week 9

8. Large-Sample Estimation

Adams Wai Kin Kong
School of Computer Science and Engineering
Nanyang Technological University, Singapore
adamskong@ntu.edu.sg

8.2 Point Estimation

Point Estimation (1 of 9)

In a practical situation, there may be several statistics that could be used as point estimators for a population parameter.

Sampling distributions provide information that can be used to select the **best estimator.** What characteristics would be valuable?

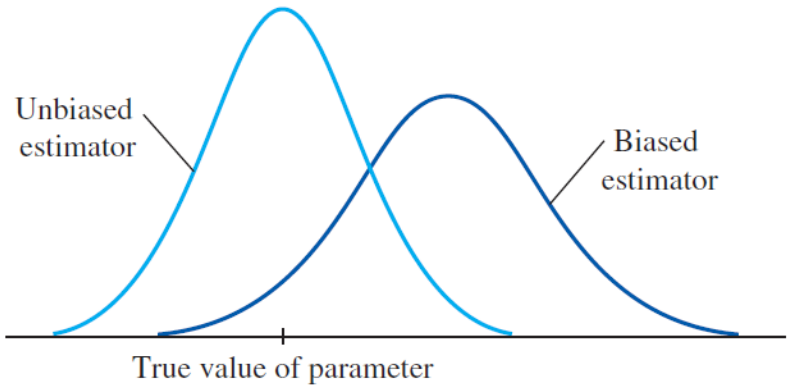
First, the **sampling distribution of the point estimator should be centered over the true value of the parameter to be estimated.**

Point Estimation (2 of 9)

Definition:

An estimator is said to be **unbiased** if the mean of its distribution is equal to the true value of the parameter being estimated. More formally, if $E[u(X_1, X_2 \cdots X_n)] = \theta$, the statistics $u(X_1, X_2 \cdots X_n)$ is called an **unbiased estimator of** θ. Otherwise, the estimator is said to be **biased.**

The sampling distributions for an unbiased estimator and a biased estimator are shown in figure.



Unbiased

biased

$X_i \sim \Rightarrow \mu, \sigma$

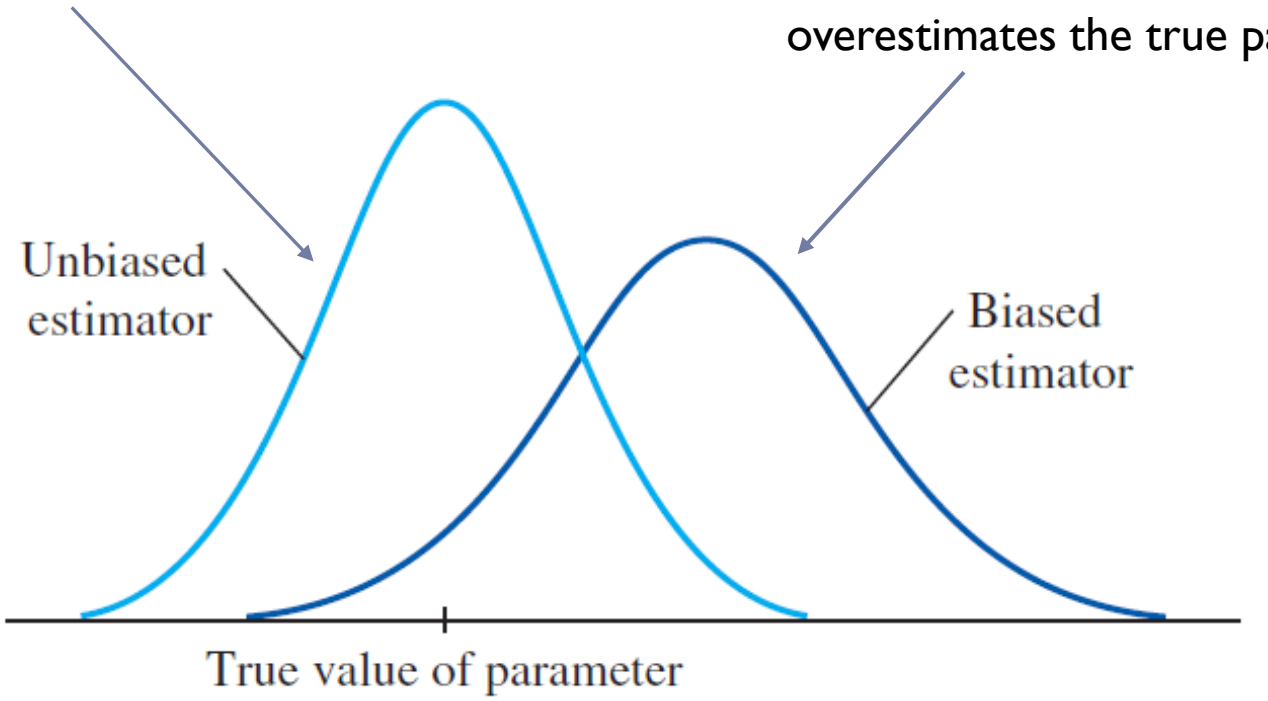
$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$

