Estimation of Means and Proportions

Large-Sample Estimation

Dr.Mamatha.H.R
Professor
Department of Computer Science and Engineering
PES University
Bangalore

Course material created using various Internet resources and text book

Real Life Examples of Confidence Intervals

- A 2008 Gallup survey found that TV ownership may be good for wellbeing. The results from the poll stated that the confidence level was 95% +/-3, which means that if Gallup repeated the poll over and over, using the same techniques, 95% of the time the results would fall within the published results. The 95% is the confidence level and the +/-3 is called a margin of error.

- For example, “For the European data, one can say with 95% confidence that the true population for wellbeing among those without TVs is between 4.88 and 5.26.” The confidence interval here is “between 4.88 and 5.26“.

- The U.S. Census Bureau routinely uses confidence levels of 90% in their surveys. One survey of the number of people in poverty in 1995 stated a confidence level of 90% for the statistics “The number of people in poverty in the United States is 35,534,124 to 37,315,094.” That means if the Census Bureau repeated the survey using the same techniques, 90 percent of the time the results would fall between 35,534,124 and 37,315,094 people in poverty. The stated figure (35,534,124 to 37,315,094) is the confidence interval.

- Example: A recent article on Rasmussen Reports states that “38% of Likely U.S. Voters now say their health insurance coverage has changed because of Obamacare”.
- “The margin of sampling error is +/- 3 percentage points with a 95% level of confidence.”

- What a 95 percent confidence level is saying is that if the poll or survey were repeated over and over again, the results would match the results from the actual population 95 percent of the time.

What about “+/- 3 percentage points”?

- The width of the confidence interval tells us more about how certain (or uncertain) we are about the true figure in the population.
- This width is stated as a plus or minus (in this case,+/- 3) and is called the confidence interval.
- When the interval and confidence level are put together, you get a spread of percentage.
- In this case, you would expect the results to be 35 (38-3) to 41 (38+3) percent, 95% of the time.

Factors that Affect Confidence Intervals (CI)

- Population size: this does not usually affect the CI but can be a factor if you are working with small and known groups of people.
- Sample Size: the smaller your sample, the less likely it is you can be confident the results reflect the true population parameter.
- Percentage: Extreme answers come with better accuracy.

For example, if 99 percent of voters are for a particular party, the chances of error are small. However, if 49.9 percent of voters are “for” and 50.1 percent are “against” then the chances of error are bigger.

0% and 100% Confidence Level

- A 0% confidence level means you have no faith at all that if you repeated the survey that you would get the same results.
- A 100% confidence level means there is no doubt at all that if you repeated the survey you would get the same results—and even then you probably couldn't be 1.

What is the Definition of a Confidence Interval?

- A confidence interval is how much uncertainty there is with any particular statistic.
- Confidence intervals are often used with a margin of error.
- It tells you how confident you can be that the results from a poll or survey reflect what you would expect to find if it were possible to survey the entire population.
- Confidence intervals are intrinsically connected to confidence levels.

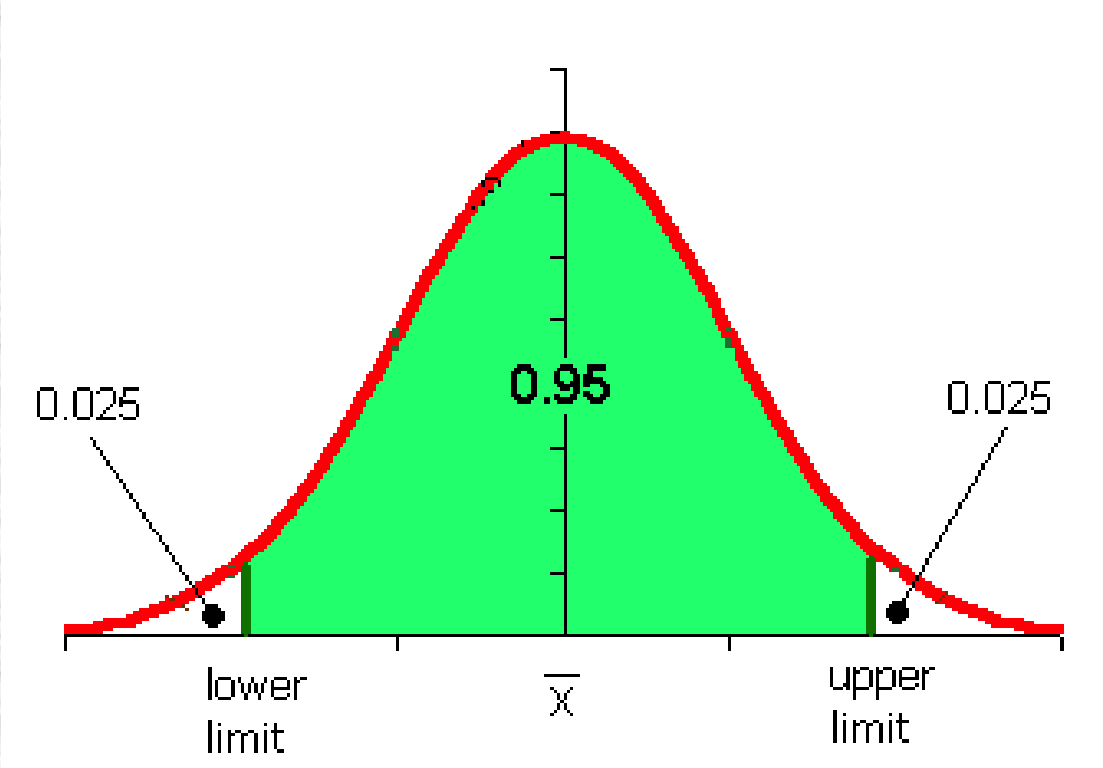
Confidence Intervals vs. Confidence Levels

- Confidence levels are expressed as a percentage (for example, a 95% confidence level).
- It means that should you repeat an experiment or survey over and over again, 95 percent of the time your results will match the results you get from a population (in other words, your statistics would be sound!).

Confidence intervals are your results…usually numbers.

For example, you survey a group of pet owners to see how many cans of dog food they purchase a year.

You test your statistics at the 99 percent confidence level and get a confidence interval of (200,300). That means you think they buy between 200 and 300 cans a year. You're super confident (99% is a very high level!) that your results are sound, statistically.

