Chapter 4
Confidence Interval of the Mean

Section 4.2
Conceptual Basis of the Confidence Interval

David W. Gerbing

The School of Business
Portland State University

4.2a
Distribution of the Sample Mean m
The z-distribution of the Sample Mean

The Sample Mean m as a Variable

Consider variation across samples

- Here we focus on the population mean for the variable of interest, denoted as μ, and its relation to the sample mean, m
- Only one m is typically observed, but each random sample generally would have a different value of m
- For multiple samples, interpret m as a variable with its own mean, standard deviation and shape
- Because only one m is typically observed, the distribution of m over many samples is a mathematical *abstraction*
- Fortunately, there is an insight that allows this abstraction to become practical to understand the variability of m
- **Key Concept**: The mean, standard deviation and shape of m over many, usually hypothetical, samples can be estimated with information from only a *single* sample

Variability of the Distribution of the Sample Mean

Standard Error – a Key Component of Inferential Statistics

- As discussed, the variability of many m's over many samples indicates how close any one m is likely to be to μ
- As usual, assess variability with the standard deviation, here applied to the statistic m defined as a variable
- **Standard Error** of a statistic: The standard deviation of the *statistic* across (usually hypothetical) multiple samples
- The standard error is a standard deviation, but applied to the variability of a statistic over (usually hypothetical) samples
- The reference to the "standard deviation" of something is usually intended to apply to the variability of the data values
- The phrase "standard error" applies to the standard deviation of a corresponding statistic

Actual Standard Error σ_m

Information about Many Samples Deduced from One Sample

- Denote the population standard deviation of m, its standard error defined over all possible samples, with σ_m
- This standard error is described with a simple expression that relates to the standard deviation of the data values
- ***Actual* (population) standard error of m**: $\sigma_m = \dfrac{\sigma}{\sqrt{n}}$
- The standard deviation of the sample means of all possible samples is the standard deviation of all possible data values divided by the square root of the size of each sample
- This discussion of the population value σ_m is theoretical as the population standard deviation, σ, on which it depends, is typically not known
- This discussion, however, provides the logical basis for what occurs with actual data analysis

Meaning of the Standard Error σ_m

The standard error is the key to statistical inference

- Statistical inference involves the consideration of the standard deviations of two different variables
  - σ: Stnd dev of the population of all potential data values
  - σ_m: The usually hypothetical standard deviation of the sample mean, m, over all possible samples
- A crucial relationship follows: $\sigma_m = \dfrac{\sigma}{\sqrt{n}} \quad < \quad \sigma$
- That is, the sample mean varies less than the data (and potential data from the entire population)
- The mean is a centering process, such that the extreme large values in a sample tend to cancel out the extreme small values in the same sample
- The result is that the sample mean fluctuates more closely around the population mean than do individual data values

Shape of the Distribution of the Sample Mean m

More explanation in the appendix

- **Central Limit Theorem**: m is at least approximately normal except for small samples from non-normal populations of data
- A "small" sample size usually means around $n = 30$, though a larger sample may be needed for a skewed distribution of data values to ensure the normality of m
- **Key Concept**: For a normally distributed m, use normal curve probabilities to calculate the range of variation of m
- Before beginning an inferential analysis of μ: Verify that m is normally distributed, so that normal probabilities can be used for inference
  - If the sample is larger than 30 or so, then assume m normal
  - If the sample size is much less than 30, inspect a histogram of the data to at least ascertain that skewness is not an issue

Mean of the Distribution of the Sample Mean m

The mean of the means

- In terms of notation, μ is the population mean of the data, and μ_m is the population mean of the distribution of all possible sample means
- There is only one logical value of the mean of the sample means, μ_m, and that is the mean of the data, μ
- Population mean of m: The mean of all possible sample means is the same mean of all the data values, $\mu_m = \mu$
- Accordingly, only the expression μ is used from here on

z-value: How Far is an Obtained m from μ?

To begin with a theoretical discussion, presume σ Known

- Because of sampling variation, each given m is some distance from the target, the true, population value, μ
- Express the usually normal m in terms of z_m, the standardized version of m, the universal metric for a normal distribution, which specifies how many standard errors separate m and μ
- **Standardized sample mean**: $z_m = \dfrac{m - \mu}{\sigma_m}$
- z_m expresses how many standard errors, the standard deviation of m over repeated samples, separate m from μ
- μ_m and σ_m are constant for *all* samples, so the variation of z_m over many samples depends only on the variation of m, so ...
- **Key Concept**: Given a normally distributed z_m, normal curve probabilities describe the extent of sampling variation

Probability Intervals

What range contains *most* of the z_m values?

