Lecture Four

Section 4.1 – Inferences About Two Portions

Objectives

Test a claim two population proportions or construct a confidence interval estimate of the difference between two population properties.

Notation for Two Proportions

For population 1, we let:

p_1 = population proportion

n_1 = size of the sample

x_1 = number of successes in the sample (the sample proportion)

The corresponding notations apply to which come from population 2.

Pooled Sample Proportion

The pooled sample proportion is denoted by p and is given by:

$$\bar{p} = \frac{x_1 + n_2}{n_1 + n_2} \qquad \bar{q} = 1 - \bar{p}$$

- We denote the complement of p by q, **so $q = 1 - p$**

Requirements

We have proportions from two independent simple random samples.
For each of the two samples, the number of successes is at least 5 and the number of failures is at least 5.

Test Statistic for Two Proportions

$$z = \frac{\left(\hat{p}_1 - \hat{p}_2\right) - \left(p_1 - p_2\right)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}} \qquad \text{where } p_1 - p_2 = 0 \text{ (assumed in the null hypothesis)}$$

$$\hat{p}_1 = \frac{x_1}{n_1} \quad \textit{and} \quad \hat{p}_2 = \frac{x_2}{n_2} \qquad \text{(sample proportions)}$$

$$\bar{p}_1 = \frac{x_1 + x_2}{n_1 + n_2} \qquad \text{(pooled sample proportion)}$$

$$\bar{q} = 1 - \bar{p}$$

P-value: Use Standard Normal Distribution Table. (Use the computed value of the test statistic *z* and find its *P*-value by following the procedure summarized by Figure 8-5 in the text.)

Critical values: Use Standard Normal Distribution Table. (Based on the significance level *a*, find critical values by using the procedures introduced in Section 8-2 in the text.)

Confidence Interval Estimate of $p_1 - p_2$

The confidence interval estimate of the difference $p_1 - p_2$ is:

$$\left(\hat{p}_1 - \hat{p}_2\right) - E < \left(p_1 - p_2\right) < \left(\hat{p}_1 - \hat{p}_2\right) + E$$

Where the margin of error *E* is given by

$$E = z_{\alpha/2}\sqrt{\frac{\hat{P}_1\hat{q}_1}{n_1} + \frac{\hat{P}_2\hat{q}_2}{n_2}}$$

Rounding: Round the confidence interval limits to three significant digits,

Example

The table below lists results from a simple random sample of front-seat occupants involved in car crashes. Use a 0.05 significance level to test the claim that the fatality rate of occupants is lower for those in cars equipped with airbags.

	Airbag Available	No Airbag Available
Occupant Fatalities	41	52
Total number of occupants	11,541	9,853

Solution

Requirements are satisfied: two simple random samples, two samples are independent; Each has at least 5 successes and 5 failures (11,500, 41; 9801, 52).
Use the *P*-value method.

Step 1: Express the claim as $p_1 < p_2$.

Step 2: If $p_1 < p_2$ is false, then $p_1 \geq p_2$.

Step 3: $p_1 < p_2$ does not contain equality so it is the alternative hypothesis. The null hypothesis is the statement of equality.

H_0: $p_1 = p_2$

H_a: $p_1 < p_2$ (original claim)

Step 4: Significance level is 0.05

Step 5: Use normal distribution as an approximation to the binomial distribution. Estimate the common values of p_1 and p_2 as follows:

Step 6: Find the value of the test statistic.

$$z = \frac{\left(\hat{p}_1 - \hat{p}_2\right) - \left(p_1 - p_2\right)}{\sqrt{\frac{\overline{pq}}{n_1} + \frac{\overline{pq}}{n_2}}}$$

$$= \frac{\left(\frac{41}{11{,}541} - \frac{52}{9{,}853}\right) - 0}{\sqrt{\frac{(0.004347)(0.995653)}{11{,}541} + \frac{(0.004347)(0.995653)}{9{,}853}}}$$

$$= -1.91$$

Left-tailed test. Area to left of $z = -1.91$ is 0.0281 (Standard Normal Distribution Table), so the *P*-value is 0.0281.

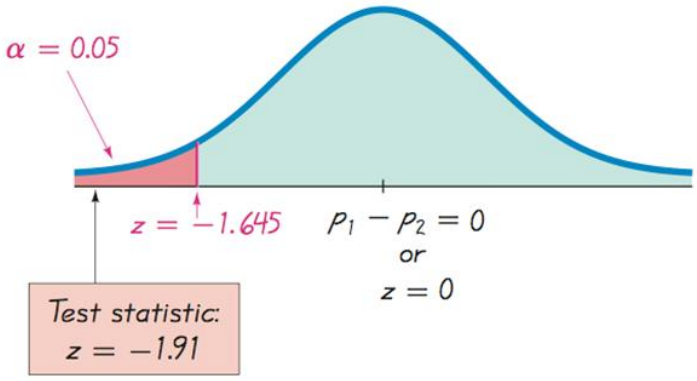
Step 7: Because the *P*-value of 0.0281 is less than the significance level of $\alpha = 0.05$, we reject the null hypothesis of $p_1 = p_2$.

Because we reject the null hypothesis, we conclude that there is sufficient evidence to support the claim that the proportion of accident fatalities for occupants in cars with airbags is less than the proportion of fatalities for occupants in cars without airbags. Based on these results, it appears that airbags are effective in saving lives.

***Example*: Using the Traditional Method**

With a significance level of $\alpha = 0.05$ in a left- tailed test based on the normal distribution, we refer to Standard Normal Distribution Table and find that an area of $\alpha = 0.05$ in the left tail corresponds to the critical value of $z = -1.645$. The test statistic of does fall in the critical region bounded by the critical value of $z = -1.645$.

We again reject the null hypothesis.



Caution

When testing a claim about two population proportions, the *P*-value method and the traditional method are equivalent, but they are *not* equivalent to the confidence interval method. If you want to test a claim about two population proportions, use the *P*-value method or traditional method; if you want to estimate the difference between two population proportions, use a confidence interval.

Example

Use the sample data given in the preceding Example to construct a 90% confidence interval estimate of the difference between the two population proportions. (As shown in Table 8-2 on page 406, the confidence level of 90% is comparable to the significance level of $\alpha = 0.05$ used in the preceding left-tailed hypothesis test.) What does the result suggest about the effectiveness of airbags in an accident?

Solution

Requirements are satisfied as we saw in the preceding example.

90% confidence interval: $z_{\alpha/2} = 1.645$ Calculate the margin of error, E

$$E = z_{\alpha/2}\sqrt{\frac{\hat{P}_1\hat{q}_1}{n_1} + \frac{\hat{P}_2\hat{q}_2}{n_2}}$$

$$= 1.645\sqrt{\frac{\frac{41}{11{,}541}\cdot\frac{11{,}500}{11{,}541}}{11{,}541} + \frac{\frac{52}{9{,}853}\cdot\frac{9801}{9{,}853}}{9{,}853}}$$

$$= 0.001507$$

Construct the confidence interval

Example

The confidence interval limits do not contain 0, implying that there is a significant difference between the two proportions. The confidence interval suggests that the fatality rate is lower for occupants in cars with air bags than for occupants in cars without air bags. The confidence interval also provides an estimate of the amount of the difference between the two fatality rates.

Why Do the Procedures of This Section Work?

The distribution of can be approximated by a normal distribution with mean p_1, standard deviation and variance p_1q_1/n_1.

The difference can be approximated by a normal distribution with mean $p_1 - p_2$ and variance

$$\sigma^2_{\left(\hat{p}_1 - \hat{p}_2\right)} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2} = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}$$

The variance of the *differences* between two independent random variables is the *sum* of their individual variances.
The preceding variance leads to

$$\sigma_{\left(\hat{p}_1-\hat{p}_2\right)}=\sqrt{\frac{\bar{p}_1\bar{q}_1}{n_1}+\frac{\bar{p}_2\bar{q}_2}{n_2}}$$

We now know that the distribution of $p_1 - p_2$ is approximately normal, with mean $p_1 - p_2$ and standard deviation as shown above, so the z test statistic has the form given earlier.
When constructing the confidence interval estimate of the difference between two proportions, we don't assume that the two proportions are equal, and we estimate the standard deviation as

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

In the test statistic $z=\dfrac{\left(\hat{p}_1-\hat{p}_2\right)-\left(p_1-p_2\right)}{\sqrt{\dfrac{\bar{p}\bar{q}}{n_1}+\dfrac{\bar{p}\bar{q}}{n_2}}}$

use the positive and negative values of z (for two tails) and solve for $p_1 - p_2$. The results are the limits of the confidence interval given earlier.

TI-83 / 84 Calculator – For Hypothesis and confidence intervals

Press **STAT**

Select **TESTS**

Choose the option of **2-PropZTest** (for hypothesis test)

Or **2-PropZInt** (for confidence test)

Result:

Calculator will display a P-value instead of critical values, so the P-value method of hypothesis is used.

Exercises *Section* 4.1 – Inferences about Two Portions

1. A Student surveyed her friends and found that among 20 males, 4 smoke and among 30 female, 6 smoke. Give two reasons why these results should not be used for a hypothesis test of the claim that the proportions of male smokers and female smokers are equal.

2. In clinical trials of the drug Zocor, some subjects were treated with Zocor and other were given a placebo. The 95% confidence interval estimate of the difference between the proportions of subjects who experienced headaches is $-0.0518 < p_1 - p_2 < 0.0194$. Write a statement interpreting that confidence interval.

3. Among 8834 malfunctioning pacemakers, in 15.8% the malfunctions were due to batteries. Find the number of successes *x*.

4. Among 129 subjects who took Chantix as an aid to stop smoking, 12.4% experienced nausea. Find the number of successes *x*.

5. Among 610 adults selected randomly from among the residents of one town, 26.1% said that they have favor stronger gun-control laws. Find the number of successes *x*.

6. A computer manufacturer randomly selects 2,410 of its computers for quality assurance and finds that 3.13% of these computer are found defective. Find the number of successes *x*.

7. Assume that you plan to use a significance level of $\alpha = 0.05$ to test the claim that $p_1 = p_2$. Use the given sample sizes and number of successes to find the pooled estimate $\bar{p}$

a) $n_1 = 288, \quad n_2 = 252, \quad x_1 = 75, \quad x_2 = 70$

b) $n_1 = 100, \quad n_2 = 100, \quad \hat{p}_1 = 0.2, \quad \hat{p}_2 = 0.18$

8. The numbers of online applications from simple random samples of college applications for 2003 and for the current year are given below.

	2003	Current Year
Number of application in sample	36	27
Number of online applications in sample	13	14

Assume that you plan to use a significance level of $\alpha = 0.05$ to test the claim that $p_1 = p_2$. Find

a) The pooled estimate $\bar{p}$

b) The *x* test statistic

c) The critical *z* values

d) The *P*–value

Assume 95% confidence interval

e) The margin of error E

f) The 95% confidence interval.

9. Chantix is a drug used as an aid to stop smoking. The numbers of subjects experiencing insomnia for each of two treatment groups in a clinical trial of the drug Chantix are given below:

	Chantix Treatment	Placebo
Number in group	129	805
Number experiencing insomnia	19	13

Assume that you plan to use a significance level of $\alpha = 0.05$ to test the claim that $p_1 = p_2$. Find

a) The pooled estimate $\overline{p}$

b) The *x* test statistic

c) The critical *z* values

d) The *P*–value

Assume 95% confidence interval

e) The margin of error E

f) The 95% confidence interval.

10. In a 1993 survey of 560 college students, 171 said that they used illegal drugs during the previous year. In a recent survey of 720 college students, 263 said that they used illegal drugs during the previous year. Use a 0.05 significance level to test the claim that the proportion of college students using illegal drugs in 1993 was less than it is now.

11. In a 1993 survey of 560 college students, 171 said that they used illegal drugs during the previous year. In a recent survey of 720 college students, 263 said that they used illegal drugs during the previous year. Construct the confidence interval corresponding to the hypothesis test conducted with a 0.05 significance level. What conclusion does the confidence interval suggest?

12. A simple random sample of front-seat occupants involved in car crashes is obtained. Among 2823 occupants not wearing seat belts, 31 were killed. Among 7765 occupants wearing seat belts, 16 were killed. Construct a 90% confidence interval estimate of the difference between the fatality rates for those not wearing seat belts and those wearing seat belts. What does the result suggest about the effectiveness of seat belts?

13. A Pew Research Center poll asked randomly selected subjects if they agreed with the statement that "It is morally wrong fir married people to have an affair/" Among the 386 women surveyed, 347 agrees with the statement. Among the 359 men surveyed, 305 agreed with the statement.

a) Use a 0.05 significance level to test the claim that the percentage of women who agree is difference from the percentage of men who agree. Does there appear to be a difference in the way women and men feel about this issue?

b) Construct the confidence interval corresponding to the hypothesis test conducted with a 0.05 significance level. What conclusion does the confidence interval suggest?

14. Tax returns include an option of designating $3 for presidential election campaigns, and it does not cost the taxpayer anything to make that designation. In a simple random sample of 250 tax returns from 1976, 27.6% of the returns designated the $3 for the campaign. In a simple random sample of 300 recent tax returns, 7.3% of the returns designated the $3 for the campaign. Use a 0.05 significance level to test the claim that the percentage of returns designating the $3 for the campaign was greater in 1973 than it is now.

