SCHOOL OF MEDICINE AND SURGERY

The point and interval estimates

Continuous variables

DIPARTIMENTO DI MEDICINA E CHIRURGIA

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Preview

This lecture presents the beginning of inferential statistics.

- The two major activities of inferential statistics are (1) to use sample data to estimate values of a population parameters, and (2) to test hypotheses or claims made about population parameters.
- We introduce methods for estimating values of these important population parameters: proportions and means.
- We also present methods for determining sample sizes necessary to estimate those parameters.

SCHOOL OF MEDICINE AND SURGERY

Estimating a Population Mean

Key Concept

This lecture presents methods for estimating a population mean. In addition to knowing the values of the sample data or statistics, we must also know the value of the population standard deviation, σ.

Here are three key concepts that should be learned in this section:

Key Concept

1. We should know that the sample mean $\bar{x}$ is the best point estimate of the population mean μ.
2. We should learn how to use sample data to construct a confidence interval for estimating the value of a population mean, and we should know how to interpret such confidence intervals.
3. We should develop the ability to determine the sample size necessary to estimate a population mean.

Point Estimate of the Population Mean

The sample mean $\bar{x}$ is the best point estimate of the population mean μ.

Example:

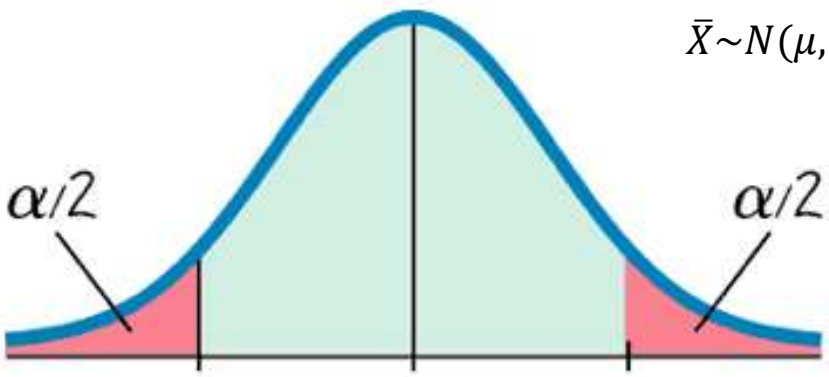
In order to estimate the mean body temperature a sample of 106 subjects was recruited. Suppose that the standard deviation is known as 0.62°F. The resulting data ended up in a mean of 98.20°F.
Can you estimate mean body temperature?

Because the sample mean is the best point estimate of the population mean, we conclude that the best point estimate of is 98.20°F.

How reliable is this estimate ?

From Central Limit Theorem …

.. the sampling distribution of the mean of a random variable with mean μ and standard deviation σ can be approximated by a normal distribution with mean μ and standard deviation σ/√n



$$z_{\alpha/2} = \frac{x_{\alpha/2} - \mu}{\sigma/\sqrt{n}}$$

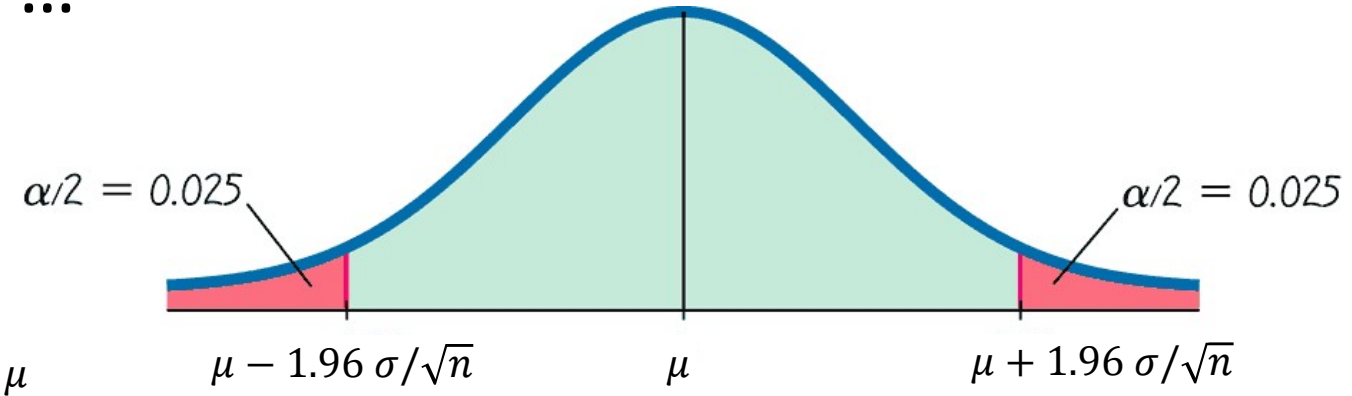
$$P\left(\mu - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

From Central Limit Theorem …

.. the sampling distribution of the mean of a random variable with mean μ and standard deviation σ can be approximated by a normal distribution with mean μ and standard deviation σ/√n

If α=0.05 …

$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$



$$1.96 = \frac{x_{0.05/2} - \mu}{\sigma/\sqrt{n}}$$

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

From the probability interval…

$$P\left(\mu - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

With simple algebraic passages on the two equations, we can express the interval in terms of the observed value

to the confidence interval…

$$P\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Definition

A confidence interval (or interval estimate) is a range (or an interval) of values used to estimate the true value of a population parameter. A confidence interval is sometimes abbreviated as CI.

Definition

A confidence level is the probability 1 – α (often expressed as the equivalent percentage value) that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times. (The confidence level is also called degree of confidence, or the confidence coefficient.)

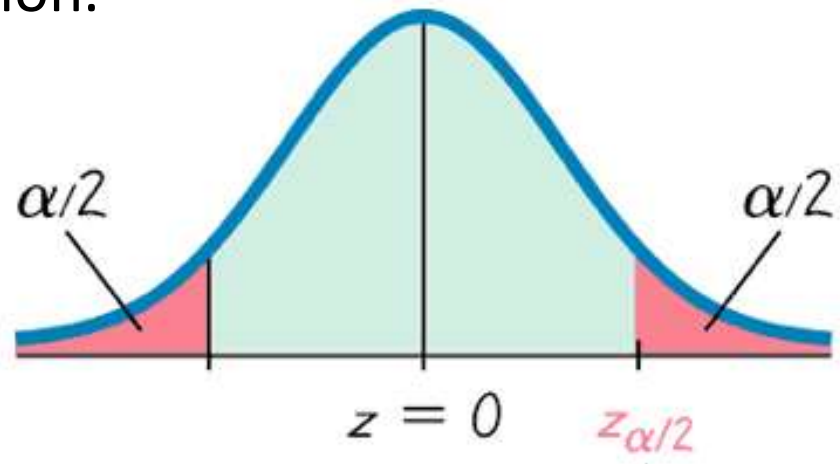
Most common choices are 90%, 95%, or 99%.