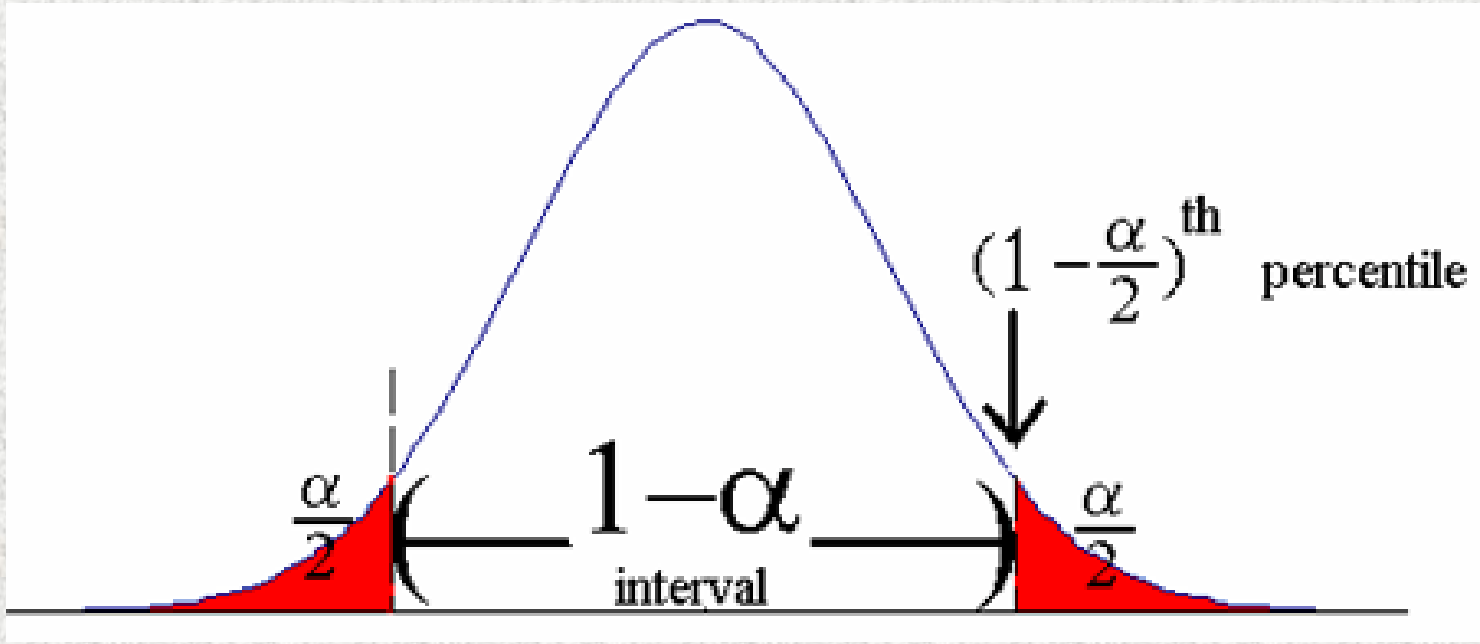

0.025
0.95
0.025
lower
limit
$\overline{x}$
upper
limit

Confidence Coefficient

- The confidence coefficient is the confidence level stated as a proportion, rather than as a percentage.
- For example, if you had a confidence level of 99%, the confidence coefficient would be .99.
- In general, the higher the coefficient, the more certain you are that your results are accurate.

- The following table lists confidence coefficients and the equivalent confidence levels.

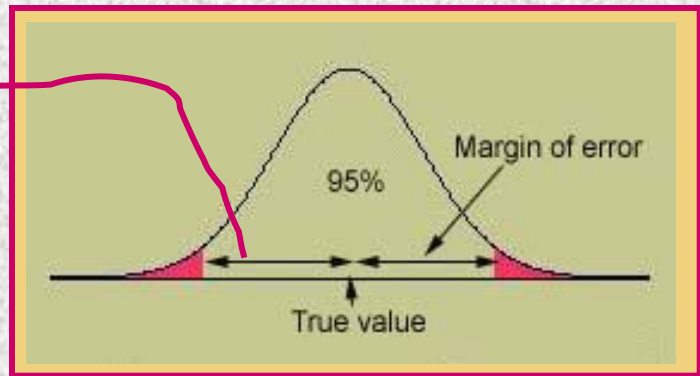
Confidence coefficient (1 – α)	Confidence level (1 – α * 100%)
0.90	90%
0.95	95%
0.99	99%


$(1-\frac{\alpha}{2})^{th}$ percentile
$\frac{\alpha}{2}$
$1-\alpha$
interval
$\frac{\alpha}{2}$

The Margin of Error

- In this note we assume that the sample sizes are large
- From the Central Limit Theorem, the sampling distributions of $\bar{x}$ and $\hat{p}$ will be **approximately normal** under certain assumptions
- For ***unbiased*** estimators with normal sampling distributions, 95% of all point estimates will lie within 1.96 standard deviations of the parameter of interest.
- **Margin of error:** provides a upper bound to the difference between a particular estimate and the parameter that it estimates. It is calculated as

$$1.96 \times \text{std error of the estimator}$$



Estimating Means and Proportions

- For a quantitative population,

Point estimator of population mean μ : $\{\bar{x}$

Margin of error $(n \geq 30)$: $\pm 1.96 \frac{s}{\sqrt{n}}$

- For a binomial population,

Point estimator of population proportion p : $\{\hat{p} = x/n$

Margin of error: $\pm 1.96 \sqrt{\frac{\hat{p}\hat{q}}{n}}$ *Assumption*: np > 5 and nq > 5; or $0 < p \pm 2 \sqrt{\frac{pq}{n}} < 1$

Example

A homeowner randomly samples 64 homes similar to her own and finds that the average selling price is $250,000 with a standard deviation of $15,000. Estimate the average selling price for all similar homes in the city.

$$\text{Point estimator of } \mu: \{\bar{x} = 250{,}000$$

$$\text{Margin of error: } \pm 1.96\frac{s}{\sqrt{n}} = \pm 1.96\frac{15{,}000}{\sqrt{64}} = \pm 3675$$

Example

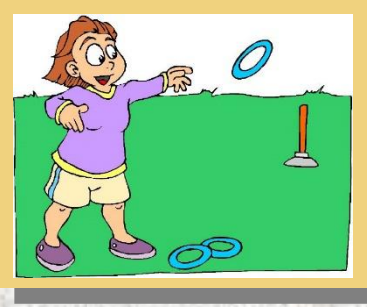
A quality control technician wants to estimate the proportion of soda cans that are underfilled. He randomly samples 200 cans of soda and finds 10 underfilled cans.

$n = 200$ $p = \text{proportion of underfilled cans}$

$\text{Point estimator of } p: \{\hat{p} = x/n = 10/200 = .05$

$$\text{Margin of error: } \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = \pm 1.96\sqrt{\frac{(.05)(.95)}{200}} = \pm .03$$

Interval Estimation/Confidence Interval



- Create an interval (a, b) so that you are fairly sure that the parameter lies between these two values.
- "Fairly sure" means "with high probability", measured using the **confidence coefficient, $1-\alpha$**.