- To express the variation of m over the usually hypothetical multiple samples, obtain the range of these normally distributed z_m distances
- **Key Question**: For the many, many hypothetical values of m, how far can a given m be from its target, μ?
- Theoretically, no limit for either $+$ or $-$ values as a normal curve never touches the horizontal axis, so choose "most"
- Usually specify *most* as the range of 95% of the values
- **Probability Interval**: Range about μ for which a randomly selected m is likely to fall within, at a specified probability
- **Cutoff (critical) Value**: A value of a distribution that isolates the upper or lower tail of the distribution, which sets the bounds of a probability interval

95% Range of Variation of m: .975 Quantile, $z_{.025} = 1.96$

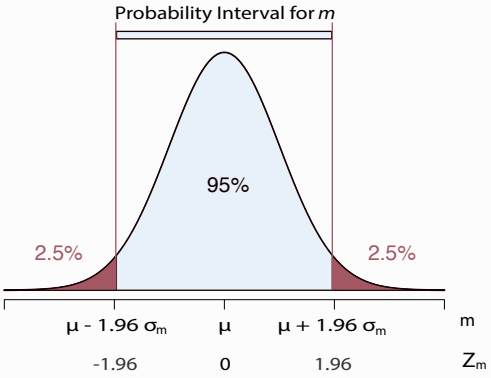


Figure: Probability interval for 95% of the values of any normal distribution, including for m, which fall within 1.96 standard deviations (errors) of the mean, μ

Illustration: 95% Probability Interval of m

Population values: $\mu = 100$, $\sigma = 28$ Sample size: $n = 90$

- $n > 30$, so distribution of m is normal regardless of the distribution of the population of the data values
- Apply normal curve probabilities to the distribution of m, so 95% of all values of m are within 1.96 standard errors of μ
- Population standard error

$$\sigma_m = \frac{\sigma}{\sqrt{n}} = \frac{28}{\sqrt{90}} = \frac{28}{9.49} = 2.95, \quad \text{and } (1.96)(2.95) = 5.78$$

- Now go 1.96 standard errors up and down from μ
- 95% range of variation of m, the probability interval

$$\text{LowerBound: } \mu - (1.96)(\sigma_m) = 100 - 5.78 = 94.22$$

$$\text{UpperBound: } \mu + (1.96)(\sigma_m) = 100 + 5.78 = 105.78$$

Illustration: Probability Interval of m (continued)

Cannot specify the value of m in advance of actually drawing the random sample from the population

However, can specify the probability interval

For m calculated from a sample of size $n = 90$ from the population of interest, with $\mu = 100$ and $\sigma = 28$:

The probability is 95% that any one m will lie between 94.22 and 105.78

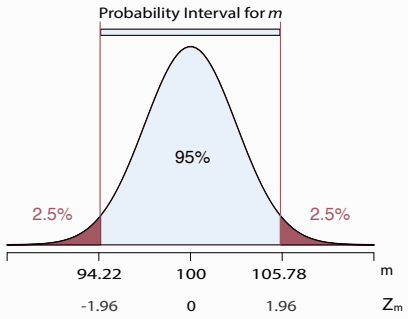


Figure: A 95% probability interval for the sample mean, m, for a specific population with $\mu = 100$ and $\sigma = 28$

Illustration: Probability Interval via Simulation

Draw some samples from this population

- In this example, for a population of any shape in which $\mu = 100$ and $\sigma = 28$, and a sample of size $n = 90$, 95% of the $m's$ vary about μ from 94.22 to 105.78
- To further illustrate, simulate the drawing of 8 samples, each of size $n = 90$ from the corresponding (normal) population

```
> simMeans(ns=8, n=90, mu=100, sigma=28)
```

	1	2	3	4	5	6	7	8
m's:	94.7	98.4	98.5	99.4	99.5	100.8	102.4	102.5

- In this particular simulation, all 8 values of m were within the 95% probability interval
- In general, about 5% of the values of m will be outside of the bounds of the probability interval

4.2b
Logic of the Confidence Interval

From probability interval to confidence interval

The basis of inference

- As discussed, the smaller the range of variability of the sample mean, m, over usually hypothetical multiple samples, the more likely that any one m is close to μ
- Define this range of variability about μ by the probability interval of m at a given level of probability, such as 95%
- This probability interval becomes the basis for the confidence interval, both of which are of the same width but centered over different values
- In actual data analysis, the confidence interval estimates an *unknown* value of μ

