Inferential Statistics - Estimation

Prof. Josefina V. Almeda
College Secretary
School of Statistics
University of the Philippines Diliman
2012

Learning Objectives

After the lesson on Estimation, the students should be able to

- Identify the two areas of Inferential Statistics;
- Distinguish the difference between point and interval estimation;
- Interpret the results of an interval estimate; and
- Interpret outputs from MS Excel/PHStat

The Logic of Statistical Inference

Key Definitions

- ***Parameters*** are numerical measures that describe the population or universe of interest. Usually donated by Greek letters; μ (mu), σ (sigma), ρ (rho), λ (lambda), τ (tau), θ (theta), α (alpha) and β (beta).

- ***Statistics*** are numerical measures of a sample

Two Areas of Statistical Inference

- Estimation
  - Point Estimation
  - Interval Estimation
- Hypothesis Testing

PROBLEMS ADDRESSED BY INFERENTIAL STATISTICS

ESTIMATION is concerned with finding a value or range of values for an unknown parameter.

TEST OF HYPOTHESIS deals with evaluating a claim or a conjecture about a parameter or distribution of the population.

Research Problem: How effective is Minoxidil in treating male pattern baldness?

This can be answered by ESTIMATION

Specific Objectives:

1. To estimate the population proportion of patients who will show new hair growth after being treated with Minoxidil.

2. To determine whether treatment using Minoxidil is better than the existing treatment that is known to stimulate hair growth among 40% of patients with male pattern baldness.

This can be answered by HYPOTHESIS TESTING

Hypothesis Testing versus Confidence Intervals

- Both confidence interval estimation and hypothesis testing deal with inferences about unknown population parameters.
- Deciding which to use in a particular application depends on the intent of the investigation. Do we need to gather information about the parameter, or do we ultimately have to make a decision concerning the parameter?
- Gathering information about the parameter involves constructing confidence intervals.
- Determining the truth of a particular conjecture about the parameter involves hypothesis testing.

FOLLOWING ARE SEVERAL EXAMPLES THAT ILLUSTRATE THE POINT:

- A politician wants to know what percent of voters in his district are in favor of his running for a second term. Here the politician will make a decision on the basis of the results of a poll, but the question at hand is, “What percent are in favor?” Thus a confidence interval is appropriate.
- The BFAD is testing a new dietary supplement to see whether it dissolves cholesterol deposits in arteries. Here they will make a decision that either the supplement is effective or it is not. Therefore a test of hypothesis is appropriate.

Following are several examples that illustrate the point:

- Government agencies that investigate such quantities as unemployment rates, inflation, gross national product, and so on, simply need estimates via confidence intervals.

- A claim is made that a greater proportion of women smoke than men. A test of hypothesis would shed light on this question.

- A textile company wishes to compare a new manufacturing process with the old one. Is the new technique an improvement over the old one? This clearly indicates that hypothesis testing is appropriate.

FOLLOWING ARE SEVERAL EXAMPLES THAT ILLUSTRATE THE POINT:

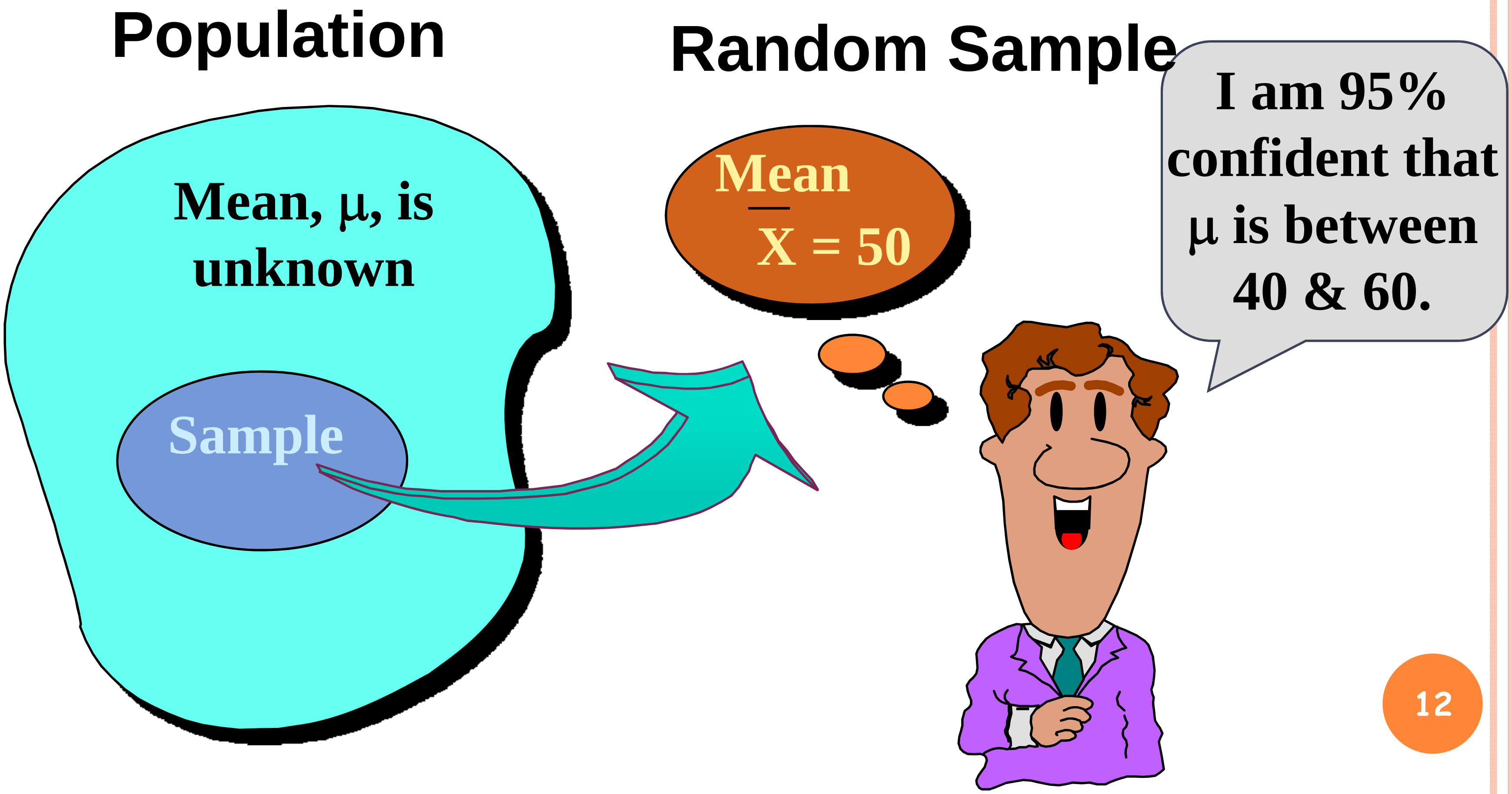
- A congressional committee investigating fatal traffic accidents caused by drinking most likely will be interested in confidence intervals.
- A business report of the inventory value of a warehouse of household carpet most likely will consist of confidence interval estimates

Some Notes:

- In general, confidence intervals give information and hypothesis testing helps make decisions.

- Although confidence intervals and hypothesis testing are different types of inference procedures, they are closely related. They are different ways of expressing the same information contained in a sample.

Basic Idea of Estimation



Population

Random Sample

Mean, μ, is unknown

Sample

Mean

$\bar{X} = 50$

I am 95% confident that μ is between 40 & 60.

ESTIMATION

How do we estimate a parameter, θ?

Specifically, how do we estimate

- **a population mean, μ ?**
- **a population standard deviation, σ?**
- **a population proportion, *p*?**

ESTIMATION

- An ***estimator*** of a parameter is a rule or a formula for computing an estimate using the sample data.
- It is usually denoted by a Greek letter with a 'hat' like $\hat{\theta}$ and $\hat{\mu}$.

ESTIMATION

- In other cases, special symbols are used like $\bar{X}$ for the sample mean as estimator of the population mean.
- An ***estimate*** is a numerical value of the estimator.

ESTIMATION

There can be several estimators for a particular parameter.

For example, a population mean can be estimated by any one of the following:

- sample mean
- sample modal value
- sample median

TWO TYPES OF ESTIMATOR

1. A **POINT ESTIMATOR** is a formula that gives a single value in estimating a parameter.

EXAMPLE: $\bar{x}$ is a point estimator of μ

s is a point estimator of σ

$\hat{p}$ is a point estimator of p

Table 1. Estimation of the Population Mean (μ) Under Simple Random Sampling

	Type of Sampling	
	Without Replacement	With Replacement
Point estimate	$\bar{x}$	$\bar{x}$
Standard error	$\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n-1}{N-1}}$	$\frac{\sigma}{\sqrt{n}}$
Estimate of standard error	$\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$	$\frac{s}{\sqrt{n}}$

Standard Error

- The standard error is used to measure how well the obtained statistic will estimate the target parameter. Its unit is the same as the unit of the data values.

- The smaller the standard error, the better the statistic estimates the parameter; that is, if the standard error is small then one can expect that the estimate will not differ substantially from the target parameter.

- The above estimate of the standard error tends to underestimate the true standard error but this becomes negligible as the sample size increases.

