Introductory Statistics: A Problem-Solving Approach

by Stephen Kokoska

Chapter 8

Confidence Intervals Based on a Single Sample

"I'm going to need a Margin of Error or I can't publish your prediction of six more weeks of winter."

Introduction

- A single value of a statistic computed from a sample conveys little information about confidence and reliability.
- Alternative method: Use a single value to construct an *interval* in which we are fairly certain the true value lies.
- A point estimate of a population parameter is a single number computed from a sample, which serves as a best guess for the parameter.
  - ✓ An **estimator** is a statistic of interest and, therefore, is a random variable. An estimator has a distribution, a mean, a variance, and a standard deviation. It is a rule used to produce a point estimate of a population parameter.
  - ✓ An **estimate** is a specific value of an estimator.

A statistic $\hat{\theta}$ is an **unbiased estimator** of a population parameter θ if $E(\hat{\theta}) = \theta$, that is, if the mean of $\hat{\theta}$ is θ.

If $E(\hat{\theta}) \neq \theta$, then the statistic $\hat{\theta}$ is a **biased estimator** of θ.

Some Unbiased Estimators

1. The sample mean $\overline{X}$ is an unbiased statistic for estimating the population mean μ because $\mathrm{E}(\overline{X}) = \mu$.

2. The sample proportion $\widehat{P}$ is an unbiased statistic for estimating the population proportion p because $\mathrm{E}(\widehat{P}) = p$.

3. The sample variance S^2 is an unbiased statistic for estimating the population variance σ^2 because $\mathrm{E}(S^2) = \sigma^2$.

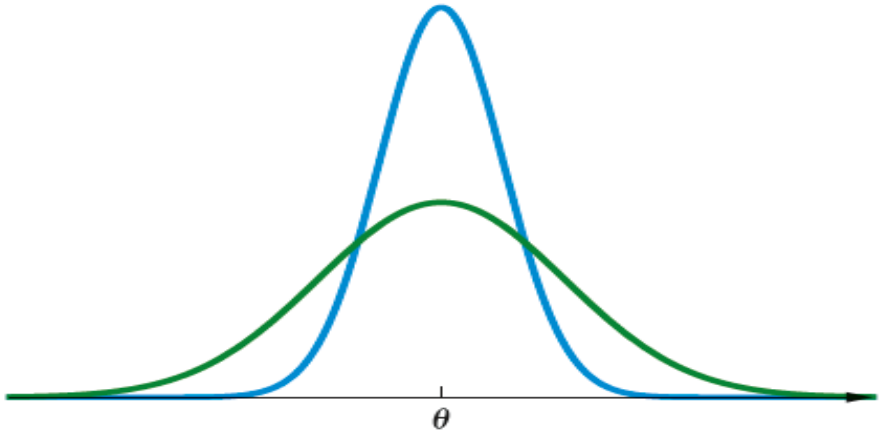
Even though S^2 is an unbiased estimator for σ^2 , the sample standard deviation S is a biased estimator for the population standard deviation σ.

$$\mathrm{E}[S] = \mathrm{E}\left(\sqrt{S^2}\right) \neq \sqrt{\mathrm{E}\left(S^2\right)} = \sqrt{\sigma^2} = \sigma$$

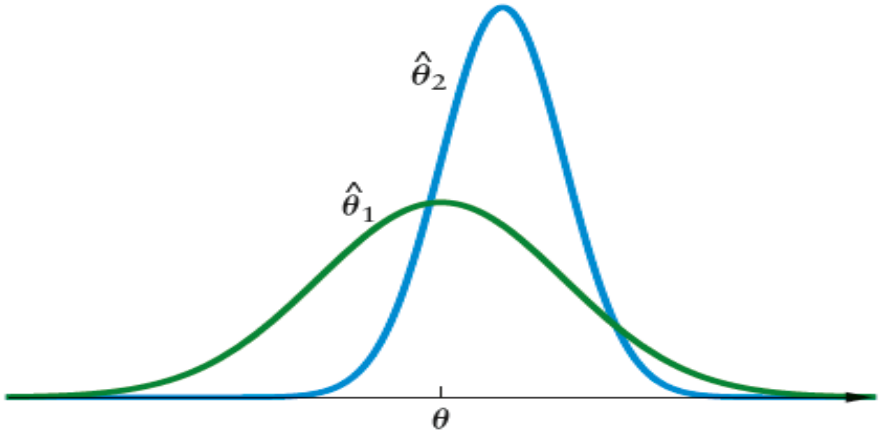
Even though S is a biased estimator for σ, it is still very important in statistical inference.

