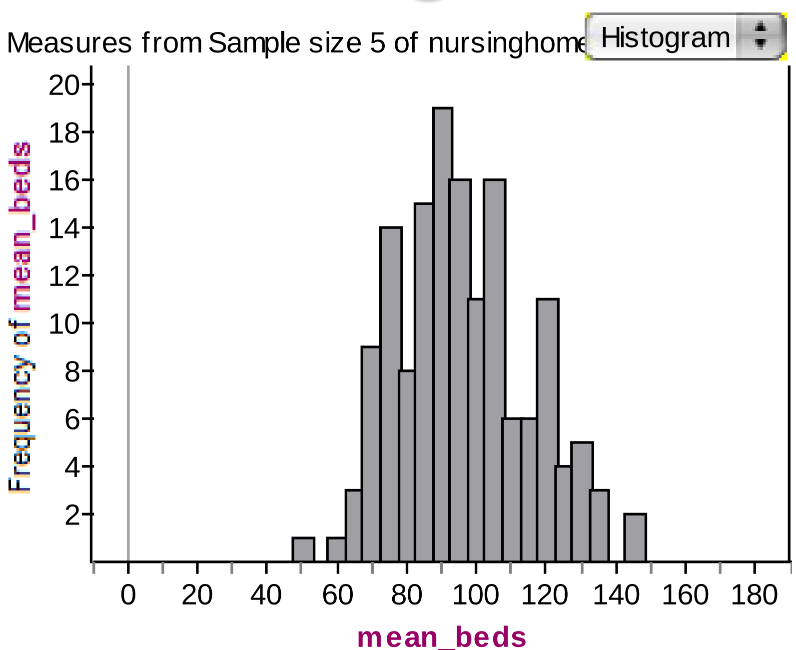
nursinghomedat

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

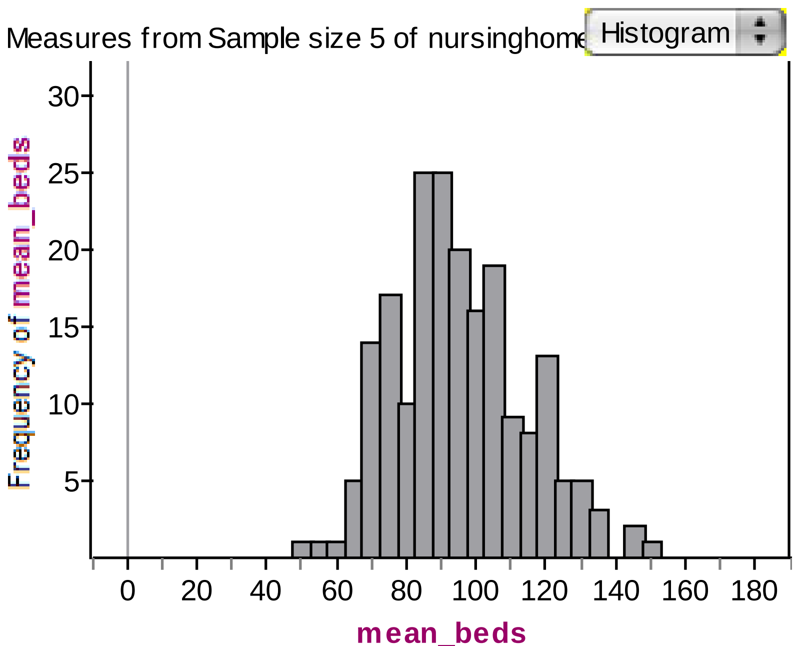
nursinghomedat

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

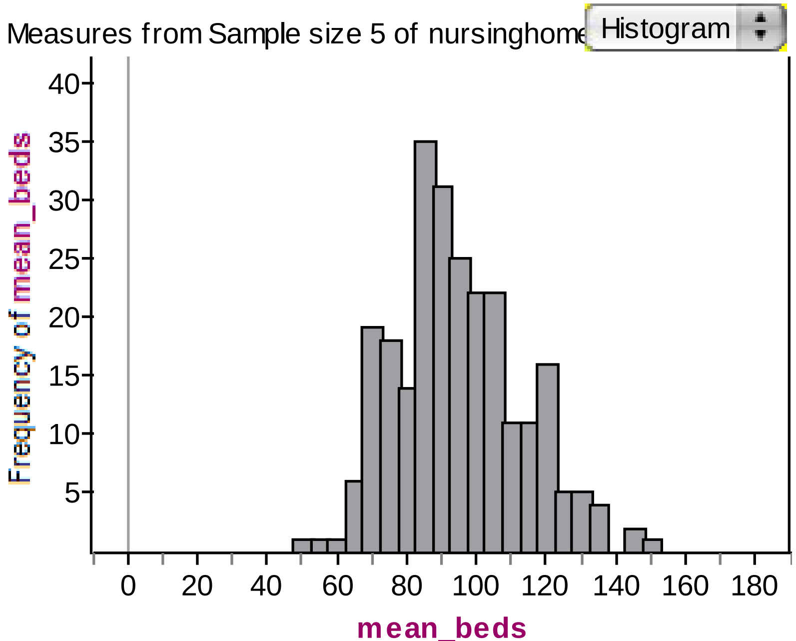
nursinghomedat

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 □ □

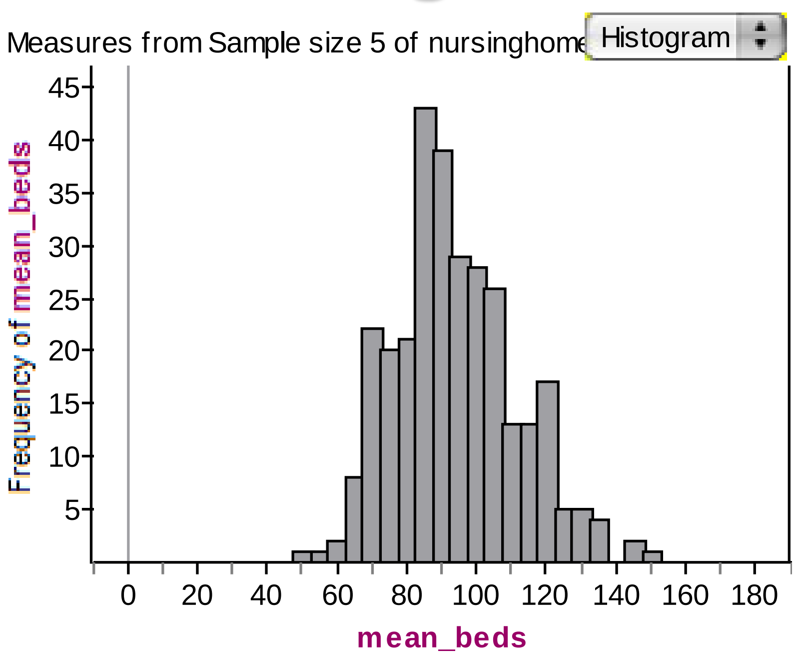
nursinghomedat

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

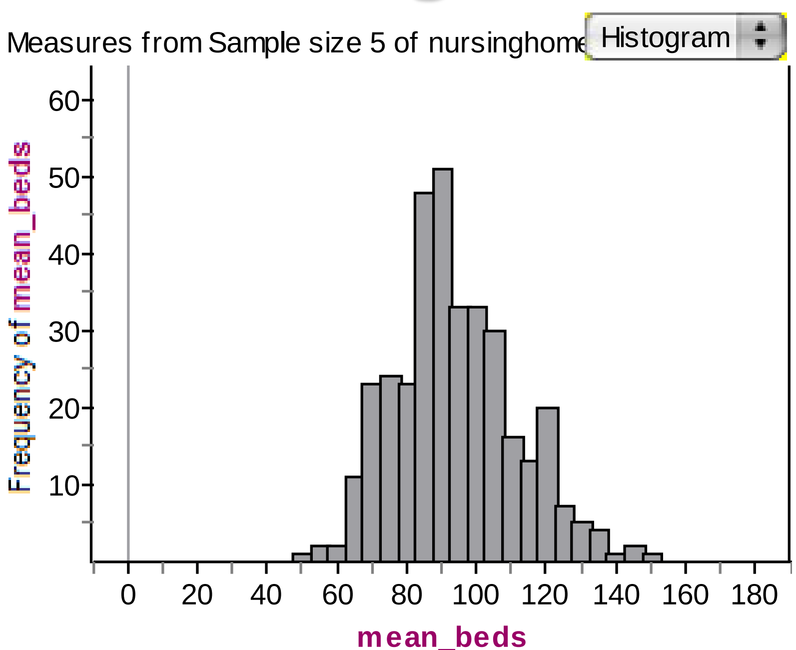
nursinghomedat

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

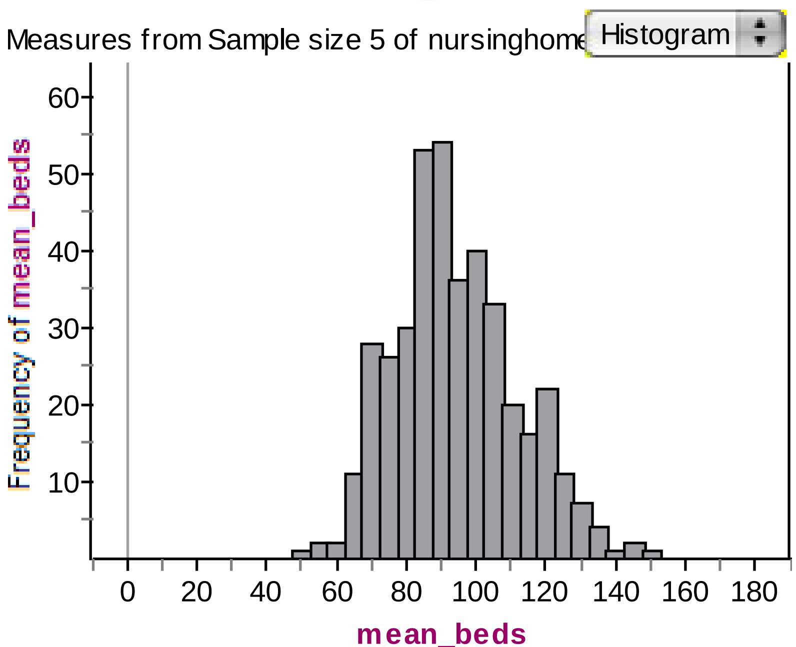
BED	93.2692

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 nursinghome



Measures from Sample size 5 of nursinghomedat

mean_beds	94.038

S1 = mean

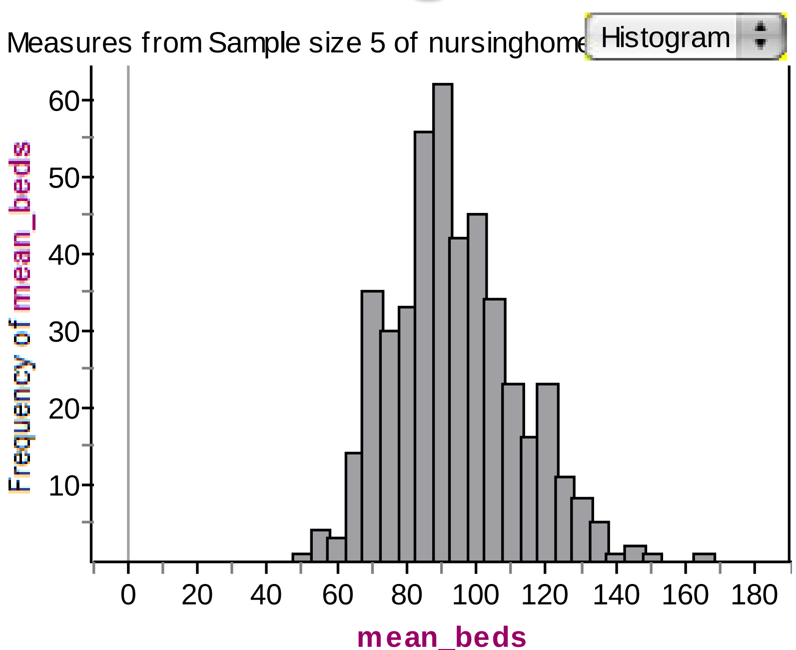
nursinghomedat

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 □ □

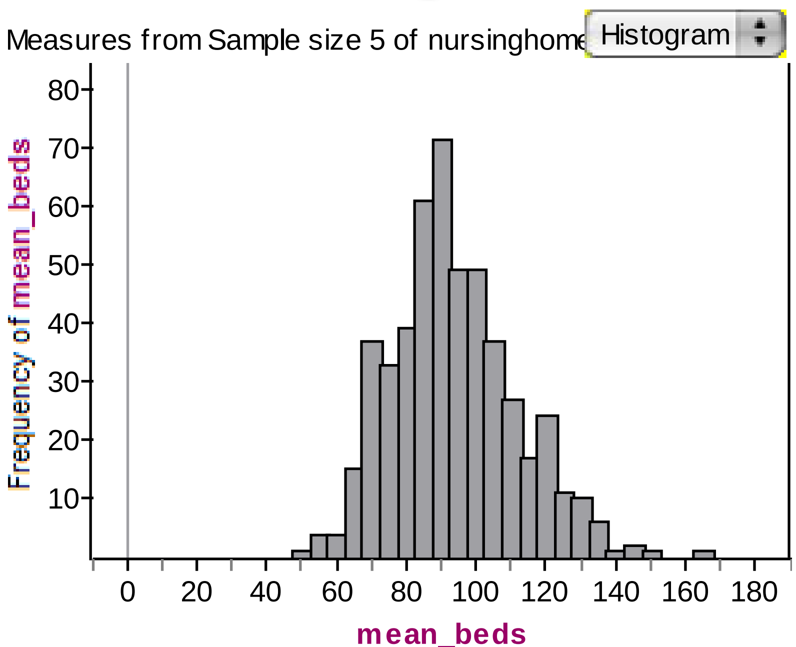
nursinghomedat

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

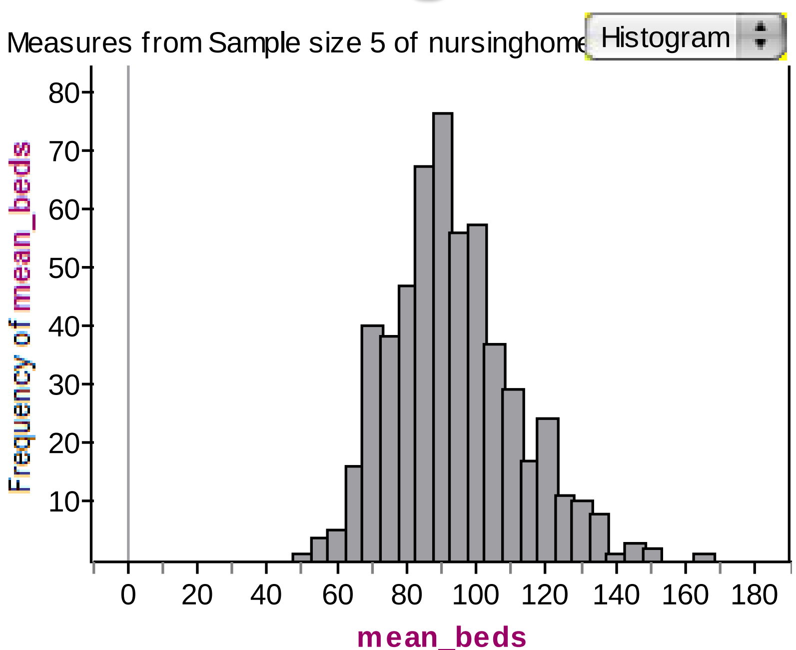
nursinghomedat

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

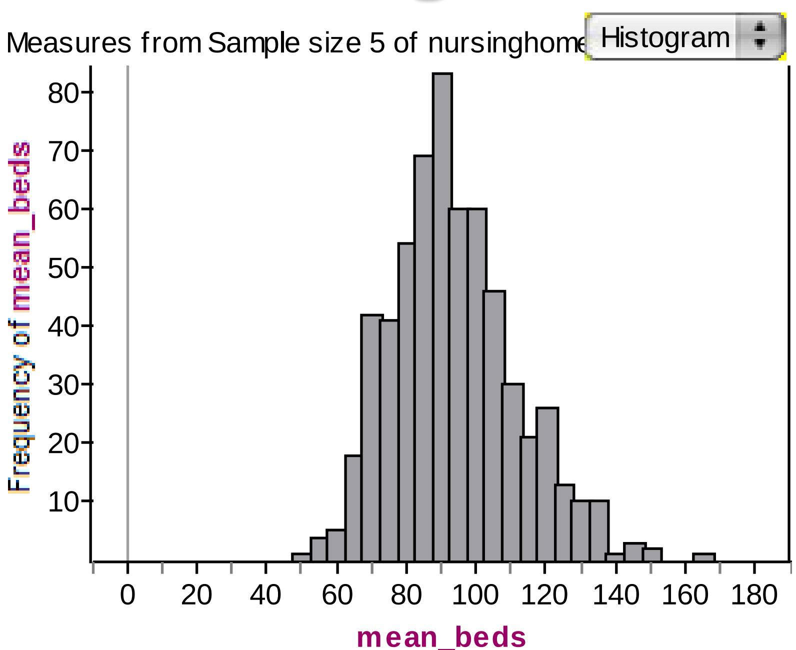
nursinghomedat

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.*

Central Limit Theorem