15. In an experiment, 16% of 734 subjects treated with Viagra experienced headaches. In the same experiment, 4% of 725 subjects given a placebo experienced headaches.

a) Use a 0.01 significance level to test the claim that the proportion of headaches is greater for those treated with Viagra. Do headaches appear to be a concern for those who take Viagra?

b) Construct the confidence interval corresponding to the hypothesis test conducted with a 0.01 significance level. What conclusion does the confidence interval suggest?

16. Two different simple random samples are drawn from two different populations. The first sample consists of 20 people with 10 having a common attribute. The second sample consists of 2000 people with 1404 of them having the same common attribute. Compare the results from a hypothesis test of $p_1 = p_2$ (with a 0.05 significance level) and a 95% confidence interval estimate of $p_1 - p_2$.

17. A report on the nightly news broadcast stated that 11 out of 142 households with pet dogs were burglarized and 21 out of 217 without pet dogs were burglarized. Find the *z* test statistic for the hypothesis test. Assume that you plan to use a significance level of $\alpha = 0.05$ to test the claim that $p_1 = p_2$.

18. Assume that the samples are independent and that they have been randomly selected. Construct a 90% confidence interval for the difference between population proportions $p_1 = p_2$

19. The sample size needed to estimate the difference between two population proportions ti within a margin of error *E* with a confidence level of $1 - \alpha$ can be found as follows:

$$E = z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}.$$

In this expression, replace n_1 and n_2 by n (assuming both samples have the same size) and replace each of p_1, q_1, p_2 *and* q_2 by 0.5 (because their values are not known). Then solve for n.

Use this approach to find the size pf each sample of you want to estimate the difference between the proportions of men and women who plan to vote in the next presidential election. Assume that you want 99% confidence that your error is no more than 0.05.

Section 4.2 – Inferences About Two Means: Independent Samples

Definitions

Two samples are ***independent*** if the sample values selected from one population are not related to or somehow paired or matched with the sample values from the other population.

Two samples are ***dependent*** if the sample values are *paired*. (That is, each pair of sample values consists of two measurements from the same subject (such as before/after data), or each pair of sample values consists of matched pairs (such as husband/wife data), where the matching is based on some inherent relationship.)

Notation

μ_1 = population mean

σ_1 = population standard deviation

n_1 = size of the first sample

= sample mean

s_1 = sample standard deviation

Corresponding notations for μ_2, σ_2, s_2, and n_2 apply to population 2.

Requirements

1. σ_1 an σ_2 are unknown and no assumption is made about the equality of σ_1 and σ_2 .
2. The two samples are independent.
3. Both samples are simple random samples.
4. Either or both of these conditions are satisfied: The two sample sizes are both large (with $n_1 > 30$ and $n_2 > 30$) or both samples come from populations having normal distributions.

Hypothesis Test for Two Means: Independent Samples

$$t = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(where $\mu_1 - \mu_2$ is often assumed to be 0)

Test Statistic for Two Means: Independent Samples

Degrees of freedom

1. In this book we use this simple and conservative estimate: d*f* = smaller of $n_1 - 1$ and $n_2 - 1$.
2. Statistically software typically use the more accurate but more difficult estimate formula

$$df = \frac{(A+B)^2}{\frac{A^2}{n_1 - 1} + \frac{B^2}{n_2 - 1}} \quad \textit{where} \quad A = \frac{s_1^2}{n_1} \quad B = \frac{s_2^2}{n_2}$$

P-values: Refer to *t* Distribution Table. Use the procedure summarized in Figure 8-5.

Critical values: Refer to *t* Distribution Table.

Confidence Interval Estimate of $\mu_1 - \mu_2$: Independent Samples

The confidence interval estimate of the difference $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E \qquad \textit{Where} \qquad E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example

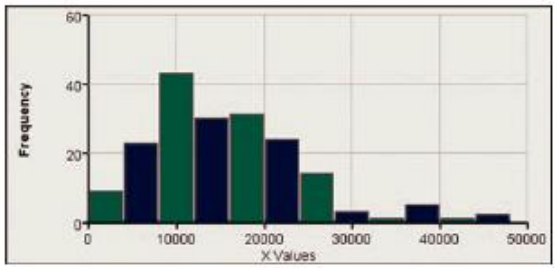
A headline in *USA Today* proclaimed that "Men, women are equal talkers." That headline referred to a study of the numbers of words that samples of men and women spoke in a day. Given below are the results from the study. Use a 0.05 significance level to test the claim that men and women speak the same mean number of words in a day. Does there appear to be a difference?

Number of Words Spoken in a Day

Men	Women
$n_1 = 186$	$n_2 = 210$
$\bar{x}_1 = 15,668.5$	$\bar{x}_2 = 16,215.0$
$s_1 = 8632.5$	$s_2 = 7301.2$

Solution

STATDISK



Requirements are satisfied: two population standard deviations are not known and not assumed to be equal, independent samples, simple random samples, both samples are large.

Step 1: Express claim as $\mu_1 = \mu_2$.

Step 2: If original claim is false, then $\mu_1 \neq \mu_2$.

Step 3: Alternative hypothesis does not contain equality, null hypothesis does.

$H_0 : \mu_1 = \mu_2$ (original claim) $H_a : \mu_1 \neq \mu_2$

Proceed assuming $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$.

Step 4: Significance level is 0.05

Step 5: Because we have two independent samples and we are testing a claim about the two population means, we use a *t* distribution.

Step 6: Calculate the test statistic

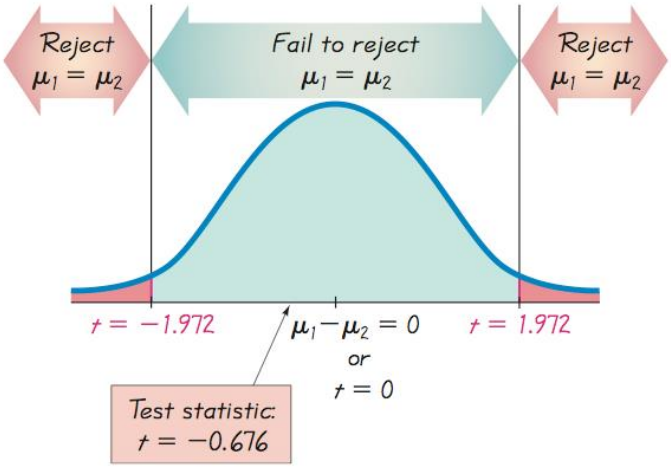
$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{(15{,}668.5 - 16{,}215.0) - 0}{\sqrt{\frac{8632.5^2}{186} + \frac{7301.2^2}{210}}}$$

$$= -0.676$$

Use *t* Distribution Table: area in two tails is 0.05, $df = 185$, which is not in the table, the closest value is $t = \pm 1.972$

Degrees of Freedom	Area in Two Tails 0.01	0.02	0.05	0.10	0.20
200	**2.601**	2.345	**1.972**	1.653	1.286



Step 7: Because the test statistic does not fall within the critical region, fail to reject the null hypothesis: $\mu_1 = \mu_2$ (or $\mu_1 - \mu_2 = 0$).

Conclusion

There is not sufficient evidence to warrant rejection of the claim that men and women speak the same mean number of words in a day. There does not appear to be a significant difference between the two means.

Example

Using the sample data given in the previous Example, construct a 95% confidence interval estimate of the difference between the mean number of words spoken by men and the mean number of words spoken by women.

Number of Words Spoken in a Day

Men	Women
$n_1 = 186$	$n_2 = 210$
$\bar{x}_1 = 15,668.5$	$\bar{x}_2 = 16,215.0$
$s_1 = 8632.5$	$s_2 = 7301.2$

Solution

Requirements are satisfied as it is the same data as the previous example.

Find the margin of Error, E; use $t_{\alpha/2} = 1.972$

$$E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= 1.972\sqrt{\frac{8632.5^2}{186} + \frac{7301.2^2}{210}}$$

$$= 1595.4$$

Construct the confidence interval use E = 1595.4 and

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$$

$$(15,668.5 - 16,215.0) - 1595.4 < (\mu_1 - \mu_2) < (15,668.5 - 16,215.0) + 1595.4$$

$$-2141.9 < (\mu_1 - \mu_2) < 1048.9$$

Step 4: Significance level is 0.05

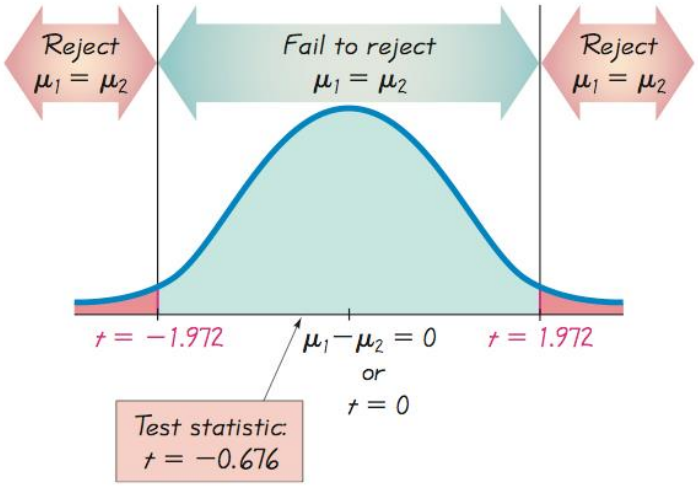
Step 5: Use a t distribution

Step 6: Calculate the test statistic

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{(15,668.5 - 16,215.0) - 0}{\sqrt{\frac{8632.5^2}{186} + \frac{7301.2^2}{210}}}$$

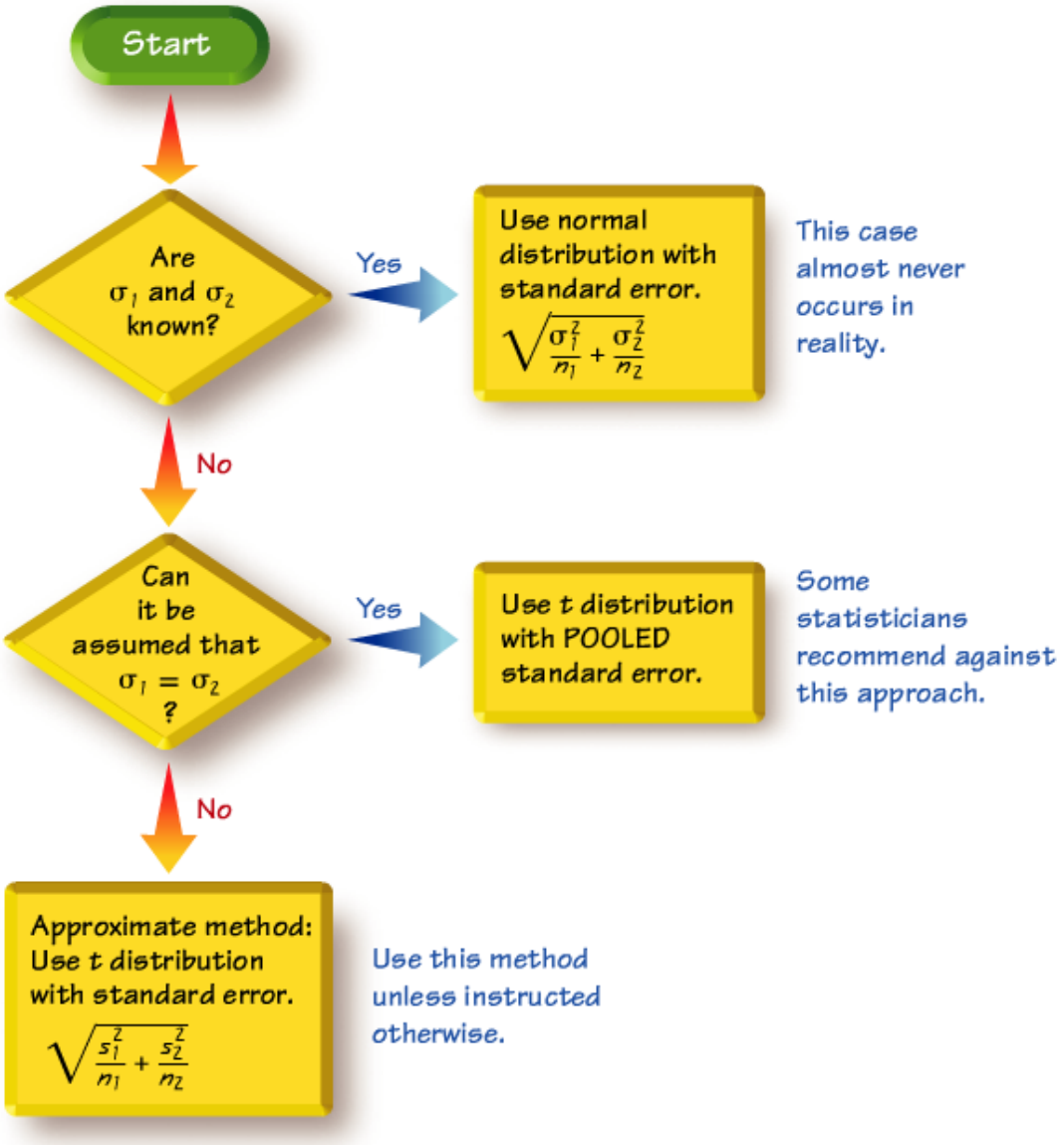
$$= -0.676$$



Use t Distribution Table: area in two tails is 0.05, df = 185, which is not in the table, the closest value is $t = \pm 1.972$

Step 7: Because the test statistic does not fall within the critical region, fail to reject the null hypothesis: $\mu_1 = \mu_2$ (or $\mu_1 - \mu_2 = 0$).