Rules for Choosing a Statistic

- The first rule for choosing a statistic is that the estimator should be unbiased.
- The second rule for choosing a statistic is that, of all unbiased statistics, the best is the one with the smallest variance.
- An unbiased statistic that has the smallest possible variance is called the **minimum-variance unbiased estimator** (MVUE).
- If the underlying population is normal, the sample mean $\bar{X}$ is the MVUE for estimating μ.



Sampling distributions of two unbiased statistics for estimating θ. Use the statistic with the smaller variance.



The statistic $\hat{\theta}_1$ is unbiased but has large variance; $\hat{\theta}_2$ is slightly biased but has small variance. The choice of an estimator is a difficult decision in this case, and there is no single, definitive right answer.

Confidence Interval

- A **confidence interval (CI)** for a population parameter is an interval of values constructed so that, with a specified degree of confidence, the value of the population parameter lies in this interval.
- The confidence coefficient is the probability that the confidence interval encloses the population parameter in repeated samplings. Typical confidence coefficients are 0.90, 0.95, and 0.99.
- The confidence level is the confidence coefficient expressed as a percentage. Typical confidence levels are therefore 90%, 95%, and 99%.

Critical Value

$Z_{\alpha/2}$ is a **critical value**. It is a value on the measurement axis in a **standard normal distribution** such that:

$$P(Z \geq Z_{\alpha/2}) = \alpha/2$$

Where, α = 0.01, 0.05, or 0.10

Typical Values for z:

Most common confidence levels:

C =	90%,	95%,	99%
z =	1.645,	1.960,	2.576

95% Confidence Interval for a Population Mean When σ is Known

1. $\bar{X} \sim \mathrm{N}\left(\mu, \sigma^2/n\right) \Rightarrow Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$

2. Use Table 3 to find a symmetric interval about 0 such that the probability that Z lies in this interval is 0.95.

$$\mathrm{P}(-1.96 < z < 1.96) = 0.95$$

3. Substitute for Z and rewrite the μ in the middle.

$$\mathrm{P}\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

$$\mathrm{P}\left(-1.96 \cdot \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\mathrm{P}\left(-\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\mathrm{P}\left(\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