$E(\bar{x}) = \mu$

Point Estimation (3 of 9)

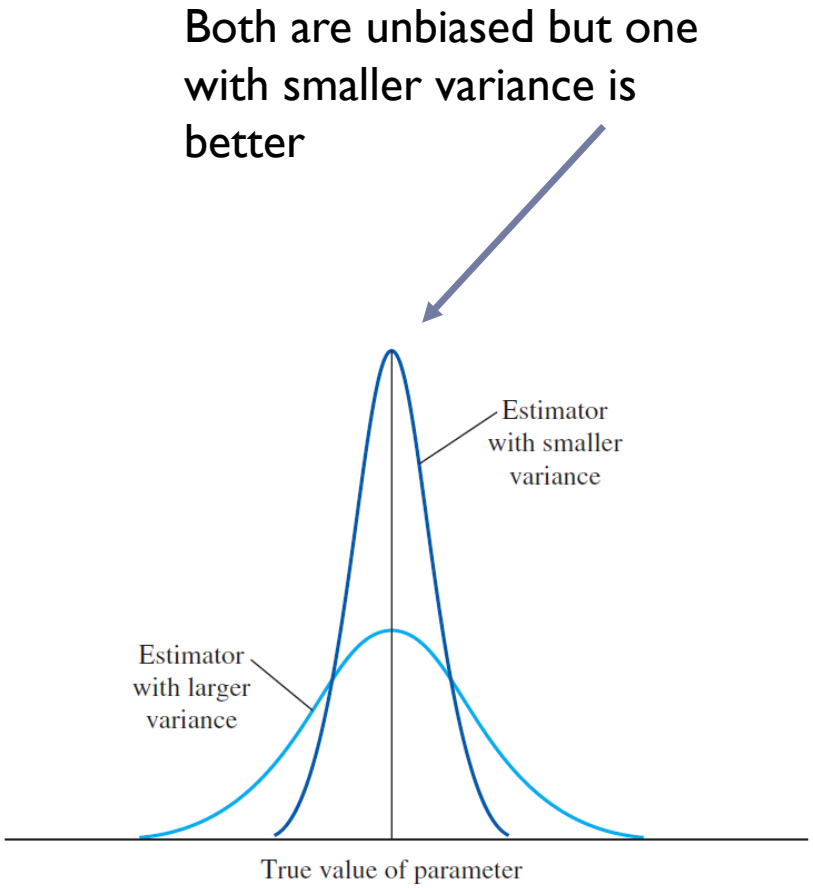
The estimated value should be around the true parameter.

This biased estimator likely overestimates the true parameter



Point Estimation (4 of 9)

The variance of the estimator's sampling distribution should be as small as possible. This ensures that, with a high probability, an individual estimate will fall close to the true value of the parameter.