We are 95% confident that the limits of -2141.9 words and 1048.9 words actually do contain the difference between the two population means. Because those limits do contain 0, there is not sufficient evidence to warrant rejection of the claim that men and women speak the same mean number of words in a day. There does not appear to be a significant difference between the two means.



Alternative Methods When σ_1 and σ_2 are Known

Requirements

1. The two population standard deviations are both known.
2. The two samples are independent.
3. Both samples are simple random samples.
4. Either or both of these conditions are satisfied: The two sample sizes are both large (with $n_1 > 30$ and $n_2 > 30$) or both samples come from populations having normal distributions.

Hypothesis Test for Two Means: Independent Samples with σ_1 and σ_2 Both Known

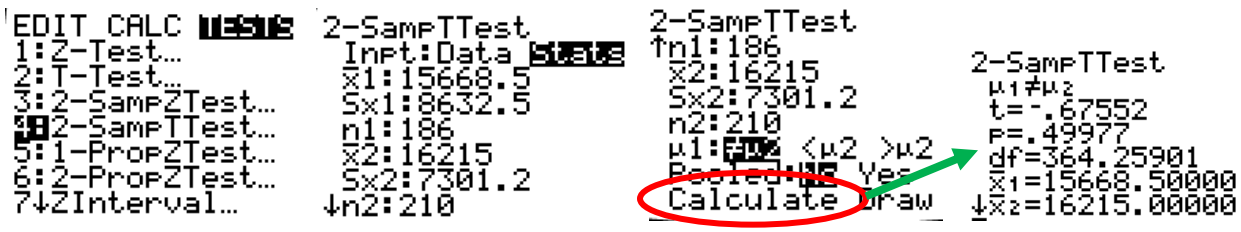
$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

P-values and critical values: Refer to Normal Distribution Table.

Confidence Interval: Independent Samples with σ_1 and σ_2 Both Known

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$$

Where $E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$



Exercises *Section* 4.2 – Inferences About Two Means: Independent Samples

1. If the pulse rates of men and women shown in the data below

Women:

76	72	88	60	72	68	80	64	68	68	80	76	68	72	96	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64

Men:

68	64	88	72	64	72	60	88	76	60	96	72	56	64	60	64	84	76	84	88
72	56	68	64	60	68	60	60	56	84	72	84	88	56	64	56	56	60	64	72

These data are used to construct 95% confidence interval for the difference between the two population means, the result is $-12.2 < \mu_1 - \mu_2 < -1.6$, where pulse rates of men correspond to population 1 and pulse rates of women correspond to population 2. Express the confidence interval with pulse rates of women being population 1 and pulse rates of men being population 2.

2. Assume that you want to use a 0.01 significance level to test the claim that the mean pulse rate of men is less than the mean pulse rate of women. What confidence level should be used if you want to test that claim using a confidence interval?

3. To test the effectiveness of Lipitor, cholesterol levels are measured in 250 subjects before and after Lipitor treatments. Determine whether this sample is independent or dependent.

4. On each of 40 different days, you measured the voltage supplied to your home and you also measured the voltage produced by the gasoline-powered generator. One sample consists of the voltages in the house and the second sample consists of the voltages produced by the generator. Determine whether this sample is independent or dependent.

5. In a randomized controlled trial conducted with children suffering from viral croup, 46 children were treated with low humidity while 46 other children were treated with high humidity. Researchers used the Westley Croup Score to assess the results after one hour. The low humidity group had a mean score of 0.98 with standard deviation of 1.22 while the high humidity group had a mean score of 1.09 with standard deviation of 1.11.

a) Use a 0.05 significance level to test the claim that the two groups are from populations with the same mean. What does the result suggest about the common treatment of humidity?
Assume that the two samples are independent simple random samples selected from normally distributed populations.

b) Assume that $\sigma_1 = \sigma_2$, how are the results affected by this additional assumption?

6. The mean tar content of a simple random sample of 25 unfiltered king size cigarettes is 21.1 mg, with a standard deviation of 3.2 mg. The mean tar content of a simple random sample of 25 filtered 100 mm cigarettes is 13.2 mg, with a standard deviation of 3.7 mg.

Assume that the two samples are independent simple random samples selected from normally distributed populations in part a and b.

a) Construct a 90% confidence interval estimate of the difference between the mean tar content of unfiltered king size cigarettes and the mean tar content of filtered 100 mm cigarettes. Does the result suggest that 100 mm filtered cigarettes have less tar than unfiltered king size cigarettes?

b) Use a 0.05 significance level to test the claim that unfiltered king size cigarettes have a mean tar content greater than that of filtered 100 mm cigarettes. What does the result suggest about the effectiveness of cigarette filters?

c) Assume that $\sigma_1 = \sigma_2$, how are the results affected by this additional assumption?

7. The heights are measured for the simple random sample of supermodels Crawford, Bundchen, Pestova, Christenson, Hume, Moss, Campbell, Schiffer, and Taylor. They have a mean of 70.0 in. and a standard deviation of 1.5 in. 40 women who are not supermodels, listed below and they have heights with means of 63.2 in. and a standard deviation of 2.7 in.

64.3	66.4	62.3	62.3	59.6	63.6	59.8	63.3	67.9	61.4	66.7	64.8	63.1	66.7	66.8
64.7	65.1	61.9	64.3	63.4	60.7	63.4	62.6	60.6	63.5	58.6	60.2	67.6	63.4	64.1
62.7	61.3	58.2	63.2	60.5	65.0	61.8	68.0	67.0	57.0					

a) Use a 0.01 significance level to test the claim that the mean height of supermodels is greater than the mean height of women who are not supermodels

b) Construct a 98% confidence interval level for the difference between the mean height of supermodels and the mean height of women who are not supermodels. What does the result suggest about those two means?

8. Many studies have been conducted to test the effects of marijuana use on mental abilities. In one such study, groups of light and heavy users of marijuana in college were tested for memory recall, with the results given below. Use a 0.01 significance level to test the claim that the population of heavy marijuana users has a lower mean than the light users. Should marijuana use be of concern to college students?

Items sorted correctly by light marijuana users: $n = 64, \quad \bar{x} = 53.3, \quad s = 3.6$

Items sorted correctly by heavy marijuana users: $n = 65, \quad \bar{x} = 51.3, \quad s = 4.5$

9. The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) for Miss America winners from two different time periods. Consider the listed values to be simple random samples selected from larger populations.

a) Use a 0.05 significance level to test the claim that recent winners have a lower mean BMI than winners from the 1920s and 1930s.

b) Construct a 90% Confidence interval for the difference between the mean BMI of recent winners and the mean BMI of winners from the 1920s and 1930s.

BMI (from recent winners):	19.5	20.3	19.6	20.2	17.8	17.9	19.1	18.8	17.6	16.8
BMI (from 1920s and 1930s):	20.4	21.9	22.1	22.3	20.3	18.8	18.9	19.4	18.4	19.1

10. Listed below are amounts of strontium-90 (in millibecquerels or mBq per gram of calcium) in a simple random sample of baby teeth obtained from Pennsylvania residents and New York residents born after 1979.

a) Use a 0.05 significance level to test the claim that the mean amount of strontium-90 from Pennsylvania residents is greater than the mean amount from New York residents.

b) Construct a 90% Confidence interval for the difference between the mean amount of strontium-90 from Pennsylvania residents and the mean amount from New York residents.

Pennsylvania:	155	142	149	130	151	163	151	142	156	133	138	161
New York:	133	140	142	131	134	129	128	140	140	140	137	143

11. Listed below are the word counts for male and female psychology students.

a) Use a 0.05 significance level to test the claim that male and female psychology students speak the same mean number of words in a day.

b) Construct a 95% Confidence interval estimate of the difference between the mean number of words spoken in a day by male and female psychology students. Do the confidence interval limits include 0, and what does that suggest about the two means?

Male	21143	17791	36571	6724	15430	11552	11748	12169	15581	23858	5269
	12384	11576	17707	15229	18160	22482	18626	1118	5319		

Female	6705	21613	11935	15790	17865	13035	24834	7747	3852	11648	25862
	17183	11010	11156	11351	25693	13383	19992	14926	14128	10345	13516
	12831	9671	17011	28575	23557	13656	8231	10601	8124		

Assume that the two samples are independent simple random samples selected from normally distributed populations. Do not assume that the population standard deviations are equal.

12. Refer to the tables below and test the claim that they contain the same amount of cola, the mean weight of cola cans of regular Coke is the same as the mean weight of cola in cans of Diet Coke. If there is a difference in the mean weights, identify the most likely explanation for that difference.

Coke	0.8192	0.815	0.8163	0.8211	0.8181	0.8247	0.8062	0.8128	0.8172	0.811
	0.8251	0.8264	0.7901	0.8244	0.8073	0.8079	0.8044	0.817	0.8161	0.8194
	0.8189	0.8194	0.8176	0.8284	0.8165	0.8143	0.8229	0.815	0.8152	0.8244
	0.8207	0.8152	0.8126	0.8295	0.8161	0.8192				

Diet	0.7773	0.7758	0.7896	0.7868	0.7844	0.7861	0.7806	0.783	0.7852	0.7879
	0.7881	0.7826	0.7923	0.7852	0.7872	0.7813	0.7885	0.776	0.7822	0.7874
	0.7822	0.7839	0.7802	0.7892	0.7874	0.7907	0.7771	0.787	0.7833	0.7822
	0.7837	0.791	0.7879	0.7923	0.7859	0.7811				

Assume that the two samples are independent simple random samples selected from normally distributed populations. Do not assume that the population standard deviations are equal.

13. An Experiment was conducted to test the effects of alcohol. Researchers measured the breath alcohol levels for a treatment group of people who drank ethanol and another group given a placebo. The results are given in the accompanying table. Use a 0.05 significance level to test the claim that the two sample groups come from populations with the same mean.

Treatment Group: $\bar{x}_1 = 0.049$ $s_1 = 0.015$

Placebo Group: $n_2 = 22$ $\bar{x}_2 = 0.000$ $s_2 = 0.000$

14. A researcher was interested in comparing the GPAs $n_1 = 22$ of students at two different colleges. Independent simple populations. Do samples of 8 students from college *A* and 13 students from college *B* yielding the following GPAs.

College *A*	3.7	3.2	3.0	2.5	2.7	3.6	2.8	3.4					
College *B*	3.8	3.2	3.0	3.9	3.8	2.5	3.9	2.8	4.0	3.6	2.6	4.0	3.6

Construct a 95% confidence interval for $\mu_1 - \mu_2$. The difference between the mean GPA of college *A* students and the mean GPA of college *B* students.

$\left(Note: \bar{x}_1 = 3.1125,\ \bar{x}_2 = 3.4385,\ s_1 = 0.4357,\ s_2 = 0.5485\right)$

15. Assume that the two samples are independent simple random samples selected from normal distributed populations. Do not assume that the population standard deviations are equal.
A researcher was interested in comparing the heights of women in two different countries. Independent simple random samples of 9 women from country A and 9 women from B yielded to the following heights (in inches).

Country A	64.1	66.4	61.7	62.0	67.3	64.9	64.7	68.0	63.6
Country B	65.3	60.2	61.7	65.8	61.0	64.6	60.0	65.4	59.0

Construct a 90% confidence interval for $\mu_1 - \mu_2$ the difference between the mean height of women in country A and the mean height of women in country B. Round to two decimal places.

$\left(Note: \bar{x}_1 = 64.744\ in,\ \bar{x}_2 = 62.556\ in,\ s_1 = 2.192\ in,\ s_2 = 2.697\ in\right)$

Section 4.3 – Inferences from Dependent Samples

Objectives

Test a claim about the mean of the differences from dependent samples or construct a confidence interval estimate of the mean of the differences from dependent samples.

Notation

d : Individual difference between the two values in a single matched pair

μ_d : Mean value of the differences d for the *population* of all pairs of data

$\bar{d}$: Mean value of the differences d for the paired *sample* data

s_d : Standard deviation of the differences d for the paired *sample* data

n : Number of all *pairs* of data

Requirements

1. The sample data are dependent.
2. The samples are simple random samples.
3. Either or both of these conditions is satisfied: The number of pairs of sample data is large ($n > 30$) or the pairs of values have differences that are from a population having a distribution that is approximately normal. (These methods are robust against departures for normality, so for small samples, the normality requirement is loose in the sense that the procedures perform well as long as there are no outliers and departures from normality are not too extreme)

Hypothesis Test Statistic for Matched Pairs

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

Where degrees of freedom = $n - 1$

P-values and Critical Values Use: t –distribution Table

Confidence Intervals for Dependent Samples

$$\bar{d} - E < \mu_d < \bar{d} + E \qquad \text{where} \quad E = t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

Critical values of $t_{\alpha/2}$: Use Table *A*-3 with $n - 1$ degrees of freedom.

