Topics:

1. Identifying and Estimating the Target Parameter
2. Confidence Interval for a Population Mean: Normal (*z*) Statistic
3. Confidence Interval for a Population Mean: Student's *t*-Statistic
4. Large-Sample Confidence Interval for a Population Proportion
5. Determining the Sample Size

Learning Objectives:

1. Estimate a population parameter (means, proportion, or variance) based on a large sample selected from the population
2. Use the sampling distribution of a statistic to form a confidence interval for the population parameter
3. Show how to select the proper sample size for estimating a population parameter

6.1 Identifying and Estimating the Target Parameter

Target parameters - NOTATION:

μ - population mean
σ^2 - population variance
p - population proportion

Introductory concepts (review)

Parameter – a numerical feature of a population

Target Parameter: population mean, population proportion, population variance – any parameter we are interested in estimating

Statistic is any numerical measure calculated from data: the proportion, mean, median, range, variance, standard deviation, etc.
Statistical inference: a method that converts the information from random samples into reliable estimates of the population parameters.

A point estimate: a single number calculated from a sample that can be regarded as an educated guess for an unknown population parameter.

A point estimator of a population parameter is a rule or formula that tells us how to use the sample data to calculate a *single* number that can be used as an *estimate* of the target parameter

Goal: Use the sampling distribution of a statistic to estimate the value of a population parameter with a known degree of certainty.

Determining the Target Parameter		
Parameter	**Key Words or Phrases**	**Type of Data**
μ	Mean; average	Quantitative
p	Proportion; percentage; fraction; rate	Qualitative

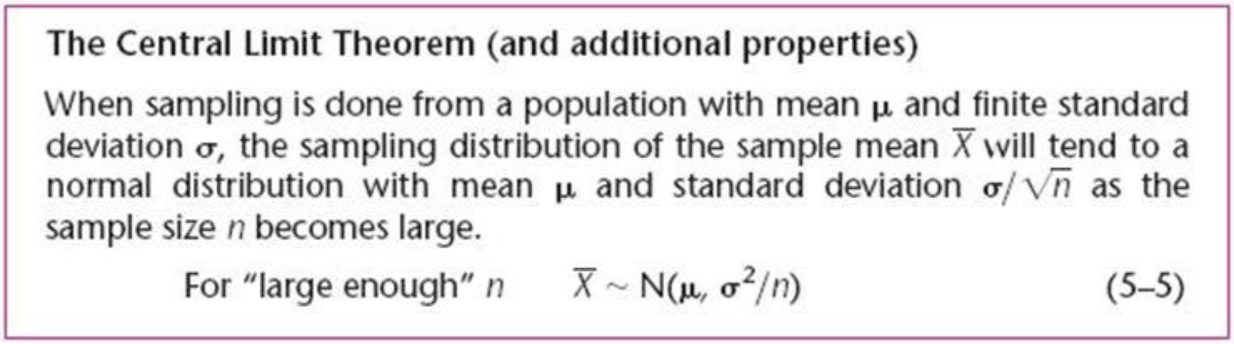
	Notation: Parameter	Estimator
Proportion	p	$\hat{p}$
Mean	μ	$\overline{X}$
Variance	σ^2	s^2

When using a sample statistic to estimate a population parameter, some statistics are good in the sense that they target the population parameter and are therefore likely to yield good results. Such statistics are called ***unbiased estimators***. **Sample mean, sample variance and sample proportion** are the examples of unbiased estimators.

Example 1: The sample mean $\overline{X}$ is an estimator of the population mean μ. The observed (computed) value $\bar{x} = 4$ is called a point estimate of μ

An **interval estimator** (or **confidence interval**) is a range of numbers that contain the target parameter with a high degree of confidence.

Recall from the last chapter



The Central Limit Theorem (and additional properties)

When sampling is done from a population with mean μ and finite standard deviation σ, the sampling distribution of the sample mean $\overline{X}$ will tend to a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as the sample size n becomes large.

For "large enough" n $\qquad \overline{X} \sim N(\mu, \sigma^2/n)$ $\qquad$ (5–5)

By the rule of thumb, n=30 is "large enough" to justify the normality of the distribution of sample means.