Usually, $1-\alpha = .90, .95, .99$

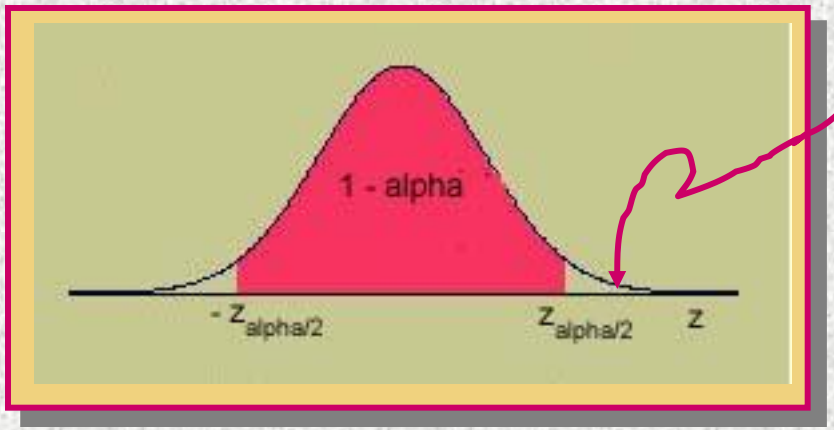
- For large-Sample size,

100(1-α)% Confidence Interval:

Point Estimator $\pm z_{\alpha/2}SE$

To Change the Confidence Level

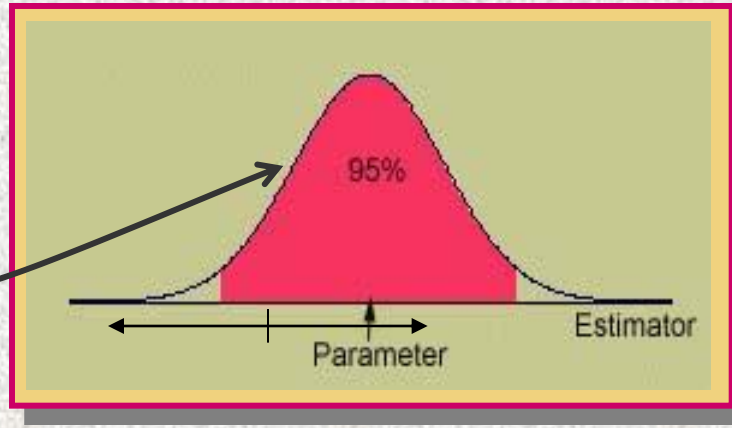
- To change to a general confidence level, 1-α, pick a value of *z* that puts area 1-α in the center of the *z* distribution.



Tail area α/2	$z_{\alpha/2}$
.05	1.645
.025	1.96
.005	2.58

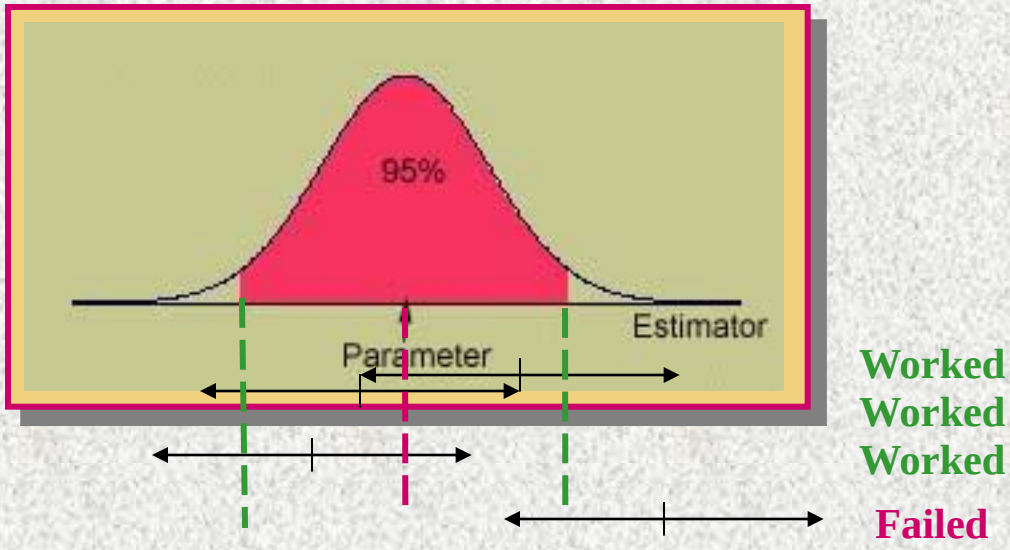
- Suppose 1-α = .95,

95% of the intervals constructed in this manner will enclose the population mean



Interval Estimation/Confidence Interval

- Since we don't know the value of the parameter, consider **Point Estimator ± 1.96SE** which has a variable center.



- Only if the estimator falls in the tail areas will the interval fail to enclose the parameter. This happens only 5% of the time.

Interpretation of A Confidence Interval

- A confidence interval is calculated from one given sample. It either covers or misses the true parameter. Since the true parameter is unknown, you'll never know which one is true.
- If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (**confidence level**) of the intervals will include the unknown population parameter.
- The **confidence level** associated with a confidence interval is the success rate of the confidence interval.

Confidence Intervals for Means and Proportions

- For a quantitative population,

Confidence interval for a population mean μ :

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

- For a binomial population,

Confidence interval for a population proportion p :

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

Example

A random sample of $n = 50$ males showed a mean average daily intake of dairy products equal to 756 grams with a standard deviation of 35 grams. Find a 95% confidence interval for the population average μ.

Example

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}} \Rightarrow 756 \pm 1.96\frac{35}{\sqrt{50}} \Rightarrow 756 \pm 9.70$$

or $746.30 < \mu < 765.70$ grams.

Example

Find a 99% confidence interval for μ, the population average daily intake of dairy products for men.

Example

$$\bar{x} \pm 2.58\frac{s}{\sqrt{n}} \Rightarrow 756 \pm 2.58\frac{35}{\sqrt{50}} \Rightarrow 756 \pm 12.77$$

$$\text{or } 743.23 < \mu < 768.77 \text{ grams.}$$

The interval must be wider to provide for the increased confidence that is does indeed enclose the true value of μ.

Example

Of a random sample of $n = 150$ college students, 104 of the students said that they had played on a soccer team during their K-12 years. Estimate the proportion of college students who played soccer in their youth with a 90% confidence interval.

Example

$$\hat{p} \pm 1.645\sqrt{\frac{\hat{p}\hat{q}}{n}} \quad \Rightarrow \quad \frac{104}{150} \pm 1.645\sqrt{\frac{.69(.31)}{150}}$$

$$\Rightarrow \quad .69 \pm .06 \text{ or } .63 < p < .75.$$

Estimating the Difference between Two Means

- Sometimes we are interested in comparing the means of two populations.
  - The average growth of plants fed using two different nutrients.
  - The average scores for students taught with two different teaching methods.
- To make this comparison,

A random sample of size n_1 drawn from population 1 with mean μ_1 and variance σ_1^2.

A random sample of size n_2 drawn from population 2 with mean μ_2 and variance σ_2^2.

Notations - Comparing Two Means

	Mean	Variance	Standard Deviation
Population 1	μ_1	σ_1^2	σ_1
Population 2	μ_2	σ_2^2	σ_2

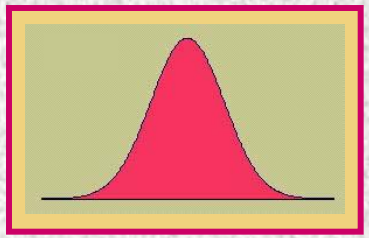
	Sample size	Mean	Variance	Standard Deviation
Sample from Population 1	n_1	$\bar{x}_1$	s_1^2	s_1
Sample from Population 2	n_2	$\bar{x}_2$	s_2^2	s_2

Estimating the Difference between Two Means

- We compare the two averages by making inferences about μ_1-μ_2, the difference in the two population averages.
  - If the two population averages are the same, then μ_1-$\mu_2 = 0$.
  - The best estimate of μ_1-μ_2 is the difference in the two sample means,

$$\bar{x}_1 - \bar{x}_2$$

The Sampling Distribution of $\bar{x}_1 - \bar{x}_2$



1. The mean of $\bar{x}_1 - \bar{x}_2$ is $\mu_1 - \mu_2$, the difference in the population means.

2. The standard deviation of $\bar{x}_1 - \bar{x}_2$ is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

3. If the sample sizes (both n_1 and n_2) are large, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal, and standard deviation can be estimated as SE $= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

Estimating μ_1-μ_2

For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal (z) distribution.