Example

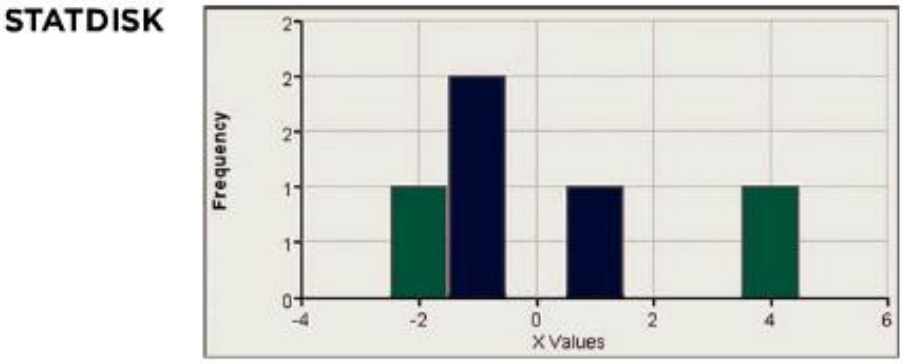
Data Set below includes measured weights of college students in September and April of their freshman year. (Here we use only a small portion of the available data so that we can better illustrate the method of hypothesis testing.) Use the sample data in Table below with a 0.05 significance level to test the claim that for the population of students, the mean change in weight from September to April is equal to 0 kg.

Weight (kg) Measurements of Students in Their Freshman Year

April weight	66	52	68	69	71
September weight	67	53	64	71	70
Difference d = (April weight) – (September weight)	–1	–1	4	–2	1

Solution

Requirements are satisfied: samples are dependent, values paired from each student; although a volunteer study, we'll proceed as if simple random sample and deal with this in the interpretation; STATDISK displays a histogram that is approximately normal



Weight gained = April weight – Sept. weight

μ_d denotes the mean of the "April – Sept." differences in weight; the claim is $\mu_d = 0\ kg$

Step 1: claim is $\mu_d = 0\ kg$

Step 2: If original claim is not true, we have $\mu_d \neq 0\ kg$

Step 3: $H_0 : \mu_d = 0\ kg$ original claim $H_1 : \mu_d \neq 0\ kg$

Step 4: significance level is $\alpha = 0.05$

Step 5: use the student t distribution

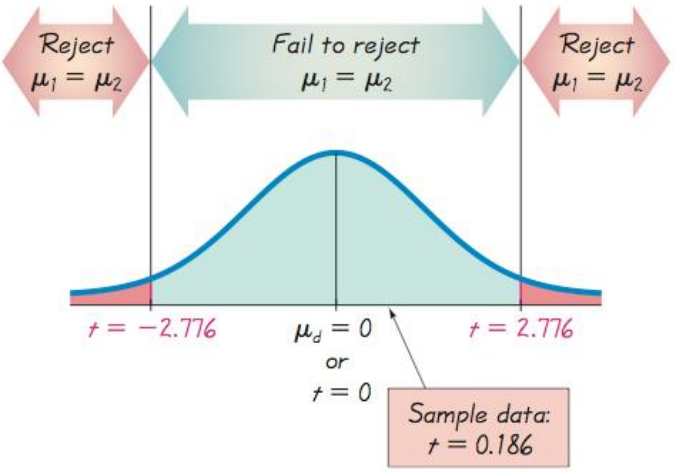
Step 6: find values of d and s_d differences are: –1, –1, 4, –2, 1 $d = 0.2$ and $s_d = 2.4$ now find the test statistic

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{0.2 - 0}{\frac{2.4}{\sqrt{5}}} = 0.186$$

From Table A-3 (t–Distribution): df = n – 1, area in two tails is 0.05, yields a critical value t = ±2.776

Degrees of Freedom	Area in Two Tails				
	0.01	0.02	0.05	0.10	0.20
4	4.604	3.747	2.776	2.132	1.533

Step 7: Because the test statistic does not fall in the critical region, we fail to reject the null *hypothesis*.



We conclude that there is not sufficient evidence to warrant rejection of the claim that for the population of students, the mean change in weight from September to April is equal to 0 kg. Based on the sample results listed in Table, there does not appear to be a significant weight gain from September to April.

The P-value method:

Using technology, we can find the *P*-value of 0.8605. (Using Table A-3 *t*–Distribution: with the test statistic of $t = 0.186$ and 4 degrees of freedom, we can determine that the *P*-value is greater than 0.20.) We again fail to reject the null hypothesis, because the *P*-value is greater than the significance level of $\alpha = 0.05$.

Confidence Interval method:

Construct a 95% confidence interval estimate of μ_d, which is the mean of the "April–September" weight differences of college students in their freshman year.

$\bar{d} = 0.2, \quad s_d = 2.4 \quad n = 5, \quad t = 2.776$

The margin error: $E = t_{\alpha/2} \dfrac{s_d}{\sqrt{n}} = 2.776 \cdot \dfrac{2.4}{\sqrt{5}} = 3.0$

The confidence interval:

$$\bar{d} - E < \mu_d < \bar{d} + E$$

$$0.2 - 3.0 < \mu_d < 0.2 + 3.0$$

$$-2.8 < \mu_d < 3.2$$

Conclusion:

We have 95% confidence that the limits of -2.8 kg and 3.2 kg contain the true value of the mean weight change from September to April. In the long run, 95% of such samples will lead to confidence interval limits that actually do contain the true population mean of the differences. Note that the confidence interval includes the value of 0 kg, so it is very possible that the mean of the weight changes is equal to 0 kg.

Exercises *Section* 4.3 – Inferences from Dependent Samples

1. Listed below are the time intervals (in minutes) before and after eruptions of the Old Faithful geyser. Find the values of $\bar{d}$ and s_d. In general, what does μ_d represent?

Time interval before eruption	98	92	95	87	96
Time interval after eruption	92	95	92	100	90

2. Listed below are measured fuel consumption amount (in miles/gal) from a sample of cars.

City fuel consumption	18	22	21	21
Highway fuel consumption	26	31	29	29

Assume that wou want to use a 0.05 significance level to test the claim that the paired sample data come from a population for which the mean differecne is $\mu_d = 0$. Find

a) $\bar{d}$

b) s_d

c) The t test statistic

d) The critical values.

3. Listed below are predicted high temperatures that were forecast different days.

Predicted high temperatures forecast 3 *days ahead*	79	86	79	83	80
Predicted high temperatures forecast 5 *days ahead*	80	80	79	80	79

Assume that wou want to use a 0.05 significance level to test the claim that the paired sample data come from a population for which the mean differecne is $\mu_d = 0$. Find

a) $\bar{d}$

b) s_d

c) The t test statistic

d) The critical values.

4. Listed below are body mass indices (BMI). The BMI of each student was measured in September and April of the freshman year.

a) Use a 0.05 significance level to tet the claim that the mean change in BMI for all students is equal to 0. Does BMI appear to change during freshman year?

b) Construct a 95% confidence interval estimate of the change in BMI during freshman year. Does the confidence interval include 0, and what does that suggest about BMI during freshman year?

April BMI	20.15	19.24	20.77	23.85	21.32
September BMI	20.68	19.48	19.59	24.57	20.96

5. Listed below are body temperature (in °F) of subjects measured at 8:00 AM and at 12:00 AM. Construct a 95% confidence interval estimate of the difference between the 8:00 AM temperatures and the 12:00 AM temperatures. Is body temperature basically the same at both times?

8:00 AM	97.0	96.2	97.6	96.4	97.8	99.2
12:00 AM	98.0	98.6	98.8	98.0	98.6	97.6

6. Listed below are systolic blood pressure measuremetns (mm Hg) taking rom the right and left arms of the same woman. Use a 0.05 significance level to test for a difference the measurements from the two arms. What do you conclude?

Right arm	102	101	94	79	79
Left arm	175	169	182	146	144

7. As part of the National Health and Nutrition Examination Survey, the Department of Health and Human Services obtained self-reported heights and measured heights for males ages 12 – 16. All measurement are in inches. Listed below are sample results

Reported height	68	71	63	70	71	60	65	64	54	63	66	72
Measured height	67.9	69.9	64.9	68.3	70.3	60.6	64.5	67.0	55.6	74.2	65.0	70.8

a) Is there sufficient evidence to support the claim that there is a difference between self-reported heights and measured heights of males? Use a 0.05 significance level.

b) Construct a 95% confidence interval estimate of the man difference between reported heights and measured heights. Interpret the resulting confidence interval, and comment on the implications of whether the confidence interval limits contain 0.

8. Listed below are combined city – highway fuel consumption ratings (in miles/gal) for different cars measured under both the old rating system and a new rating system introducing in 2008. The new ratings were implemented in response to complaints that the old ratings were too high. Use a 0.01 significance level to test the claim the old ratings are higher than the new ratings.

Old rating	16	18	27	17	33	28	33	18	24	19	18	27	22	18	20	29	19	27	20	21
New rating	15	16	24	15	29	25	29	16	22	17	16	24	20	16	18	26	17	25	18	19

9. Listed below are 2 tables. Construct a 95% confidence interval estimate of the mean of the differences between weights of discarded paper and weights of discarded plastic. Which seems to weigh more: discarded paper or discarded plastic?

Paper

2.41	7.57	9.55	8.82	8.72	6.96	6.83	11.42	16.08	6.38	13.05	11.36	15.09
2.80	6.44	5.86	11.08	12.43	6.05	13.61	6.98	14.33	13.31	3.27	6.67	17.65
12.73	9.83	16.39	6.33	9.19	9.41	9.45	12.32	20.12	7.72	6.16	7.98	9.64
8.08	10.99	13.11	3.26	1.65	10.00	8.96	9.46	5.88	8.26	12.45	10.58	5.87
8.78	11.03	12.29	20.58	12.56	9.92	3.45	9.09	3.69	2.61			

Plastic

0.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05	3.42	2.10	2.93	2.44	2.17
1.41	2.00	0.93	2.97	2.04	0.65	2.13	0.63	1.53	4.69	0.15	1.45	2.68
3.53	1.49	2.31	0.92	0.89	0.80	0.72	2.66	4.37	0.92	1.40	1.45	1.68
1.53	1.44	1.44	1.36	0.38	1.74	2.35	2.30	1.14	2.88	2.13	5.28	1.48
3.36	2.83	2.87	2.96	1.61	1.58	1.15	1.28	0.58	0.74			

10. Suppose you wish to test the claim that μ_d, the mean value of the differences d for a population of paired data, is different from 0. Given a sample of $n = 23$ and a significance level of $\alpha = 0.05$, what criterion would be used for rejecting the null hypothesis?

11. Assume that he paired data came from a population that is normally distributed. Using a 0.05 significance level, find $\bar{d}$, s_d, the *t* test statistic, and the critical values to test the claim that $\mu_d = 0$

x	14	8	4	14	3	12	4	13
y	15	8	7	13	5	11	6	15

12. Assume that he paired data came from a population that is normally distributed. Using a 0.05 significance level, find $\bar{d}$, s_d, the *t* test statistic, and the critical values to test the claim that $\mu_d = 0$

x	12	5	1	20	3	16	12	8
y	7	10	5	15	7	14	10	13

Section 4.4 – Correlation

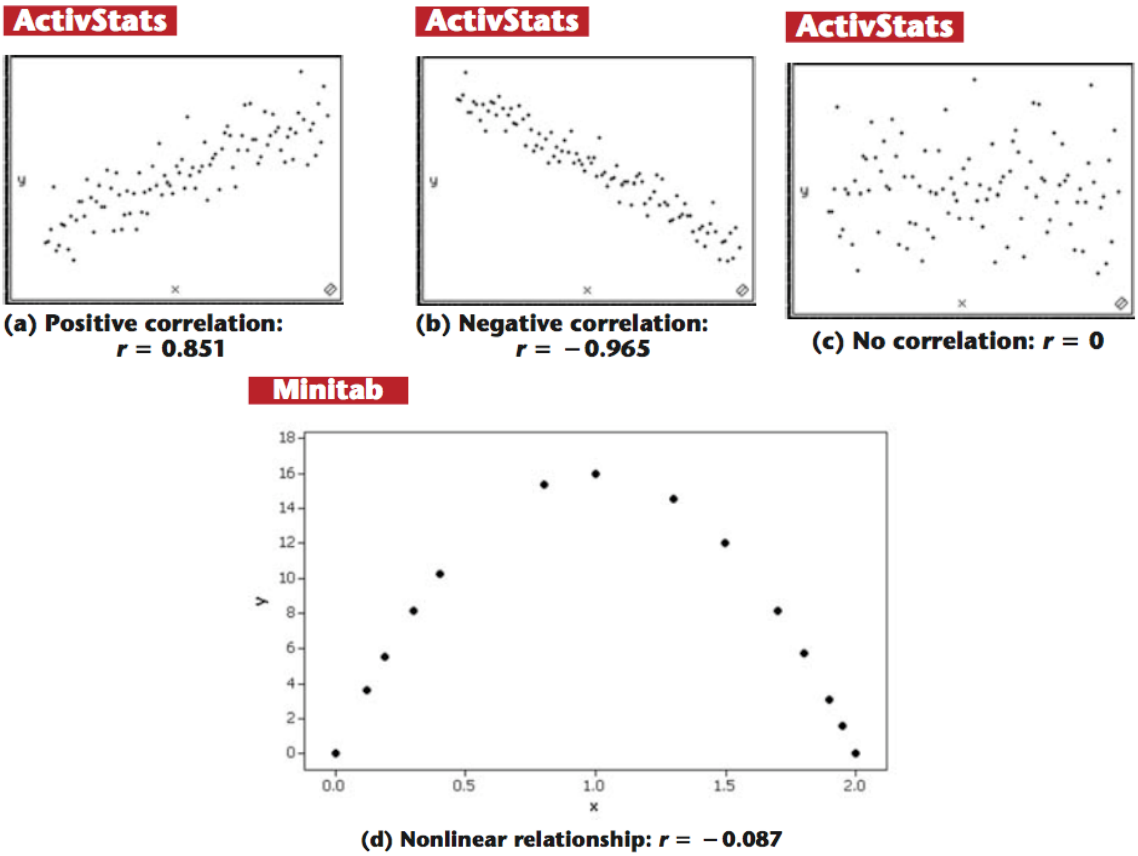
Definition

A ***correlation*** exists between two variables when the values of one are somehow associated with the values of the other in some way.