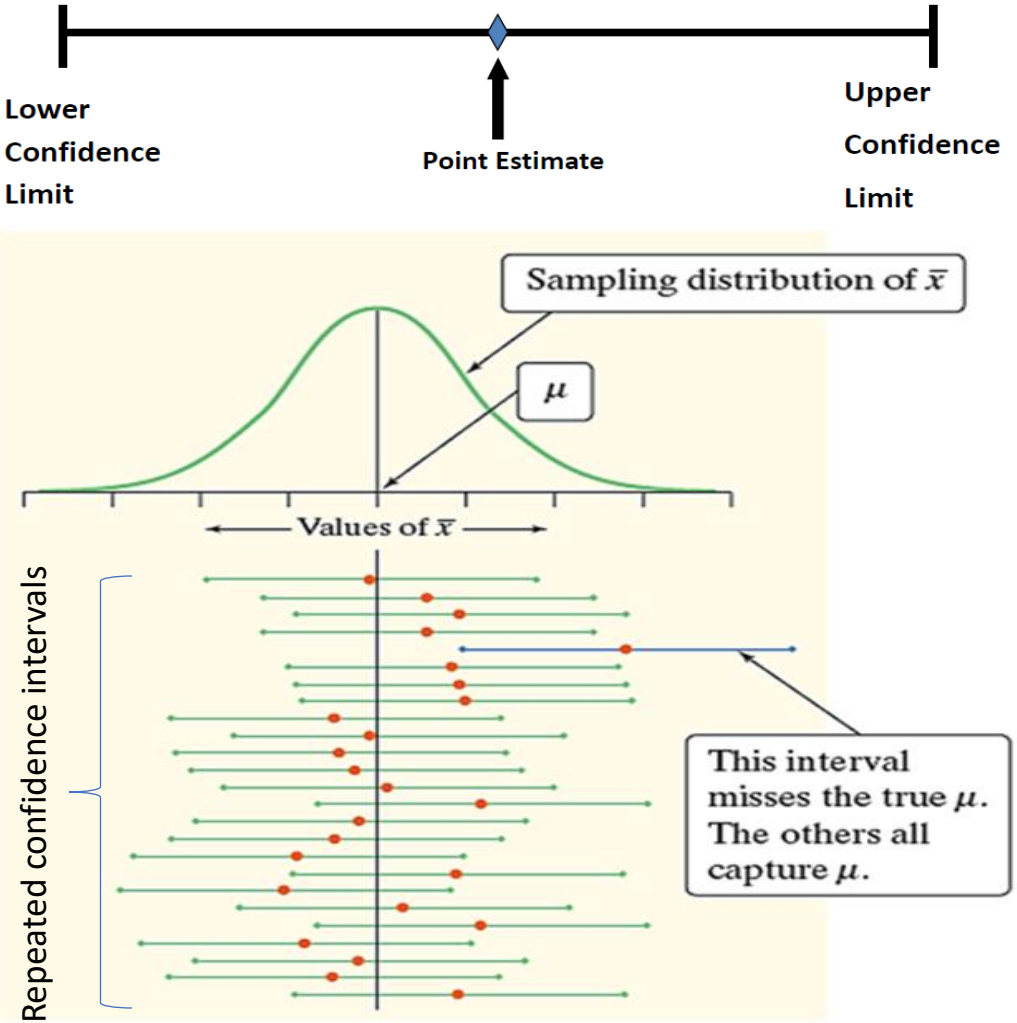
Finding a 100(1− α)% CI for μ When σ is Known

A confidence interval (C.I.) is an alternative to reporting a single value (point estimate) for the parameter being estimated is to calculate and report an entire interval (interval estimate) of plausible values.

Given a random sample of size *n* from a population with mean μ, if

1. The underlying population distribution is normal and/or *n* is large, and
2. The population standard deviation σ is known, then the endpoints for a 100(1−α)% confidence interval for μ have the values

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$



General Form of CI

- A confidence interval ("C.I.") can be expressed in a general form as:

$$\textbf{point estimate } (\hat{\theta}) \pm \textbf{margin of error } (m)$$

where, m = (critical value) x (standard deviation of statistic (point estimate))

- The margin (it is called Bound, B in your textbook) of error measures the maximum amount by which the sample results are expected to differ from those of the actual population.
- m expresses the maximum expected difference between the true population parameter and a sample estimate of that parameter.

Ideas to Remember

- The population parameter μ is fixed. The confidence interval varies from sample to sample.
- It is correct to say, "We are 95% confident that the interval captures the true mean μ."
- As the confidence coefficient increases (with σ and n constant), $z_{\alpha/2}$ increases, and the CI gets wider.

Example: Cell Phone Weight

Suppose the weight of a typical cell phone is normally distributed with a standard deviation of 18 g. In a random sample of 15 cell phones, the sample mean weight was 155.7 g . Find a 95% confidence interval for the true mean weight of cell phones.

$\bar{x} = 155.7,\ \sigma = 18$, and $n = 15$

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = \bar{x} \pm z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}$$

$$= 155.7 \pm (1.96) \cdot \frac{18}{\sqrt{15}}$$

= 155.7 ± 9.11

= (146.59, 164.81)

Necessary Sample Size

Selecting the Sample Size for C.I.