Comparison of estimator variability

Figure 8.3

Point Estimation (5 of 9)

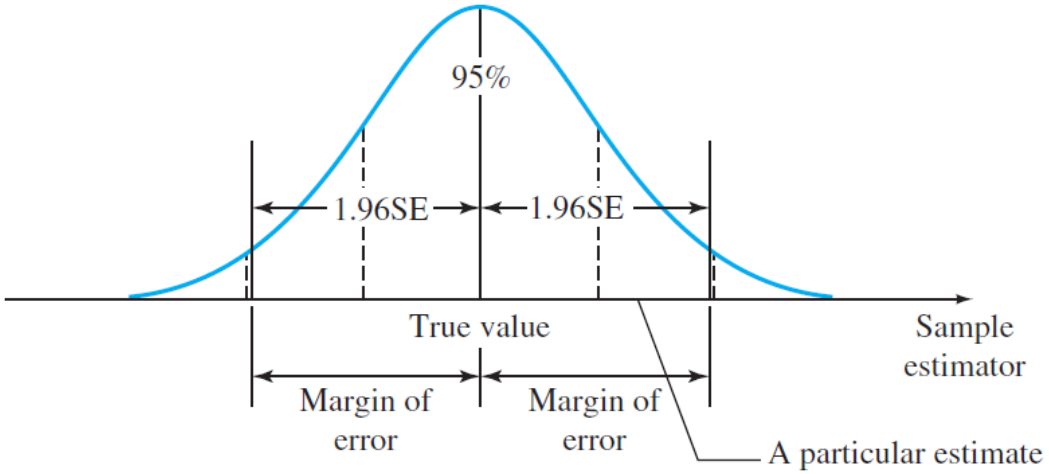
Definition:

The distance between an estimate and the true value of the parameter is called the **error of estimation**.

We normally measure this error by number of SE.

Point Estimation (6 of 9)

This quantity, called the **95% margin of error** (or simply the **"margin of error"**), provides a practical upper bound for the error of estimation.



Sampling distribution of an unbiased estimator

Figure 8.4

Point Estimation (7 of 9)

Point Estimation of a Population Parameter

- Point estimator: a statistic calculated using sample measurements
- 95% Margin of error: 1.96×Standard error of the estimator

Point Estimation (8 of 9)

How to Estimate a Population Mean or Proportion

- To estimate the population mean μ for a quantitative population, the point estimator $\bar{x}$ is *unbiased* with standard error estimated as

$$\text{SE} = \frac{s}{\sqrt{n}}$$

- The 95% margin of error when $n \geq 30$ is estimated as

$$\pm 1.96\left(\frac{s}{\sqrt{n}}\right)$$

Point Estimation (9 of 9)

- To estimate the population proportion p for a binomial population, the point estimator $\hat{p} = x/n$ is *unbiased*, with standard error estimated as

$$\text{SE} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- The 95% margin of error is estimated as $\pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$

Assumptions: $n\hat{p} > 5$ and $n\hat{q} > 5$.

Example 8.4 (1 of 2)

A scientist is studying a species of polar bear, found in and around the Arctic Ocean. Their range is limited by the availability of sea ice, which they use as a platform to hunt seals, the mainstay of their diet.

The destruction of its habitat on the Arctic ice, which has been attributed to global warming, threatens the bear's survival as a species; it may become extinct within the century.

Example 8.4 (2 of 2)

$\sigma \approx 105$
$n = 50$

A random sample of $n = 50$ polar bears produced an average weight of 980 pounds with a standard deviation of 105 pounds.

Use this information to estimate the average weight of all Arctic polar bears.

$1.96\sqrt{\frac{105}{50}} \approx$

Solution:

The random variable measured is weight, a quantitative random variable best described by its mean μ.

Example 8.4 – Solution

The point estimate of μ, the average weight of all Arctic polar bears, is $\bar{x}=980$ pounds. The margin of error is estimated as

$$1.96\,\text{SE} = 1.96\left(\frac{s}{\sqrt{n}}\right) = 1.96\left(\frac{105}{\sqrt{50}}\right) = 29.10 \approx 29 \text{ pounds}$$

You can be fairly confident that the sample estimate of 980 pounds is within ±29 pounds of the population mean.