From Probability to Inferential Statistics

Deduction and induction (inference) compared

- **Key Concept**: The purpose of statistical inference is to estimate the value of a population characteristic for a variable of interest, such as its population mean μ
  - Last section focused on *deduction* to understand the extent of random fluctuation about a *known* value of μ
  - Here turn the situation around to *induction*, inference

Logic	Direction	Purpose
Deduction	from a model of the population to the data	*deduce* probabilities of data values from known population values
Induction	from data to a model of the population	*infer* unknown population values from data

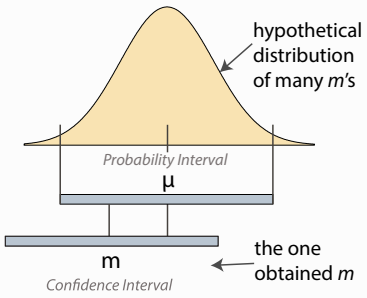
Presenting the Confidence Interval

Basic definitions

- **Confidence Level**: Specified percentage of values that defines the typical range of variation of the statistic of interest over repeated samples about the corresponding population value
- Most widely used confidence level is 95%, a "nice" number that gets "most" of the values
- **Confidence interval** for the population mean: Range of values that likely contains the population mean, μ, at a specified confidence level
- The confidence interval for the mean specifies the range of plausible values of the constant population mean, μ
- The sample mean generally does not equal the population mean, $m \neq \mu$
  - Just knowing m by itself does not inform us as to μ
  - So construct an interval around m that likely contains μ

Logic of the Confidence Interval

- The probability interval around the true mean μ contains 95% of all m's
- Get one sample, so one m, here a little less than μ
- **Key Concept**: If the interval constructed about μ contains m, then an interval of the *same width* about m contains μ
- The interval about m is the confidence interval, in practice constructed *without* knowledge of μ



From Probability Interval to Confidence Interval

The interval centered over the sample value m

- As seen, the range of variation of the sample statistic m is the key to constructing the confidence interval
- The remarkable Central Limit Theorem specifies that the sample mean m is typically normally distributed over multiple samples, so normal curve probabilities provide a specified range
- Once the size of the range is known, the confidence interval is the centering of this range over the sample value m
- **95% confidence interval** calculated with knowledge of the population standard deviation, σ: $m \pm (1.96)(\sigma_m)$
- Although a true statement, the problem of employing this expression in practice is that if μ is not known, then almost always neither is σ, in which case σ_m cannot be calculated
- In the analysis of data, typically replace σ with its sample estimate, s, discussed next

Appendix
Shape of the Distribution of the Sample Mean
The Central Limit Theorem

Show the Shape of the Distribution of the Sample Mean

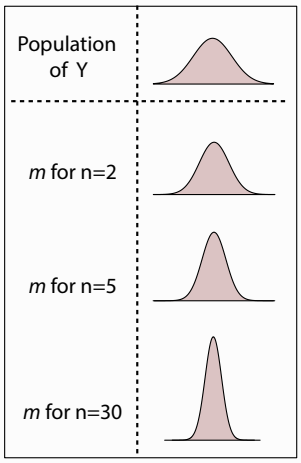
Use computer simulation to illustrate the distribution of *m*

- `lessR` function `simCLT()`: To investigate if *m* is normal for a given population and sample size, simulate its distribution
- To run a simulation, specify the shape of the population of all of the data with the `dist` parameter
- *Required*: Four possible values of `dist`:
  `"normal"`, `"uniform"`, `"antinormal"`, `"lognormal"`
- *Required*: `ns`, number of samples
  `n`, size of each sample
- For example, to calculate *m* from each of 1000 samples, each with two data values, from a uniform distribution:

```
> simCLT(ns=1000, n=2, dist="uniform")
```


Central Limit Theorem: Normal Data

- The normal population, from which the data are sampled
- Take many, many different samples, each sample of the smallest possible size, $n = 2$
- The many, many means of each of these samples for $n = 2$ has a distribution: *m* is also normal
- If the sample size is $n = 5$ for each of the many, many samples, *m* is also normal
- *m* also normal for many, many samples of $n = 30$ and larger



Example of All Possible Sample Means of Size 2

Uniform distribution, 5 equally probable values: 0, 1, 2, 3, 4

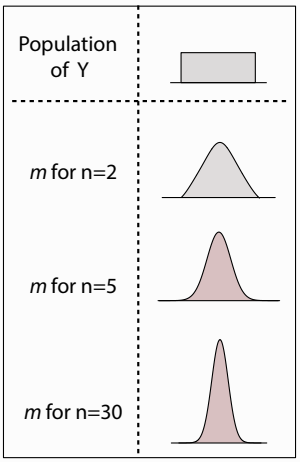
- There are only 25 possible samples of size 2