- The distribution for the sample mean, $\overline{X}$, becomes approximately normal as the sample size n increases.

Central Limit Theorem

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

Central Limit Theorem

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

$$n \geq 30$$

What about the Sample Proportion?

Sample Proportion

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

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

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

Distribution for Sample Proportion

- The participants must be *independent*.

Distribution for Sample Proportion

- The participants must be *independent*.
  - the sample size must be *no more than 5% of the population size*
  - $n \leq 0.05N$

Distribution for Sample Proportion

- For a simple random sample of size n for which the sample size is less than 5% of the population size ($n \leq 0.05N$),
  - the *distribution for the sample proportion* is approximately normal provided

$$n\hat{p}\left(1-\hat{p}\right) \geq 10.$$

Confidence Interval

- *A confidence interval estimate of a parameter consists of*
  - an interval of numbers
    - Lower estimate (lower bound)
    - Upper estimate (upper bound)

together with
  - a measure of the likelihood

that the interval contains the unknown parameter

Confidence Interval

- *A confidence interval estimate of a parameter consists of*
  - an interval of numbers
    - Lower estimate (lower bound)
    - Upper estimate (upper bound)

together with

  - a Level of Confidence

that the interval contains the unknown parameter

Confidence Interval

- *A confidence interval estimate of a parameter consists of*
  - an interval of numbers
    - Lower estimate (lower bound)
    - Upper estimate (upper bound)

together with

  - a Level of Confidence

that the unknown parameter lies between the two estimates

Level of Confidence

- The level of confidence in a confidence interval is the percentage of intervals that will contain population mean μ if a large number of repeated samples is taken.

Confidence interval for the mean, μ

Level of Confidence

- The level of confidence in a confidence interval is the percentage of intervals that will contain population proportion p if a large number of repeated samples is taken.

Confidence interval

for the proportion, p

Interpretation of Confidence Interval

- *A* 90% confidence interval tells us that *if* we were to obtain many simple random samples of size n from a population for which the population mean μ is *unknown then* approximately 90% of the intervals would contain the value of the *population mean* μ.

Interpretation of Confidence Interval

- A 95% confidence interval tells us that *if* we were to obtain many simple random samples of size n from a population for which the population mean μ is *unknown then* approximately 95% of the intervals would contain the value of the *population mean* μ.

Interpretation of Confidence Interval

- A 98% confidence interval tells us that *if* we were to obtain many simple random samples of size n from a population for which the population mean μ is *unknown then* approximately 98% of the intervals would contain the value of the *population mean* μ.

Interpretation of Confidence Interval

- A 99% confidence interval tells us that *if* we were to obtain many simple random samples of size n from a population for which the population mean μ is *unknown then* approximately 99% of the intervals would contain the value of the *population mean* μ.

Interpretation of Confidence Interval

- *A* $(1 - \alpha)\cdot 100\%$ confidence interval tells us that *if* we were to obtain many simple random samples of size n from a population for which the population mean μ is *unknown then* approximately $(1 - \alpha)\cdot 100\%$ of the intervals would contain the value of the *population mean* μ.