It is likely this calculated value for n will not be an integer. **Always** round up to the nearest whole number.

$$n = \left(\frac{\sigma \cdot z}{m}\right)^2 \qquad OR \qquad n = \left(\frac{s \cdot z}{m}\right)^2$$

Example1: You want to estimate the average wait time at the doctors' office. You wish to estimate within 4 minutes from the true average with 99% level of confidence. You were aware that the standard deviation for the population is 12 minutes. What minimum sample size would be required?

$$n = \left(\frac{z\sigma}{m}\right)^2 = \left(\frac{(2.576)(12)}{4}\right)^2 \approx 59.72$$

after rounding up, $n \geq 60$

Example2: A National Park Service researcher would like to find a 95% confidence interval for the mean height of the Castle Geyser eruption with a bound on the error of estimation of 5 ft. Previous experience suggests that the population standard deviation for the height is approximately 15 ft. How large a sample is necessary to achieve this accuracy?

$$n = \left(\frac{z\sigma}{m}\right)^2 = \left(\frac{(1.96)(15)}{5}\right)^2 \approx 34.57$$

The necessary sample size is $n \geq 35$

Finding a 100(1− α)% CI for μ When σ is Unknown

What if σ is unknown 

Replace the unknown population parameter, σ by its estimate, S, the sample standard deviation.

- If the distribution is normal and n is large ($n \geq 30$). This produces an *approximate* confidence interval for μ :

$$\overline{x} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$$

- If the distribution is normal and n is small ($n < 30$). This produces an *exact* confidence interval for μ:
  - ✓ S is no longer close to σ.
  - ✓ The variability in the distribution of Z arises from randomness in both ($\overline{x}$, S)
  - ✓ The probability distribution of $\frac{\overline{X} - \mu}{S}$ will be more spread out than the standard normal distribution.
  - ✓ The *one-sample t-statistic* is $T = \frac{\overline{X} - \mu}{S}$ has a t-distribution with *df* = *n-1*
  - ✓ So, a confidence interval for μ is given by

$$\overline{x} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

A Confidence Interval for a Population Mean When σ is Unknown

Confidence intervals for μ are based on a standard normal, or Z, distribution, and are valid only when σ is known (and either the underlying population is normal or the sample size is sufficiently large).

It is unrealistic to assume that the population standard deviation is always known.

Example: A college admissions director wishes to estimate the mean age of all students currently enrolled. In a random sample of 50 students, the mean age is found to be 22.9 years and the standard deviation is known to be 1.5 years, and the population is Normally distributed. Construct a 90% confidence interval of the population mean age.

$$\bar{x} \pm z \frac{S}{\sqrt{n}}$$

$$22.9 \pm 1.645\left(\frac{1.5}{\sqrt{50}}\right)$$

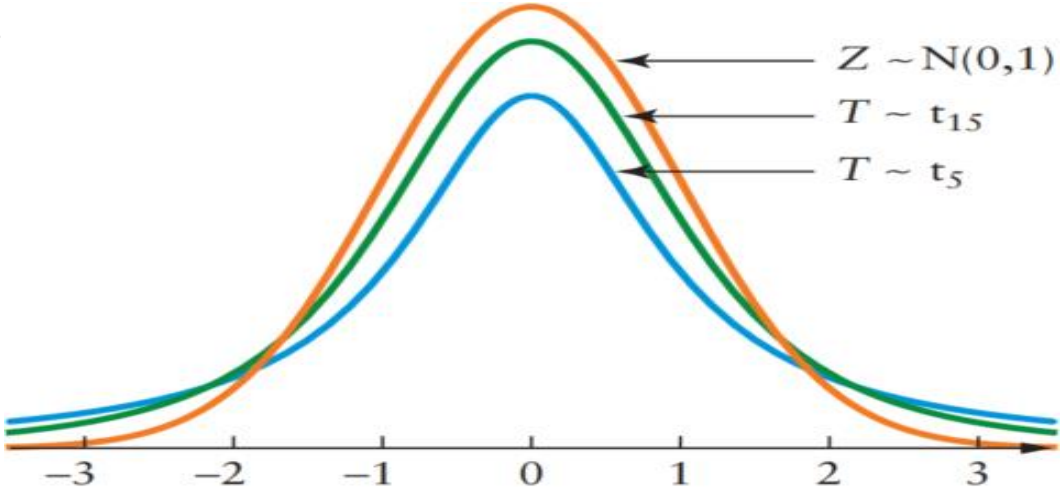
$$22.9 \pm 0.34896$$

$$(22.55,\ 23.25)$$

With 90% confidence, we can say that the mean age of **all** students is between 22.55 and 23.25 years.