6.2 Confidence Interval for a Population Mean: Normal (*z*) Statistic

For an approximately normal distribution we expect 95% of all data to stay within 2 standard deviations from the mean.

When using the calculator, the more accurate number of standard deviations separating 95% of central data from those lowest and greatest is 1.96 (check!) Forming an interval of the numbers within 1.96 standard deviations from the mean we build a **95% confidence interval**:

$$\bar{x} \pm 1.96\sigma_{\bar{x}} = \bar{x} \pm \frac{1.96\sigma}{\sqrt{n}}$$

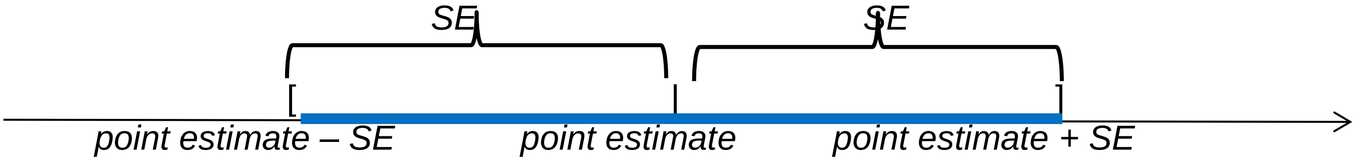
We form a 95% interval with the endpoints at the points on data line located 1.96 standard deviations below the sample mean and 1.96 standard deviations above the mean.

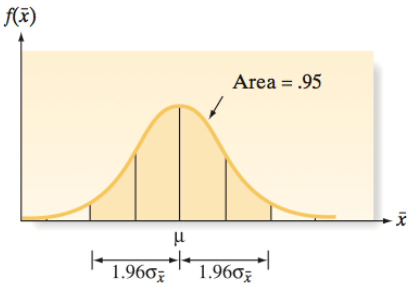
The expression $1.96\sigma/\sqrt{n}$ is an example of a **sampling error (margin of error).**

Confidence intervals introduced in your textbook has often the form

point estimate ± margin of error

This means that the interval has the endpoints built as follows:
[*point estimate – margin of error, point estimate + margin of error*]



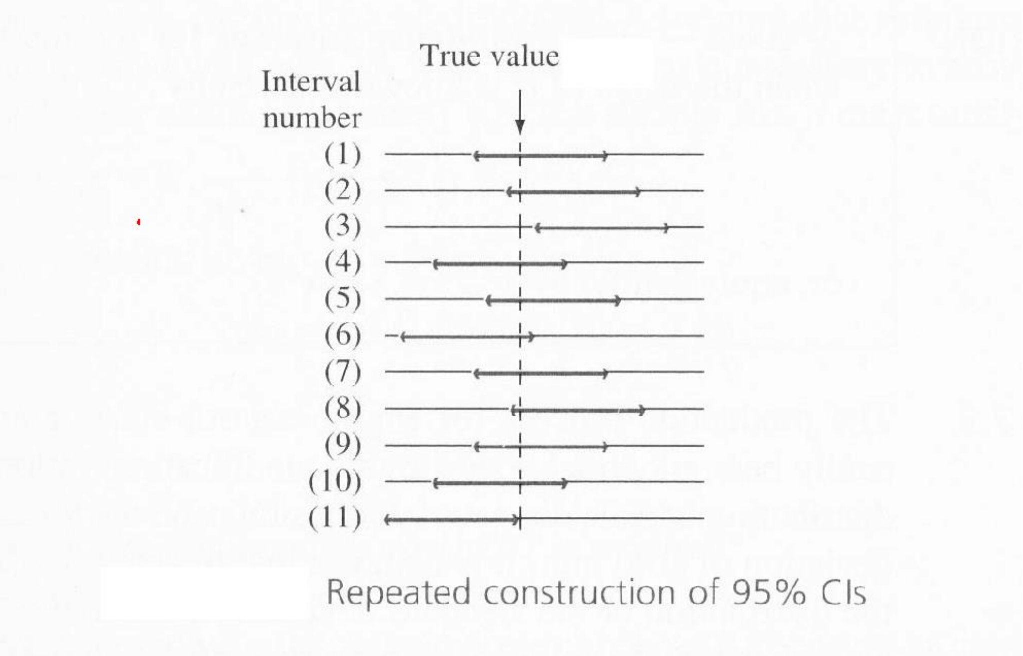


In our case, the area under the normal curve between these two boundaries
($\bar{x}$-1.96σ/sqrt(n), $\bar{x}$+1.96σ/sqrt(n))
is exactly .95.
Thus, the probability that a randomly selected interval will contain μ is equal to 0.95.