(α = 10%), (α = 5%), (α = 1%)

Critical Values

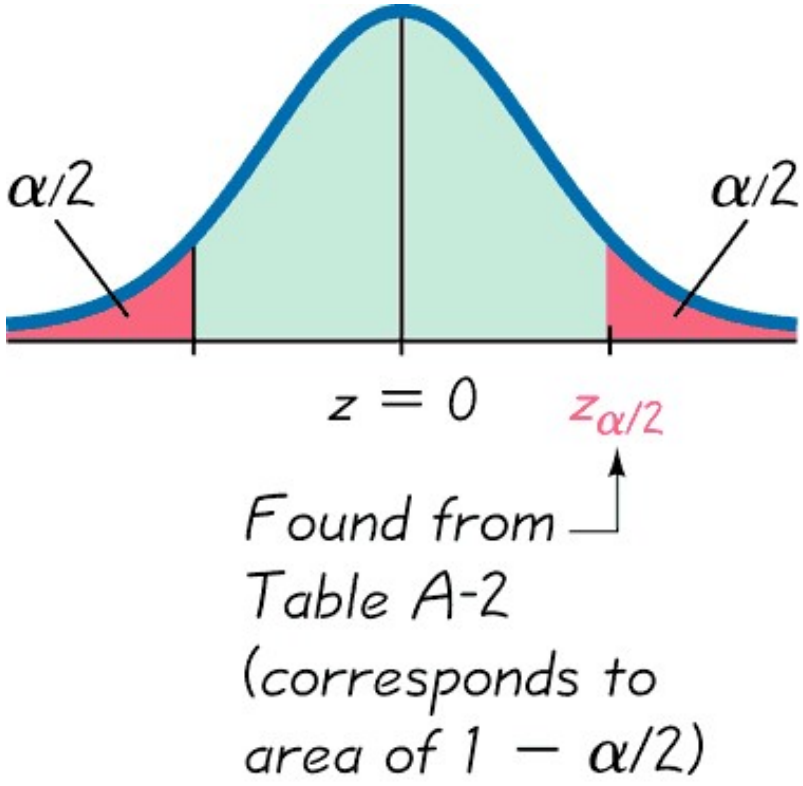
A standard *z* score can be used to distinguish between sample statistics that are likely to occur and those that are unlikely to occur. Such a *z* score is called a critical value. Critical values are based on the following observations:

1. Under certain conditions, the sampling distribution of sample mean can be approximated by a normal distribution.

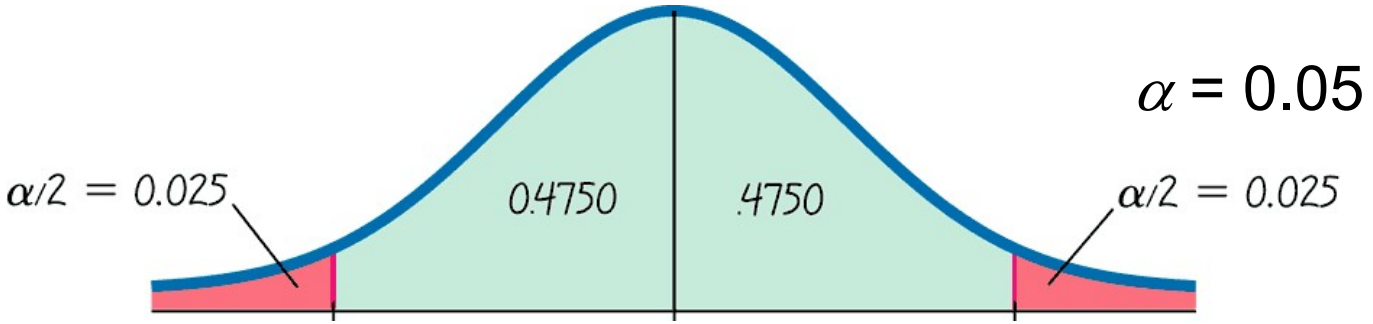


Critical Values

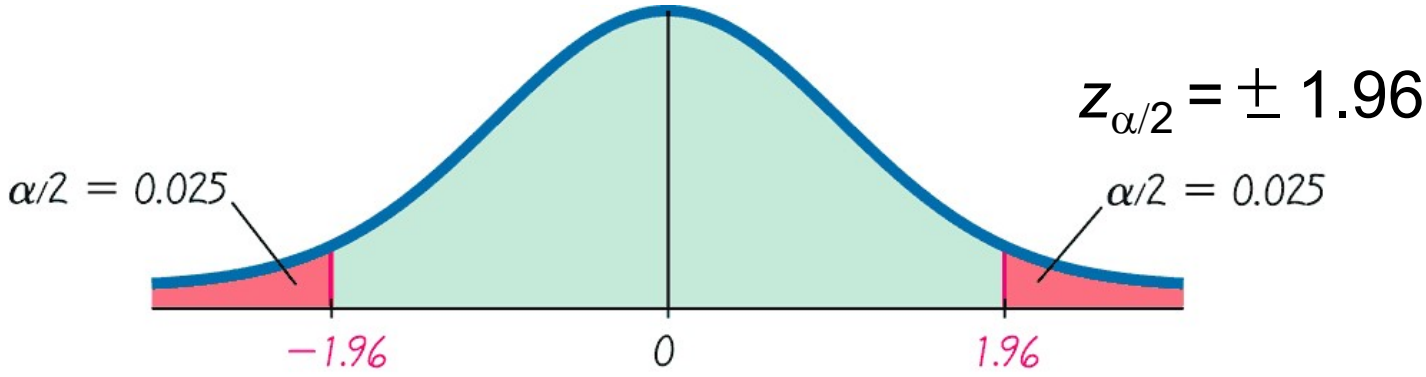
2. A *z* score associated with a sample mean has a probability of $\alpha/2$ of falling in the right tail.