Example 8.5

$1.96\sqrt{\frac{pq}{n}} \approx 0.196\sqrt{0.73 \cdot 0.27} = 0.087 \approx 0.09$

$\hat{p}$: 0.64 to 0.82

In addition to the average weight of the Arctic polar bear, the scientist in the previous example is also interested in the opinions of adults on the subject of global warming.

if increase n we can lower the range

He selects a random sample of $n = 100$ adults, and finds that 73% of the sampled adults think global warming is a very serious problem.

Estimate the true population proportion of adults who believe that global warming is a very serious problem, and find the margin of error for the estimate.

$1.96\sqrt{\frac{0.73 \cdot 0.27}{100}}$

Example 8.5 – Solution (1 of 2)

The parameter of interest is now p, the proportion of adults in the population who believe that global warming is a very serious problem.

The best estimator of p is the sample proportion $\hat{p}$ which for this sample is $\hat{p}$=0.73. In order to find the margin of error, you can approximate the value of p with its estimate $\hat{p}$=0.73:

$$1.96\,\text{SE} = 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96\sqrt{\frac{.73(.27)}{100}} = .09$$

Example 8.5 – Solution (2 of 2)

With this margin of error, you can be fairly confident that the estimate of .73 is within ±.09 of the true value of p.

Hence, you can conclude that the true value of p could be as small as .64 or as large as .82.

This margin of error is quite large when compared to the estimate itself and reflects the fact that large samples are required to achieve a small margin of error when estimating p.

8.3 Interval Estimation

Interval Estimation (1 of 1)

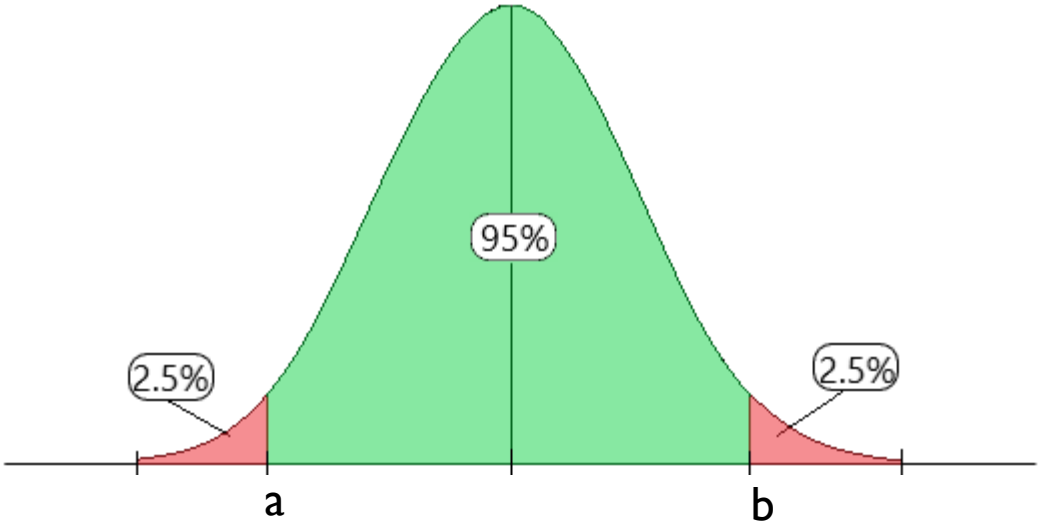
An *interval estimator* is a rule for calculating two numbers—say, a and b—to create an interval that you are fairly certain contains the parameter of interest. The concept of "fairly certain" means "with high probability."

We measure this probability using the **confidence coefficient**, designated by $\mathbf{1 - \alpha}$.

Definition:

The probability that a confidence interval will contain the estimated parameter is called the **confidence coefficient**.