Exploring the Data

We can often see a relationship between two variables by constructing a scatterplot.

Scatterplots with different characteristics



(a) Positive correlation: $r = 0.851$

(b) Negative correlation: $r = -0.965$

(c) No correlation: $r = 0$

(d) Nonlinear relationship: $r = -0.087$

Definition

The ***linear correlation*** coefficient r measures the strength of the linear relationship between the paired quantitative x- and y-values in a sample.

Requirements

1. The sample of paired (x, y) data is a simple random sample of quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.
3. The outliers must be removed if they are known to be errors. The effects of any other outliers should be considered by calculating r with and without the outliers included.

Notation for the Linear Correlation Coefficient

n number of pairs of sample data

$\sum$ denotes the addition of the items indicated.

$\sum x$ denotes the sum of all x-values.

$\sum x^2$ indicates that each x-value should be squared and then those squares added.

$\left(\sum x\right)^2$ indicates that the x-values should be added and then the total squared

$\sum xy$ indicates that each x-value should be first multiplied by its corresponding y-value. After obtaining all such products, find their sum.

r = linear correlation coefficient for ***sample*** data.

ρ = linear correlation coefficient for **population** data.

Formula

The linear correlation coefficient r measures the strength of a linear relationship between the paired values in a sample.

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2} \cdot \sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

- Computer software or calculators can compute r

Interpreting *r*

Using Table – Critical Values of Spearman's Rank Correlation Coefficient rs: If the absolute value of the computed value of r, denoted $|r|$, exceeds the value in Table A-6, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

Using Software: If the computed P-value is less than or equal to the significance level, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

- ✓ *Know that the methods of this section apply to a linear correlation. If you conclude that there does not appear to be linear correlation, know that it is possible that there might be some other association that is not linear.*

Rounding the Linear Correlation Coefficient *r*

- Round to three decimal places so that it can be compared to critical values in Table Critical Values of Spearman's Rank Correlation Coefficient *r*'s.
- Use calculator or computer if possible.

Properties of the Linear Correlation Coefficient *r*

1. The value of r is always between –1 and 1 inclusive. That is $-1 \le r \le 1$
2. If all values of either variable are converted to a different scale, the value of *r* does not change.
3. The value of *r* is not affected by the choice of *x* and *y*. Interchange all *x*- and *y*-values and the value of *r* will not change.
4. *r* measures strength of a linear relationship.
5. *r* is very sensitive to outliers, they can dramatically affect its value.

Example

The paired pizza/subway fare costs are shown in the table below. Use computer software with these paired sample values to find the value of the linear correlation coefficient *r* for the paired sample data.

Table *– Cost of a Slice of Pizza, subway Fare, and the CPI*						
Year	**1960**	**1973**	**1986**	**1995**	**2002**	**2003**
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00
CPI	30.2	48.3	112.3	162.2	191.9	197.8

Solution

x (Pizza)	y (Subway)	x^2	y^2	xy
0.15	0.15	0.0225	0.0225	0.0225
0.35	0.35	0.1225	0.1225	0.1225
1.00	1.0	1.0	1.0	1.0
1.25	1.35	1.5625	1.8225	1.6875
1.75	1.50	3.0625	2.250	2.6250
2.0	2.0	4.0000	4.0000	4.0000
$\sum x = 6.50$	$\sum y = 6.35$	$\sum x^2 = 9.77$	$\sum y^2 = 9.2175$	$\sum xy = 9.4575$

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2} \cdot \sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

$$= \frac{6(9.4575) - (6.5)(6.35)}{\sqrt{6(9.77) - (6.5)^2} \cdot \sqrt{6(9.2175) - (6.35)^2}}$$

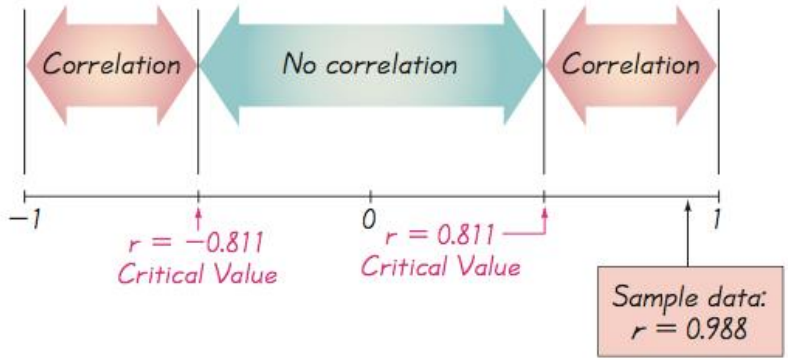
$$= 0.988$$

Interpreting the Linear Correlation Coefficient *r*

We can base our interpretation and conclusion about correlation on a *P*-value obtained from computer software or a critical value from Table *Critical Values of Spearman's Rank Correlation Coefficient* r_s.

Using Computer Software to Interpret *r*:
If the computed *P*-value is less than or equal to the significance level, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

Using Table Critical Values of Spearman's Rank Correlation Coefficient r_s to Interpret *r*:
If $|r|$ exceeds the value in Critical Values of Spearman's Rank Correlation Coefficient Table (A-6), conclude that there is a linear correlation.
Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.



Example

Using a 0.05 significance level, interpret the value of $r = 0.117$ found using the 62 pairs of weights of discarded paper and glass listed below. When the paired data are used with computer software, the *P*-value is found to be 0.364. Is there sufficient evidence to support a claim of a linear correlation between the weights of discarded paper and glass?

Paper

2.41	7.57	9.55	8.82	8.72	6.96	6.83	11.42	16.08	6.38	13.05	11.36	15.09
2.80	6.44	5.86	11.08	12.43	6.05	13.61	6.98	14.33	13.31	3.27	6.67	17.65
12.73	9.83	16.39	6.33	9.19	9.41	9.45	12.32	20.12	7.72	6.16	7.98	9.64
8.08	10.99	13.11	3.26	1.65	10.00	8.96	9.46	5.88	8.26	12.45	10.58	5.87
8.78	11.03	12.29	20.58	12.56	9.92	3.45	9.09	3.69	2.61			

Plastic

0.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05	3.42	2.10	2.93	2.44	2.17
1.41	2.00	0.93	2.97	2.04	0.65	2.13	0.63	1.53	4.69	0.15	1.45	2.68
3.53	1.49	2.31	0.92	0.89	0.80	0.72	2.66	4.37	0.92	1.40	1.45	1.68
1.53	1.44	1.44	1.36	0.38	1.74	2.35	2.30	1.14	2.88	2.13	5.28	1.48
3.36	2.83	2.87	2.96	1.61	1.58	1.15	1.28	0.58	0.74			

Solution

Requirements are satisfied: simple random sample of quantitative data; scatterplot approximates a straight line; no outliers

Using Software to Interpret *r*:
The *P*-value obtained from software is 0.364. Because the *P*-value is not less than or equal to 0.05, we conclude that there is not sufficient evidence to support a claim of a linear correlation between weights of discarded paper and glass.

Using Table Critical Values of Spearman's Rank Correlation Coefficient r_s to Interpret *r*:
If we refer to Table Critical Values of Spearman's Rank Correlation Coefficient r_s with $n = 62$ pairs of sample data, we obtain the critical value of 0.254 (approximately) for $\alpha = 0.05$. Because |0.117| does not exceed the value of 0.254 from Table Critical Values of Spearman's Rank Correlation Coefficient r_s, we conclude that there is not sufficient evidence to support a claim of a linear correlation between weights of discarded paper and glass.

Interpreting *r*: Explained Variation

The value of r^2 is the proportion of the variation in *y* that is explained by the linear relationship between *x* and *y*.

Example

Using the pizza subway fare costs in Table below, we have found that the linear correlation coefficient is $r = 0.988$. What proportion of the variation in the subway fare can be explained by the variation in the costs of a slice of pizza?

Table *– Cost of a Slice of Pizza, subway Fare, and the CPI*

Year	**1960**	**1973**	**1986**	**1995**	**2002**	**2003**
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00
CPI	30.2	48.3	112.3	162.2	191.9	197.8

Solution

With $r = 0.988$, we get $r^2 = 0.976$.

We conclude that 0.976 (or about 98%) of the variation in the cost of a subway fares can be explained by the linear relationship between the costs of pizza and subway fares. This implies that about 2% of the variation in costs of subway fares cannot be explained by the costs of pizza.

Common Errors Involving Correlation

1. ***Causation***: It is wrong to conclude that correlation implies causality.
2. ***Averages***: Averages suppress individual variation and may inflate the correlation coefficient.
3. ***Linearity***: There may be some relationship between *x* and *y* even when there is no linear correlation.

Formal Hypothesis Test

We wish to determine whether there is a significant linear correlation between two variables.

Hypothesis Test for Correlation Notation

n = number of pairs of sample data
r = linear correlation coefficient for a *sample* of paired data
ρ = linear correlation coefficient for a *population* of paired data

Hypothesis Test for Correlation Requirements

1. The sample of paired (x, y) data is a simple random sample of quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.
3. The outliers must be removed if they are known to be errors. The effects of any other outliers should be considered by calculating r with and without the outliers included.

Hypothesis Test for Correlation Hypotheses

$H_0 : r = 0$ (There is no linear correlation.)

$H_1 : r \neq 0$ (There is a linear correlation.)

Test Statistic: *r*

Critical Values: Refer to Table A-6

Hypothesis Test for Correlation Conclusion

If $|r|$ > critical value from Table A-6, reject H_0 and conclude that there is sufficient evidence to support the claim of a linear correlation.

If $|r| \leq$ critical value from Table A-6, fail to reject H_0 and conclude that there is not sufficient evidence to support the claim of a linear correlation.

Example

Use the paired pizza subway fare data in table below to test the claim that there is a linear correlation between the costs of a slice of pizza and the subway fares. Use a 0.05 significance level.

Table *– Cost of a Slice of Pizza, subway Fare, and the CPI*

Year	**1960**	**1973**	**1986**	**1995**	**2002**	**2003**
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00
CPI	30.2	48.3	112.3	162.2	191.9	197.8

Solution

Requirements are satisfied as in the earlier example.

$H_0: r = 0$ (There is no linear correlation.)

$H_1: r \neq 0$ (There is a linear correlation.)

The test statistic is $r = 0.988$ (from an earlier Example). The critical value of $r = 0.811$ is found in Table A-6 with $n = 6$ and $\alpha = 0.05$. Because $|0.988| > 0.811$, we reject $H_0: r = 0$. (Rejecting "no linear correlation" indicates that there is a linear correlation.)

We conclude that there is sufficient evidence to support the claim of a linear correlation between costs of a slice of pizza and subway fares.

Hypothesis Test for Correlation *P*-Value from a *t* Test

$H_0: \rho = 0$ (There is no linear correlation.)

$H_1: \rho \neq 0$ (There is a linear correlation.)

Test Statistic: *t*

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Hypothesis Test for Correlation Conclusion

P-value: Use computer software or use *t* Distribution Table (A-3) with $n - 2$ degrees of freedom to find the *P*-value corresponding to the test statistic *t*.

If the *P*-value is less than or equal to the significance level, reject H_0 and conclude that there is sufficient evidence to support the claim of a linear correlation.

If the *P*-value is greater than the significance level, fail to reject H_0 and conclude that there is not sufficient evidence to support the claim of a linear correlation.

Example

Use the paired pizza subway fare data in below table and use the *P*-value method to test the claim that there is a linear correlation between the costs of a slice of pizza and the subway fares. Use a 0.05 significance level.

Table *– Cost of a Slice of Pizza, subway Fare*						
Year	**1960**	**1973**	**1986**	**1995**	**2002**	**2003**
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00

Solution

Requirements are satisfied as in the earlier example.

$H_0: \ \rho = 0$ (There is no linear correlation.)

$H_1: \ \rho \neq 0$ (There is a linear correlation.)