Finding $z_{\alpha/2}$ for a 95% Confidence Level - cont



Use Table A-2 to find a *z* score of 1.96



Example:

In order to estimate the mean body temperature a sample of 106 subjects was recruited. Suppose that the standard deviation is known as 0.62°F. The resulting data ended up in a mean of 98.20°F.

Construct a 95% confidence interval estimate of the mean body temperature:

$$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

$$E = 1.96\frac{0.62}{\sqrt{106}} = 0.1180$$

$$[98.20 - 0.1180; 98.20 + 0.1180]$$

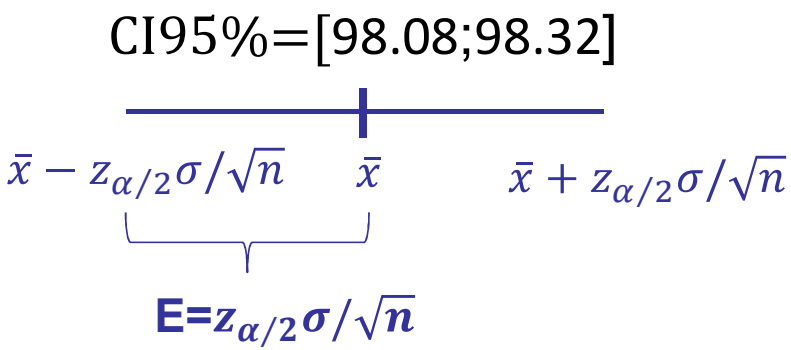
$$CI95\% = [98.08; 98.32]$$

$$CI95\% = [98.08; 98.32]$$

"We are 95% confident that the interval from 98.08 and 98.32 actually does contain the true value of the mean body temperature"

98.20 **mean body temperature**

Margin of error (E)



When data from a simple random sample are used to estimate a population mean, the **difference between the sample mean and the population mean is an error.**
The maximum likely amount of that error is the margin of error, denoted by E.
There is a probability of 1 - α (such as 0.95) that the difference between μ and $\bar{x}$ is E or less.

The margin of error E is also called the maximum error of the estimate and can be found by multiplying the critical value and the standard error

Interpretation :

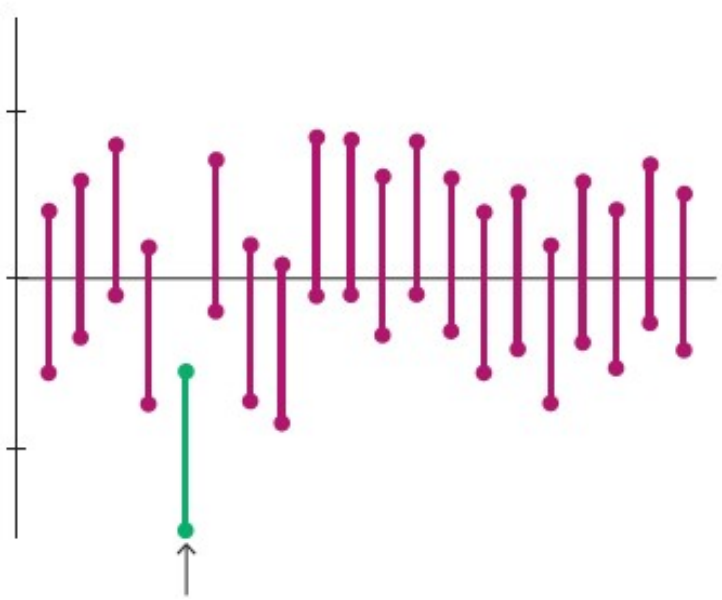
Correct:
“We are 95% confident that the interval from 98.08 and 98.32 actually does contain the true value of the mean body temperature.”

This is a short and acceptable way of saying that if we were to select many different random samples of size 106 and construct the corresponding confidence intervals, 95% of them would contain the population mean μ.
In this correct interpretation, the confidence level of 95% refers to the success rate of the process used to estimate the population mean.”

Wrong:
“There is a 95% chance that the true value of μ will fall between 98.08 and 98.32 .” This is wrong because μ is a population parameter with a fixed value; it is not a random variable with values that vary.
Wrong:
“95% of sample means will fall between 98.08 and 98.32 .”
This is wrong because the values of 98.08 and 98.32 result from one sample; they are not parameters describing the behavior of all samples.

Confidence interval: The Process Success Rate

A confidence level of 95% tells us that the process we are using should, in the long run, result in confidence interval limits that contain the true population proportion 95% of the time.



This confidence interval does not contain the true μ

19 out of 20 (or 95%) different confidence intervals contain μ.

With a 95% confidence level, we expect about 19 out of 20 confidence intervals (or 95%) to contain the true value of μ.

Confidence interval (CI) length

The length of the confidence interval $2* \ z_{\alpha/2} \, {}^{\sigma}/_{\sqrt{n}}$ expresses the uncertainty with which the value of μ is known.

- The amplitude does not depend on the value of μ.
- At the same level of confidence level 1- α: the larger the CI is, the less precise is the sample estimate of μ,
- the amplitude depends on the standard error (σ / √n), which in turn depends on the size (n) of the sample and therefore decreases as n increases, that is, as n increases the precision increases
- As the confidence level increases (eg 99% instead of 95%) the CI width increases and the precision decreases

Example:

In order to estimate the mean body temperature a sample of 106 subjects was recruited. Suppose that the standard deviation is known as 0.62°F. The resulting data ended up in a mean of 98.20°F.
Construct a 90% confidence interval estimate of the mean body temperature for the entire population.

$$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] \quad [\bar{x} - E; \bar{x} + E]$$

$$E = 1.645\frac{0.62}{\sqrt{106}} = 0.09906$$

$$[98.20 - 0.09906; 98.20 + 0.09906]$$

$$\text{CI}90\% = [98.10; 98.30]$$

"We are 90% confident that the interval from 98.10 to 98.30 actually does contain the true value of the mean body temperature for the entire population"

$$\text{CI}95\% = [98.08; 98.32]$$

$$\text{CI}90\% = [98.10; 98.30]$$

Sample size determination

By knowing the standard deviation σ

↓

It is possible to calculate the sample size needed to get a confidence interval (1-α) with a certain length 2E

$$E = z_{\alpha/2}\sigma/\sqrt{n}$$

If we solve the formula for the margin of error for the sample size n, we get :

$$n = \left[\frac{z_{\alpha/2}\sigma}{E}\right]^2$$

NOTE: If the computed sample size n is not a whole number, round the value of n up to the next larger whole number.