T Random Variables

Comparison of Density Curves



Properties of a *t* Distribution:

1. A *t* distribution is completely determined by only one parameter *v*, called the number of degrees of freedom (*df*). There is a different *t* distribution corresponding to each value of *v*.
2. If *T* has a *t* distribution with *v* degrees of freedom, ($T \sim t_v$), then

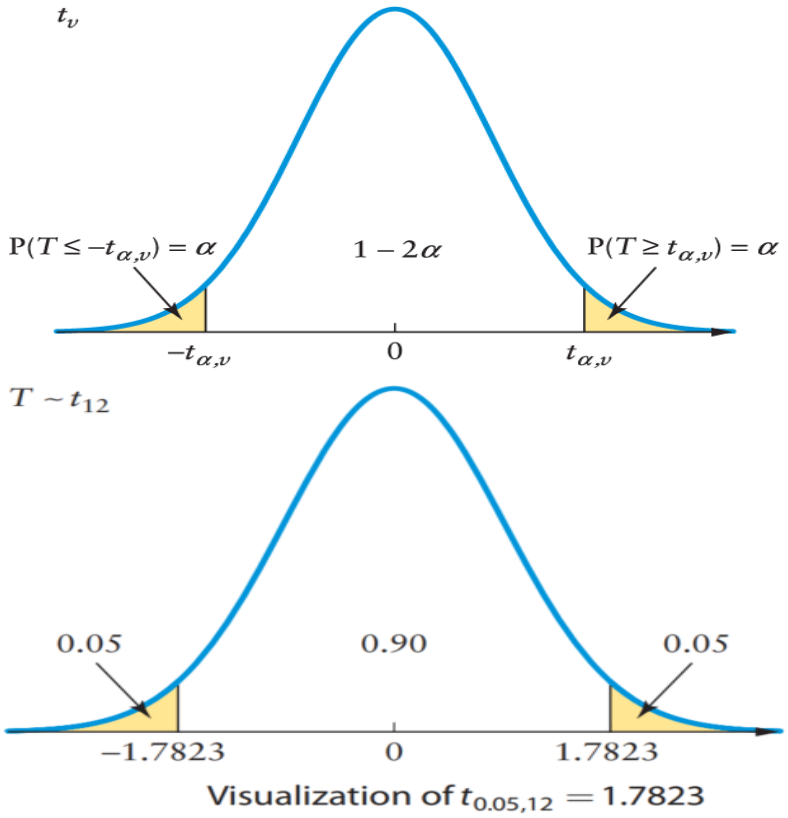
$$\mu_T = 0 \text{ and } \sigma_T^2 = \frac{v}{v-2} \quad (v \geq 3)$$

3. The density curve for every *t* distribution is bell-shaped and centered at 0, but more spread out than the density curve for a standard normal random variable *Z*. As *v* increases, the density curve for *T* becomes more compact and closer to the density curve for *Z*.

T Random Variables

- $t_{\alpha,v}$ is a **critical value** related to a t distribution with v degrees of freedom. If T has a t distribution with v degrees of freedom, then $P(T \geq t_{\alpha,v}) = \alpha$.
  - ✓ For any t distribution, $t_{\alpha,v}$ is simply a t-value such that α of the area (probability) lies to the right of $t_{\alpha,v}$. The negative critical value is $-t_{\alpha,v}$.
  - ✓ Critical values are always defined in terms of right-tail probability.
  - ✓ As v increases, the t critical values approach the corresponding Z critical values.