Table 2. Estimation of the Proportion of Elements in the Population Belonging in Class A (P_A) Under Simple Random Sampling

	Type of Sampling	
	Without Replacement	With Replacement
Point estimate	$p_a = \dfrac{\text{no.of obsvns.in sample that belong in Class } A}{n}$	p_a
Standard error	$\sqrt{\dfrac{P_A(1-P_A)}{n}\left(1-\dfrac{n-1}{N-1}\right)}$	$\sqrt{\dfrac{P_A(1-P_A)}{n}}$
Estimate of standard error	$\dfrac{p_a(1-p_a)}{1-p_a}\left(1-\dfrac{n}{N}\right)$	$\sqrt{\dfrac{p_a(1-p_a)}{n-1}}$

Example:

A simple random sample of 20 households was drawn from a barangay containing 250 households. The numbers of persons per household in the sample were as follows:

5	6	3	3	2	5	4	7	8	10
4	4	5	8	9	6	6	7	6	6

a) Estimate the mean number of persons per household of the barangay.

b) Estimate the proportion of households in the barangay that have more than 5 members.

TWO TYPES OF ESTIMATOR

2. An **INTERVAL ESTIMATOR** is a formula that gives a range of values for estimating a parameter.

EXAMPLE:

$\bar{X} \pm d$ is an interval estimator of μ where *d* is a specified half-width of the interval.

Basic Idea

- Typically, when we estimate a parameter (such as the population mean) by a single value (a statistic such as the sample mean).
- Since we know the behavior of these sample estimates due to their sampling distribution, perhaps it may be of interest to give interval estimates for the parameter rather than point estimates.
- Interval Estimate Provides Range of Values
- Gives Information about Closeness (to Unknown Population Parameter)

Level of Confidence

- Denoted by $(1 - \alpha)100\%$
- Typical levels: 90%, 95%, 99%
- A relative frequency interpretation
- In the long run, $(1 - \alpha)100\%$ of all the confidence intervals that can be constructed will contain the unknown parameter
- Wrong to say: that a specific interval will either have 95% probability of containing the parameter

INTERPRETING CONFIDENCE INTERVALS

- The 95% confidence intervals are different for 100 samples

- For about 95 of them the interval covers the parameter, but about 5 do not!

