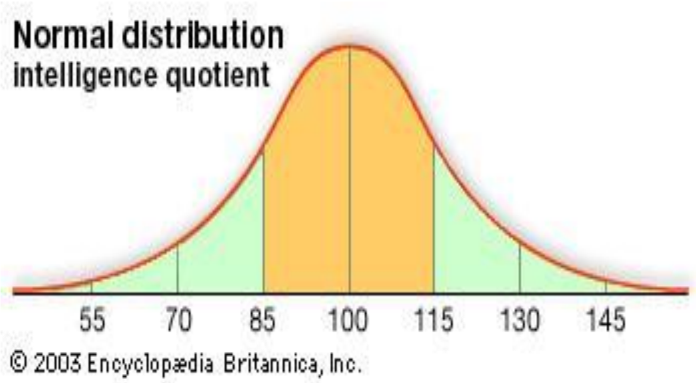
From Descriptive Statistics to Inferential Statistics

Z-score

Normal Curve

Standard Error of the Mean

Estimation with Confidence Intervals



Rossi A. Hassad, PhD, MPH

What you can expect in this module:

- So far we have covered descriptive statistics, which is summary information about the sample: The center, shape and spread of the distribution of a particular variable.
- Good logic will suggest that when we work with a sample (a subset of a population), our goal is to learn about the population (the larger group from which the sample was taken
- Using information from a sample to make statements about or generalize to the population is referred to as inferential statistics
- Inferential statistics involves making inferences about the population based on the sample
- Because we are using a sample statistic to generalize to, or estimate a population value (a parameter), this can result in some degree of error (referred to as standard error)
- Hence when it comes to inferential statistics, we do not speak with certainty (which is the absence of error), rather, we report our estimates with a stated level of confidence, hence the 95% and 99% confidence intervals of the population mean, which we will calculate in this module
- The theory of the normal distribution or bell curve, generally referred to as the empirical rules of the normal curve, is the basis for estimation in statistics, when the distribution is normal
- In this module (see separate PowerPoint presentation), you will also learn about the z-score, a type of data transformation, which allows for meaningful comparison of scores within and between distributions

The Normal Curve

- The word "normal" suggests that this is a typical shape for a variety of real-world measurements, such as height, weight, IQ, exam scores, etc.
- As to whether this shape is indeed typical, is debatable, as normal distributions come from large random samples, and for many research domains such as where sensitive information is being obtained, samples sizes are generally "small".

Normal Curve Characteristics

- Shape looks like a bell
- Mean, median, mode are equal to one another
- The mean represents the center of the distribution
  - The shape is symmetrical
  - 50% of the values lie above the mean and 50% of the values lie below this point
  - Tails are asymptotic—tails come closer and closer to axis but never touch it

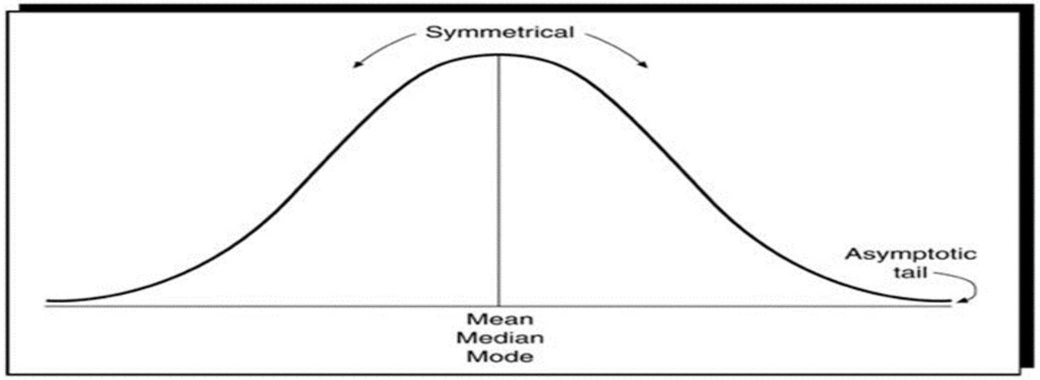


Figure 7.1. The Normal, or Bell-Shaped, Curve

The Normal Distribution – The Empirical Rules

- If data are normally distributed, then we can use the mean and standard deviation to approximate the proportion of the data (scores or observations) that will fall within a certain range or interval
- The Empirical Rules of the normal curve are used for this
- It is called empirical because this pattern is repeatedly observed for data with this shape
- The Empirical Rules of the normal curve are as follows:

✓ Approximately 68% of the data will fall within one standard deviation of the mean (Mean +/- 1SD)

✓ 95% will fall within two standard deviations of the mean (Mean +/- 2SD)

✓ 99% will fall within three standard deviations of the mean (Mean +/- 3SD)

❑ These empirical rules (95% and 99%) are the basis for estimation in research

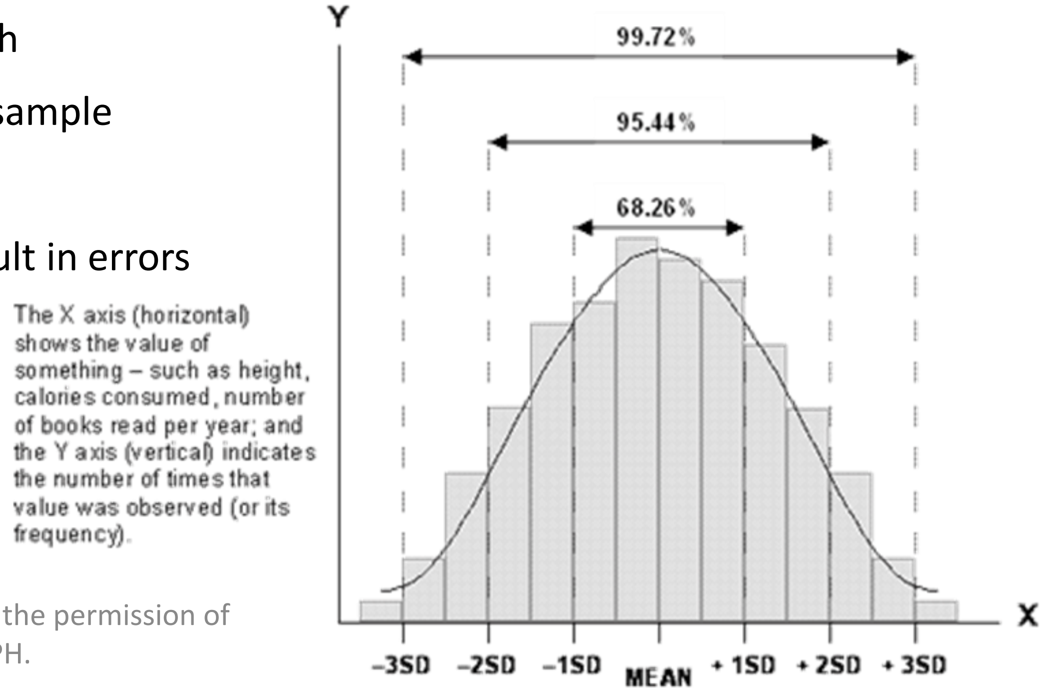
❑ Estimation refers to making inferences about the population based on the sample