Interval Estimation (Illustration)

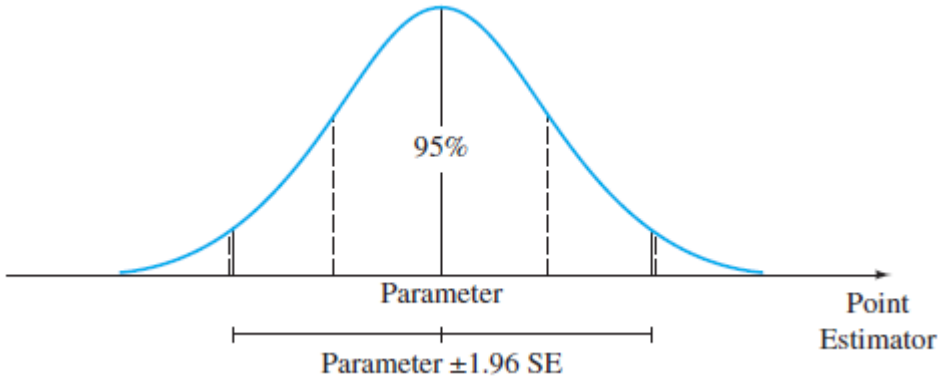


Interval estimator [a b]. Its **confidence coefficient is 0.95 and α is 0.05.**

Constructing a Confidence Interval

Constructing a Confidence Interval (1 of 3)

When the sampling distribution of a point estimator is approximately normal, an interval estimator or **confidence interval** can be constructed using the following reasoning. For simplicity, assume that the confidence coefficient is 0.95 and refer to figure.



Parameter ±1.96 SE

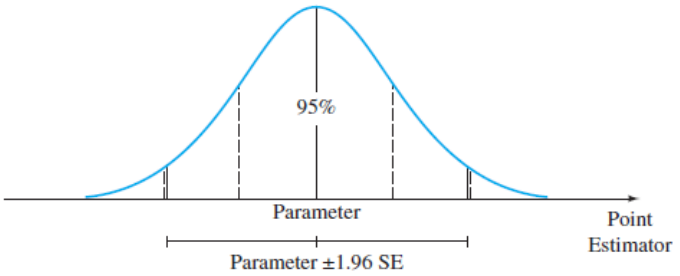
Figure 8.6

Constructing a Confidence Interval (2 of 3)

- We know that, of all possible values of the point estimator that we might select, 95% of them will be in the interval

  Parameter ± 1.96 SE

  shown in figure.



- Since the value of the parameter is unknown, consider constructing the interval

  Point Estimator ± 1.96SE

  which has the same width as the first interval, but has a variable center.

Constructing a Confidence Interval (3 of 3)

A $(1 - \alpha)100\%$ Large-Sample Confidence Interval

(Point estimator) $\pm z_{\alpha/2} \times$ (Standard error of the estimator)

where $z_{\alpha/2}$ is the z-value with an area $\alpha/2$ in the right tail of a standard normal distribution.

This formula generates two values: the **lower confidence limit (LCL)** and the **upper confidence limit (UCL)**.

Large-Sample Confidence Interval for a Population Mean μ

Large-Sample Confidence Interval for a Population Mean μ (1 of 6)

When the sample size n is large, the sample mean $\bar{x}$ is the best point estimator for the population mean μ.

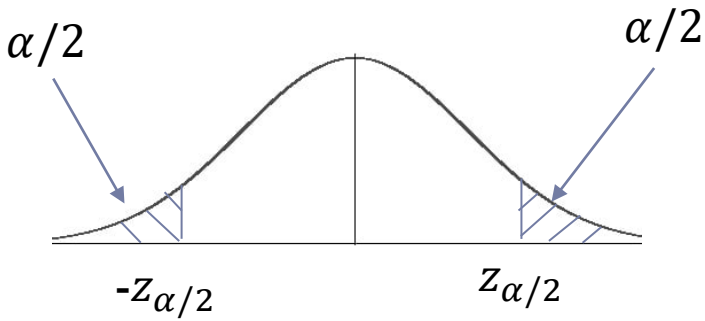
Since its sampling distribution is approximately normal, it can be used to construct a confidence interval according to the general approach given earlier.