Interpretation of a 95% confidence interval:
We are 95% confident that the population mean of this distribution is between… and ….

Confidence coefficient is the probability that a confidence interval constructed from a random sample contains the target population parameter (or, the relative frequency with which similarly constructed intervals enclose the population parameter when the estimator is used repeatedly a very large number of times).

The confidence level is the confidence coefficient expressed as a percentage:

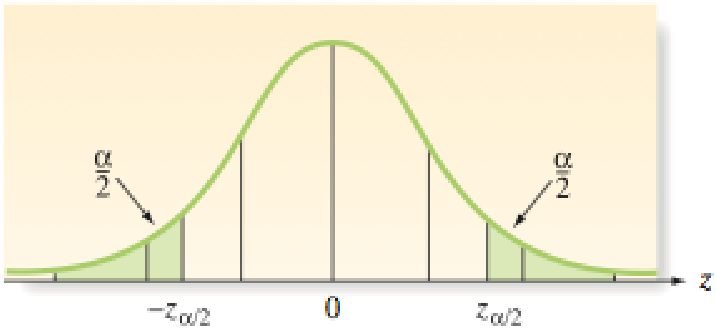


If our confidence level is 95%, then in the long run, 95% of our confidence intervals will contain µ and 5% will not.

We can select any confidence level we like. Typical sizes are below.

Notation: The confidence level is expressed in percent and marked 100(1-α)%, so α = (1 – confidence coefficient)

Example: Confidence Level	Confidence Coefficient α	tail size α/2
99% (.99)	.01	0.005
98% (.98)	.02	0.01
95% (.95)	.05	0.025
90% (.90)	.10	0.05



Critical Value

A critical value is a number of standard deviations separating likely location of the population parameter from the unlikely locations on a number line representing all data of standard distribution. For instance, z=1.96 is the number of standard deviations separating 2.5% of the highest data under normal distribution.

We will deal with two standard distributions: normal and t-distribution.

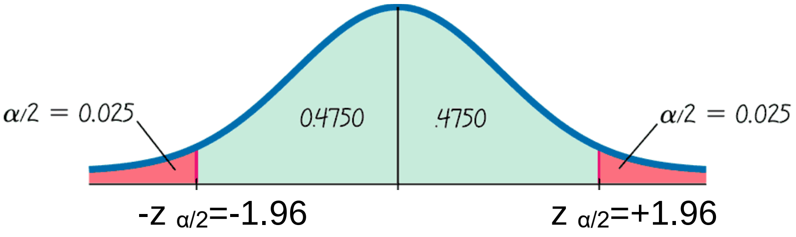
For critical value in confidence intervals we will use the notation $z_{\alpha/2}$ representing standard normal distribution. (The $t_{\alpha/2}$ represents t-distribution which will be introduced later)

$\boldsymbol{z_{\alpha/2}}$ = critical value for the standard normal distribution (z-distribution)

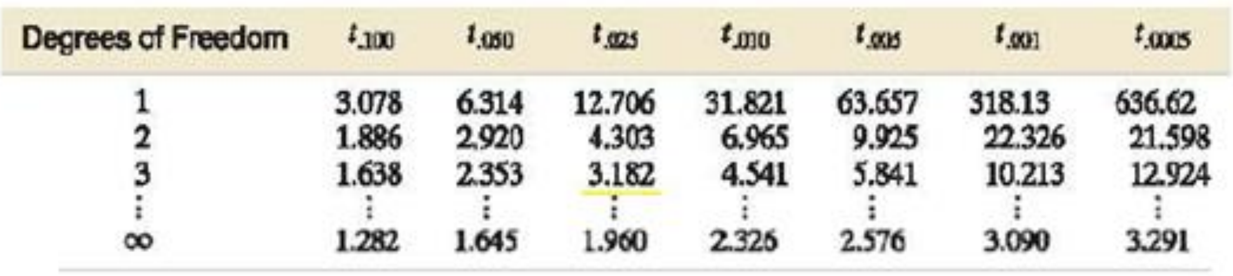
Example: Find critical value $z_{\alpha/2}$ separating 95% of most likely scores from the remaining 5% of the least likely scores under Standard Normal curve.