	α								
v	0.20	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	0.0001
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	0.8791	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869	5.6938
11	0.8755	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247	4.4370	5.4528
12	0.8726	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296	4.3178	5.2633
13	0.8702	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520	4.2208	5.1106
14	0.8681	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874	4.1405	4.9850
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮



Example: Orange Blossom Perfume

When heat is applied to a mixture, the substance that evaporates and is collected as it cools is called the distillate. Oil obtained from orange blossoms through distillation is used in perfume. Suppose the oil yield is normally distributed. In a random sample of 11 distillations, the sample mean oil yield was 980.2 g with standard deviation $s = 27.6$ g . Find a 95% confidence interval for the true mean oil yield per batch.

$$\bar{x} = 980.2, \quad s = 27.6, \quad \text{and } n = 11$$

$1 - \alpha = 0.95 \Rightarrow \alpha \Rightarrow 0.05 \Rightarrow \alpha/2 = 0.025$

$t_{\alpha/2,\, n-1} = t_{0.025,10} = 2.2281$

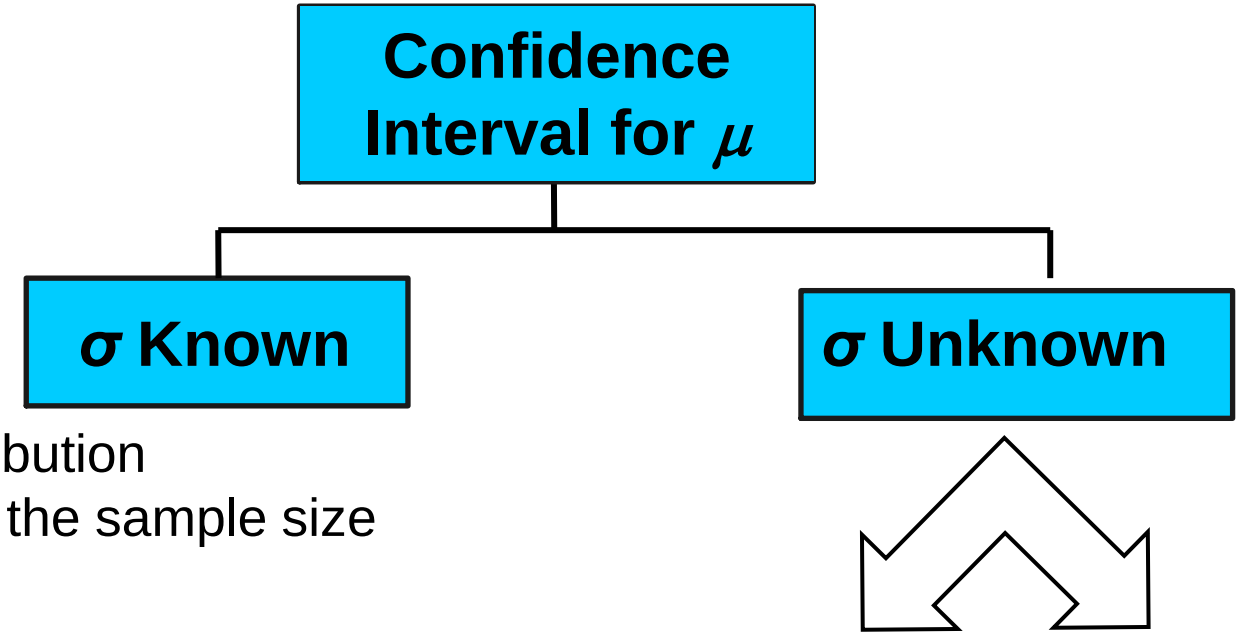
$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t_{0.025,10} \cdot \frac{s}{\sqrt{n}} = 980.2 \pm (2.2281) \cdot \frac{27.6}{\sqrt{11}}$$

$= 980.2 \pm 18.54$

$= (961.66, 998.74)$

Summary



use z-distribution
regardless the sample size

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

$n < 30$
use t-distribution

$$\bar{x} \pm t \frac{S}{\sqrt{n}}$$

$n \geq 30$
use z-distribution

$$\bar{x} \pm z \frac{S}{\sqrt{n}}$$

A Large-Sample CI for a Population Proportion

- Let *p* = true population proportion, the fraction of individuals or objects with a specific characteristic (the probability of a success).
- It is reasonable to use the sample proportion to construct a CI for *p*.
- In a sample of n objects, let X be the number of successes in the sample.