Large-Sample Confidence Interval for a Population Mean μ (2 of 6)

A $(1-\alpha)100\%$ Large-Sample Confidence Interval for a Population Mean

95%

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



Standard normal distribution

where $z_{\alpha/2}$ is the z-value corresponding to an area $\alpha/2$ in the upper tail of a standard normal z distribution, and

n = Sample size

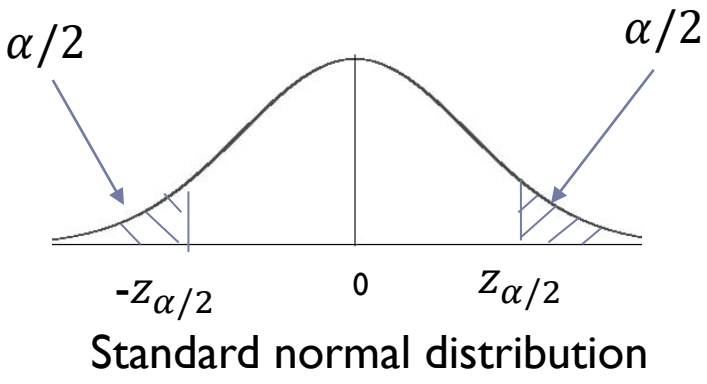
σ = Standard deviation of the sampled population

Large-Sample Confidence Interval for a Population Mean μ (3 of 6)

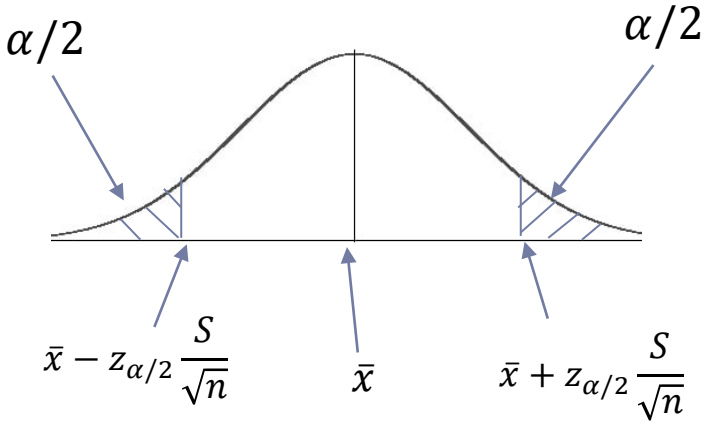
If σ is unknown, it can be approximated by the sample standard deviation s when the sample size is large ($n \geq 30$) and the approximate confidence interval is

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

! if dun have sigma use S
if hv sigma use sigma



Standard normal distribution



Deriving a Large-Sample Confidence Interval

Another way to find the large-sample confidence interval for a population mean μ is to begin with the statistic

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

which has a standard normal distribution. If you write $z_{\alpha/2}$ as the value of z with area $\alpha/2$ to its right, then you can write

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

Large-Sample Confidence Interval for a Population Mean μ (5 of 6)

You can rewrite this inequality as

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

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

so that

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

Large-Sample Confidence Interval for a Population Mean μ (6 of 6)

Both $\bar{x} - z_{\alpha/2}(\sigma/\sqrt{n})$ and $\bar{x} + z_{\alpha/2}(\sigma/\sqrt{n})$, the lower and upper confidence limits, are actually random quantities that depend on the sample mean $\bar{x}$.

Therefore, in repeated sampling, the random interval, $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ will contain the population mean μ with probability $(1 - \alpha)$.

$(1-\alpha) \uparrow \equiv \alpha \downarrow \equiv z_{\alpha/2} \uparrow$

Interpreting the Confidence Interval

Interpreting the Confidence Interval (1 of 1)

A good confidence interval has two desirable characteristics:

- It is as **narrow** as possible. The narrower the interval, the more exactly you have located the estimated parameter.