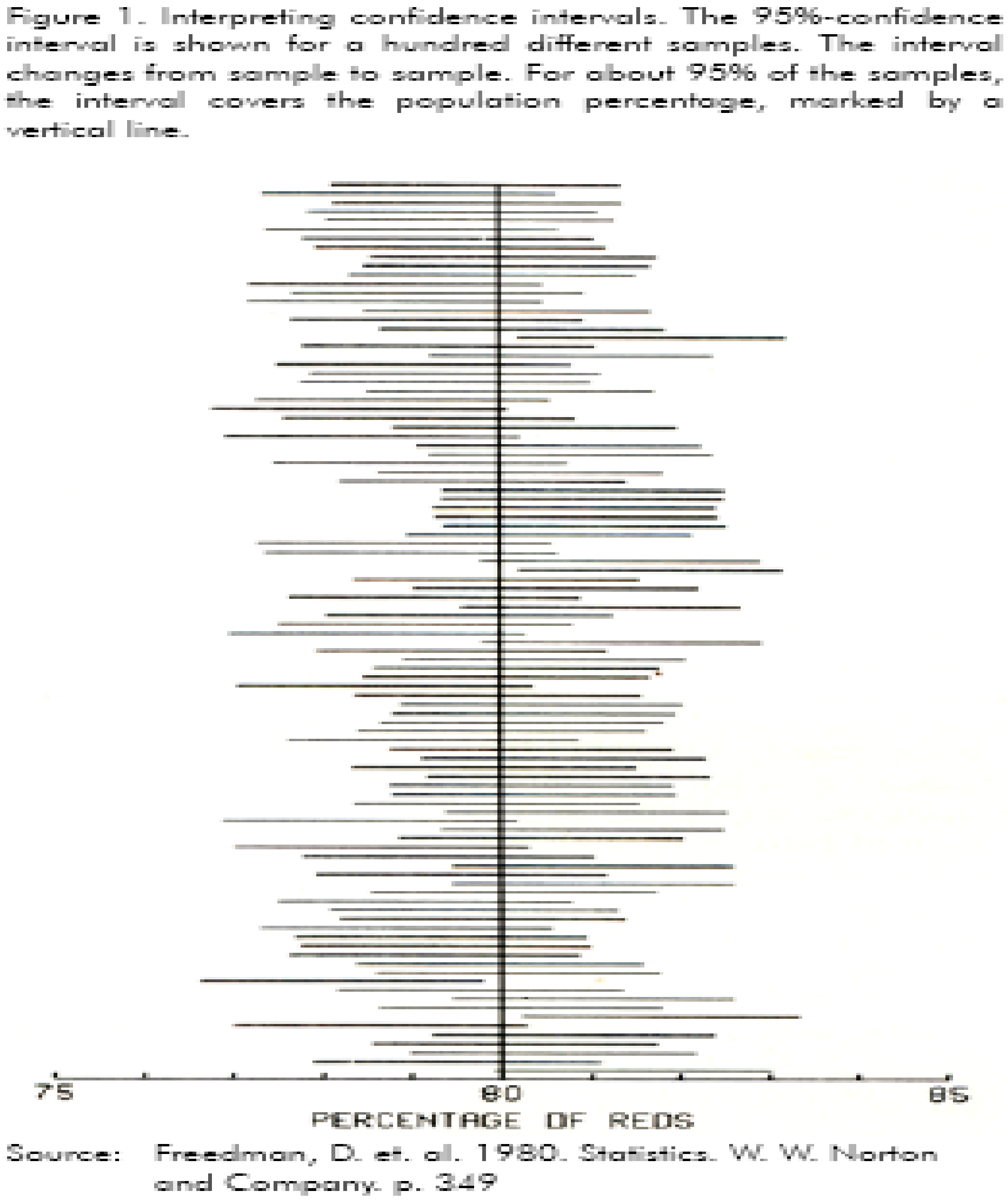


Figure 1. Interpreting confidence intervals. The 95%-confidence interval is shown for a hundred different samples. The interval changes from sample to sample. For about 95% of the samples, the interval covers the population percentage, marked by a vertical line.

Source: Freedman, D. et. al. 1980. Statistics. W. W. Norton and Company. p. 349

$(1-\alpha)100\%$ Confidence Interval Estimation (where $0 < \alpha < 1$)

- Formulas used to construct the $(1-\alpha)100\%$ confidence interval of a parameter were derived in such a way that if the experiment were repeated many times and the confidence interval is computed each time then $(1-\alpha)100\%$ of these intervals are expected to contain the true value of the parameter and the remaining $\alpha 100\%$ are not expected to contain it.

Definition: Confidence interval

- The fraction in a 100 % confidence interval estimate is called the ***confidence coefficient***, and the endpoints are called the ***lower*** and ***upper confidence limits***.

- The ***length of the interval*** is defined as the difference between the upper and lower confidence limits.

Considerations in Setting α

- a smaller α means more possible samples of size n will yield intervals that contain the true value of the parameter

- a larger α means shorter intervals

Common Choices of α: 0.01, 0.05, 0.10

Confidence Interval for the parameter, μ

Assumption: we will assume that we have a random sample (X_1, X_2, …, X_n) of size n taken from a normal population with mean, μ, and variance, σ^2, unless otherwise specified.

$(1-\alpha)$100% Confidence Interval for the Population Mean

- Case 1: when σ is known

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

where $z_{\alpha/2}$ is the 100(1- α/2)th percentile of the Standard Normal distribution. The length of this interval is given by $2z_{\alpha/2}\sigma/\sqrt{n}$

Common values of $z_{\alpha/2}$

α	0.01	0.05	0.10
$z_{\alpha/2}$	2.576	1.96	1.645

EXAMPLE:

- The mean for the quality point indices (QPI) of a random sample of 36 students is 2.6. Find the 99% CI for the mean. Assume the population standard deviation to be 0.3.