Point estimate for $\mu_1 - \mu_2$: $\{\bar{x}_1 - \bar{x}_2$

Margin of Error : $\pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Confidence interval for $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Assumption :

Both $n_1 \geq 30$ and $n_2 \geq 30$

Example

Avg Daily Intakes	Men	Women
Sample size	50	50
Sample mean	756	762
Sample Std Dev	35	30

Compare the average daily intake of dairy products of men and women using a 95% confidence interval.

Example

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\Rightarrow (756 - 762) \pm 1.96\sqrt{\frac{35}{50} + \frac{30}{50}} \quad \Rightarrow \quad -6 \pm 12.78$$

or $-18.78 < \mu_1 - \mu_2 < 6.78.$

Example, continued

$$-18.78 < \mu_1 - \mu_2 < 6.78$$

- Could you conclude, based on this confidence interval, that there is a difference in the average daily intake of dairy products for men and women?
- The confidence interval contains the value $\mu_1 - \mu_2 = 0$. Therefore, it is possible that $\mu_1 = \mu_2$. You would not want to conclude that there is a difference in average daily intake of dairy products for men and women.

Summary

Let $X_1, \ldots, X_{nX}$ be a *large* random sample of size n_X from a population with mean μ_X and standard deviation σ_X, and let $Y_1, \ldots, Y_{nY}$ be a *large* random sample of size n_Y from a population with mean μ_Y and standard deviation σ_Y. If the two samples are independent, then a level $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is

$$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \qquad (5.16)$$

When the values of σ_X and σ_Y are unknown, they can be replaced with the sample standard deviations s_X and s_Y.

Estimating the Difference between Two Proportions

- Sometimes we are interested in comparing the proportion of "successes" in two binomial populations.
  - The germination rates of untreated seeds and seeds treated with a fungicide.
  - The proportion of male and female voters who favor a particular candidate for governor.
- To make this comparison,

A random sample of size n_1 drawn from binomial population 1 with parameter p_1.

A random sample of size n_2 drawn from binomial population 2 with parameter p_2.

Notations - Comparing Two Proportions

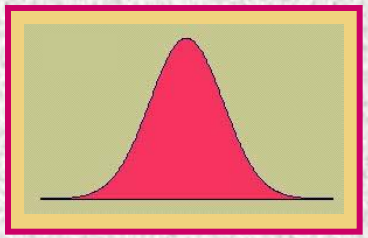
	Sample size	Sample Proportion	Sample Variance	Standard Deviation
Sample from Population 1	n_1	$\hat{p}_1 = \frac{x_1}{n_1}$	$\frac{\hat{p}_1\hat{q}_1}{n}$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n}}$
Sample from Population 2	n_2	$\hat{p}_2 = \frac{x_2}{n_2}$	$\frac{\hat{p}_2\hat{q}_2}{n}$	$\sqrt{\frac{\hat{p}_2\hat{q}_2}{n}}$

Estimating the Difference between Two Proportions

- We compare the two proportions by making inferences about p_1-p_2, the difference in the two population proportions.
  - If the two population proportions are the same, then p_1-$p_2 = 0$.
  - The best estimate of p_1-p_2 is the difference in the two sample proportions,

$$\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$$

The Sampling Distribution of $\hat{p}_1 - \hat{p}_2$



1. The mean of $\hat{p}_1 - \hat{p}_2$ is $p_1 - p_2$, the difference in the population proportions.

2. The standard deviation of $\hat{p}_1 - \hat{p}_2$ is $\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$.

3. If the sample sizes (both n_1 and n_2) are large, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal, and standard deviation can be estimated as

$$SE = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}.$$

Estimating p_1-p_2

For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal (z) distribution.

Point estimate for $p_1 - p_2$: $\{\hat{p}_1 - \hat{p}_2$

$$\text{Margin of Error}: \pm 1.96\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Confidence interval for $p_1 - p_2$:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Assumption : both n_1 and n_2 are sufficiently large so that

$$-1 \le \hat{p}_1 - \hat{p}_2 \pm 2SE \le 1$$

Example

Youth Soccer	Male	Female
Sample size	80	70
Played soccer	65	39

Compare the proportion of male and female college students who said that they had played on a soccer team during their K-12 years using a 99% confidence interval.

Example

Youth Soccer	Male	Female
Sample size	80	70
Played soccer	65	39

Compare the proportion of male and female college students who said that they had played on a soccer team during their K-12 years using a 99% confidence interval.

$$(\hat{p}_1 - \hat{p}_2) \pm 2.58\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\Rightarrow \left(\frac{65}{80} - \frac{39}{70}\right) \pm 2.58\sqrt{\frac{.81(.19)}{80} + \frac{.56(.44)}{70}} \quad \Rightarrow \quad .25 \pm .19$$

or $.06 < p_1 - p_2 < .44.$

Example, continued

$$.06 < p_1 - p_2 < .44$$

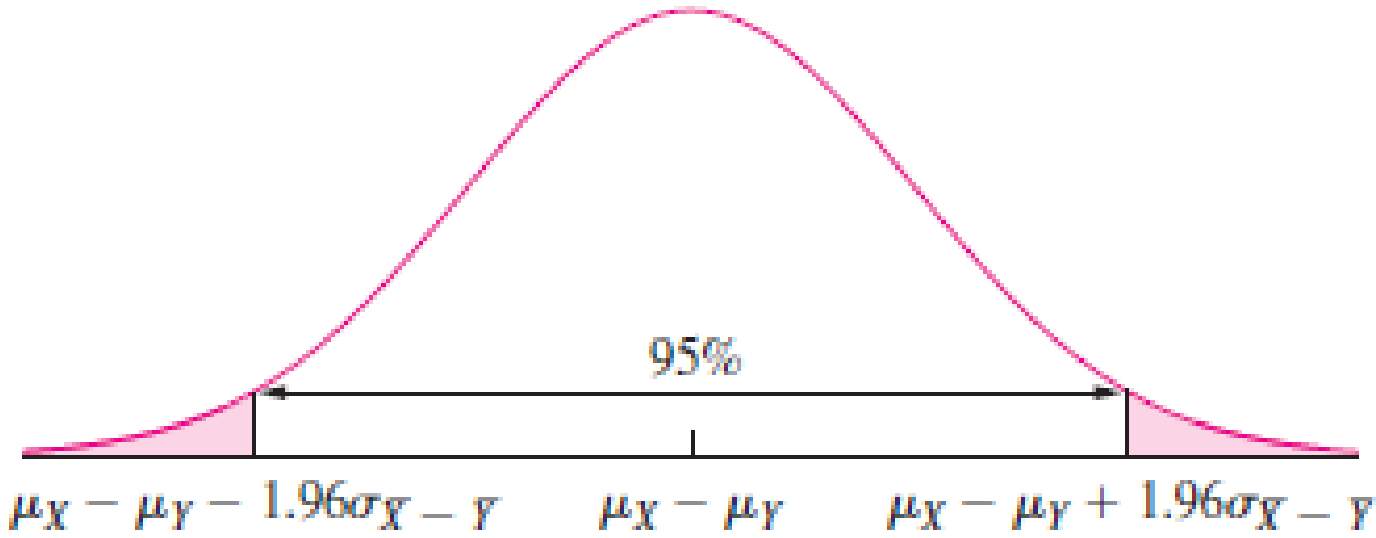
- Could you conclude, based on this confidence interval, that there is a difference in the proportion of male and female college students who said that they had played on a soccer team during their K-12 years?
- The confidence interval does not contains the value $p_1 - p_2 = 0$. Therefore, it is not likely that $p_1 = p_2$. You would conclude that there is a difference in the proportions for males and females.