.

Solution: On the illustration below we see that 95% of the most likely scores are represented by green area under the curve. Five percent of least likely scores are represented by two red areas, "the tails".



Critical values for *z* distribution can be found in the last row (marked ∞) of Table Crit. Values of t: (in the end of the text).



Degrees of Freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$	$t_{.0005}$
1	3.078	6.314	12.706	31.821	63.657	318.13	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	21.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Exercise: Finding **Commonly Used Critical Values** $z_{\alpha/2}$ for (1-α) Confidence Level:
Confidence level 90%, α=10%, α/2=5%, and $z_{\alpha/2}$=.............
Confidence level 80%, α=20%, α/2=10%, $z_{\alpha/2}$=.............

Confidence level 99%, α=1%, α/2=0.5%, $z_{\alpha/2}$=.............

Derivation of the formula for confidence intervals for the large sample:
By CLT, for large *n* $\bar{x}$ is approximately normal with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ is approx. standard normal; the z-scores of the most likely sample means will stay within $\pm z_{\propto/2}$; that is, $-z_{\alpha/2} \le \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \le z_{\alpha/2}$

Isolate μ:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

100(1-α)% CI for μ is $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

The above formula is valid under the following assumptions

- **Random sample**
- No assumption about distribution of target population
- **<u>Large sample size</u> (n ≥ 30)**

Large-Sample (1 – α)% Confidence Interval for μ

$$\bar{x} \pm (z_{\alpha/2})\sigma_{\bar{x}} = \bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

where $z_{\alpha/2}$ is the *z*-value with an area $\alpha/2$ to its right.
The parameter σ is the standard deviation of the sampled population.
n is the sample size.

Note: When σ is unknown and *n* is large (*n* > 30), the confidence interval is only approximately equal to

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

s is the sample standard deviation. For smaller samples the critical value z is not good enough. More exact answer can be given by using t-distribution (next section).

Exercise: for sample mean = 5, standard deviation = 2 and sample size = 100 find:

a) 80% CI
b) 95% CI
c) 99% CI

Compare the sizes of the intervals. What can be concluded?

Exercise: for sample mean = 5, standard deviation = 2 and confidence level 95%
Find:

a) CI for sample size = 100
b) CI for sample size = 200
c) CI for sample size = 1000

Compare the sizes of the intervals. What can be concluded?

Classwork:

6.2 What is the confidence level of each of the following confidence intervals for μ?

d. $\bar{x} \pm 1.282\left(\frac{\sigma}{\sqrt{n}}\right)$

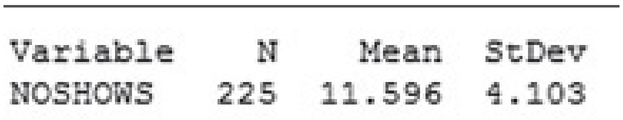
e. $\bar{x} \pm .99\left(\frac{\sigma}{\sqrt{n}}\right)$

6.3 A random sample of *n* measurements was selected from a population with unknown mean μ and known standard deviation σ. Calculate a 95% confidence interval for μ for each of the following situations:

a. $n = 75, \bar{x} = 28, \sigma^2 = 12$

b. $n = 200, \bar{x} = 102, \sigma^2 = 22$

Exercise:
Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its average number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected, and the number of unoccupied seats is noted for each of the sampled flights.
Use enclosed printout to estimate μ, the mean number of unoccupied seats per flight during the past year with a 90% confidence. Interpret the result.



```
Variable    N     Mean   StDev
NOSHOWS   225   11.596   4.103
```

