5: Inference About a Mean: Estimation

Introduction to Statistical Inference | Confidence Interval for μ, σ Known | Student's *t* Distribution
Confidence Interval for μ, σ Estimated | Sample Size Requirements

Introduction to Statistical Inference

Statistical inference is the act of generalizing from a sample to a population with calculated degree of certainty. The two primary forms of statistical inference are:

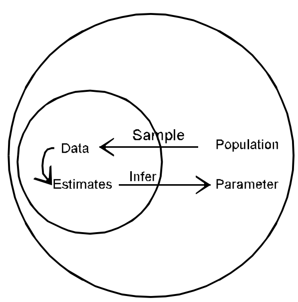
- Estimation
- Hypothesis testing

Estimation provides the most likely location of a population parameter, often with a built-in "margin of error." **Hypothesis testing** provides a way to judge the non-chance occurrence of a finding. Examples will help illustrate each:

Suppose you want to learn about the prevalence of a condition in a population—smoking for instance—based on the prevalence of the condition in a sample. IN a given study, th efinal inference may be "25% of the adult population smokes" (**point estimation**). Alternatively, the infernce might be "between 20% and 30% of the population smokes" (**interval estimation**). Finally, the epidemiologists might simply want to test whether the prevalence of smoking has changed over time (**hypothesis testing**). This chapter introduces estimation; the next chapter introduces hypothesis testing.

Parameters and Estimates

Regardless of the type of statistical inference you pursue, it is important to distinguish between the parameter you are trying to learn about the estimate used to infer it. The **parameter** a the numeric characteristic of the population you want to learn about. The **estimates** is a numerical characteristic of the sample that you have. **Although the two are related, they are not interchangeable.**



For example, a population mean is a parameter. You might then use a sample mean as the estimate for this parameter. To clearly distinguish these two mean, we refer to them with different symbols: μ and $\bar{x}$. In general, we will use **Greek characters to denote parameters** and Roman characters to denote estimates.

Sampling Distributions of Means

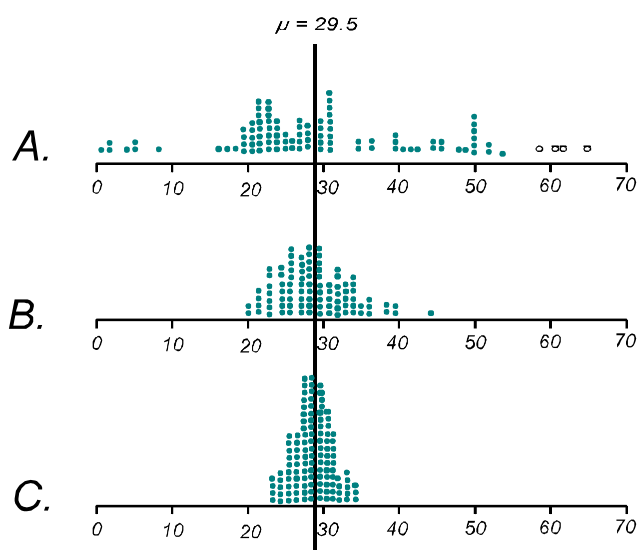
The **key** to understanding statistical inference is in viewing any single sample mean as an example of a mean from the population. This idea forms the basis of what is known as a **sampling distribution** of means. (All statistics, not just means, have sampling distributions. For now, let us focus on sampling distributions of means.)

A **sampling distribution of a mean** is the *hypothetical* frequency distribution of sample means that would occur from repeated, independent samples of size *n* taken from the population. It takes imagination to understand the concept of a sampling distribution. (Imagination? Statistics? Yes!) We imagine taking all possible samples of size *n* from a given population. We then take the mean of these many samples, and the means are arranged to form a distribution. This is what we mean by a sampling distribution.

Some students confuse this hypothetical sampl*ing* distributions of means with the distribution of the actual data. To avoid this confusion, let us consider three experiments based in a population of $N = 600$. The mean of the population (μ) is 29.5 with a standard deviation (σ) of 13.6. Three separate **sampling experiments** are considered.

- **Sampling experiment A** takes 100 samples of $n = 1$ from the population. We calculate the mean of each sample. (The mean of each sample is the value of the data point itself.)
- **Sampling experiment B** takes 100 samples of $n = 10$ from the population. We calculate the mean of each sample.
- **Sampling experiment C** takes 100 samples of $n = 30$ from the population. We calculate the mean of each sample.

We plot each batch of sample means and find:



The above distributions are **inceptions** of [three separate] sampling distributions of means. The true sampling distribution would have all possible samples from the population, but this would take too long to show.