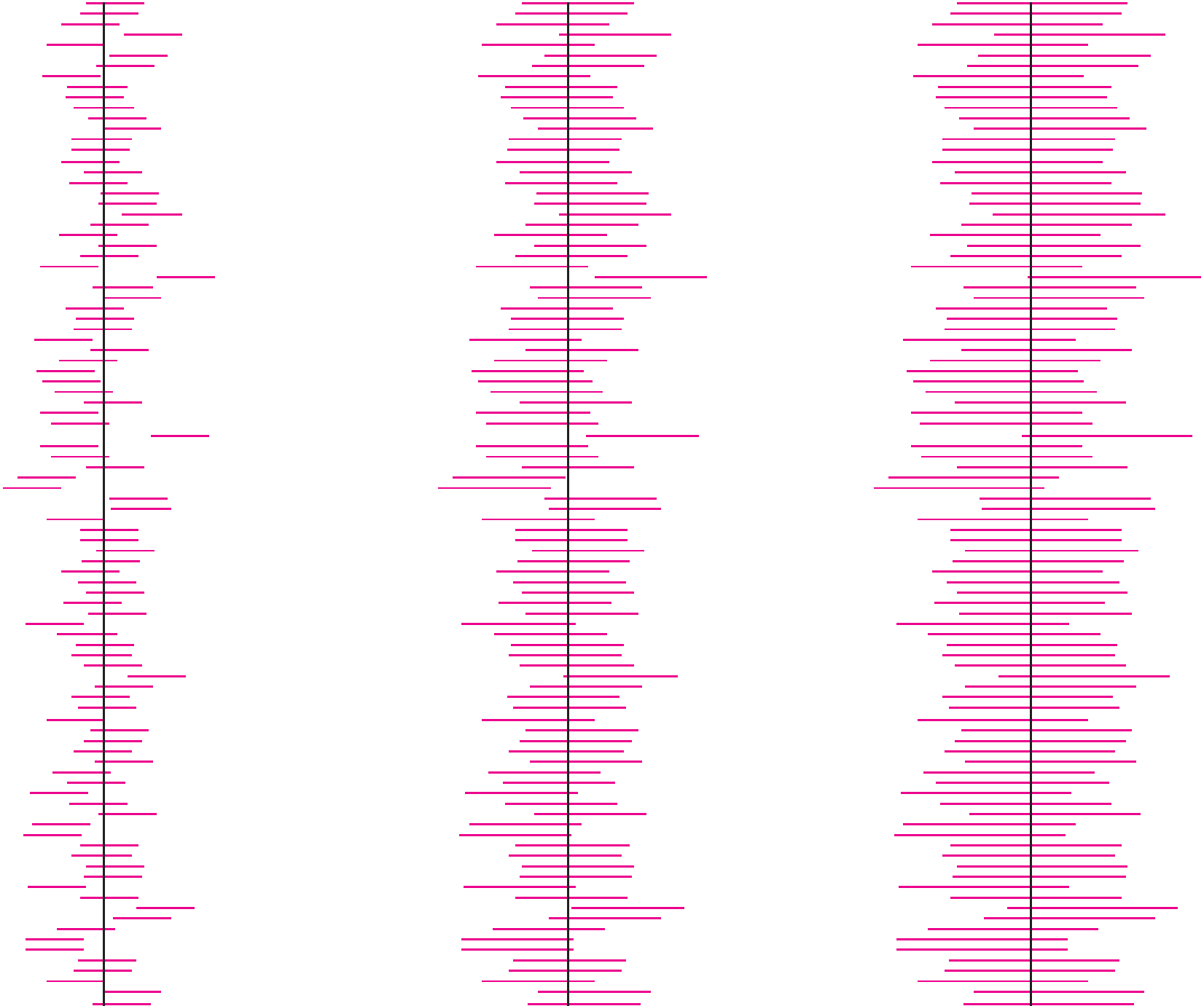
A higher proportion of males than females played soccer in their youth.

Let X and Y be independent, with $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$. Then

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \tag{5.14}$$

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2) \tag{5.15}$$





Probability versus Confidence

a 95% confidence interval for the population mean μ was computed to be (12.304, 12.696).

It is tempting to say that the probability is 95% that μ is between 12.304 and 12.696.

The term *probability* refers to random events, which can come out differently when experiments are repeated.

The numbers 12.304 and 12.696 are fixed, not random.

The population mean is also fixed. The mean is either in the interval 12.304 to 12.696, or it is not.

There is no randomness involved. Therefore we say that we have 95% *confidence* (not probability) that the population mean is in this interval.

A 90% confidence interval for the mean diameter (in cm) of steel rods manufactured on a certain extrusion machine is computed to be (14.73, 14.91). True or false: The probability that the mean diameter of rods manufactured by this process is between 14.73 and 14.91 is 90%.

Solution

False. A specific confidence interval is given. The mean is either in the interval or it isn't. We are 90% confident that the population mean is between 14.73 and 14.91.

The term *probability* is inappropriate.

An engineer plans to compute a 90% confidence interval for the mean diameter of steel rods. She will measure the diameters of a large sample of rods, compute X and s, and then compute the interval $X \pm 1.645s/\sqrt{n}$. True or false: The probability that the population mean diameter will be in this interval is 90%.

Solution

True. What is described here is a method for computing a confidence interval, rather than a specific numerical value. It is correct to say that a method for computing a 90% confidence interval has probability 90% of covering the population mean.

The sample mean and standard deviation for the fill weights of 100 boxes are $X = 12.05$ and $s = 0.1$ oz. How many boxes must be sampled to obtain a 99% confidence interval of width ±0.012 oz?

Solution

The level is 99%, so $1 - \alpha = 0.99$. Therefore $\alpha = 0.01$ and $z\alpha/2 = 2.58$. The value of σ is estimated with $s = 0.1$. The necessary sample size is found by solving $(2.58)(0.1)/\sqrt{n} = 0.012$. We obtain $n \approx 463$.

One-Sided Confidence Intervals

Summary

Let $X_1, \ldots, X_n$ be a *large* ($n > 30$) random sample from a population with mean μ and standard deviation σ, so that $\overline{X}$ is approximately normal. Then level $100(1-\alpha)\%$ lower confidence bound for μ is

$$\overline{X} - z_\alpha \sigma_{\overline{X}} \tag{5.2}$$

and level $100(1-\alpha)\%$ upper confidence bound for μ is

$$\overline{X} + z_\alpha \sigma_{\overline{X}} \tag{5.3}$$

where $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. When the value of σ is unknown, it can be replaced with the sample standard deviation s.