❑ For example, estimating the population mean based on the sample mean

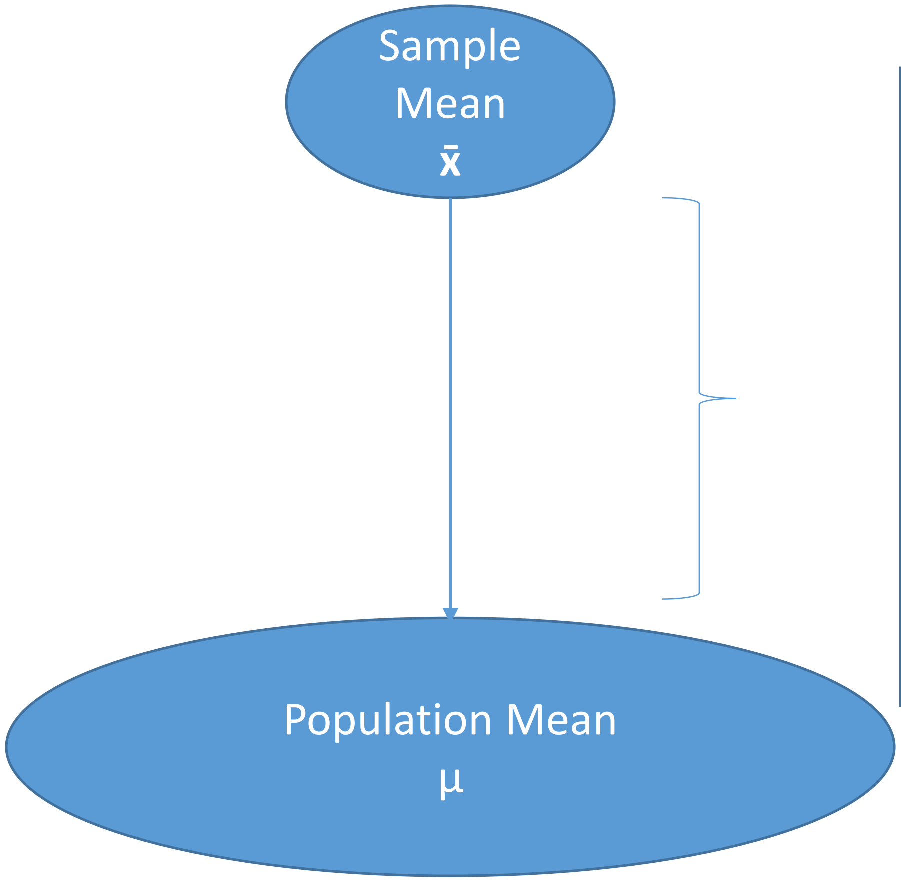
❑ Indeed, since we are generalizing or inferring from the sample, this can result in errors

❑ Hence, we will next address standard error of the mean (SE)

❑ SE is required for estimation



Standard Error (SE) of the Sample Mean ($\bar{x}$)



This distance represents the Standard Error (SE) of the Mean of the Sample. SE is a measure of variability between the sample and population means - also referred to as a measure of uncertainty. In other words how uncertain you are about your estimation of the population mean based on the sample mean.

Indeed a shorter distance (or smaller SE) is desirable.

Understanding the Formula for Standard Error (SE) of the Sample Mean

- This formula shows an inverse relationship between the sample size (n) and Standard Error (SE) of the sample mean (**$\bar{x}$**)
- As the size of the sample (m) increases, the SE (standard error) of the sample mean (**$\bar{x}$**) decreases
- The logic is, as the sample gets larger, we have more information about the population and hence our estimation of the population mean will have less error, and we can be more confident in our estimation
- It is also important to ensure that the sample is random and representative of the population

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Let's Calculate and Apply the Standard Error (SE) of the Sample Mean

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Calculate the Standard Error (SE) of the mean of the sample, given the following:

Mean height of students (**x̄**) = 65 ins

S (SD or Standard Deviation) = 6 ins

n (size of sample) = 16

SE = 6 ins / $\sqrt{16}$

SE = 6 ins/4 = 1.5ins (This value by itself is not helpful. It has to be used to estimate the population mean)

- So we will next address estimation

Estimation

- Okay, so now that we know the Standard Error (SE) of the sample mean (**x̄**), our task is to estimate the population mean (μ)

- However, a scientific estimation requires a level of confidence (generally, either 95% or 99%)

- Merely making a statement about the population without indicating how confident you are about that statement, is not helpful

- So let's clearly state what is it we want to estimate

➢ Estimate the population mean (for heights of students), based on the sample mean, with 95% confidence