The linear correlation coefficient is $r = 0.988$ (from an earlier Example) and $n = 6$ (six pairs of data), so the test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$= \frac{0.988}{\sqrt{\frac{1-0.988^2}{6-2}}}$$

$$= 12.793$$

With $df = 4$, Table A-6 yields a *P*-value that is less than 0.01.

Computer software generates a test statistic of $t = 12.692$ and *P*-value of 0.00022.

Using either method, the *P*-value is less than the significance level of 0.05 so we reject $H_0: \ \rho = 0$.

We conclude that there is sufficient evidence to support the claim of a linear correlation between costs of a slice of pizza and subway fares.

One-Tailed Tests

One-tailed tests can occur with a claim of a positive linear correlation or a claim of a negative linear correlation. In such cases, the hypotheses will be as shown here.

Claim of Negative Correlation **(Left-tailed test)**	***Claim of Positive Correlation*** **(Right-tailed test)**
$H_0 : \rho = 0$ $H_1 : \rho < 0$	$H_0 : \rho = 0$ $H_1 : \rho > 0$

TI-83/84 PLUS Enter the paired data in lists L1 and L2, then press **STAT** and select **TESTS.** Using the option of **LinRegTTest** will result in several displayed values, including the value of the linear correlation coefficient *r*. To obtain a scatterplot, press **2nd,** then **Y =** (for STAT PLOT). Press **Enter** twice to turn Plot 1 on, then select the first graph type, which resembles a scatterplot. Set the *X* list and *Y* list labels to L1 and L2 and press the **ZOOM** key, then select **ZoomStat** and press the **Enter** key.

Exercises *Section* 4.4 – Correlation

1. For each of several randomly selected years, the total number of points scored in the Super Bowl football game and the total number of new cars sold in The U.S. are recorded. For this sample of paired data

a) What does r represent?

b) What does ρ represent?

c) With our doing any research or calculations, estimate the value of r.

2. The heights (in inches) of a sample of eight mother/daughter pairs of subjects measured. Using Excel with the paired mother/daughter heights, the linear correlation coefficient is found to be 0.693. Is there sufficient evidence to support the claim that there is a linear correlation between the heights of mothers and the heights of their daughters? Explain.

3. The heights and weights of a sample of 9 supermodels were measured. Using a TI calculator, the linear correlation coefficient is found to be 0.360. Is there sufficient evidence to support the claim that there is a linear correlation between the heights and weights of supermodels? Explain.

4. Given the table below

x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

a) Construct a scatterplot

b) Find the value of linear correlation coefficient r and then determine whether there is sufficient evidence to support the claim of a linear correlation between the 2 variables.

c) Identify the feature of the data that would be missed if part (*b*) was completed without constructing the scatterplot.

5. Given the table below

x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

a) Construct a scatterplot

b) Find the value of linear correlation coefficient r and then determine whether there is sufficient evidence to support the claim of a linear correlation between the 2 variables.

c) Identify the feature of the data that would be missed if part (*b*) was completed without constructing the scatterplot.

6. The paired values of the Consumer Price Index (CPI) and the cost of a slice of pizza are shown below

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00

a) Construct a scatterplot

b) Find the value of linear correlation coefficient *r* and find the critical values if *r*, using $\alpha = 0.05$.

c) Determine whether there is sufficient evidence to support the claim of a linear correlation between the CPI and the cost of a slice of pizza?

7. Listed below are systolic blood pressure measurements (in mm HG) obtained from the same woman.

Right Arm	102	101	94	79	79
Left Arm	175	169	182	146	144

a) Construct a scatterplot

b) Find the value of linear correlation coefficient *r* and find the critical values if *r*, using $\alpha = 0.05$.

c) Determine whether there is sufficient evidence to support the claim of a linear correlation between the right and left arm systolic blood pressure measurements?

8. Listed below are costs (in dollars) of air fares for different airlines from NY to San Francisco. The costs are based on tickets purchased 30 days in advance and one day in advance.

30 *Days*	244	260	264	264	278	318	280
One Day	456	614	567	943	628	1088	536

a) Construct a scatterplot

b) Find the value of linear correlation coefficient *r* and find the critical values if *r*, using $\alpha = 0.05$.

c) Determine whether there is sufficient evidence to support the claim of a linear correlation between costs of tickets purchased 30 days in advance and those purchased one day in advance?

9. Listed below are repair costs (in dollars) for cars crashed at 6 mi/h in full-front crash tests and the same cars crashed at 6 mi/f in full-rear crash tests.

Front	936	978	2252	1032	3911	4312	3469
Rear	1480	1202	802	3191	1122	739	2767

a) Construct a scatterplot

b) Find the value of linear correlation coefficient *r* and find the critical values if *r*, using $\alpha = 0.05$.

c) Determine whether there is sufficient evidence to support the claim of a linear correlation between costs from full-front crashes and full-rear crashes?

Section 4.5 – Regression

Basic Concept of Regression

Two variables are sometimes related in a ***deterministic*** way, meaning that given a value for one variable, the value of the other variable is exactly determined from a given equation.
The regression equation expresses a relationship between *x* (called the ***explanatory*** variable, ***predictor*** variable or ***independent*** variable), and *y* (called the ***response*** variable or ***dependent*** variable).

The typical equation of a straight line
$y = mx + b$ is expressed in the form
$\hat{y} = b_0 + b_1 x$, where b_0 is the *y*-intercept and b_1 is the slope.

Definitions

❖ **Regression Equation**
Given a collection of paired data, the regression equation algebraically describes the relationship between the two variables.

$$\hat{y} = b_0 + b_1 x$$

❖ **Regression Line**
The graph of the regression equation is called the regression line (or line of best fit, or least squares line).

$$b_1 = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} \qquad b_0 = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

Notation for Regression Equation

	Population Parameter	***Sample Statistic***
y-intercept of regression equation	β_0	b_0
Slope of regression equation	β_1	b_1
Equation of the regression line	$y = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

Requirements

1. The sample of paired (*x*, *y*) data is a random sample of quantitative data.
2. Visual examination of the scatterplot shows that the points approximate a straight-line pattern.
3. Any outliers must be removed if they are known to be errors. Consider the effects of any outliers that are not known errors.

Formulas for b_0 and b_1

Slope: $b_1 = r\dfrac{s_y}{s_x}$

y-intercept: $b_0 = \bar{y} - b_1\bar{x}$

Where r is the linear correlation coefficient,

s_y is the standard deviation of the y values, and

s_x is the standard deviation of the x values

***Special Property*:** The regression line fits the sample points best.

Rounding the *y*-intercept b_0 and the Slope b_1

- Round to three significant digits.
- If you use the formulas 10-3 and 10-4, do not round intermediate values.

Example

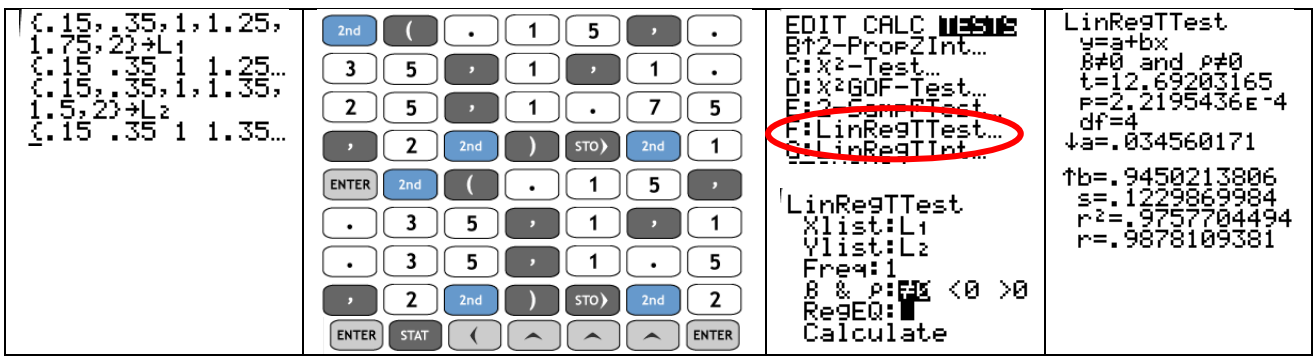
Using the pizza subway fare costs in Table below, Use technology to find the equation of the regression line in which the explanatory variable (or x variable) is the cost of a slice of pizza and the response variable (or y variable) is the corresponding cost of a subway fare. What proportion of the variation in the subway fare can be explained by the variation in the costs of a slice of pizza?

Table *– Cost of a Slice of Pizza, subway Fare, and the CPI*

Year	**1960**	**1973**	**1986**	**1995**	**2002**	**2003**
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00
CPI	30.2	48.3	112.3	162.2	191.9	197.8

Solution

Requirements are satisfied: simple random sample; scatterplot approximates a straight line; no outliers
Here are results from four different technologies



Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9878109
R Square	0.9757704
Adjusted R	0.9697131
Standard E	0.122987
Observatic	6

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	2.436580126	2.4365801	161.08767	0.000222
Residual	4	0.060503207	0.0151258		
Total	5	2.497083333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Jpper 95.0%*
Intercept	0.0345602	0.095012806	0.3637422	0.7344608	-0.229238	0.298358	-0.229238	0.298358
X Variable	0.9450214	0.074457849	12.692032	0.000222	0.7382932	1.1517495	0.7382932	1.1517495

All of these technologies show that the regression equation can be expressed as $y = 0.0346 + 0.945x$, where y is the predicted cost of a subway fare and x is the cost of a slice of pizza. We should know that the regression equation is an estimate of the true regression equation. This estimate is based on one particular set of sample data, but another sample drawn from the same population would probably lead to a slightly different equation.

Using the Formula

$$b_1 = r\frac{s_y}{s_x} = 0.987811 \cdot \frac{0.706694}{0.738693} = \underline{0.945}$$

$$b_0 = \bar{y} - b_1\bar{x} = 1.058333 - (0.945)(1.083333) = \underline{0.0346}$$

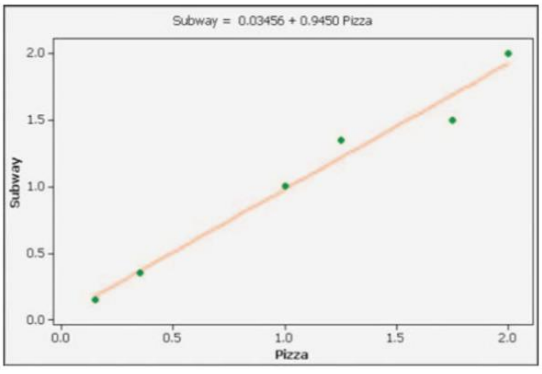
$$\hat{y} = b_0 + b_1 x = \underline{0.0346 + 0.945x}$$

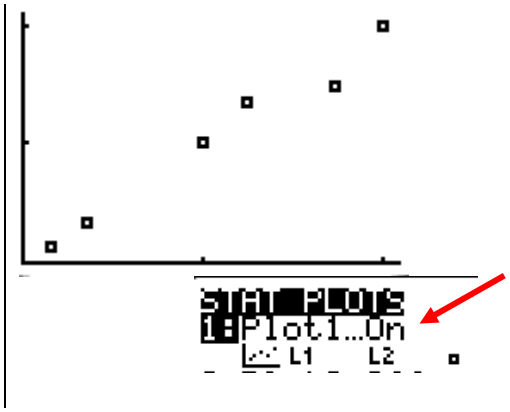
Example

Graph the regression equation $\hat{y} = 0.0346 + 0.945x$ (from the preceding Example) on the scatterplot of the pizza/subway fare data and examine the graph to subjectively determine how well the regression line fits the data.

Solution

MINITAB





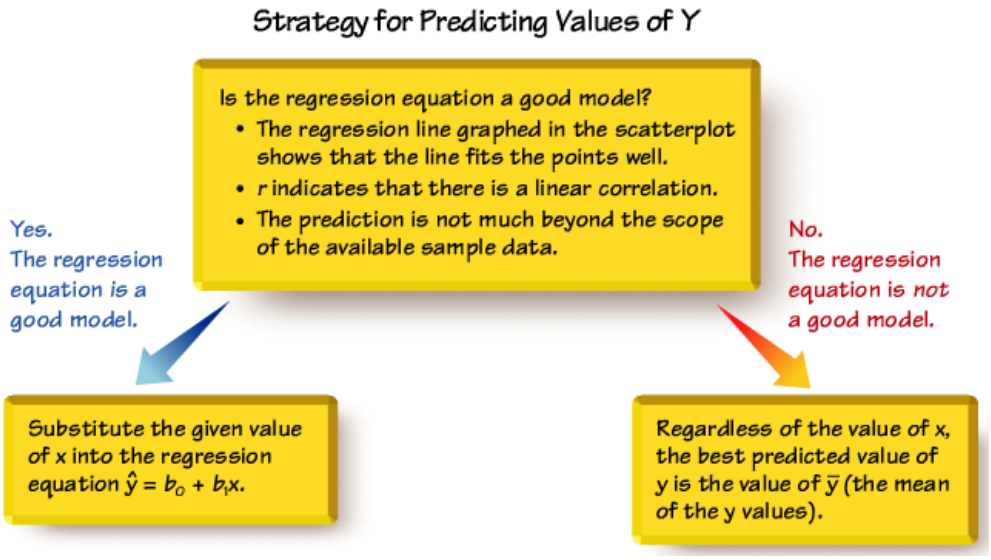
Using the Regression Equation for Predictions

1. Use the regression equation for predictions only if the graph of the regression line on the scatterplot confirms that the regression line fits the points reasonably well.
2. Use the regression equation for predictions only if the linear correlation coefficient *r* indicates that there is a linear correlation between the two variables.
3. Use the regression line for predictions only if the data do not go much beyond the scope of the available sample data. (Predicting too far beyond the scope of the available sample data is called *extrapolation*, and it could result in bad predictions.)
4. If the regression equation does not appear to be useful for making predictions, the best predicted value of a variable is its point estimate, which is its sample mean.