Interpretation of Confidence Interval

- *A* 90% confidence interval tells us that *if* we were to obtain many simple random samples of size n from a population for which the population proportion p is *unknown then* approximately 90% of the intervals would contain the value of the *population proportion* p.

Interpretation of Confidence Interval

- A 95% confidence interval tells us that *if* we were to obtain many simple random samples of size n from a population for which the population proportion p is *unknown* *then* approximately 95% of the intervals would contain the value of the *population proportion* p.

Interpretation of Confidence Interval

- A 98% confidence interval tells us that *if* we were to obtain many simple random samples of size n from a population for which the population proportion p is *unknown* *then* approximately 98% of the intervals would contain the value of the *population proportion* p.

Interpretation of Confidence Interval

- *A* 99% confidence interval tells us that *if* we were to obtain many simple random samples of size n from a population for which the population proportion p is *unknown then* approximately 99% of the intervals would contain the value of the *population proportion* p.

Interpretation of Confidence Interval

- A $(1 - \alpha) \cdot 100\%$ confidence interval tells us that *if* we were to obtain many simple random samples of size n from a population for which the population proportion p is *unknown* *then* approximately $(1 - \alpha) \cdot 100\%$ of the intervals would contain the value of the *population proportion* p.

Constructing a (1 - α)·100% Confidence Interval for the Population Proportion p

- Suppose a simple random sample size n is taken from a population.

  A (1 - α)·100% confidence interval for p is determined by

  - Lower bound: $$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$
  - Upper bound: $$\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

  for which both n $\leq$ 0.05N and $n\hat{p}\left(1-\hat{p}\right) \geq 10$ must be satisfied.

Constructing a (1 - α)·100% Confidence Interval for the Population Proportion p

- For a simple random sample size n,

  A (1 - α)·100% confidence interval for p is determined by

  - Lower bound: $\hat{p} - z_{\alpha/2}\sqrt{\dfrac{\hat{p}\left(1-\hat{p}\right)}{n}}$
  - Upper bound: $\hat{p} + z_{\alpha/2}\sqrt{\dfrac{\hat{p}\left(1-\hat{p}\right)}{n}}$

  for which both n $\leq$ 0.05N and $n\hat{p}\left(1-\hat{p}\right) \geq 10$ must be satisfied.

  The margin for error is $E = z_{\alpha/2}\sqrt{\dfrac{\hat{p}\left(1-\hat{p}\right)}{n}}$.

Constructing a $(1 - \alpha)\cdot 100\%$ Confidence Interval about μ and σ unknown

- Suppose a simple random sample size n is taken from a population with unknown population mean μ and an unknown population standard deviation σ.
- A $(1 - \alpha)\cdot 100\%$ confidence interval for μ is determined by
  - Lower bound: $\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}$
  - Upper bound: $\bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$

for $t_{\alpha/2}$ computed with n-1 degrees of freedom

Constructing a $(1 - \alpha)\cdot 100\%$ Confidence Interval about μ and σ unknown

- Suppose a simple random sample size n is taken from a population for which μ and σ are unknown.
- A $(1 - \alpha)\cdot 100\%$ confidence interval for μ is determined by
  - Lower bound: $\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}$
  - Upper bound: $\bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$

  The margin for error is $E = t_{\alpha/2}\frac{s}{\sqrt{n}}$.

  $t_{\alpha/2}$ computed with n-1 degrees of freedom

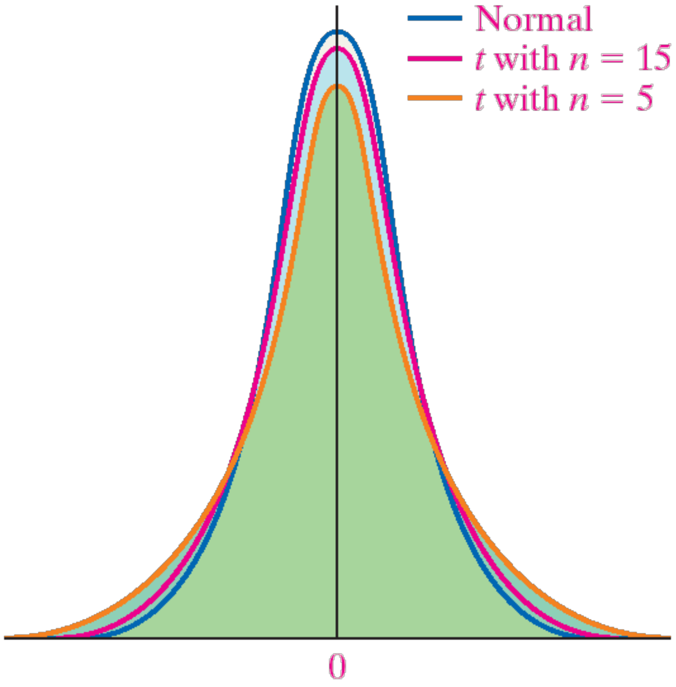
Choosing between the Normal Distribution and the t-Distribution for Confidence Intervals

Use Normal Distribution (z)	Proportion $n \le 0.05N$ and $n\hat{p}\left(1-\hat{p}\right) \ge 10$ must be satisfied
Use t-Distribution (t)	Mean $n \le 0.05N$ and data from population which is normally distributed OR $n \geq 30$

What is the t-Distribution?