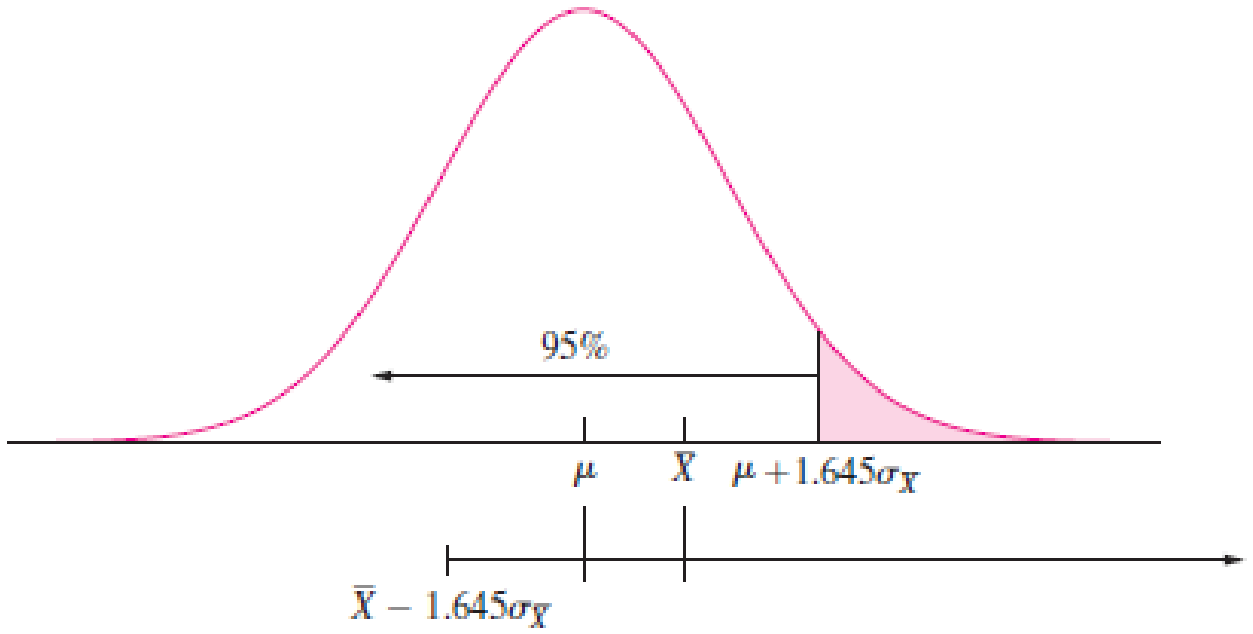


FIGURE 5.6 The sample mean $\overline{X}$ is drawn from a normal distribution with mean μ and standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. For this particular sample, $\overline{X}$ comes from the lower 95% of the distribution, so the 95% one-sided confidence interval $(\overline{X} - 1.645\sigma_{\overline{X}}, \infty)$ succeeds in covering the population mean μ.

a sample of 50 microdrills drilling a low-carbon alloy steel, the average lifetime (expressed as the number of holes drilled before failure) was 12.68 with a standard deviation of 6.83. Find both a 95% lower confidence bound and a 99% upper confidence bound for the mean lifetime of the microdrills.

Solution

The sample mean and standard deviation are $X = 12.68$ and $s = 6.83$, respectively. The sample size is $n = 50$. We estimate $\sigma X \approx s/\sqrt{n} = 0.9659$.

The 95% lower confidence bound is $X - 1.645\sigma X = 11.09$, and the 99% upper confidence bound is $X + 2.33\sigma X = 14.93$.

Confidence Intervals Must Be Based on Random Samples

A chemical engineer wishes to estimate the mean yield of a new process. The process is run 100 times over a period of several days. Figure presents the 100 yields plotted against time. Would it be appropriate to compute a confidence interval for the mean yield by calculating X and s for the yields.

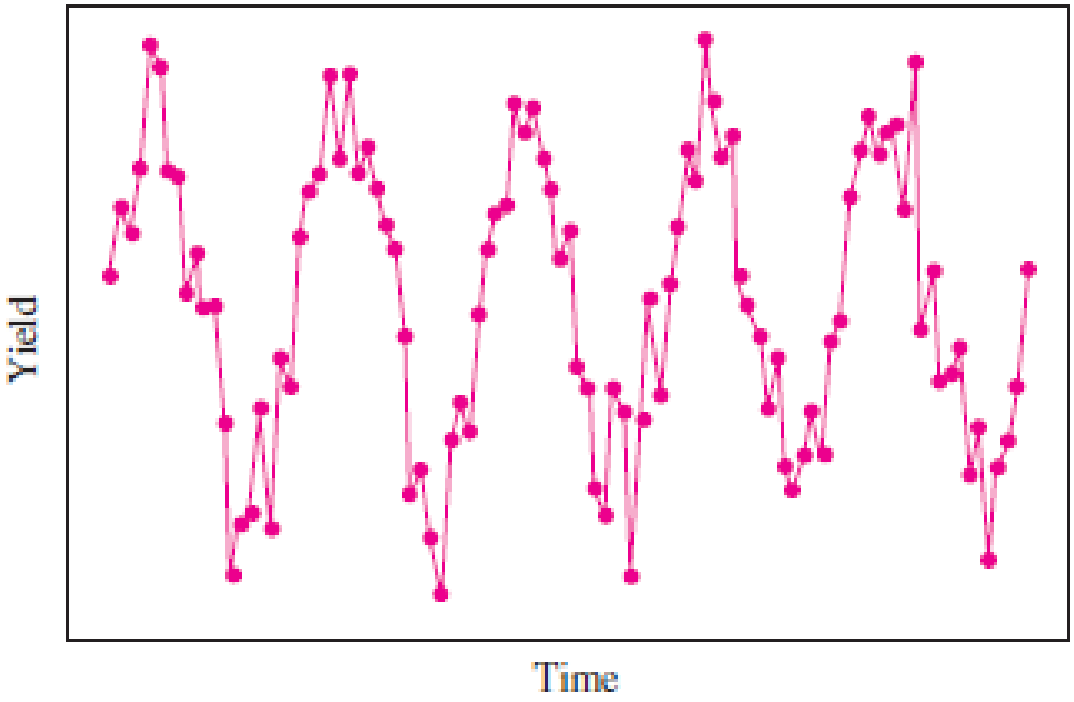


FIGURE 5.7 Yields from 100 runs of a chemical process, plotted against time. There is a clear pattern, indicating that the data do not constitute a random sample.

The engineer in Example is studying the yield of another process. Figure presents yields from 100 runs of this process, plotted against time.

compute a confidence interval for the mean yield of this process?