If the regression equation is not a good model, the best predicted value of *y* is simply $\bar{y}$, the mean of the *y* values.

Remember, this strategy applies to linear patterns of points in a scatterplot.

If the scatterplot shows a pattern that is not a straight-line pattern, other methods apply.



Definitions

In working with two variables related by a regression equation, the marginal change in a variable is the amount that it changes when the other variable changes by exactly one unit. The slope b_1 in the regression equation represents the marginal change in *y* that occurs when *x* changes by one unit.

In a scatterplot, an outlier is a point lying far away from the other data points.

Paired sample data may include one or more influential points, which are points that strongly affect the graph of the regression line.

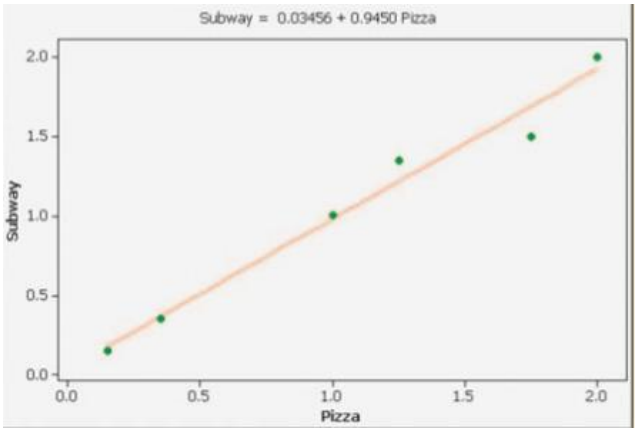
Example

Consider the pizza subway fare data. The scatterplot located to the left on the next slide shows the regression line. If we include this additional pair of data: $x = 2.00$, $y = -20.00$ (pizza is still \$2.00 per slice, but the subway fare is \$–20.00 which means that people are paid \$20 to ride the subway), this additional point would be an influential point because the graph of the regression line would change considerably, as shown by the regression line located to the right.

Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00

Solution

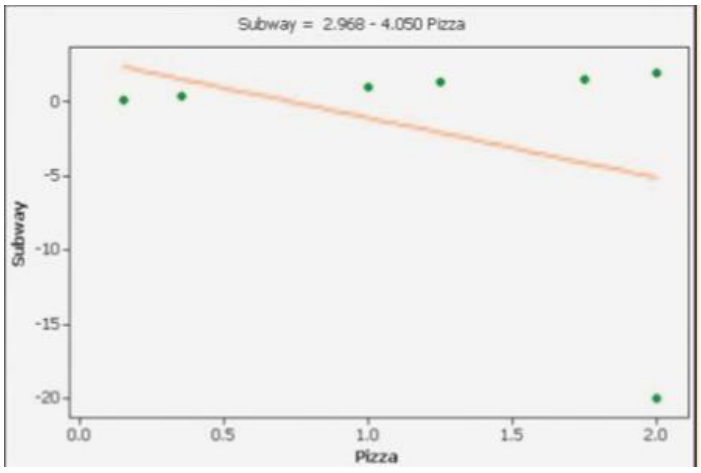
a) Pizza / subway fare data to predict subway fare when pizza costs \$2.25



$r = 0.988$, which suggests that there is a linear correlation between pizza costs and subway fares
The pizza cost of \$2.25 is not too far beyond the scope of the available data.

$\hat{y} = 0.0346 + 0.945(2.25)$ $= \$2.16$

b) runs / subway fare data to predict subway fare when 33 runs are scored in the World Series



The regression line does not fit the points well.

$r = -0.332$, which suggests that there is a linear correlation between World Series runs and subway fares, the *P*-value is 0.520. $\bar{y}$ $= \$1.06$

Compare the two graphs and you will see clearly that the addition of that one pair of values has a very dramatic effect on the regression line, so that additional point is an influential point. The additional point is also an outlier because it is far from the other points.

Beyond the Basics of Regression

Definition

In working with two variables related by a regression equation, the marginal change in a variable is the amount that it changes when the other variable changes by exactly one unit. The slope b_1 in the regression equation represents the marginal change in y that occurs when x changes by one unit.

Definition

In a scatterplot, an ***outlier*** is a point lying far away from the other data points. Paired sample data include one or more ***influential points***, which are points that strongly affect the graph of the regression line.

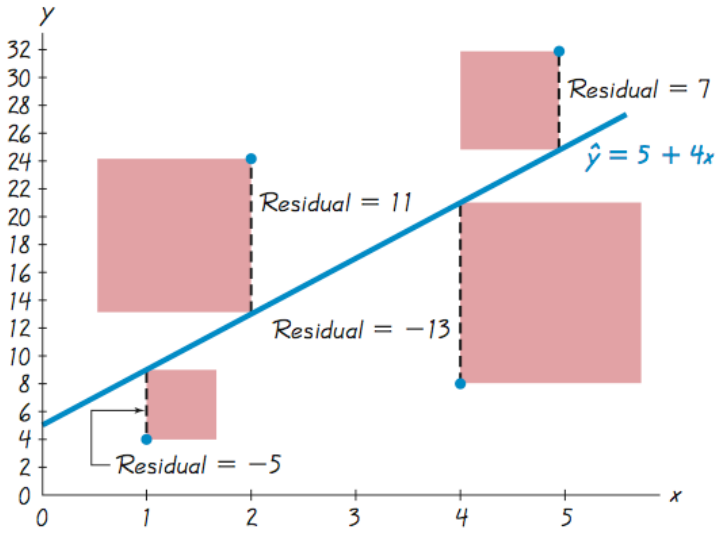
Residuals and the Least-Squares Property

Definition

For a pair of sample x and y values, the residual is the difference between the *observed* sample value of y and the y-value that is *predicted* by using the regression equation. That is,

$$residual = observed\ y - predicted\ y = y - \hat{y}$$

Residuals



Definitions

A straight line satisfies the least-squares property if the sum of the squares of the residuals is the smallest sum possible.

A residual plot is a scatterplot of the (x, y) values after each of the y-coordinate values has been replaced by the residual value $y - \hat{y}$ (where y denotes the predicted value of y). That is, a residual plot is a graph of the points (x, $y - \hat{y}$).

Residual Plot Analysis

When analyzing a residual plot, look for a pattern in the way the points are configured, and use these criteria:

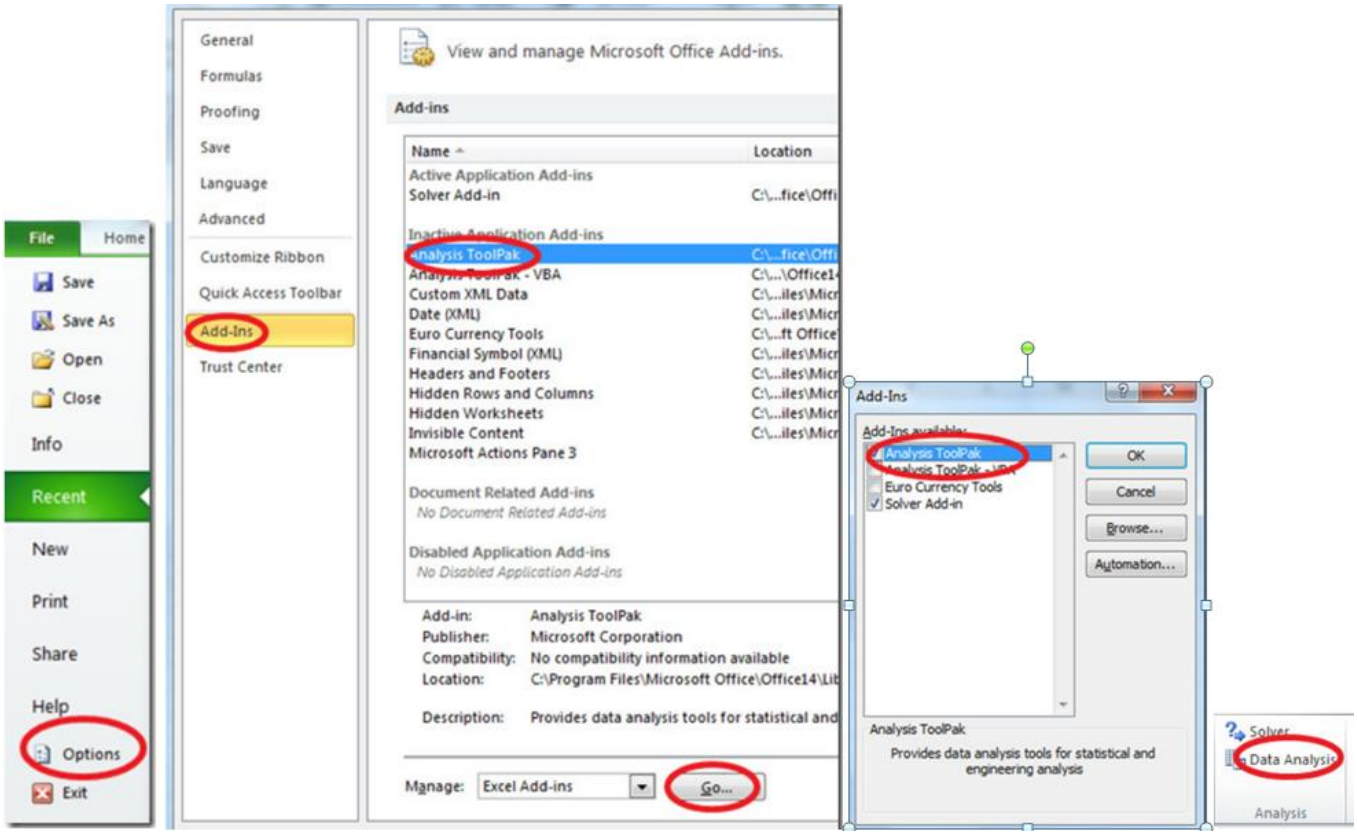
The residual plot should not have an obvious pattern that is not a straight-line pattern.

The residual plot should not become thicker (or thinner) when viewed from left to right.

Complete Regression Analysis

1. Construct a scatterplot and verify that the pattern of the points is approximately a straight-line pattern without outliers. (If there are outliers, consider their effects by comparing results that include the outliers to results that exclude the outliers.)
2. Construct a residual plot and verify that there is no pattern (other than a straight-line pattern) and also verify that the residual plot does not become thicker (or thinner).
3. Use a histogram and/or normal quantile plot to confirm that the values of the residuals have a distribution that is approximately normal.
4. Consider any effects of a pattern over time.

Installation analysis package to use regression



Exercises *Section* 4.5 – Regression

1. A physician measured the weights and cholesterol levels of a random sample of men. The regression equation is $\hat{y} = -116 + 2.44x$, where x represents weight (in pounds). What does the symbol $\hat{y}$ represent? What does the predictor variable represent? What does the response variable represent?

2. In what sense is the regression line the straight line that "best" fits the points in a scatterplot?

3. In a study, the total weight (in pounds) of garbage discarded in one week and the household size were recorded for 62 households. The linear correlation coefficient is $r = 0.759$ and the regression equation $\hat{y} = 0.445 + 0.119x$, where x represents the total weight of discarded garbage. The mean of the 62 garbage weights is 27.4 lb. and the 62 households have a mean size of 3.71 people. What is the best predicted number of people in a household that discards 50 lb. of garbage?

4. A sample of 8 mother/daughter pairs of subjects was obtained, and their heights (in inches) were measured. The linear correlation coefficient is 0.693 and the regression equation $\hat{y} = 69 - 0.0849x$, where x represents the height of the mother. The mean height of the mothers is 63.1 in. and the mean height of the daughters is 63.3 in. Find the best predicted height of a daughter given that the mother has a height of 60 in.

5. A sample of 40 women is obtained, and their heights (in inches) and pulse rates (in beats per minute) are measured. The linear correlation coefficient is 0.202 and the equation of the regression line is $\hat{y} = 18.2 + 0.920x$, where x represents height. The mean of the 40 heights is 63.2 in. and the mean of the 40 pulse rates is 76.3 beats per minute. Find the best predicted pulse rate of a woman who is 70 in. tall.

6. Heights (in inches) and weights (in pounds) are obtained from a random sample of 9 supermodels. The linear correlation coefficient is 0.360 and the equation of the regression line is $\hat{y} = 31.8 + 1.23x$, where x represents height. The mean of the 9 heights is 69.3 in. and the mean of the 9 weights is 117 lb. Find the best predicted weight of a supermodel with a height of 72 in.?