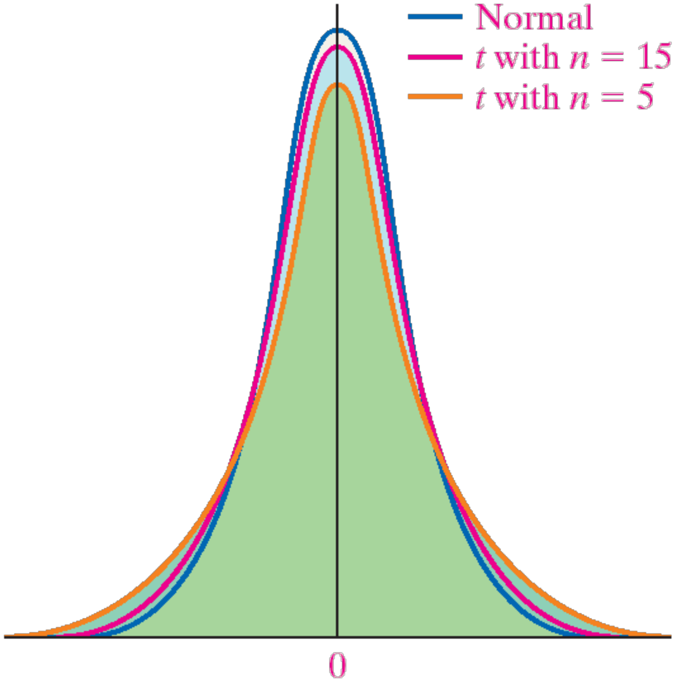
Properties of t-Distribution

- The t-Distribution changes based on the number of degrees of freedom



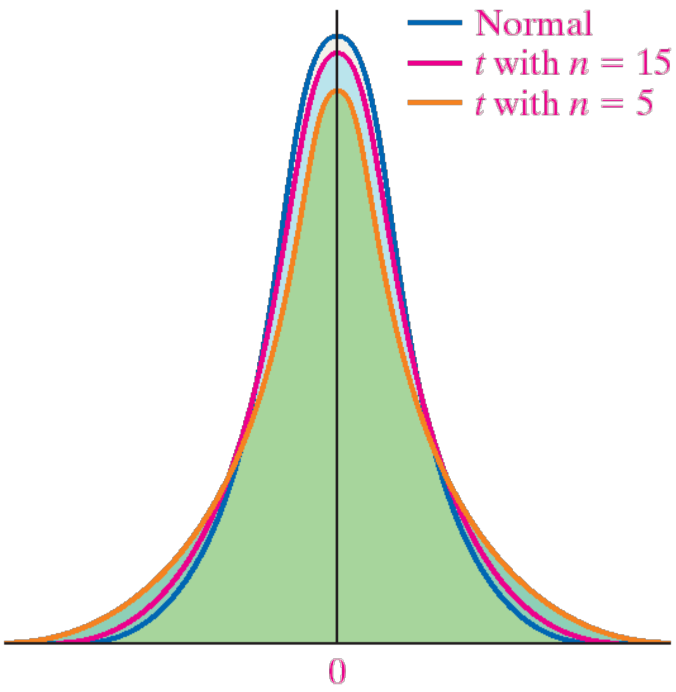
Properties of t-Distribution

- The t-Distribution is centered at 0 and is symmetric about 0.



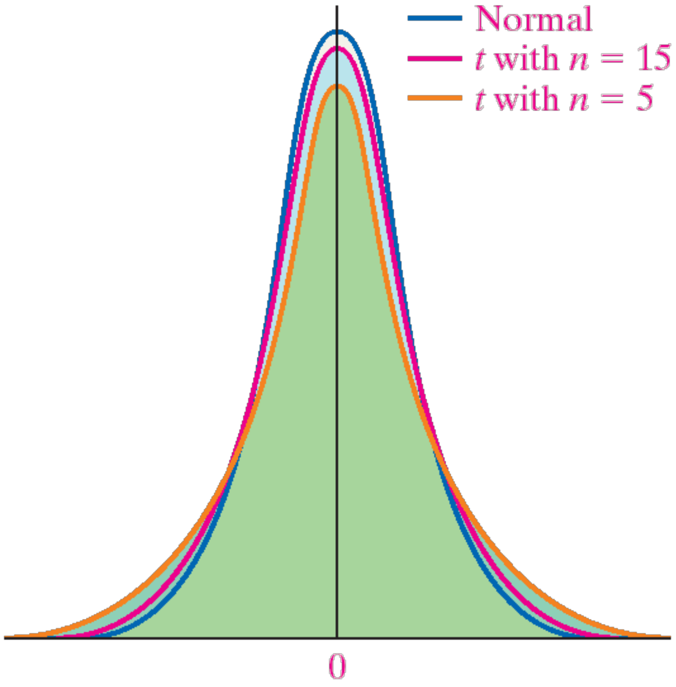
Properties of t-Distribution

- The area under the curve is 1.