$$\hat{P} = \frac{X}{n} = \frac{\#\text{ individuals with the characteristic}}{\text{sample size}}$$

- From Ch.7, if *n* is large and both $np \geq 5$ and $n(1-p) \geq 5$, then the random variable $\hat{P}$ is approximately normal with mean *p* and variance $p(1-p)/n$. Since $\hat{P}$ is approximately normal, we can standardize to obtain:

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathrm{N}(0,1)$$

- A 100(1 – α)% confidence interval for the true proportion *p* is

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example: Magnesium Levels

Recent research suggests that many Americans have deficient magnesium levels. Vitamin D, which increases calcium and phosphate levels, cannot be metabolized without magnesium, but this study suggests that many Americans are consuming only half the recommended daily allowance of magnesium. A random sample of 1200 Americans was obtained, and each was tested for magnesium level. In total, 540 were found to have deficient magnesium levels.

a. Find a 95% confidence interval for the true proportion of Americans who have deficient magnesium levels.

b. The American Medical Association claims that 50% of Americans are magnesium deficient. Is there any evidence to suggest that the percentage of Americans who have deficient magnesium levels is different from this value?

Check the nonskewness criterion to confirm that the distribution of $\hat{P}$ is approximately normal.

$$\hat{p} = \frac{x}{n} = \frac{540}{1200} = 0.45$$

$$n\hat{p} = (1200)(0.45) = 540 \geq 5$$

$$n(1 - \hat{p}) = (1200)(0.55) = 660 \geq 5$$

Example: Magnesium Levels

a. Find a 95% confidence interval for the true proportion of Americans who have deficient magnesium levels.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.45 \pm 1.96\sqrt{\frac{(0.45)(1-0.45)}{1200}}$$

$$= 0.45 \pm 0.0281$$

$$= (0.4219, 0.4781)$$

A 95% confidence interval for the true proportion of Americans who are magnesium deficient is: (0.4219, 0.4781)

b. The American Medical Association claims that 50% of Americans are magnesium deficient. Is there any evidence to suggest that the percentage of Americans who have deficient magnesium levels is different from this value?

Using the usual four-step inference procedure:

Claim: $p = 0.5$

Experiment: $\hat{p} = 0.45$

Likelihood: The likelihood is a 95% confidence interval, an interval of likely values for p: (0.4219, 0.4781).

Conclusion: The claimed value, 0.50, does not lie in this confidence interval. There is evidence to suggest that p is different from (less than) 0.50.

Necessary Sample Size

The margin of error (the bound on the error estimation:

$$m = z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

After solving for *n*, the formula of sample size is:

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{m}\right)^2$$

We do not know $\hat{P}$ until we have the sample size *n* (and *x*). There are two possible solutions to this problem:

1. Use a reasonable estimate for the sample proportion based on previous experience.
2. If no prior information is available, use 0.5 as the estimate of the sample proportion. This produces a very conservative, large value of *n*.

Example: Negative Equity

During the housing crisis, more than 30% of homeowners owed lenders more than the actual value of their homes, meaning they had negative equity. To help banks set interest rates, a company plans to estimate the proportion of American homeowners with negative equity. A 95% confidence interval for *p* with bound on the error of estimation of 0.02 is needed. How large a sample size is necessary in each of the following cases?

a) Prior experience suggests $\hat{p} \approx 0.2$.

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{m}\right)^2$$

$1-\alpha = 0.95 \Rightarrow \alpha/2 = 0.05 \Rightarrow z_{0.025} = 1.96$

$$= 0.2(0.8)\left(\frac{1.96}{0.02}\right)^2 = 576.24$$

The necessary sample size is $n \geq 577$

b) There is no prior information about the value of $\hat{p}$.

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{m}\right)^2$$

$$= 0.5(0.5)\left(\frac{1.96}{0.02}\right)^2 = 2401$$

The necessary sample size is $n \geq 2401$