The experiments sampling experiments demonstrate three important **sampling phenomena**:

(1) With increasing sample size, the sampling distribution becomes increasingly bell-shaped. This phenomenon is known as the **central limit theorem**. The central limit theorem states the sampling distribution of means tends toward normality as *n* becomes large. This justifies use of statistical procedures based on normal distribution theory even when working with non-normal data.

(2) Each distribution is centered on population mean μ. We say "the expected value of the sample mean *is* the population mean" or "the sample mean is an **unbiased (valid) estimator** of the population mean."

(3) With increasing sample size the sampling distribution tends to cluster more tightly around the population mean. The larger the sample, the more likely the sample mean precisely reflects the population mean. Precision of the sample mean can be quantified in terms of the standard deviation of the sampling distribution, which is now called the **standard error of the mean**, or ***SEM*** for short.

When the population standard deviation (σ) is known, the standard error of the mean is:

$$SEM = \frac{s}{\sqrt{n}} \qquad \textbf{(5.1)}$$

For example, the standard error of the mean for a variable with σ = 13.586 based on *n* = 10 has *SEM* = 13.586 / √10 = 4.296 .

When the population standard deviation is not known, it is estimated by sample standard deviation *s* and the standard error of the mean (now denoted ***sem***) is:

$$sem = \frac{s}{\sqrt{n}} \qquad \textbf{(5.2)}$$

For example, for a sample with a standard deviation (*s*) of 16 and sample size of 10 has *sem* = 16/√10 = 5.060.

Notes about the standard error of the mean:

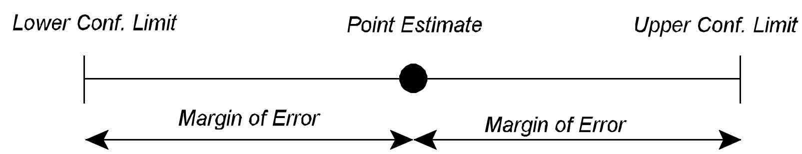
- The standard error of the mean quantifies the **precision** of the sample mean as an estimator of the population mean.
- The precision of the sample mean is inversely proportion to $\sqrt{n}$. We refer to as the **square root law**. For example. quadrupling the sample size will double the precision of the estimate.
- Since the sampling distribution of the mean tends toward normality, 68% of sample means will fall within ±1 standard error of the population mean and **95% will fall within ±2 standard errors of the**

population mean.

Confidence Interval for μ (σ Known)

95% CI for μ

Point estimators are formulas that provide single value that help locate the parameter. For example, $\bar{X}$ is the point estimator of μ. However, point estimates provide no information about the precision of the estimate. To quantify precision, we surround the **point estimate** with a **margin of error** so that there is an **upper confidence limit** and **lower confidence limit**:



This interval is called a **confidence interval**. Confidence intervals are usually constructed at the 95% level of confidence, although other levels of confidence are possible (see next page).

Because sampling distributions of means tend toward normality, 95% of sample means will lie within 1.96 standard errors of μ. Therefore, $\Pr(\mu - (1.96)(SEM) < \bar{x} < \mu + (1.96)(SEM)) = .95$. From algebra, it follows $\Pr(\bar{x} + (1.96)(SEM) > \mu > \bar{x} - (1.96)(SEM)) = .95$. This can be written:

$$\bar{x} \pm (1.96)(SEM) \qquad \textbf{(5.3)}$$

where $\bar{x}$ represents the sample mean and *SEM* represents the standard error of the mean (formula 5.1).

Illustrative example. Suppose a sample mean is 29.0. The sample has of $n = 10$. We know ahead of time $\sigma = 13.586$. Therefore, $SEM = \frac{13.586}{\sqrt{10}} = 4.30$ and a 95% confidence interval for μ is

= 29.0 ± (1.96)(4.30)
= 29.0 ± 8.4
= (20.6, 37.4)

Interpretation. A 95% confidence interval for a mean is constructed so that 95% of "like intervals" will capture μ. This is what we mean by "95% confident." Therefore, we are 95% confident the above interval of 20.6 – 37.4 captures μ (the parameter). Another good way to think of this is "the sample mean of 29.0 has a margin of error of ± 8.4.

Other Levels of Confidence

Let α represent the chance the researcher is willing to take of *not* capturing μ. We will refer to **$(1-\alpha)100\%$ confidence intervals**. For example, when $\alpha = .05$, $1-\alpha = .95$. When $\alpha = .10$, confidence $(1-\alpha) = .90$. When $\alpha = .01$, $1-\alpha = .99$.