6.13 Budget lapsing at army hospitals. Budget lapsing occurs when unspent funds do not carry over from one budgeting period to the next. Refer to the *Journal of Management*

Accounting Research (Vol. 19, 2007) study on budget lapsing at U.S. Army hospitals. Because budget lapsing often leads to a spike in expenditures at the end of the fiscal year, the researchers recorded expenses per full-time equivalent employee for each in a sample of 1,751 army hospitals. The sample yielded the following summary statistics: x ̅=$6,563 and s=$2,484. Estimate the mean expenses per full-time equivalent employee of all U.S. Army hospitals using a 90% confidence interval. Interpret the result.

Chapter 6.3 Confidence Interval for a Population Mean: Student t-Statistic

Key concepts: *t-statistic, t-distribution, degrees of freedom (df)*

Derivation of confidence interval for large sample was based on the fact, that if the sample size *n* is large and if we replace σ by *s,* then both statistics

$\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ and $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

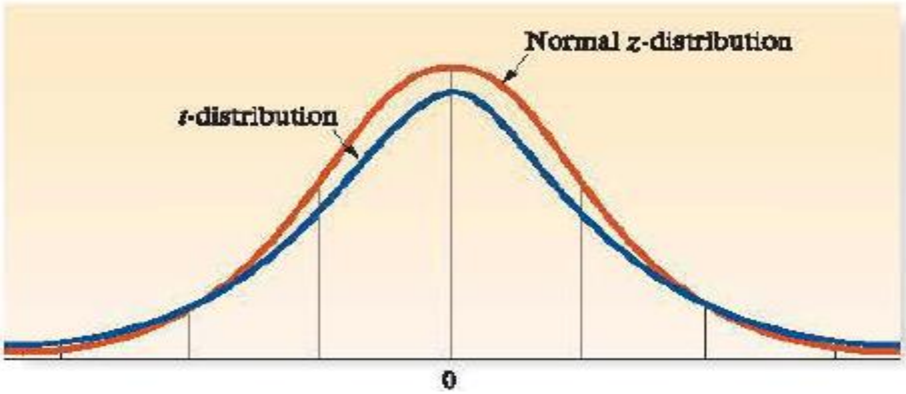
have approximately the same distribution (*z-distribution*). This is **not true** for small *n.*

t-Distribution. If we are sampling from normal population, then the sampling distribution of sample means for small samples is not exactly normal. The shape is also bell shape, but the thickness of the tails varies with sample size n.

the statistics (***t*-statistic)**

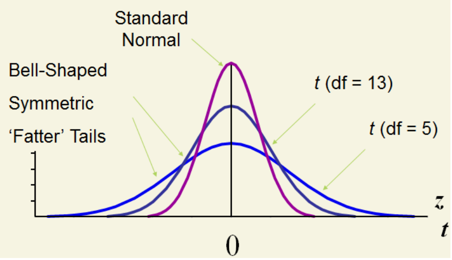
$$t=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

has so called ***t*-distribution** with *df* = *n*-1 **degrees of freedom.**

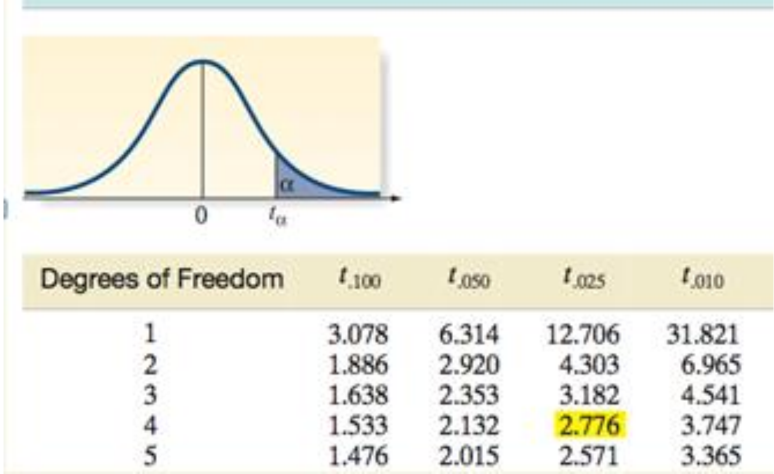


Properties:

- *t*-distribution depends on *df*, and hence on the sample size *n*, ($df = n - 1$)
- *t*-distribution is symmetric about 0,
- if $df \to \infty$, then the *t*-distribution approaches the standard normal distribution (i.e., *z*-distribution),
- critical values t_α of *t*-distribution are in Table (last pages of the text),
- critical values t_α of *t*-distribution are bigger than the corresponding z_α values



Example: find the critical value $t_{\alpha/2}$for 95% CI and n=5.
Solution:



Degrees of Freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$
1	3.078	6.314	12.706	31.821
2	1.886	2.920	4.303	6.965
3	1.638	2.353	3.182	4.541
4	1.533	2.132	2.776	3.747
5	1.476	2.015	2.571	3.365

Small Sample Confidence Interval for Population Mean

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

where $t_{a/2}$ is based on (n – 1) degrees of freedom.

Assumptions:

- Random sample (small)
- The distribution of target population is normal
- Population standard deviation σ is unknown
- No assumption about a sample size *n*

Example

You're a time study analyst in manufacturing. You've recorded the following task times (min.):

3.6, 4.2, 4.0, 3.5, 3.8, 3.1. What is the **90%** confidence interval estimate of the population **mean** task time? Assume normality of distribution of the times.

a) Solve by hand
b) Check with the calculator

Class Exercises

# 6.25 p. 317
Let t_0 be a particular value of t. Use Table to find t_0 values such that the following statements are true.

a. $P(-t_0 < t < t_0) = .90$ where n=11
b. $P(t \le t_0) = .05$ where n=16.

Exercise 1 [**6.28, p. 318**] The following sample of 16 measurements was selected from a population that is approximately normally distributed:
91 80 99 110 95 106 78 121 106 100 97 82 100 83 115 104

a. Compute the sample mean and the sample standard deviation

b. Construct an 80% confidence interval for the population mean by hand, and then repeat using TI-83

c. Construct a 95% confidence interval for the population mean and compare it with that of part b.

d. Carefully interpret each of the confidence intervals and explain why the 80% confidence interval is narrower.

e. What assumption is necessary to ensure the validity of this confidence interval?

Exercise To help consumers assess the risks they are taking, the Food and Drug Administration (FDA) publishes the amount of nicotine found in all commercial brands of cigarettes. A new cigarette has recently been marketed. The DA tests on this cigarette yielded mean nicotine content of 26.7 milligrams and standard deviation of 2.4 milligrams for a sample of 9 cigarettes. Construct a 98% confidence interval for the mean nicotine content of this brand of cigarette. What assumption do you have to make to solve the problem?

6.4 Large Sample Confidence Interval for the Population Proportion p

Suppose that *p* is an unknown population proportion of elements of certain type *S.* The estimator of p is the sample proportion

$$\hat{P} = \frac{x}{n}$$

where *x* is the number of elements of type S in the sample.

In Chapter 5 we studied sampling distribution of sample proportions $\hat{P}$

By CLT, **for large random samples (np≥15 and nq≥15),** the distribution is approximately normal with the mean p and standard deviation $\sqrt{pq/n}$

Large Sample (1-α)100% Confidence Interval for the Population Proportion p

Estimated parameter: population proportion: *p*

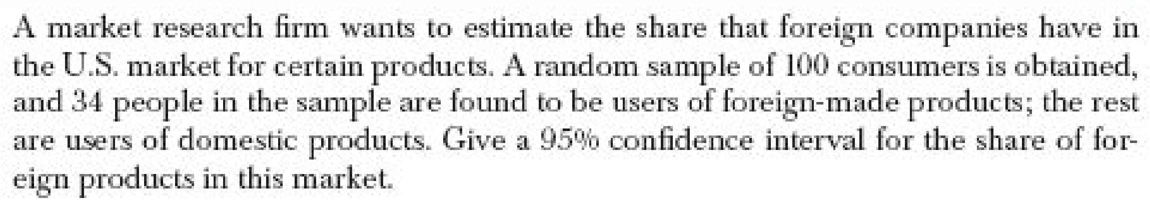
Assumptions: large sample size: (*n* $\hat{p}$ *≥ 15 and*) *n* $\hat{q}$ *≥ 15*
Random sample