Example

We want to estimate the mean value of uricemia in a male population: it is known that in this population the dispersion of the uricemia is σ = 1.1 mg / dl. It is required that the confidence is 95% and that the indeterminacy E does not exceed 0.35 mg / dl.

$$n = \left[\frac{z_{\alpha/2}\sigma}{E}\right]^2 = \left[\frac{1.96*1.1}{0.35}\right]^2 = 37.9 \sim 40$$

Dealing with unknown σ when finding sample size

Assume that we want to estimate the mean IQ score for the medical students. How many students must be randomly selected for IQ tests if we want 95% confidence that the sample mean is within 3 IQ points of the population mean?

$$n = \left[\frac{z_{\alpha/2}\sigma}{E}\right]^2 = \left[\frac{1.96\sigma}{3}\right]^2$$

? σ

For example by using the results of some other earlier study:

Wechsler IQ tests are designed so that the standard deviation is 15. Medical students have IQ scores with a standard deviation less than 15, because they are a more homogeneous group than people randomly selected from the general population. We do not know the specific value of σ for Medical students, but we can be safe by using σ = 15. Using a value for σ that is larger than the true value will make the sample size larger than necessary, but using a value for σ that is too small would result in a sample size that is inadequate. **When determining the sample size n, any errors should always be conservative in the sense that they make the sample size too large instead of too small.**

Triola, Marc M. *Biostatistics for the Biological and Health Sciences, Global Edition, 2nd Edition*. Pearson (Intl),

Example

Assume that we want to estimate the mean IQ score for the population of adults who smoke. How many smokers must be randomly selected for IQ tests if we want 95% confidence that the sample mean is within 3 IQ points of the population mean?

$$n = \left[\frac{z_{\alpha/2}\sigma}{E}\right]^2 = \left[\frac{1.96 * 15}{3}\right]^2 = 96.70 \sim 97$$

Triola, Marc M. *Biostatistics for the Biological and Health Sciences, Global Edition, 2nd Edition*. Pearson (Intl),

Example:

Listed below are weights (hectograms, or hg) of randomly selected girls at birth, based on data from the National Center for Health Statistics.

How would you estimate the mean birth weight of girls?
Here are the summary statistics:

33 28 33 37 31 32 31 28 34 28 33 26 30 31 28

n = 15
$\bar{x}$ = 30.9 hg
s = 2.9 hg.

Confidence interval for estimating a (population) mean μ

$$P\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$or\ [\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}};\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

However σ is not usually known, so we have to estimate it by:

$$s = \hat{\sigma} = \sqrt{\sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{n-1}}$$

Confidence interval for estimating a (population) mean μ

When σ is estimated the standardised difference:

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

becomes $$\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim Tstudent_{df=n-1}$$

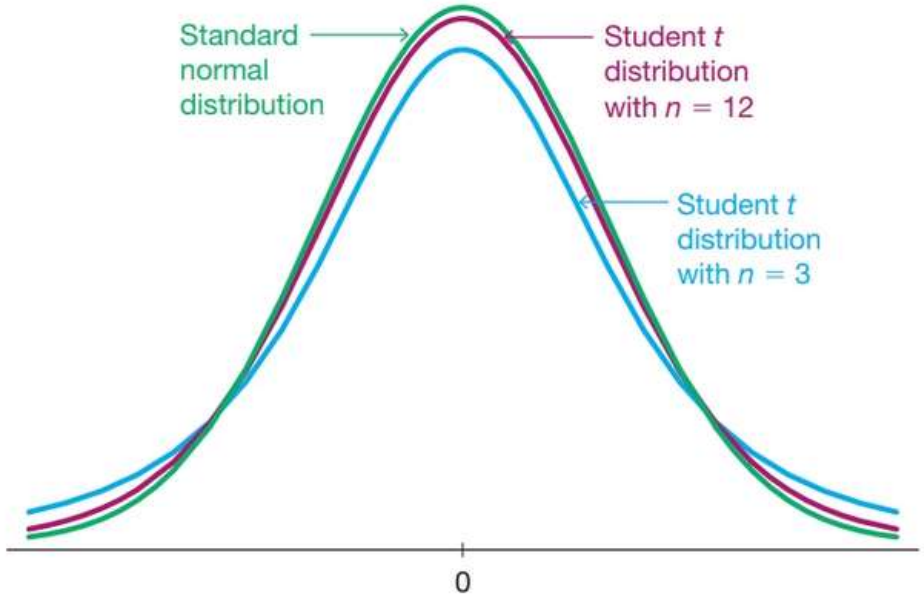
df: number of degrees of freedom (number of sample values that can vary after certain restrictions have been imposed on all data values)
df=n-1

For large sample size t-student ~N(0,1)

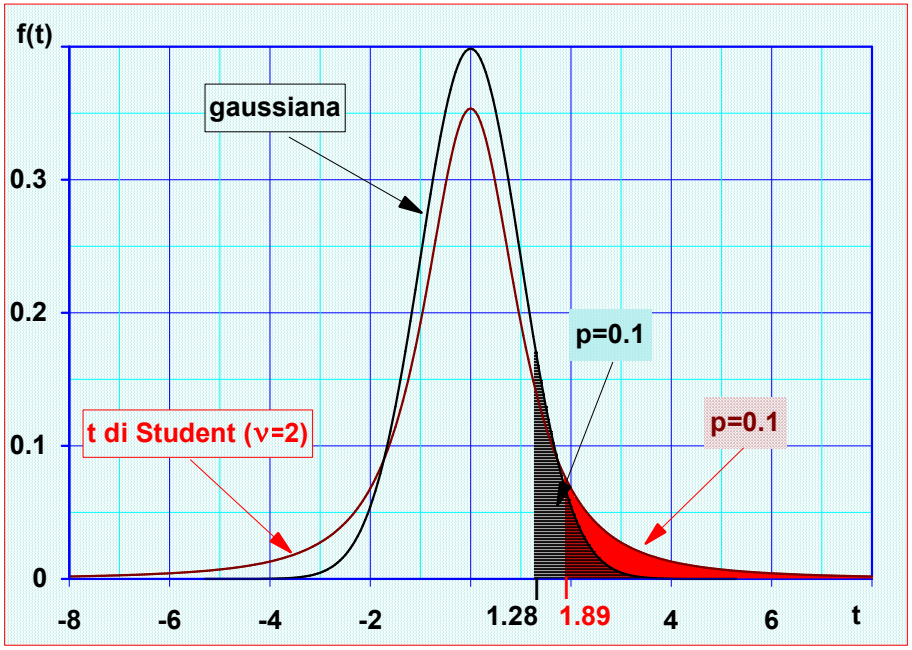
Student t distribution - William Gosset (1876–1937)