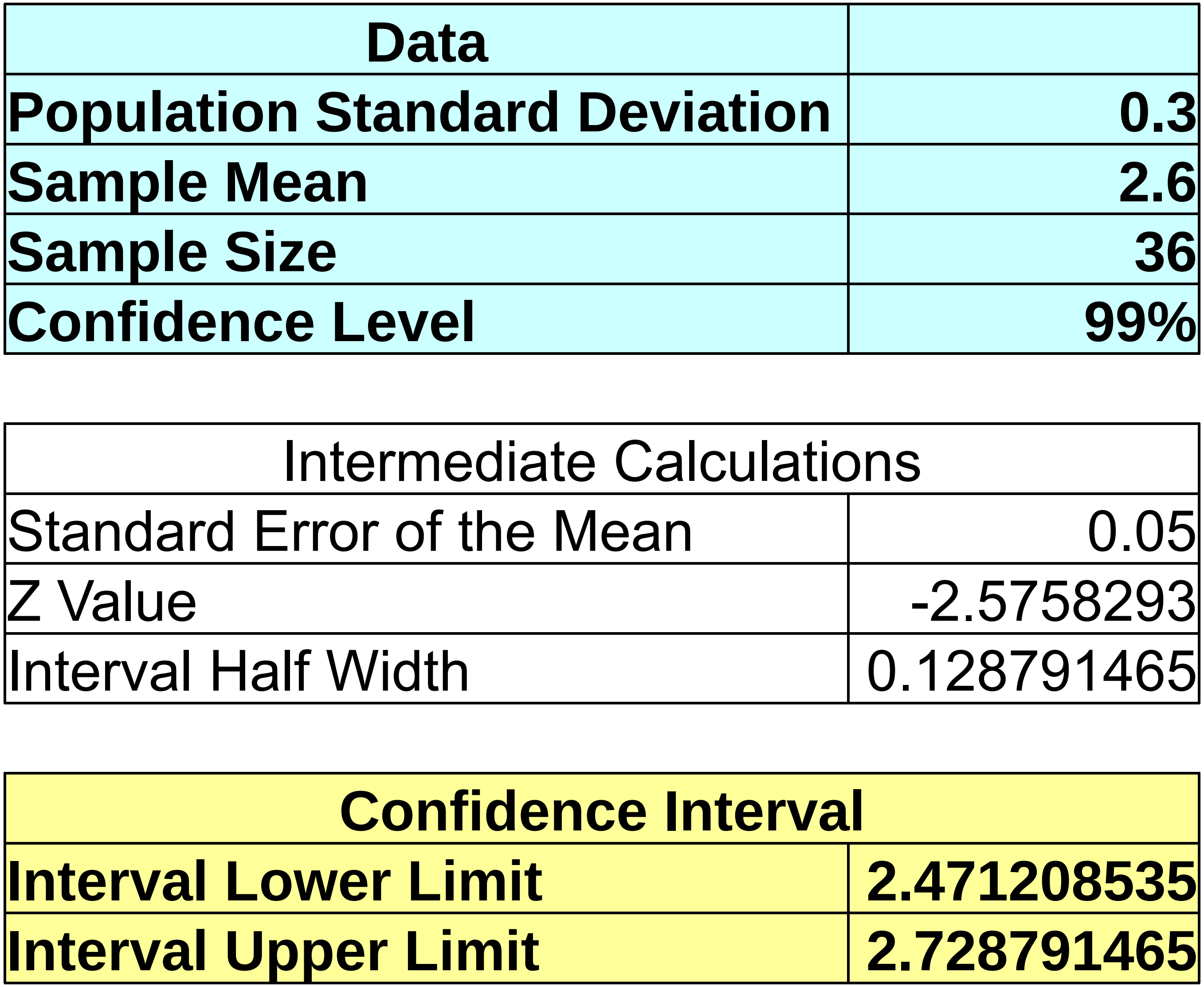
Soln: 99% CI for the mean

$$2.6 - 2.576\frac{0.3}{\sqrt{36}} \leq \mu \leq 2.6 + 2.576\frac{0.3}{\sqrt{36}}$$

or equivalently, 2.47 to 2.73

Confidence Interval Estimate for the Mean

USING PHStat



Data	
Population Standard Deviation	**0.3**
Sample Mean	**2.6**
Sample Size	**36**
Confidence Level	**99%**

Intermediate Calculations	
Standard Error of the Mean	0.05
Z Value	-2.5758293
Interval Half Width	0.128791465

Confidence Interval	
Interval Lower Limit	**2.471208535**
Interval Upper Limit	**2.728791465**

- Case 2: when σ is unknown

$$\left(\bar{x} - t_{(\alpha/2,v=n-1)}\frac{s}{\sqrt{n}}, \bar{x} + t_{(\alpha/2,v=n-1)}\frac{s}{\sqrt{n}}\right)$$

where $t_{(\alpha/2,v=n-1)}$ is the 100(1- ɑ)th percentile of the t-distribution with v=n-1 degrees of freedom.

What is the length of the interval?

Remarks:

- The above formulas hold strictly if the population from where the sample was taken is Normally distributed. However, they provide good approximate (1-α)100% confidence intervals when the distribution is not Normal provided the sample size is large, that is, n>30.

EXAMPLE

The contents of 8 similar bottles of acetic acid are 110, 112, 111, 109, 107, 113, 110, and 109 milliliters. Find a 95% confidence interval for the mean of all such bottles, assuming an approximate normal distribution for the population of the acetic acid contents.

Solution:

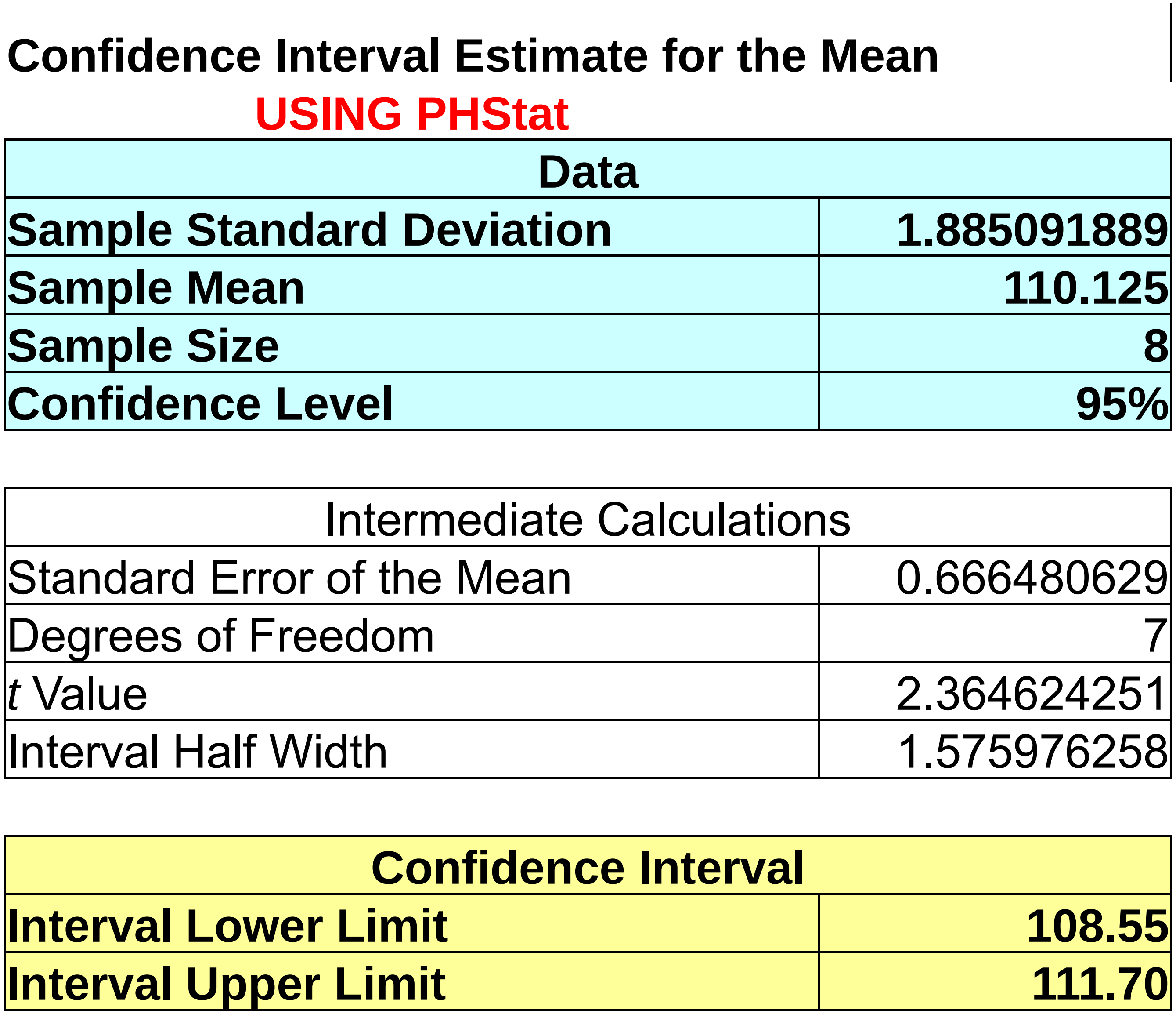
$$\bar{x} = 881/8 = 110.125$$

$$s = \sqrt{3.554} = 1.89$$

Thus the 95% confidence interval is:

$$110.125 \pm 2.365 \frac{1.89}{\sqrt{8}}$$

or, equivalently, from 108.545 to 111.705.



Confidence Interval Estimate for the Mean

USING PHStat

Data	
Sample Standard Deviation	**1.885091889**
Sample Mean	**110.125**
Sample Size	**8**
Confidence Level	**95%**

Intermediate Calculations	
Standard Error of the Mean	0.666480629
Degrees of Freedom	7
t Value	2.364624251
Interval Half Width	1.575976258

Confidence Interval	
Interval Lower Limit	**108.55**
Interval Upper Limit	**111.70**

Case 3: If σ is unknown and n>30, use

$$\left(\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

Example:

Forty-seven sedentary men were studied over a one-year period. These men tried to lose weight through exercise alone. The results of the study showed that the mean weight loss of these men was 0.7 kilograms with a standard deviation of 4.8. Find the 95% confidence interval for the population mean and interpret the result.

DEFINITION: MARGIN OF ERROR

- The **margin of *error***, denoted by *e*, is the upper bound on the absolute difference between the estimator and the parameter called the **error of estimation**

Notes on Margin of Error

- Clearly, the margin of error is . $e = \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}$

- Note that we can also view the margin of error as one-half of the length of the interval.

Notes on Margin of Error

- For this reason, some define the margin of error as: *estimate ± margin of error*, in the confidence interval estimate .

- Some researchers report the margin of error without mentioning the size of the associated risk, α. In such cases, it is understood that α=0.05.