Estimation Cont’d

- Estimate the population mean (for heights of students), based on the sample mean, with 95% confidence
- The percentage 95% should ring a bell (yes, refer to the empirical rules of the normal curve)
- So we will use an adaptation of that rule for our estimation
- Recall: 95% of the data will fall within 2SD of the Mean
- The formula for the 95% confidence interval (CI) is the same except that SD (standard deviation) becomes SE (standard error), given that we are estimating the population mean from the sample mean, and sampling error must be factored into the calculation

➢ See calculation on the next slide

Calculating the 95% Confidence Interval (CI) of the Population Mean

- So let's write the formula for the 95% CI of the population mean

  95% confidence interval (CI) of the population mean = sample mean +/- 2 SE (note SE and not SD)

- Recall: Mean = 65ins, SE = 1.5ins

  95% confidence interval (CI) of the population mean = 65 ins +/- 2 (1.5ins)

  95% confidence interval (CI) of the population mean = 65 ins +/- 3ins

- This will result in a range of values or an interval (with lower and upper limits) rather that a single value or point estimate
- Lower limit: 65ins – 3 = 62ins
- Upper limit: 65ins + 3 = 68ins
- Hence the 95% confidence interval of the population mean is: 62 ins to 68ins

➢ See next slide for interpretation

Interpretation of the 95% CI of the population mean

- 95% Confidence Interval (CI) of the population mean is: 62 ins to 68ins

- Interpretation:

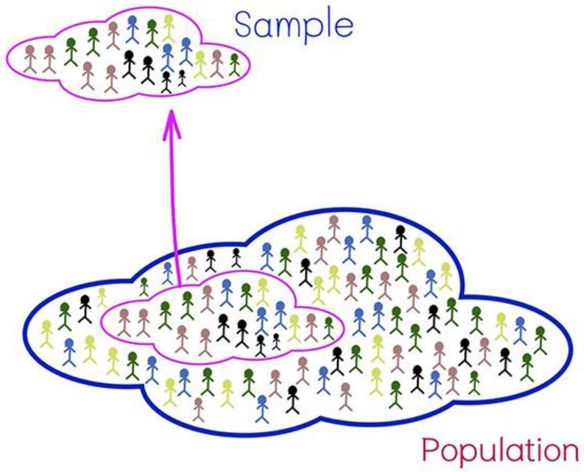
  In the sample we obtained a mean height of 65 ins, however, we are 95% confident that the true mean height of the population will fall between 62 ins and 68ins.

- Now let's calculate the 99% confidence interval for the population mean, instead.
- Recall from the empirical rules of the normal curve that 99% of the data will fall within 3 SD of the mean
- The 99% confidence interval of the population mean is: sample mean +/- 3 SE
- Recall: sample mean = 65 ins, SE = 1.5ins

  95% confidence interval (CI) of the population mean = 65 ins +/- 3 (1.5ins)

  95% confidence interval (CI) of the population mean = 65 ins +/- 4.5ins

99% CI of the Population Mean



- Lower limit: 65 ins – 4.5 ins = 60.5 ins
- Upper limit: 65 ins + 4.5 ins = 69.5ins
- 99% Confidence Interval (CI) of the population mean is.

  60.5 ins to 69.5 ins

- Interpretation:

In the sample we obtained a mean height of 65 ins, however, we are 99% confident that the true mean height of the population will fall between 60.5ins and 69.5ins.

Precision of Confidence Intervals

- 95% Confidence Interval (CI) of the population mean is: 62 ins to 68 ins
- 99% Confidence Interval (CI) of the population mean is: 60.5 ins to 69.5ins

❑ NOTE: For the same distribution, the 99% confidence interval (CI) is wider, than the 95% CI, and this should be obvious since the range or distance around the mean is 3 SE (rather than 2 SE for the 95% CI)

❑ IMPLICATIION: A wider confidence interval is less precise. A narrower range of values as is obtained for the 95% CI is more exact and hence more precise, and desirable for decision-making purposes

❑ Also, larger samples will result in a smaller standard error (SE) and hence a narrower and more precise confidence interval (CI)

❑ In summary, standard error (SE) can be used for estimating population values (such as the mean) using confidence intervals