- The Student t distribution has the same general symmetric bell shape as the standard normal distribution, but has more variability (with wider distributions), as we expect with small samples.
- The Student t distribution has a mean of t = 0 (just as the standard normal distribution has a mean of z = 0)
- The standard deviation of the Student t distribution varies with the sample size, but it is greater than 1 (unlike the standard normal distribution, which has s = 1).
- The Student t distribution is different for different sample sizes.
- As the sample size n gets larger, the Student t distribution gets closer to the standard normal distribution.

Student t distribution - William Gosset (1876–1937)



Student t distribution - William Gosset (1876–1937)



Student's t distribution has percentiles with an absolute value that is higher than that of the corresponding Gaussian percentiles, the lesser the number of degrees of freedom.

For example, the 90th percentile of the Gaussian standard is 1.282, while the corresponding percentiles of Student's with 1, 2, 3 and 9 g.d.l. are respectively 3.078, 1.886, 1.638 and 1.383.

Confidence interval for estimating a (population) mean μ with unknown σ

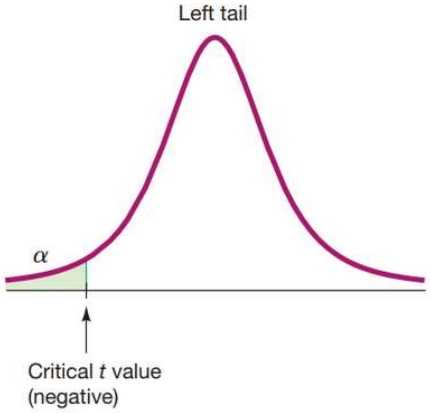
$$P\left(\bar{x} - t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

$$or$$

$$\left[\bar{x} - t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \quad ; \quad \bar{x} + t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$$

If normally distributed population or n>30
otherwise other nonparametric methods should be used

Student t table



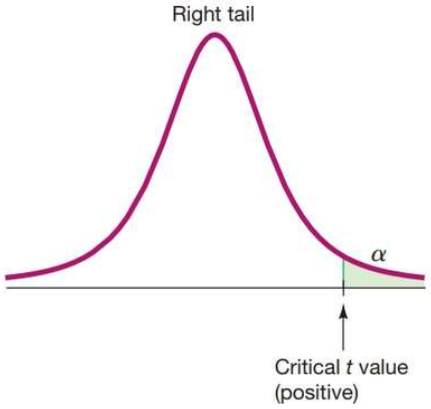


TABLE A-3 *t* Distribution: Critical *t* Values

	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
Degrees of Freedom	**Area in Two Tails**				
	0.01	0.02	0.05	0.10	0.20
1	63.657	31.821	12.706	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.132	1.533
5	4.032	3.365	2.571	2.015	1.476
6	3.707	3.143	2.447	1.943	1.440
7	3.499	2.998	2.365	1.895	1.415
8	3.355	2.896	2.306	1.860	1.397
9	3.250	2.821	2.262	1.833	1.383
10	3.169	2.764	2.228	1.812	1.372
11	3.106	2.718	2.201	1.796	1.363
12	3.055	2.681	2.179	1.782	1.356
13	3.012	2.650	2.160	1.771	1.350
14	2.977	2.624	2.145	1.761	1.345
15	2.947	2.602	2.131	1.753	1.341
16	2.921	2.583	2.120	1.746	1.337
17	2.898	2.567	2.110	1.740	1.333
18	2.878	2.552	2.101	1.734	1.330
19	2.861	2.539	2.093	1.729	1.328
20	2.845	2.528	2.086	1.725	1.325
21	2.831	2.518	2.080	1.721	1.323
22	2.819	2.508	2.074	1.717	1.321
23	2.807	2.500	2.069	1.714	1.319
24	2.797	2.492	2.064	1.711	1.318
25	2.787	2.485	2.060	1.708	1.316
26	2.779	2.479	2.056	1.706	1.315
27	2.771	2.473	2.052	1.703	1.314
28	2.763	2.467	2.048	1.701	1.313
29	2.756	2.462	2.045	1.699	1.311

T student table

Using Table A-3 To find the critical value using Table A-3, use the column with 0.05 for the "Area in Two Tails" (or use the same column with 0.025 for the "Area in One Tail"). The number of degrees of freedom is df $= n - 1 = 14$. We get $t_{\alpha/2} = t_{0.025} = 2.145$.

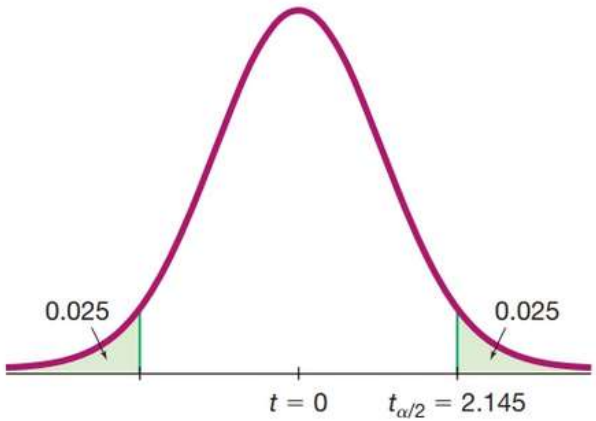


FIGURE 7-5 Critical Value $t_{\alpha/2}$

Example:

Listed below are weights (hectograms, or hg) of randomly selected girls at birth, based on data from the National Center for Health Statistics.