Confidence Interval for the Proportion

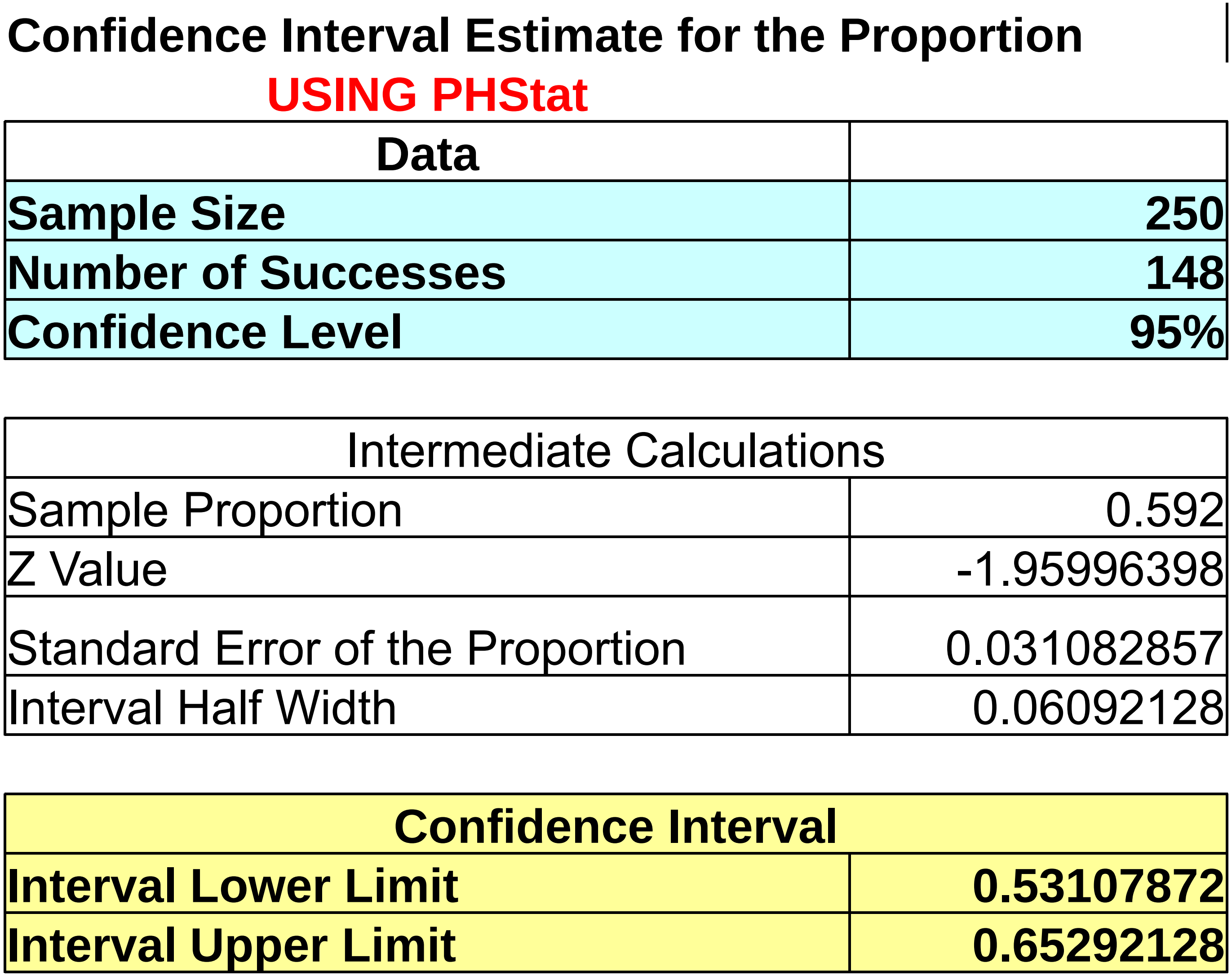
- If the population proportion is not expected to be too close to 0 or 1 and the sample size n is large, then an approximate 100(1-α)% confidence interval estimator for the population proportion, p, is given by:

$$\hat{p} \mp z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

EXAMPLE:

- In a random sample of 250 persons who took the civil service exam, 148 passed the exam. Find an approximate 99% confidence interval for the population proportion of persons who passed the exam.