(1-α)*100% Confidence Interval:

$$\boxed{\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}}$$

where $q = 1 – p$, $\hat{q} = 1 - \hat{p}$, and $z_{\alpha/2}$ is the **critical value** for the **standard normal distribution**

Example



A market research firm wants to estimate the share that foreign companies have in the U.S. market for certain products. A random sample of 100 consumers is obtained, and 34 people in the sample are found to be users of foreign-made products; the rest are users of domestic products. Give a 95% confidence interval for the share of foreign products in this market.

Solution:
First, check the assumptions:
Random sample, and
$n\hat{p} = 34 > 15$, $n\hat{q} = 66 > 15$ → *assumptions are satisfied*

$\hat{p} = 34/100 = 0.34$, $\hat{q} = 1-.0.34 = 0.66$
$1-\alpha = 0.95$, $\alpha = 0.05$, $\alpha/2 = 0.025$, $z_{\alpha/2} = z_{0.025} = 1.960$ *(the Table)*

Margin of error: $$z_{\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}} = 1.96\times\sqrt{\frac{0.34\times 0.66}{100}} = 0.093$$

95% Confidence Interval: 0.340 ± 0.093 or [0.247, 0.433]

Answer: *The share of foreign products in this market is between 24.7% and 43.3%*

TI-83:

- *Confidence interval for the population mean μ when σ is known*
  STAT → TESTS → 7: Z Interval…

- *Confidence interval for the population mean μ when σ is unknown*
  STAT → TESTS → 8: T Interval…

- *Large sample confidence interval for the population proportion p*
  STAT → TESTS → A: 1-PropZInt…

Example 2: Find an 80% confidence interval for μ given that n=25, x̄ = 122, and σ =20 (known)

Answer: (116.87, 127.13)

Example 3: Find an 80% confidence interval for μ given that n=25, x̄ = 122, and s =20
(σ unknown) *Answer: (116.73, 127.27)*

Example 4: Find a 95% confidence interval for p given that n=100, x = 34 (p̂ = 0.34)
Answer: (0.247, 0.433)

Note:
If p is very small or very close to 1.0, then in order to satisfy condition 2. we need to take a very large sample; for example if p = 0.001 then n must be at least ≥ 15/0.001 = 15,000.

*In the case when is very small or very close to 1, authors suggest an alternative procedure called **Wilson's adjusted confidence interval***

Adjusted (1 – α)100% Confidence Interval for a Population Proportion, *p*

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$ Where x=# of successes, n=sample size and $$\tilde{p} = \frac{x+2}{n+4}$$

Class Exercises:

6.42 A random sample of size n=121 yielded p^=.88.

a. Is the sample size large enough to use the methods of this section to construct a confidence interval for *p*? Explain.

b. Construct a 90% confidence interval for *p*.

c. What assumption is necessary to ensure the validity of this confidence interval?

6.50 Nannies who work for celebrities. The International Nanny Association reports that in a sample of 528 in-home child care providers (nannies), 20 work for either a nationally known, locally known, or internationally known celebrity (*2011 International Nanny Association Salary and Benefits Survey*). Use Wilson's adjustment to find a 95% confidence interval for the true proportion of all nannies who work for a celebrity. Interpret the resulting interval.

6.54 Interviewing candidates for a job. The costs associated with conducting interviews for a job opening have skyrocketed over the years. According to a Harris Interactive survey, 211 of 502 senior human resources executives at U.S. companies believe that their hiring managers are interviewing too many people to find qualified candidates for the job (*Business Wire*, June 8, 2006).

a. Describe the population of interest in this study.

b. Identify the population parameter of interest, *p.*

c. Is the sample size large enough to provide a reliable estimate of *p*?