Here are the summary statistics:

33 28 33 37 31 32 31 28 34 28 33 26 30 31 28

n = 15
$\bar{x}$ = 30.9 hg
s = 2.9 hg.

Use the sample data to construct a 95% confidence interval for the mean birth weight of girls.

Example:

n = 15
$\bar{x}$ = 30.9 hg
s = 2.9 hg.

Using *t* Distribution Table Using Table A-3, the critical value is $t_{0.025} = 2.145$ as shown in Example 1. We now find the margin of error E as shown here:

$$E = t_{\alpha/2}\frac{s}{\sqrt{n}} = 2.145 \cdot \frac{2.9}{\sqrt{15}} = 1.606126$$

With $\bar{x} = 30.9$ hg and $E = 1.606126$ hg, we construct the confidence interval as follows:

$$\bar{x} - E < \mu < \bar{x} + E$$
$$30.9 - 1.606126 < \mu < 30.9 + 1.606126$$
$$29.3 \text{ hg} < \mu < 32.5 \text{ hg} \quad \text{(rounded to one decimal place)}$$

The lower confidence interval limit of 29.3 hg is actually 29.2 hg if we use technology or if we use summary statistics with more decimal places than the one decimal place used in the preceding calculation.

INTERPRETATION

We are 95% confident that the limits of 29.2 hg and 32.5 hg actually do contain the value of the population mean μ. If we were to collect many different random samples of 15 newborn girls and find the mean weight in each sample, about 95% of the resulting confidence intervals should contain the value of the mean weight of all newborn girls.

Choosing the appropriate distribution

TABLE 7-1 Choosing Between Student *t* and *z* (Normal) Distributions

Conditions	Method
σ not known and normally distributed population *or* σ not known and $n > 30$	Use Student *t* distribution.
σ known and normally distributed population *or* σ known and $n > 30$ (In reality, σ is rarely known.)	Use normal (*z*) distribution.
Population is not normally distributed and $n \leq 30$.	Use the bootstrapping method (Section 7-4) or a nonparametric method.

Triola, Marc M. *Biostatistics for the Biological and Health Sciences, Global Edition, 2nd Edition*. Pearson (Intl),

SCHOOL OF MEDICINE AND SURGERY

Estimating a Population Proportion

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Key Concept

In this lecture we present methods for using a sample proportion to estimate the value of a population proportion.

- The sample proportion is the best point estimate of the population proportion.
- We can use a sample proportion to construct a confidence interval to estimate the true value of a population proportion, and we should know how to interpret such confidence intervals.
- We should know how to find the sample size necessary to estimate a population proportion.

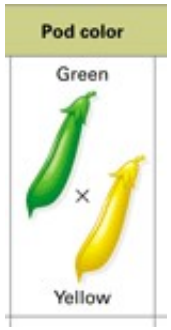
Definition

The sample proportion *p* is the best point estimate of the population proportion *π*.

Example

In one of Mendel's famous hybridization experiments,
Mendel found 152 out of 580 yellow peas.

By looking only at experiment results, could you estimate the true proportion of yellow peas?

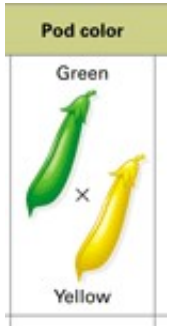


Because the sample proportion is the best point estimate of the population proportion,
we conclude that the best point estimate of π is 152/580=0.2620

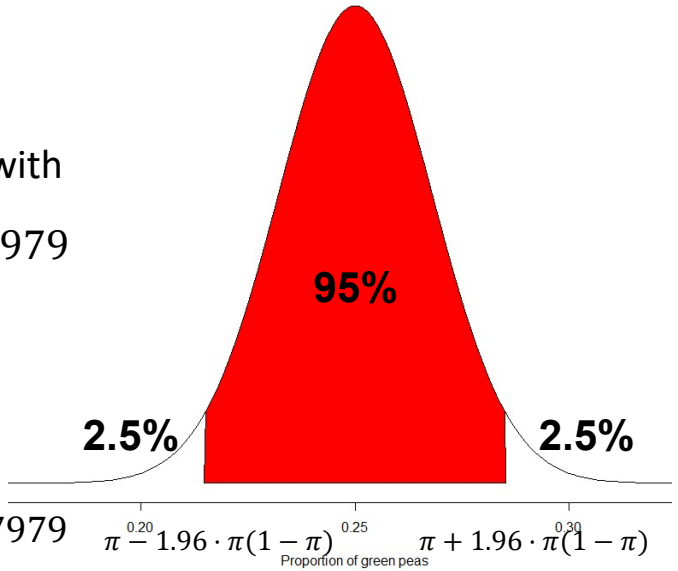
How reliable is this estimate ?

Example

In the Mendel's experiments, he expected that among 580 offspring peas, 145 of them (or 25%) would be yellow.
Find the interval around the true proportion (25%) in which we expect to find 95% of the sampling proportion.



The sampling distribution of the sample probability, if n π ≥5 and n(1- π) ≥5, can be approximate by a Gaussian distribution with mean 0.25 and standard error $\sqrt{\frac{0.25*0.75}{580}}=0.017979$



z1=-1.96 ;z2=+1.96

$$z1 = \frac{x1 - 0.25}{0.017979}; z2 = \frac{x2 - 0.25}{0.017979}$$

X1=0.25-1.96* 0.017979; X2=0.25+1.96* 0.017979

[0.2178;0.2852]

The interval [21.8%;28.5%] is called **probability interval** and is an interval centered on the true proportion with length 2*0.017979

Example

At one experiment Mendel found 152 out of 580 yellow peas.
By looking only at experiment results (and not at his expectation), could you estimate the true proportion of yellow peas?

The best point estimate if 152/580=0.2620
How reliable is this estimate ?