A $(1-\alpha)100\%$ confidence interval for μ is given by:

$$\bar{x} \pm (z_{1-\alpha/2})(SEM) \tag{5.4}$$

where $\bar{x}$ represents the sample mean, $z_{1-\alpha/2}$ represents the $(1-\alpha/2)100^{th}$ percentile on a standard normal curve (use z table), and *SEM* represents the standard error of the mean as calculated by formula 5.1.

Illustrative Example (90% confidence interval). Using the same data as above, we want a 90% confidence interval for μ. For 90% confidence, use $\alpha = .10$ and **confidence coefficient for *z*** is $z_{1-\alpha/2} = z_{1-.10/2} = z_{.95} = 1.645$. Thus, a 90% confidence interval for μ is

$= 29.0 \pm (1.645)(4.30)$
$= 29.0 \pm 7.1$
$= (21.9, 36.1)$

We are 90% confident this interval will capture μ.

Illustrative Example (99% confidence interval). Using the same data, we want to construct a 99% confidence interval. For 99% confidence, $\alpha = .01$ and the confidence coefficient for z is $z_{1-\alpha/2} = z_{1-.01/2} = z_{.995} = 2.58$. Thus, a 99% confidence interval for μ is

$= 29.0 \pm (2.58)(4.30)$
$= 29.0 \pm 11.1$
$= (17.9, 40.1)$.

We are 99% confident this interval will capture μ.

Notice the length of the confidence interval increases as we go from 90% to 95% to 99% confidence.

Student's *t* Distribution

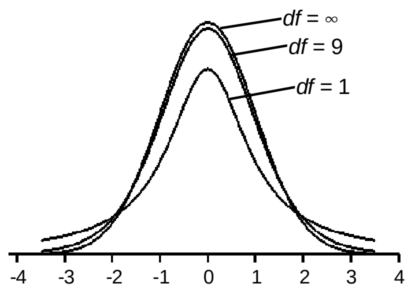
Confidence interval formulas based on normal distribution theor take the form:

(point estimate) ± (confidence coefficient)(standard error)

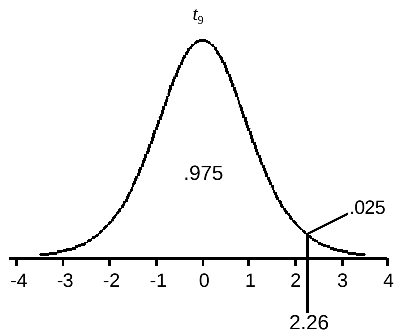
When σ is known, the confidence coefficient comes from the standard normal distribution. When σ is not known we use sample standard deviation *s* to estimate σ and adopt Student's *t* distribution to determine the confidence coefficient.

Student's *t* distribution is a family of symmetrical, unimodal, and bell-shaped distribution. They are slightly flatter with broader tails than normal distribution.

Student *t* distributions are identified by **degrees of freedom (df).** The *df* determines extent the broadness of the distribution's tails: *t* distributions with little degrees are flatter with broad tails. A *t* distribution becomes increasingly mesokurtotic as its *df* increases so that a *t* distribution with many degrees of freedom *is* essentially a standard normal (*Z*) distribution:



The next figure shows a *t* distribution with 9 degrees of freedom with its 97.5th percentile marked:



In this instance, the 97.5th percentile ("*t* value") is 2.26. Notice that the 97.5th percentile has a tail region of .025.

Notation. Let $t_{df,p}$ represent a *t* value with *df* degrees of freedom and a cumulative probability of *p*. For example, $t_{9,.975}$ = 2.26, is the 97.5th percentile on t_9. Values for *t* percentiles are found in *t* tables (**Appendix 2**).

Confidence Interval for μ (σ Estimated by *s*)

When the population standard deviation is *not* known, we use the *sample* standard deviation *s* as an estimate of σ and calculate a (1 − α)100% confidence interval for μ as:

$$\bar{x} \pm (t_{n-1,1-a/2})(sem) \qquad \textbf{(5.5)}$$

where $\bar{x}$ represents the sample mean, $t_{n\text{-}1,1\text{-}\alpha/2}$ represents the (1 − α/2)100% percentile on a *t* distribution with $n-1$ degrees of freedom, and *sem* represents the estimated standard error of the mean as calculated by formula 5.2.