Confidence Interval Estimate for the Proportion

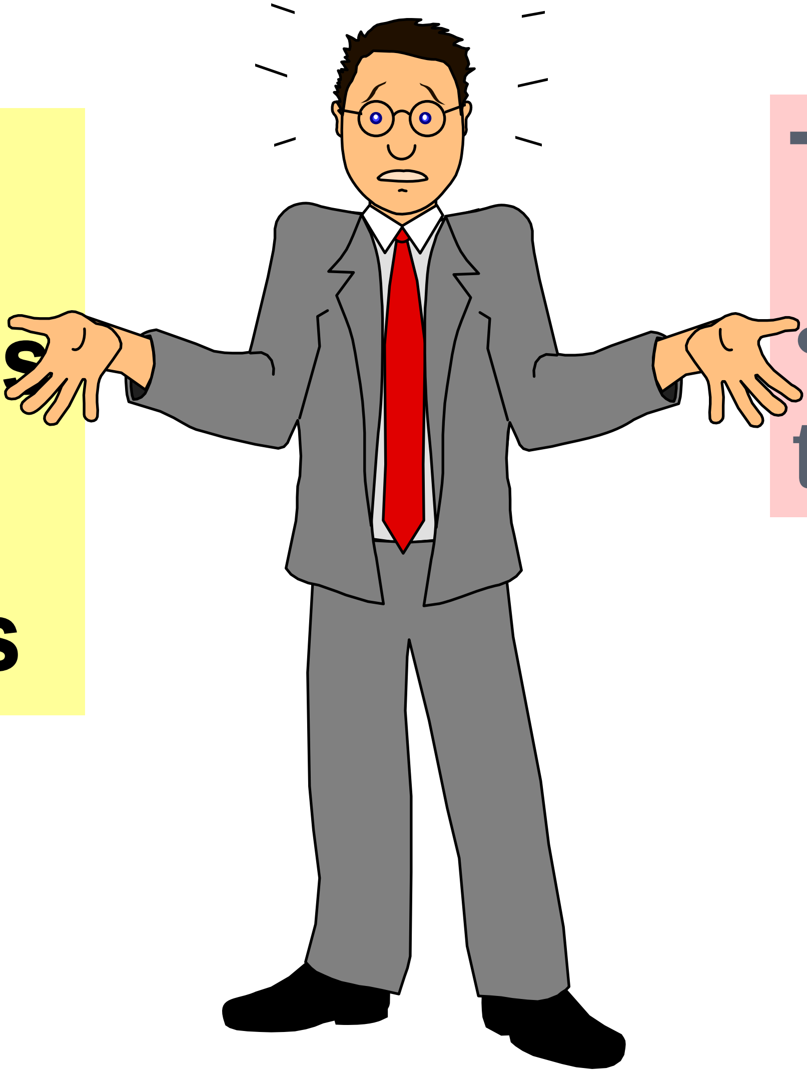
USING PHStat

Data	
Sample Size	**250**
Number of Successes	**148**
Confidence Level	**95%**

Intermediate Calculations	
Sample Proportion	0.592
Z Value	-1.95996398
Standard Error of the Proportion	0.031082857
Interval Half Width	0.06092128

Confidence Interval	
Interval Lower Limit	**0.53107872**
Interval Upper Limit	**0.65292128**

Determining Sample Size (Cost)



Too Big:

- **Requires too much resources**

Too small:

- Won’t do the job

Determining Sample Size for Mean

What sample size is needed to be 90% confident of being correct within ± 5? A pilot study suggested that the standard deviation s is 45.

$$n = \frac{Z^2\sigma^2}{\text{Error}^2} = \frac{1.645^2\left(45^2\right)}{5^2} = \underset{\uparrow}{219.2} \cong 220$$

Round Up

EXAMPLE:

- An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with a standard deviation of 40 hours. How large a sample is needed if we wish to be 95% confident that the sample mean will be within 10 hours of the true mean?

Sample Size Formula for Population proportion, P

- $n = z^2/4e^2$

- If you wanted to estimate the proportion of defective roads to within ± 0.005 with 95% confidence, how many roads would you have to test?

EXERCISES

1. An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a standard deviation of 40 hours. If a random sample of 30 bulbs has an average life of 780 hours,

   a) Find a 95% confidence interval for the population mean of all bulbs produced by this firm.

   b) How large a sample is needed if we wish to be 95% confident that our sample mean will be within 10 hours of the true mean?

EXERCISES

2. A machine is producing metal pieces that are cylindrical in shape. A sample of pieces is taken and the diameters are 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, and 1.03 centimeters.

   a) Find a 99% confidence interval for the mean diameter of pieces from this machine, assuming an approximate normal distribution.

   b) How large a sample is needed if we wish to be 99% confident that the sample mean will be within 0.05 centimeters of the true mean?

EXERCISES

3. In a random sample of 1000 homes in a certain city, it is found that 228 are heated by oil.

   a) Find the 99% confidence interval for the proportion of homes in this city that are heated by oil.

   b) How large a sample is needed if we wish to be 99% confident that our sample proportion will be within 0.5 of the true proportion of homes in this city that are heated by oil?

4. a) Compute a 95% confidence interval for the proportion of defective items in a process when it is found that a sample size of 100 yields 8 defectives.

   b) How large a sample is needed if we wish to be 95% confidence that our sample proportion will be within 0.05 of the true proportion of defectives?

Thank You