As we normally do not know the true π we cannot use the probability interval, but we can get an interval estimate to get the precision of the estimate as

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

$$0.2620 \pm z_{\alpha/2} * \sqrt{\frac{0.2620(1-0.2620)}{580}}$$

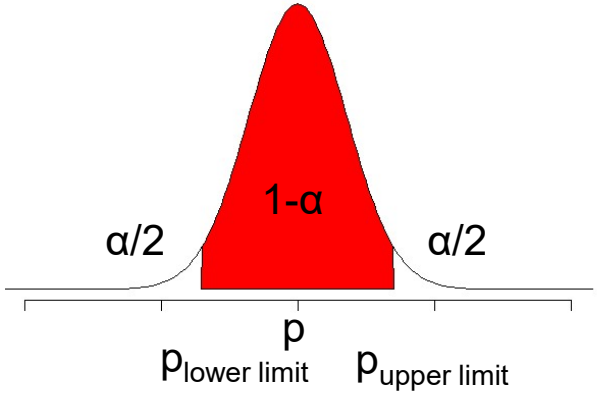
$$[0.2262;0.2978]$$

A confidence interval (or interval estimate) is a range (or an interval) of values used to estimate the true value of a population parameter. A confidence interval is sometimes abbreviated as CI.

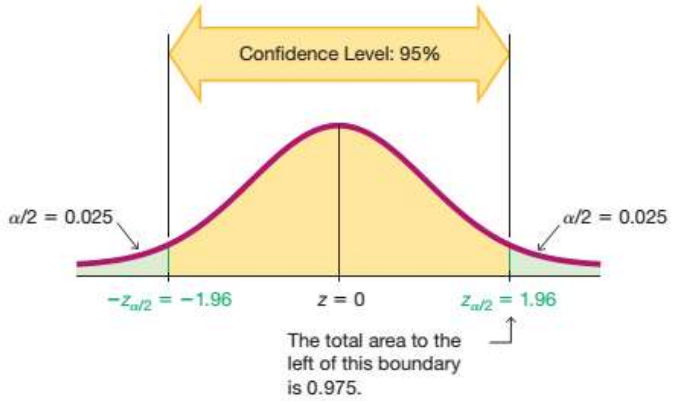
Confidence interval for a sample proportion

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

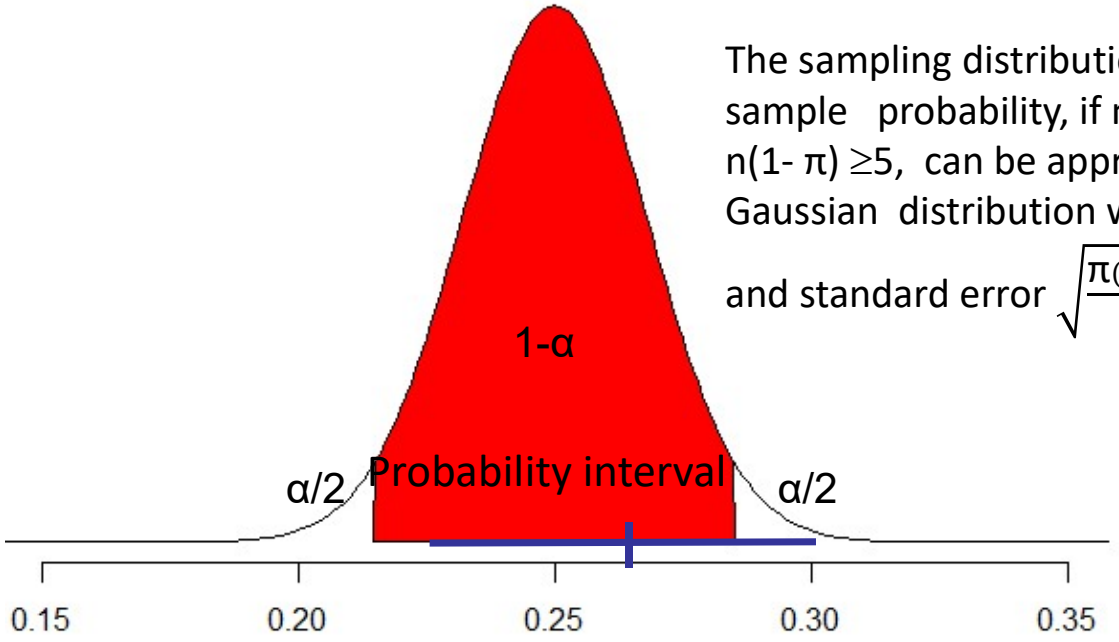
$$[p_{lower\ limit}; p_{upper\ limit}]$$



The **confidence level** is the probability 1 – α (such as 0.95, or 95%) that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times. (The confidence level is also called the degree of confidence, or the confidence coefficient).



Probability and confidence intervals



The sampling distribution of the sample probability, if n π ≥5 and n(1- π) ≥5, can be approximate by a Gaussian distribution with mean π and standard error $\sqrt{\frac{\pi(1-\pi)}{n}}$

$$\pi - z_{\alpha/2}\sqrt{\pi(1-\pi)/n} \quad \pi \quad \pi + z_{\alpha/2}\sqrt{\pi(1-\pi)/n}$$

$$p - z_{\alpha/2}se(p) \quad p \quad p + z_{\alpha/2}se(p)$$

Confidence interval

Example

At one experiment Mendel found 152 out of 580 yellow peas.
By looking only at experiment results (and not at his expectation), could you estimate the true proportion of yellow peas?
The best point estimate if 152/580=0.2620
The interval estimate (95% confidence interval is):

$$0.2620 \pm z_{\alpha/2} * \sqrt{\frac{0.2620(1-0.2620)}{580}}$$

$$[0.2262;0.2978]$$

"We are 95% confident that the interval from 0.2262 to 0.2978 actually does contain the true value of the population proportion π."
If we were to select many different random samples of size 580 and construct the corresponding confidence intervals, 95% of them would contain the population proportion π.

In this case we know the true π=0.25 thus we are able to say that this confidence interval does contain the true value 0.25!
This is not the case usually!

Example: Does Touch Therapy Work?

SCHOOL OF MEDICINE AND SURGERY

Touch Therapy: Structured and standardized healing practice performed by practitioners trained to be sensitive to the receiver's energy field that surrounds the body...no touching is required.

Emily' science fair project:

Each touch therapist would put both hands through the two holes, and Emily would place her hand just above one of the therapist's hands; then the therapist was asked to identify the hand that Emily had selected. Emily used a coin toss to randomly select the hand to be used. This test was repeated 280 times.

Among the 280 trials, the touch therapists identified the correct hand 123 times.