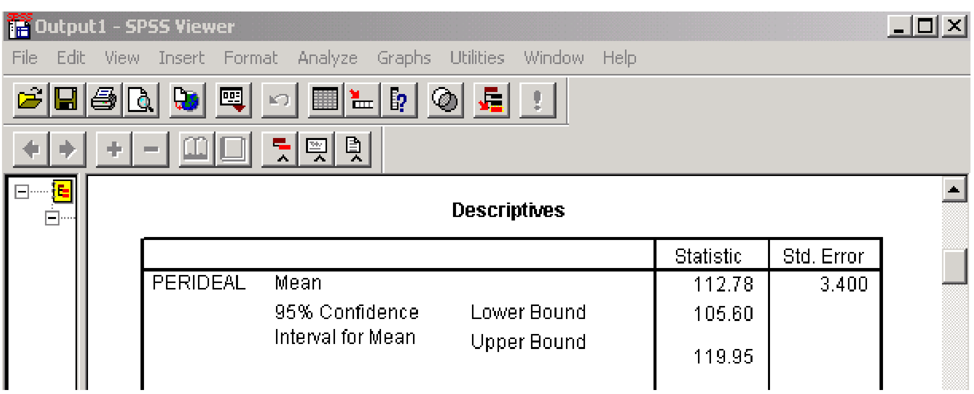
Illustrative Example (`%ideal.sav`). In this data set, body weight is reported as a percentage of ideal. Thus, 100 represents 100% of ideal body weight, 120% represents 20% over ideal, and so on. A sample of 18 people reveals $\bar{x}$ = 112.778 and *s* = 14.424. Thus, *sem* = $\frac{14.424}{\sqrt{18}}$ = 3.400 (by formula 5.2). For 95% confidence, α = .05. Thus, the we use a confidence coefficient of $t_{18\text{-}1,1\text{-}.05/2}$ = $t_{17,.975}$ = 2.11 (from *t* table).

The 95% CI for μ is

= 112.78 ± (2.11)(3.400)
= 112.78 ± 7.17
= (105.61, 119.95)

We are 95% confident the population mean lies between 105.61 and 120.95.

SPSS. SPSS calculates confidence intervals for means with its `Analyze > Summary Statistics > Explore` command. Output for the `%ideal.sav` illustrative example (above) is:



			Statistic	Std. Error
PERIDEAL	Mean		112.78	3.400
	95% Confidence Interval for Mean	Lower Bound	105.60	
		Upper Bound	119.95	

Sample Size Requirements

One of the questions a statistician often faces before doing a study is "How much data should be collected?" Collecting too much data is a waste of time, and collecting too little data renders a study too imprecise to be useful. Therefore, it is important to determine the sample size requirements of a study before collecting data.

To derive a reasonably precise estimate of a mean, let ***d*** represent the **margin of error**, or "wiggle room" around the sample mean when drawing a confidence interval. The confidence interval for $\mu = \bar{x} \pm d$ where $d \approx (2)(SEM)$. Since $SEM = \frac{s}{\sqrt{n}}$, $d \approx \frac{2s}{\sqrt{n}}$. Solving for *n*, we get

$$n = \frac{4s^2}{d^2} \qquad \textbf{(5.6)}$$

Illustrative examples: For example, to estimate μ with a margin of error of 5 for a variable with a standard deviation of 15, $n = \frac{(4)(15^2)}{5^2} = 36$. In contrast, estimating the same population mean with a margin of error of 2.5 requires $n = (4)(15^2)/(2.5)^2 = 144$.

Selected Vocabulary

This chapter introduces many new words and ideas. Here is a list of selected term and their definitions.

Sampling distribution of a mean: the hypothetical frequency distribution of sample means that would occur from repeated independent samples of size *n* from the population.

Central Limit Theorem: an axiom that states that distributions of sample means will tend toward normality, even if the initial variable is non-normal, especially when *n* is large.

Unbiased estimate: an estimate from a sample which has an expected value that is equal to the parameter it is trying to estimate. For example, $\bar{x}$ is an unbiased estimate of μ, because the expected value of the sample mean is the population.

Law of large numbers: the law of large numbers states that the larger the sample, the more likely it is to represeent the population from which it was drawn -- specifically, the more likely it is that the sample mean will equal the population mean.

Standard error of the mean (*SEM* or *sem*): a statistic that indicates how greatly a particular sample mean is likely to differ from the mean of the population. When the population standard deviation is known: $SEM = \sigma$ / sqrt(*n*) When the population standard deviation is not known, $sem = s$ / sqrt(*n*)

Margin of error (*d*): the "plus-or-minus wiggle-room" that the statistician draws around the estimate in order to locate the probable location of the parameter being estimated; half the confidence interval length.