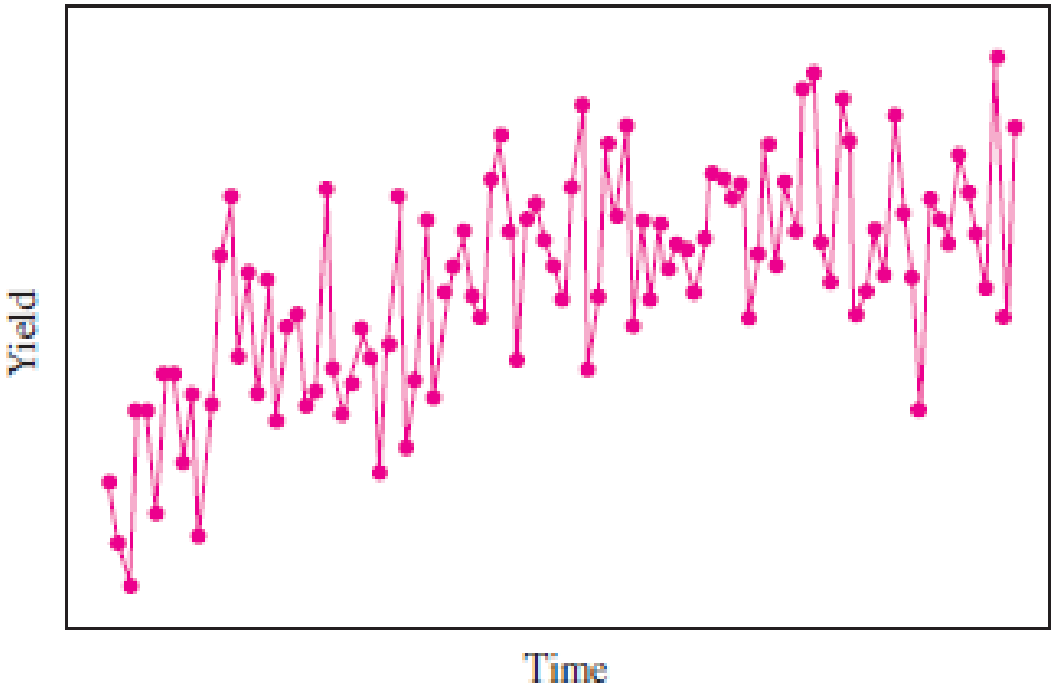


FIGURE 5.8 Yields from 100 runs of a chemical process, plotted against time. There is an increasing trend with time, at least in the initial part of the plot, indicating that the data do not constitute a random sample.

Summary

The Traditional Method for Computing Confidence Intervals for a Proportion (widely used but not recommended)

Let $\widehat{p}$ be the proportion of successes in a *large* number n of independent Bernoulli trials with success probability p. Then the traditional level $100(1-\alpha)\%$ confidence interval for p is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \qquad (5.8)$$

The method cannot be used unless the sample contains at least 10 successes and 10 failures.

Agresti Coull confidence interval

The traditional method of computing confidence interval was justified on the basis of the Central Limit Theorem,

which requires n to be large.

However, this method of computing confidence intervals is

appropriate for any sample size n.

When used with small samples, it may occasionally happen that the lower limit is less than 0 or that the upper limit is greater than 1.

Since $0 < p < 1$, a lower limit less than 0 should be replaced with 0, and an upper limit greater than 1 should be replaced with 1.

The confidence interval given by expression (5.5) is sometimes called the *Agresti.Coull* interval, after Alan Agresti and Brent Coull, who developed it.

Summary

Let X be the number of successes in n independent Bernoulli trials with success probability p, so that $X \sim \text{Bin}(n, p)$.

Define $\tilde{n} = n + 4$, and $\tilde{p} = \dfrac{X + 2}{\tilde{n}}$. Then a level $100(1 - \alpha)\%$ confidence interval for p is

$$\tilde{p} \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}} \qquad (5.5)$$

If the lower limit is less than 0, replace it with 0. If the upper limit is greater than 1, replace it with 1.

Agresti–Coull interval [edit]

The Agresti–Coull interval is also another approximate binomial confidence interval.[10]

Given X successes in n trials, define

$$\tilde{n} = n + z^2$$

and

$$\tilde{p} = \frac{1}{\tilde{n}}\left(X + \frac{z^2}{2}\right)$$

Then, a confidence interval for p is given by

$$\tilde{p} \pm z\sqrt{\frac{\tilde{p}}{\tilde{n}}(1-\tilde{p})}$$

where z is the $1 - \frac{\alpha}{2}$ quantile of a standard normal distribution, as before. For example, for a 95% confidence interval, let $\alpha = 0.05$, so $z = 1.96$ and $z^2 = 3.84$. If we use 2 instead of 1.96 for z, this is the "add 2 successes and 2 failures" interval in Agresti and Coull's 1998 paper "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions." [10]

Arcsine transformation [edit]

Further information: Cohen's h

Let X be the number of successes in n trials and let $p = X/n$. The variance of p is

$$\operatorname{var}(p) = \frac{p(1-p)}{n}.$$

Using the arc sine transform the variance of the arcsine of $p^{1/2}$ is[11]

$$\operatorname{var}\left(\arcsin\left(\sqrt{p}\right)\right) \approx \frac{\operatorname{var}(p)}{4p(1-p)} = \frac{p(1-p)}{4np(1-p)} = \frac{1}{4n}.$$

So, the confidence interval itself has the following form:

Summary

Let X be the number of successes in n independent Bernoulli trials with success probability p, so that $X \sim \text{Bin}(n, p)$.

Define $\tilde{n} = n + 4$, and $\tilde{p} = \dfrac{X + 2}{\tilde{n}}$. Then a level $100(1 - \alpha)\%$ lower confidence bound for p is

$$\tilde{p} - z_\alpha \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}} \tag{5.6}$$

and level $100(1 - \alpha)\%$ upper confidence bound for p is

$$\tilde{p} + z_\alpha \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}} \tag{5.7}$$

If the lower bound is less than 0, replace it with 0. If the upper bound is greater than 1, replace it with 1.

The method made "large" errors (errors whose magnitude was above a commonly accepted threshold) at 26 of the 74 sample test locations. Find a 90% confidence interval for the proportion of locations at which this method will make large errors.

Solution

The number of successes is $X = 26$, and the number of trials is $n = 74$. We therefore compute $\tilde{n} = 74 + 4 = 78$, $\tilde{p} = (26 + 2)/78 = 0.3590$, and

$\sqrt{\tilde{p}(1 - \tilde{p})/\tilde{n}} = (0.3590)(0.6410)/78 = 0.0543$.

For a 90% confidence interval, the value of $\alpha/2$

is 0.05, so $z\alpha/2 = 1.645$.

The 90% confidence interval is therefore $0.3590 \pm (1.645)(0.0543)$, or $(0.270, 0.448)$.

what sample size is needed to obtain a 95% confidence interval with width ±0.08?

Solution

A 95% confidence interval has width ±1.96 $\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}$, where $\tilde{n} = n + 4$.

Therefore we determine

the sample size n by solving the equation $1.96\sqrt{\tilde{p}(1-\tilde{p})/(n+4)} = 0.08$.

$\tilde{p} = 0.3590$. Substituting this value for $\tilde{p}$ and solving, we obtain $n \approx 135$.

how large a sample is needed to guarantee that the width of the 95% confidence interval will be no greater than ±0.08, if no preliminary sample has been taken?

Solution

A 95% confidence interval has width $\pm 1.96\sqrt{\tilde{p}(1-\tilde{p})/(n+4)}$.