Find the best point estimate of the proportion of correct responses

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

A Close Look at Therapeutic Touch" Journal of the American Medical Association, Vol. 279, No. 13

Example: Does Touch Therapy Work?

N=280
x=123
p=123/280=0.439

0.439 is our best point estimate of the population proportion π, but we have no indication of how good that best estimate is.

A confidence interval gives us a much better sense of how good an estimate is.

A Close Look at Therapeutic Touch" Journal of the American Medical Association, Vol. 279, No. 13

Exercise:

N=280
x=123
p=123/280=0.439

a. Find the 95% confidence interval estimate of the population proportion p.
b. Based on the results, can we safely conclude that the touch therapists had a success rate equivalent to tossing a coin?

a. 95% confidence interval: $0.439 \pm 1.96\sqrt{\frac{0.439(1-0.439)}{280}}$

0.439 ± 0.05812

$[0.381; 0.497]$

b. Based on the confidence interval obtained, it appears that fewer than 50% of the touch therapist responses are correct (because the interval of values from 0.381 to 0.497 is an interval that is completely below 0.50).

A Close Look at Therapeutic Touch" Journal of the American Medical Association, Vol. 279, No. 13

Interpretation :

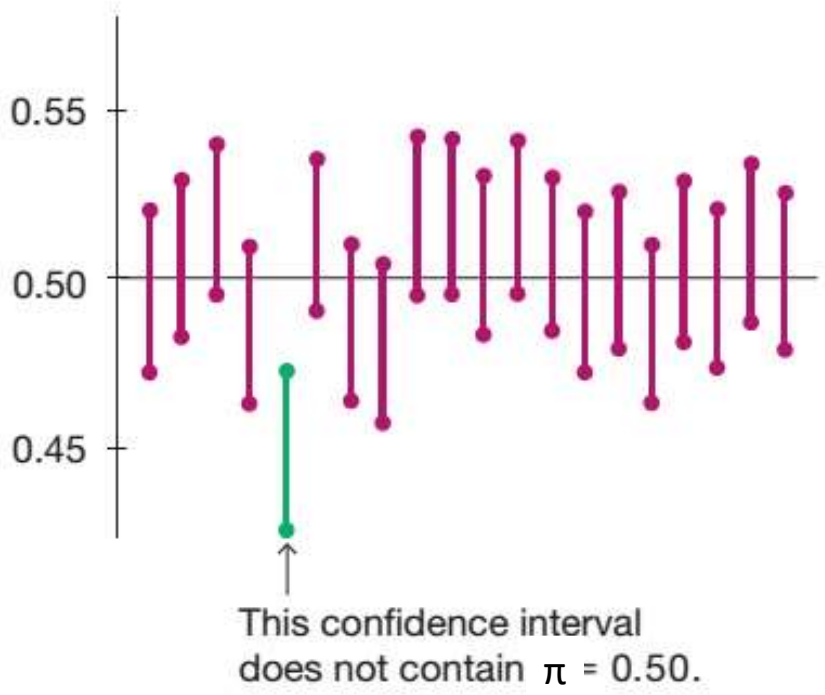
Correct:
“We are 95% confident that the interval from 0.381 to 0.497 actually does contain the true value of the population proportion.”

This is a short and acceptable way of saying that if we were to select many different random samples of size 280 and construct the corresponding confidence intervals, 95% of them would contain the population proportion π.
In this correct interpretation, the confidence level of 95% refers to the success rate of the process used to estimate the population proportion.”

Wrong:
“There is a 95% chance that the true value of **π** will fall between 0.381 and 0.497.” This is wrong because **π** is a population parameter with a fixed value; it is not a random variable with values that vary.
Wrong:
“95% of sample proportions will fall between 0.381 and 0.497.”
This is wrong because the values of 0.381 and 0.497 result from one sample; they are not parameters describing the behavior of all samples.

Confidence interval: The Process Success Rate

A confidence level of 95% tells us that the process we are using should, in the long run, result in confidence interval limits that contain the true population proportion 95% of the time.



19 out of 20 (or 95%) different confidence intervals contain the assumed value of $\pi = 0.50$.

With a 95% confidence level, we expect about 19 out of 20 confidence intervals (or 95%) to contain the true value of π.

Example: What Percentage of Children Have Received Measles Vaccinations?

If we were to conduct a survey to determine the percentage of children (older than 1 year) who have received measles vaccinations, how many children must be surveyed in order to be 95% confident that the sample percentage is in error by no more than three percentage points?

a. Assume that a recent survey showed that 90% of children have received measles vaccinations.

b. Assume that we have no prior information suggesting a possible value of the population proportion.

Sample size determination

If we plan to collect sample data in order to estimate some population proportion, how do we know how many sample units we must get?

If we solve the formula for the margin of error for the sample size n, we get :

$$E = z_{\alpha/2} se(p)$$

$$E = z_{\alpha/2} \sqrt{p(1-p)/n}$$

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{E^2}$$

p: an estimate of the population proportion π (if no such estimate is known we replace p by 0.5 resulting in the largest possible sample size)

E: desired margin of error of our estimate

Example: What Percentage of Children Have Received Measles Vaccinations?

a. With a 95% confidence level, we have $\alpha = 0.05$, so $z_{\alpha/2} = 1.96$. Also, the margin of error is $E = 0.03$, which is the decimal equivalent of "three percentage points." The prior survey suggests that $\hat{p} = 0.90$, so $\hat{q} = 0.10$ (found from $\hat{q} = 1 - 0.90$). Because we have an estimated value of $\hat{p}$, we use Formula 7-2 as follows:

$$n = \frac{[z_{\alpha/2}]^2 \hat{p}\hat{q}}{E^2} = \frac{[1.96]^2 (0.90)(0.10)}{0.03^2}$$

$$= 384.16 = 385 \text{ (rounded } up\text{)}$$

We must obtain a simple random sample that includes at least 385 children.

b. With no prior knowledge of $\hat{p}$ (or $\hat{q}$), we use Formula 7-3 as follows:

$$n = \frac{[z_{\alpha/2}]^2 \cdot 0.25}{E^2} = \frac{[1.96]^2 \cdot 0.25}{0.03^2}$$

$$= 1067.11 = 1068 \text{ (rounded up)}$$

We must obtain a simple random sample that includes at least 1068 children.