Sum	Mean	Possible Samples	Count	Prob
0	0.0	0,0	1	1/25=0.04
1	0.5	0,1 1,0	2	2/25=0.08
2	1.0	0,2 1,1 2,0	3	3/25=0.12
3	1.5	0,3 1,2 2,1 3,0	4	4/25=0.16
4	2.0	0,4 1,3 2,2 3,1 4,0	5	5/25=0.20
5	2.5	1,4 2,3 3,2 4,1	4	4/25=0.16
6	3.0	2,4 3,3 4,2	3	3/25=0.12
7	3.5	3,4, 4,3	2	2/25=0.08
8	4.0	4,4	1	1/25=0.04
Total			25	1.00

- More ways to get $m = 2$ than $m = 0$ or $m = 4$, so values of m tend to converge toward $\mu = 2$

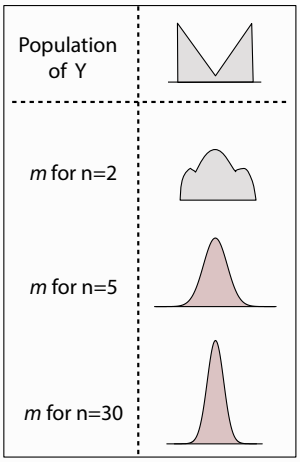
Central Limit Theorem: Uniform Data

- The uniform population Y, from which the data are sampled, small, medium and large values of Y *equally* likely
- Take many, many different samples of Y, each sample of the smallest possible size, $n = 2$
- The distribution of the sample means for $n = 2$ is *almost* normal
- If the sample size is $n = 5$ for each of the many, many samples, m is approximately normal
- m also normal for many, many samples of $n = 30$ and larger



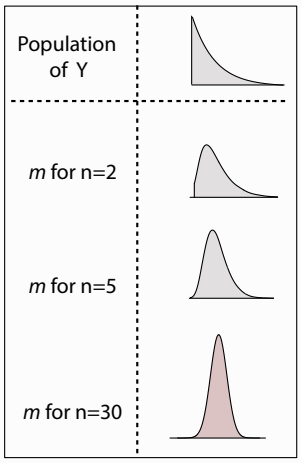
Central Limit Theorem: "Anti-Normal" Data

- The population Y from which the data are sampled, with values in the middle *less* likely than tail values
- Take many, many different samples of Y, each sample of the smallest possible size, $n = 2$
- The distribution of the sample means for $n = 2$ is almost normal
- If the sample size is $n = 5$ for each of the many, many samples, m is approximately normal
- m also normal for many, many samples of $n = 30$ and larger



Central Limit Theorem: Skewed Data

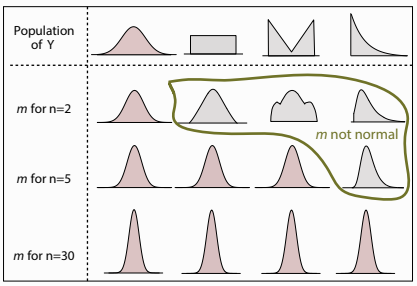
- A skewed population Y from which the data are sampled
- Take many, many different samples of Y, each sample of the smallest possible size, $n = 2$
- The distribution of the many, many means of each of these samples is also skewed
- If $n = 5$ distribution of many m's still retains some skew
- m also normal for many, many samples of $n = 30$ and larger



Central Limit Theorem: Summary

The sample mean m is . . .

- exactly normal when the population of Y is normal for any sample size n from 2 onward
- approximately normal for a symmetric population of Y *except* for very small sample sizes n less than 5
- approximately normal for even a skewed population of Y when the sample size n is at least 30



Conclusion: The sample mean m follows a normal distribution *except* when the sample size is small and the population of the data is not normal, particularly when skewed

CLT: Conclusion and Practical Consequences

Rules to establish normality of the sample mean, m

- m must be at least approximately normal to apply normal (or related) distribution probabilities for the confidence interval
- IF sample size $n > 30$, then the sample mean, m, is at least approximately normally distributed *unless* the data are sampled from a severely skewed population
- Generally, and always with smaller sample sizes, evaluate the shape of the population distribution of Y with an analysis of the sample data
  - IF the population for the sample data is normal, the population distribution of m is exactly normal
  - IF the population for the sample data is not skewed, n can be quite small and m will be at least approximately normal
  - IF the population is skewed, then $n > 30$ or more are needed for m to be approximately normal

- The End

Index

Subtract 2 from each listed value to get the Slide #