- It has a **large confidence coefficient,** near 1. The larger the confidence coefficient, the more likely it is that the interval will contain the estimated parameter.

Example 8.7

The average daily intake of dairy products for a random sample of $n = 50$ adult males was $\bar{x} = 756$ grams per day with a standard deviation of $s = 35$ grams per day. Construct a 99% confidence interval for the mean daily intake of dairy products for adult men.

Solution:

To change the confidence level to .99, you must find the appropriate value of the standard normal z that puts area $(1 - \alpha) = .99$ in the center of the curve.

Example 8.7 – Solution (1 of 4)

This value, with tail area $\alpha/2 = .005$ to its right, is found from table to be $z = 2.58$.

$(1-\alpha) \cdot 100\% = 99\%$ $\alpha = 0.01$

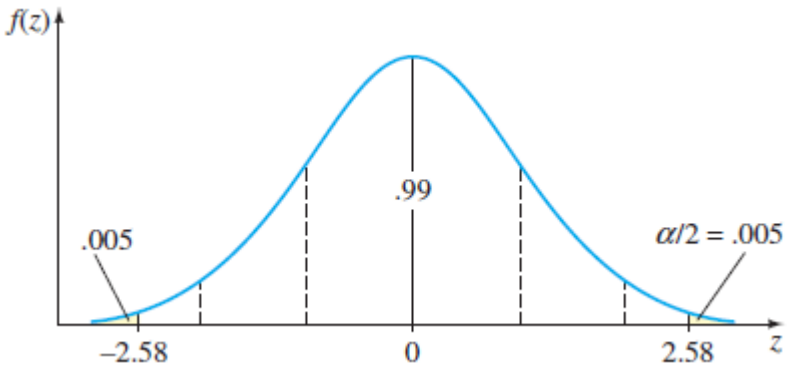
Confidence Coefficient, $(1 - \alpha)$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	1.645
.95	.05	.025	1.96
.98	.02	.01	2.33
.99	.01	.005	2.58

Values of z Commonly Used for Confidence Intervals

Table 8.2

Example 8.7 – Solution (2 of 4)

See figure.



Standard normal values for a 99% confidence interval

Figure 8.10

Example 8.7 – Solution (3 of 4)

The 99% confidence interval is then

$$\bar{x} \pm 2.58\left(\frac{s}{\sqrt{n}}\right)$$

$$756 \pm 2.58(4.95)$$

$$756 \pm 12.77$$

or 743.23 to 768.77 grams per day. This confidence interval is *wider* than the 95% confidence interval.

Example 8.7 – Solution (4 of 4)

The increased width is necessary to increase the confidence.

The only way to *increase the confidence* without increasing the width of the interval is to *increase the sample size, n*.

Large-Sample Confidence Interval for a Population Proportion p

Large-Sample Confidence Interval for a Population Proportion p (1 of 5)

The objective of many research experiments or sample surveys is to estimate the proportion of people or objects in a large group that possess a certain characteristic. Here are some examples:

- The proportion of sales in a large number of customer contacts
- The proportion of "likely" voters who plan to vote for a particular political candidate

Large-Sample Confidence Interval for a Population Proportion p (2 of 5)

Each of these is an example of the binomial experiment, and the parameter to be estimated is the binomial proportion p.

When the sample size is large, the sample proportion,

$$\hat{p} = \frac{x}{n} = \frac{\text{Total number of successes}}{\text{Total number of trials}}$$

is the best point estimator for the population proportion p.

Large-Sample Confidence Interval for a Population Proportion p (3 of 5)

Since its sampling distribution is approximately normal, with mean p and standard error $SE = \sqrt{pq/n}$, $\hat{p}$ can be used to construct a confidence interval.