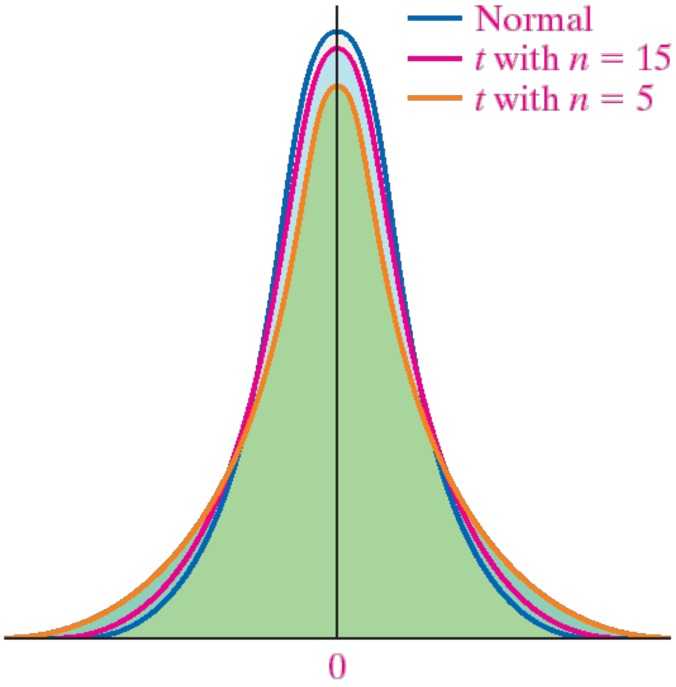
Properties of t-Distribution

- The area under the curve is to the left of 0 is ½.



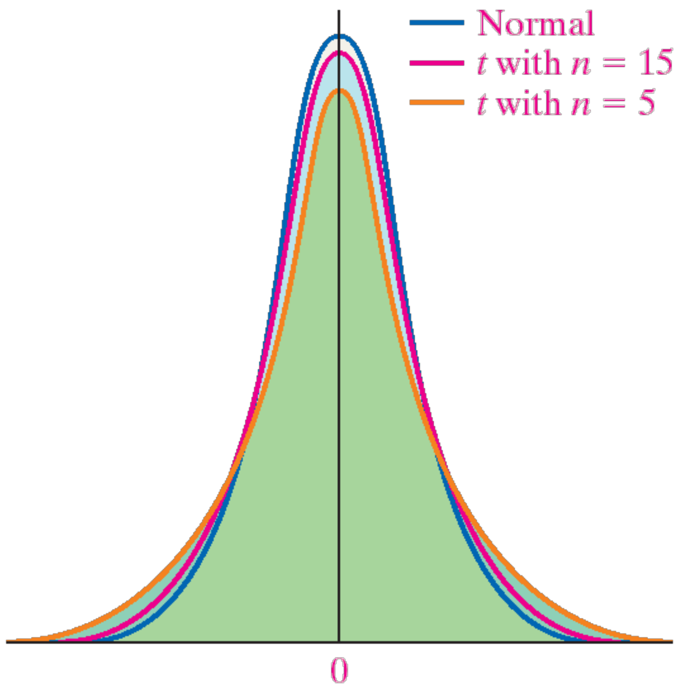
Properties of t-Distribution

- The area under the curve is to the right of 0 is ½.



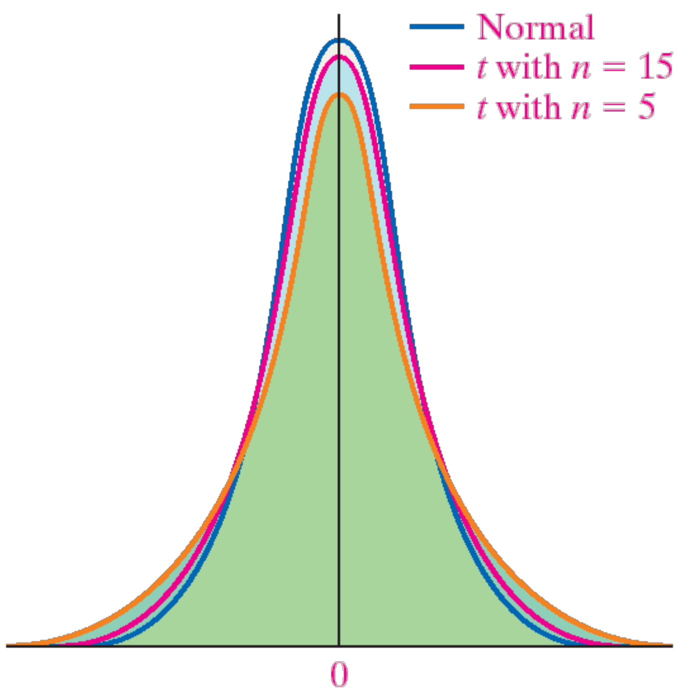
Properties of t-Distribution

- The curve never touches the horizontal axis.