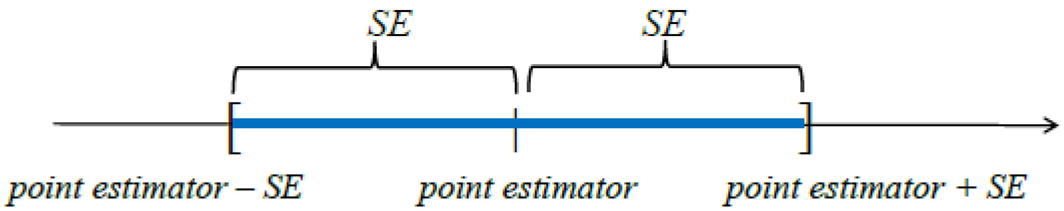
d. Find and interpret an interval estimate for the true proportion of senior human resources executives who believe that their hiring managers interview too many candidates during a job search. Use a confidence level of 98%.

e. If you had constructed a 90% confidence interval, would it be wider or narrower?

6.5 Determining the Sample Size

Recall that a confidence interval is of the form

point estimator ± margin of error, called by the author Sampling Error (SE)



Example If a confidence interval is [33.9, 35.1] or 34.5 ± 0.6, then margin of error = 0.6

(a half of the width of the interval)

Sampling Errors:

- Confidence interval for μ: $SE = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- Confidence interval for p: $SE = z_{\alpha/2}\sqrt{\frac{pq}{n}}$

Sample Size Determination for 100(1 – *α*) % Confidence Interval for *μ*

To estimate population mean with given sampling error, confidence level and known standard deviation, the required sample size can be found by the formula derived from the equation above for SE by isolating n:

$$n = \frac{z_{\alpha/2}^2\sigma^2}{SE^2}$$

Round always UP!

Example: The manufacturer wishes to estimate the mean inflation pressure to within .025 pound of its true value with a 99% confidence interval. The standard deviation of inflation pressure is about 0.1 (pound). What sample size should be used?

Note: If sigma is unknown, some researchers use sample standard deviation s, or even a quarter of the range instead.

Sample Size Determination for 100(1 – α) % Confidence Interval for *p*

In order to estimate *p* with a sampling error SE and with 100(1 – α)% confidence, the required sample size is found by the formula derived from the equation above for SE by isolating n:

$$\boldsymbol{n = \frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{SE^2}}$$

Use p-hat and q-hat from pilot study or from other research.
If sample proportions are not available, use (conservatively) 0.5 for both.
(Why “conservatively 0.5”?)

Example: Suppose a candidate wants to estimate voter support *p* within 3% with 95% confidence. How large sample does she need,

a) if she knows from pilot study that $\hat{p}$ is about .6 ?

b) if she has no idea about *p*?

Classwork:

Exercise
A 98% confidence interval for *p* is [0.23, 0.37].

a. What is p-hat?

b. What is sampling error SE?

c. What is *n*?

Exercise [6.60, p. 331].
If you wish to estimate a population mean with a margin of error of ME=.3 using a 95% confidence interval, and you know from prior sampling that σ2 is approximately equal to 7.2, how many observations would have to be included in your sample?

Exercise [6.62] In each case, find the approximate sample size required to construct a 95% confidence interval for *p* that has sampling error of SE = .08.

a. Assume *p* is near .2.
b. Assume you have no prior knowledge about *p,* but you wish to be certain that your sample is large enough to achieve the specified accuracy for the estimate.

One More Exercise

The countries of Europe report that 46% of the labor force is female. The United Nations wonders if the percentage of females in the labor force is the same in the United States. Representatives from the United States Department of Labor plan to check a random sample of over 10,000 employment records on file to estimate a percentage of females in the United States labor force.

a). The representatives from the Department of Labor want to estimate a percentage of females in the United States labor force to within ±5%, with 90% confidence. How many employment records should they sample?

b) They actually select a random sample of 525 employment records, and find that 229 of the people are females. Create the confidence interval. Show steps: find standard error and margin of error, then write the interval in the interval notation) *Find Standard Error, Critical value and Margin of error, then Confidence Interval.*

c) Should the representatives from the Department of Labor conclude that the percentage of females in their labor force is lower than Europe's rate of 46%? Explain.