The widest the confidence interval could be, for a sample of size n, is $\pm 1.96\sqrt{(0.5)(1-0.5)/(n+4)}$, or $\pm 0.98/\sqrt{n+4}$.

Solving the equation $0.98/\sqrt{n+4} = 0.08$ for n, we obtain $n \approx 147$.

Confidence intervals for paired data

We have seen how to compare two independent samples

Now we will see how to compare two samples that are paired „

In other words the two samples are not independent, Y1 and Y2 are linked in some way, usually by a direct relationship „

For example, measure the weight of subjects before and after a six month diet

A paired t-test is used to compare two population means where you have two samples in which observations in one sample can be paired with observations in the other sample.

Examples of where this might occur are:

- Before-and-after observations on the same subjects (e.g. students' diagnostic test results before and after a particular module or course).
- A comparison of two different methods of measurement or two different treatments where the measurements/treatments are applied to the same subjects (e.g. blood pressure measurements using a stethoscope and a dynamap).

Suppose a sample of n students were given a diagnostic test before studying a particular module and then again after completing the module. We want to find out if, in general, our teaching leads to improvements in students' knowledge/skills (i.e. test scores). We can use the results from our sample of students to draw conclusions about the impact of this module in general.

Let x = test score before the module, y = test score after the module .To find that the true mean difference is zero, the procedure is as follows:

1. Calculate the difference $(d_i = y_i - x_i)$ between the two observations on each pair, making sure you distinguish between positive and negative differences.

2. Calculate the mean difference, $\bar{d}$.

3. Calculate the standard deviation of the differences, s_d, and use this to calculate the standard error of the mean difference, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$

4. Calculate the t-statistic, which is given by $T = \frac{\bar{d}}{SE(d)}$. Under the null hypothesis, this statistic follows a t-distribution with $n - 1$ degrees of freedom.

5. Use tables of the t-distribution to compare your value for T to the t_{n-1} distribution. This will give the p-value for the paired t-test.

NOTE: For this test to be valid the differences only need to be approximately normally distributed. Therefore, it would not be advisable to use a paired t-test where there were any extreme outliers.

Student	Pre-module score	Post-module score	Difference
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

Calculating the mean and standard deviation of the differences gives:

$\bar{d} = 2.05$ and $s_d = 2.837$. Therefore, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{2.837}{\sqrt{20}} = 0.634$

So, we have:

$$t = \frac{2.05}{0.634} = 3.231 \qquad \text{on 19 df}$$

Looking this up in tables gives $p = 0.004$. Therefore, there is strong evidence that, on average, the module does lead to improvements.

Confidence interval for the true mean difference

Summary

Let $D_1, \ldots, D_n$ be a *small* random sample ($n \leq 30$) of differences of pairs. If the population of differences is approximately normal, then a level $100(1-\alpha)\%$ confidence interval for the mean difference μ_D is given by

$$\overline{D} \pm t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}} \tag{5.24}$$

where s_D is the sample standard deviation of $D_1, \ldots, D_n$. Note that this interval is the same as that given by expression (5.9).

If the sample size is large, a level $100(1-\alpha)\%$ confidence interval for the mean difference μ_D is given by

$$\overline{D} \pm z_{\alpha/2}\sigma_{\overline{D}} \tag{5.25}$$

In practice $\sigma_{\overline{D}}$ is approximated with $s_D/\sqrt{n}$. Note that this interval is the same as that given by expression (5.1).

We have a mean difference of 2.05.

The 2.5% point of the t-distribution with 19 degrees of freedom is 2.093.

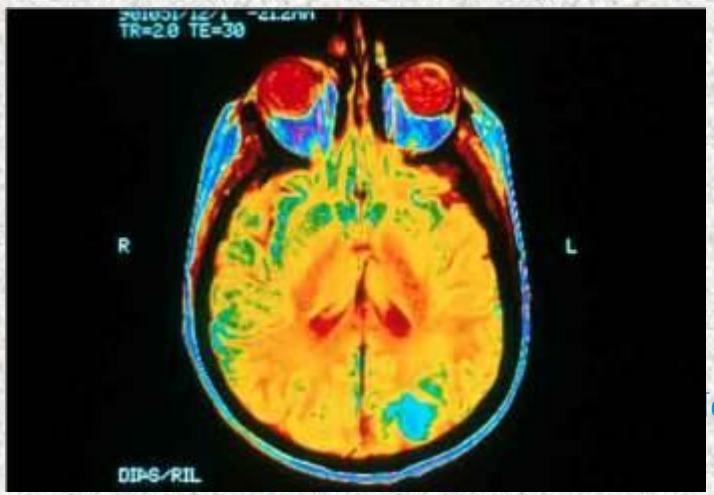
The 95% confidence interval for the true mean difference is therefore:

$2.05 \pm (2.093 \times 0.634) = 2.05 \pm 1.33 = (0.72, 3.38)$

This confirms that, although the difference in scores is statistically significant, it is actually relatively small. We can be 95% sure that the true mean increase lies somewhere between just under one point and just over 3 points.

Example

Are there physiological indicators associated with schizophrenia? In a 1990 article, researchers reported the results of a study that controlled for genetic and socioeconomic differences by examining 15 pairs of identical twins, where one of the twins was schizophrenic and the other not. The researchers used magnetic resonance imaging to measure the volumes (in *cubic centimeters*) of several regions and subregions inside the twins' brains. The following data came from one of the subregions, the left hippocampus:



What is the magnitude of the difference in the volumes of the left hippocampus between (all) unaffected and affected individuals?

Pair	Unaffect	Affect
1	1.94	1.27
2	1.44	1.63
3	1.56	1.47
4	1.58	1.39
5	2.06	1.93
6	1.66	1.26
7	1.75	1.71
8	1.77	1.67
9	1.78	1.28
10	1.92	1.85
11	1.25	1.02
12	1.93	1.34
13	2.04	2.02
14	1.62	1.59
15	2.08	1.97

we can be 95% confident that the mean size for unaffected individuals is between 0.067 and 0.331 cubic centimeters larger than the mean size for affected individuals.

Key Concepts

I. Large-Sample Point Estimators

To estimate one of four population parameters when the sample sizes are large, use the following point estimators with the appropriate margins of error.

Parameter	Point Estimator	Margin of Error
μ	$\bar{x}$	$\pm 1.96\left(\frac{s}{\sqrt{n}}\right)$
p	$\hat{p} = \frac{x}{n}$	$\pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) = \left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)$	$\pm 1.96\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

Key Concepts

II. Large-Sample Interval Estimators

To estimate one of four population parameters when the sample sizes are large, use the following interval estimators.

Parameter	$(1-\alpha)100\%$ Confidence Interval
μ	$\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$
p	$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$
$\mu_1 - \mu_2$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

Key Concepts

1. All values in the interval are possible values for the unknown population parameter.
2. Any values outside the interval are unlikely to be the value of the unknown parameter.
3. To compare two population means or proportions, look for the value 0 in the confidence interval. If 0 is in the interval, it is possible that the two population means or proportions are equal, and you should not declare a difference. If 0 is not in the interval, it is unlikely that the two means or proportions are equal, and you can confidently declare a difference.