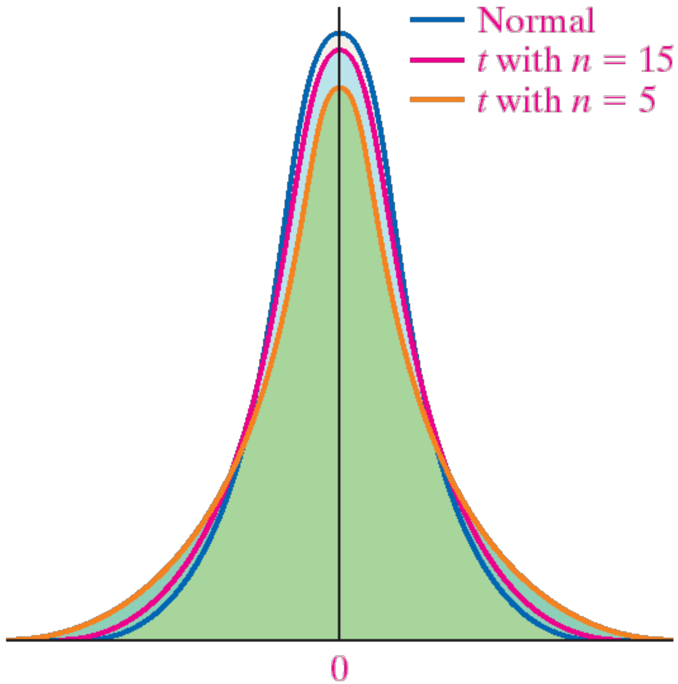
Properties of t-Distribution

- The t-distribution is similar to the standard normal distribution.



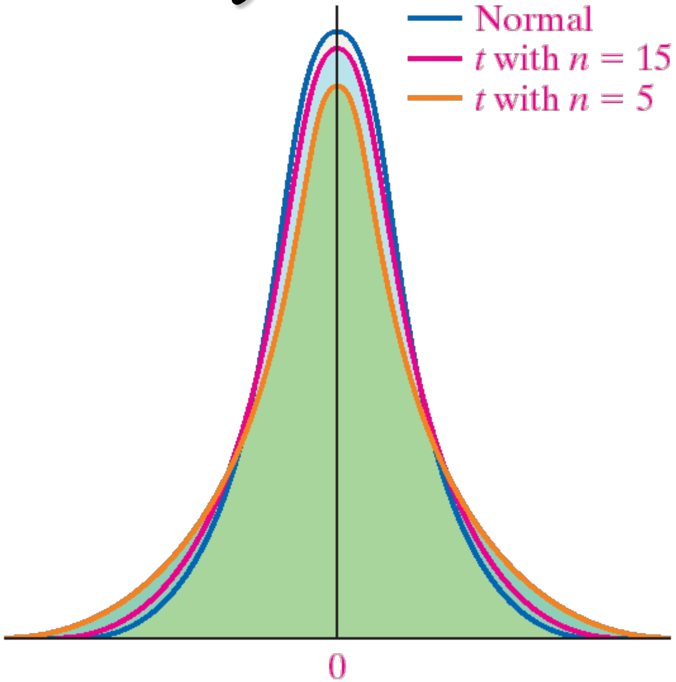
Properties of t-Distribution

- The area in the "tails" of the t-Distribution is a little greater than the area in the tails of the standard normal distribution.



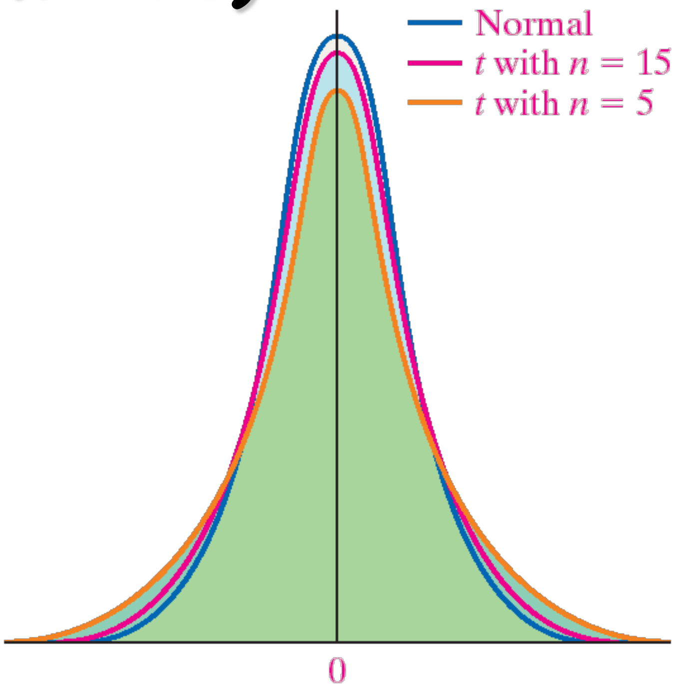
Properties of t-Distribution

- As the sample size n increases, the density curve of t gets closer to the standard normal density curve. (Law of Large Numbers)



Properties of t-Distribution

- As the sample size n increases, the density curve of t gets closer to the standard normal density curve. (Law of Large Numbers)



Round-off Rule for Confidence Intervals used to Estimate μ

- When using *original sample data* to construct a confidence interval, round the endpoints of the confidence interval to *one more decimal place than the original data*.
- When using given values of $\bar{x}$, n, and s with the sample data unknown, round the endpoints of the confidence interval to the *same number of decimal places as the sample mean*.