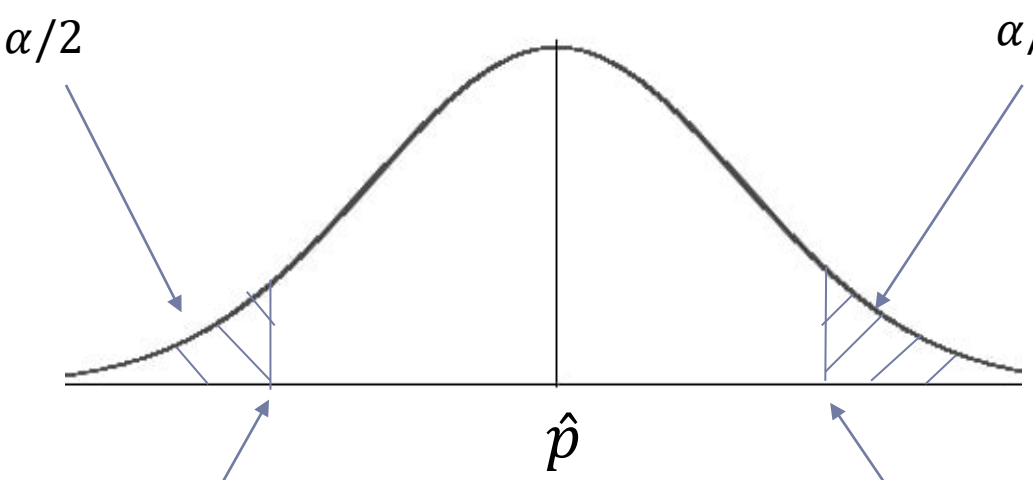
A $(1-\alpha)100\%$ Large-Sample Confidence Interval for a Population Proportion

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

where $z_{\alpha/2}$ is the z-value corresponding to an area $\alpha/2$ in the right tail of a standard normal z distribution. Since p and q are unknown, they are estimated using the best point estimators: $\hat{p}$ and $\hat{q}$.

Assume that $n\hat{p} > 5$ and $n\hat{q} > 5$

Large-Sample Confidence Interval for a Population Proportion p (5 of 5)



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

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

Example 8.8

$n = 985$

$\hat{p} = \frac{592}{985} = 0.601$

estimation $= \sqrt{\frac{0.2398}{985}}$

$= 0.0156$

A random sample of 985 "likely" voters—those who are likely to vote in the upcoming election—were polled by the Republican Party. Of those surveyed, 592 indicated that they intended to vote for the Republican candidate in the upcoming election.

0.1

$0.601 \pm z_{0.05} \sqrt{\frac{0.601 \cdot 0.399}{985}}$

Construct a 90% confidence interval for p, the proportion of likely voters in the population who intend to vote for the Republican candidate. Based on this information, can you conclude that the candidate will win the election?

$\alpha = 0.1$

$\alpha/2 = 0.05$

$0.601 \pm$

Example 8.8 – Solution (1 of 4)

The point estimate for p is

$$\hat{p} = \frac{x}{n} = \frac{592}{985} = .601$$

and the estimated standard error is

$$\sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(.601)(.399)}{985}} = .016$$

Example 8.8 – Solution (2 of 4)

The z-value for a 90% confidence interval is the value that has area $\alpha/2 = .05$ in the upper tail of the z distribution, or $z_{0.5} = 1.645$ from table.

Confidence Coefficient, $(1 - \alpha)$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	1.645
.95	.05	.025	1.96
.98	.02	.01	2.33
.99	.01	.005	2.58

Values of z Commonly Used for Confidence Intervals

Table 8.2

Example 8.8 – Solution (3 of 4)

The 90% confidence interval for p is

$$\hat{p} \pm 1.645\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$.601 \pm .026$$

or $.575 < p < .627$. You estimate that the percentage of likely voters who intend to vote for the Republican candidate is between 57.5% and 62.7%.

Example 8.8 – Solution (4 of 4)

Will the candidate win the election? Assuming that she needs more than 50% of the vote to win, and because both the upper and lower confidence limits exceed this minimum value, you can say with 90% confidence that the candidate will win.