7. Find the equation of the regression line for the given data below

x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

Examine the scatterplot and identify a characteristic of the data that is ignored by the regression line

8. Find the equation of the regression line for the given data below

x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Examine the scatterplot and identify a characteristic of the data that is ignored by the regression line

9. Find the equation of the regression line for the given data below

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00

Let the first variable be the predictor (x) variable. Find the best indicated predicted cost of a slice of pizza when the Consumer Price Index (CPI) is 182.5 (in the year 2000).

10. Find the equation of the regression line for the given data below

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Subway fare	0.15	0.35	1.00	1.35	1.5	2.00

Let the first variable be the predictor (x) variable. Find the best indicated predicted cost of a slice of pizza when the Consumer Price Index (CPI) is 182.5 (in the year 2000).

11. Listed below are systolic blood pressure measurements (in mm HG) obtained from the same woman.

Right Arm	102	101	94	79	79
Left Arm	175	169	182	146	144

Find the best predicted systolic blood pressure in the left arm given that the systolic blood pressure in the right arm is 100 mm Hg.

12. Find the best predicted height of runner-up Goldwater, given that the height of the winning presidential candidate is 75 in. Is the predicted height of Goldwater close to his actual height of 72 in.?

Winner	69.5	73	73	74	74.5	74.5	71	71
Runner-Up	72	69.5	70	68	74	74	73	76

13. Find the best predicted amount of revenue (in millions of dollars), given that the amount has a size 87 thousand ft^2. How does the result compare to the actual revenue of $65.1 million?

Size	160	227	140	144	161	147	141
Revenue	189	157	140	127	123	106	101

14. Find the best predicted new mileage rating of a jeep given that old rating is 19 mi/gal. Is the predicted value close to the actual value of 17 mi/gal?

Old	16	27	17	33	28	24	18	22	20	29	21
New	15	24	15	29	25	22	16	20	18	26	19

15. Find the best predicted temperature for a recent year in which the concentration (in parts per million) of CO_2 is 370.9. Is the predicted temperature close to the actual temperature of 14.5° C??

CO_2	314	317	320	326	331	339	346	354	361	369
Temperature	13.9	14.0	13.9	14.1	14.0	14.3	14.1	14.5	14.5	14.4

16. Find the best predicted IQ score of someone with a brain size of 1275 cm^3

Brain Size	965	1029	1030	1285	1049	1077	1037	1068	1176	1105
IQ	90	85	86	102	103	97	124	125	102	114

17. Listed below are the word counts for men and women.

Male

27531	15684	5638	27997	25433	8077	21319	17572	26429	21966	11680	10818
12650	21683	19153	1411	20242	10117	20206	16874	16135	20734	7771	6792
26194	10671	13462	12474	13560	18876	13825	9274	20547	17190	10578	14821
15477	10483	19377	11767	13793	5908	18821	14069	16072	16414	19017	37649
17427	46978	25835	10302	15686	10072	6885	20848				

Female

20737	24625	5198	18712	12002	15702	11661	19624	13397	18776	15863	12549
17014	23511	6017	18338	23020	18602	16518	13770	29940	8419	17791	5596
11467	18372	13657	21420	21261	12964	33789	8709	10508	11909	29730	20981
16937	19049	20224	15872	18717	12685	17646	16255	28838	38154	25510	34869
24480	31553	18667	7059	25168	16143	14730	28117				

Find the best predicted word count of a woman given that her male partner speaks 6,000 words in a day.

18. According the least-squares property, the regression line minimizes the sum of the squares of the residuals. Listed below are the paired data consisting of the first 6 pulse and the first systolic blood pressures of males.

Pulse (***x***)	68	64	88	72	64	110
Systolic (***y***)	125	107	126	110	72	107

a) Find the equation of the regression line.

b) Identify the residuals, and find the sum of squares of the residuals.

c) Show that the equation $\hat{y} = 70 + 0.5x$ results in a larger sum of squares of residuals.

Section 4.6 – Variation and Prediction Intervals

Method for constructing a prediction interval, which is an interval estimate of a predicted value of ***y***

Unexplained, Explained, and Total Deviation

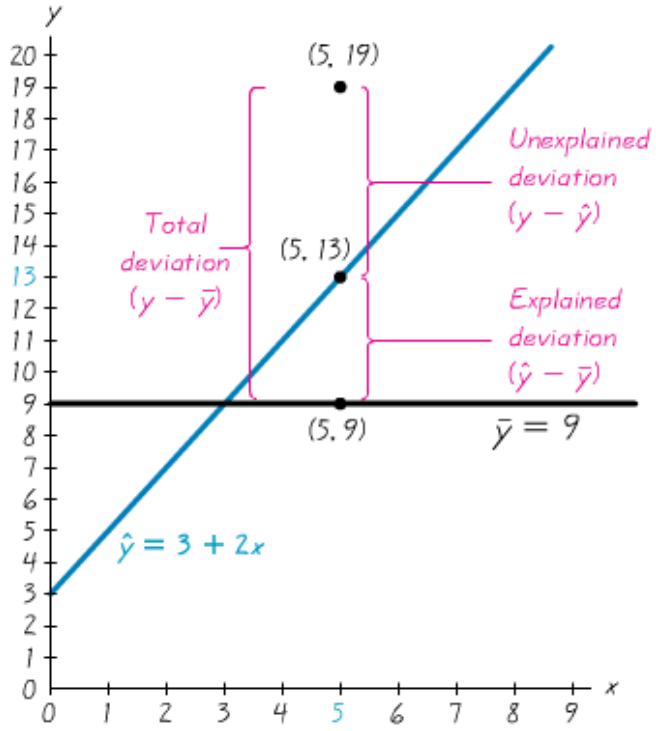
Definitions

Assume that we have a collection of paired data containing the sample point (*x*, *y*), that $\hat{y}$ is the predicted value of *y* (obtained by using the regression equation), and that the mean of the sample *y*-values is $\bar{y}$.

The ***total deviation*** of (x, y) is the vertical distance $y - \bar{y}$, which is the distance between the point (x, y) and the horizontal line passing through the sample mean $\bar{y}$.

The ***explained deviation*** is the vertical distance $\hat{y} - \bar{y}$, which is the distance between the predicted *y*-value and the horizontal line passing through the sample mean $\bar{y}$.

The ***unexplained deviation*** is the vertical distance $y - \hat{y}$, which is the vertical distance between the point (*x*, *y*) and the regression line. (The distance $y - \hat{y}$ is also called a ***residual***.



- ✓ The mean of the *y*-value is given by $\bar{y} = 9$
- ✓ One of the pairs of sample data is $x = 5$ *and* $y = 19$
- ✓ The point (5, 13) is one of the points on the regression line, because substituting *x* = 5 into the regression equation of $\hat{y} = 3 + 2x$ yields $\hat{y} = 3 + 2(5) = 13$

Formula

$$(\textit{total variation}) = (\textit{explained variation}) + (\textit{unexplained variation})$$

or $$\sum(y-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y-\hat{y})^2$$

Definition

The coefficient of determination is the amount of the variation in *y* that is explained by the regression line.

$$r^2 = \frac{\textit{explained variation}}{\textit{total variation}}$$

The value of r^2 is the proportion of the variation in *y* that is explained by the linear relationship between *x* and *y*.

Example

We used the paired pizza/subway fare costs to get $r = 0.988$. Find the coefficient of determinant. Also, find the percentage of the total variation in y (subway fare) that can be explained by the linear relationship between the cost of a slice pizza and the cost of a subway fare.

Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00

Solution

The coefficient of determinant is $r^2 = 0.988^2 = 0.976$

Because r^2 is the proportion of total variation that is explained, we conclude that 97.6% of the total variation in subway fares can be explained by the cost of a slice of pizza. This means that 2.4% of the total variation in cost of subway fares can be explained by factors other than the cost of a slice of pizza.

Definition

A ***prediction interval***, is an interval estimate of a predicted value of *y*.

Definition

The ***standard error of estimate***, denoted by s_e is a measure of the differences (or distances) between the observed sample *y*-values and the predicted values $\hat{y}$ that are obtained using the regression equation.

$$s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} \qquad (\textit{where } \hat{y} \textit{ is the predicted y-value})$$

Or

$$s_e = \sqrt{\frac{\sum y^2 - b_0\sum y - b_1\sum xy}{n-2}}$$

Example

Find the standard error of estimate s_e for the paired pizza/subway fare data listed below.

Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00

Solution

x	y	xy	x^2	y^2
0.15	0.15	0.0225	0.0225	0.0225
0.35	0.35	0.1225	0.1225	0.1225
1	1	1	1	1
1.25	1.35	1.6875	1.5625	1.8225
1.75	1.5	2.625	3.0625	2.25
2	2	4	4	4
6.5	6.35	9.4575	9.77	9.2175

$n = 6, \quad b_0 = 0.034560171, \quad b_1 = 0.94502138$

	Coefficients
Intercept	0.0345602
X Variable	0.9450214

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

$$= \sqrt{\frac{9.2175 - (0.034560171)(6.35) - (0.94502138)(9.4575)}{6-2}}$$

$$\approx 0.123$$

Prediction Interval for an Individual *y*

Given the fixed value x_0 the prediction interval for an individual *y* is

$$\hat{y} - E < y < \hat{y} + E$$

Where the margin of error *E* is

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}}$$

And x_0 represents the given value of *x*

$t_{\alpha/2}$ has $n - 2$ degrees of freedom

s_e is given in the previous formula

Example

For the paired pizza/subway fare costs from the Chapter Problem, we have found that for a pizza cost of $2.25, the best predicted cost of a subway fare is $2.16. Construct a 95% prediction interval for the cost of a subway fare, given that a slice of pizza costs $2.25 (so that $x = 2.25$).

Solution

$s_e = 0.122987$

$n = 6, \quad \bar{x} = \frac{6.5}{6} = 1.083333 \quad \sum x = 6.5 \quad \sum x^2 = 9.77$

$\alpha = 0.05$ (2-tails)

$t_{\alpha/2} = 2.776 \quad df = 6 - 2 = 4$

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n\left(x_0 - \bar{x}\right)^2}{n\left(\sum x^2\right) - \left(\sum x\right)^2}}$$

$$= (2.776)(0.122987)\sqrt{1 + \frac{1}{6} + \frac{6(2.25 - 1.0833333)^2}{6(9.77) - (6.5)^2}}$$

≈ 0.441

With $\hat{y} = 2.16$ *and* $E = 0.441$

$\hat{y} - E < y < \hat{y} + E$

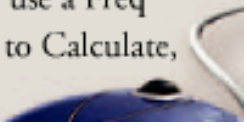
$2.16 - 0.441 < y < 2.16 + 0.441$

$1.72 < y < 2.60$

If the cost of a slice of pizza is $2.25, we have 95% confidence that the cost of a subway fare is between $1.72 and $2.60. That is a fairly large range of possible values, and one major factor contributing to the large range is that the sample size is very small with $n = 6$.

y = nicotine in menthol cigarettes

TI-83/84 PLUS The TI-83/84 Plus calculator can be used to find the linear correlation coefficient r, the equation of the regression line, the standard error of estimate s_e, and the coefficient of determination (labeled r^2). Enter the paired data in lists L1 and L2, then press **STAT** and select **TESTS,** and then choose the option **LinRegTTest.** For Xlist enter L1, for Ylist enter L2, use a Freq (frequency) value of 1, and select $\neq 0$. Scroll down to Calculate, then press the **ENTER** key.



Exercises *Section* 4.6 – Variation and Prediction Intervals

1. A height of 70 in. is used to find the predicted weight is 180 lb. In your own words, describe a prediction interval in this situation.

2. A height of 70 in. is used to find the predicted weight is 180 lb. What is the major advantage of using a prediction interval instead of the predicted weight of 180 lb.? Why is the terminology of prediction interval used instead of confidence interval?

3. Use the value of the linear correlation $r = 0.873$ to find the coefficient of determination and the percentage of the total variation that can be explained by the linear relationship between the 2 variables

$x = tar\ in\ menthol\ cigarettes$

$y = movie\ gross$														
16	13	16	9	14	13	12		.14	13	13	16	13	13	18
9	19	2	13	14	14	15	16	6.						

1.1	0.8	1	0.9	0.8	0.8	0.8	0.8	0.9	0.8	0.8	1.2	0.8	0.8	1.3
0.7	1.4	0.2	0.8	1	0.8	0.8	1.2	0.6	0.7					

4. Use the value of the linear correlation

5. $r = 0.744$ to find the coefficient of determination and the percentage of the total variation that can be explained by the linear relationship between the 2 variables

$x = movie\ budget$

41	20	116	70	75	52	120	65	6.5	60	125	20	5	150
4.5	7	100	30	225	70	80	40	70	50	74	200	113	68
72	160	68	29	132	40								

117	5	103	66	121	116	101	100	55	104	213	34	12	290
47	10	111	100	322	19	117	48	228	47	17	373	380	118
114	120	101	120	234	209								

6. Use the value of the linear correlation $r = -0.865$ to find the coefficient of determination and the percentage of the total variation that can be explained by the linear relationship between the 2 variables

$x = car\ weight, \quad y = city\ fuel\ consumption\ in\ mi/gal$

7. Use the value of the linear correlation $r = -0.488$ to find the coefficient of determination and the percentage of the total variation that can be explained by the linear relationship between the 2 variables

$x = age\ of\ home, \quad y = home\ selling\ price$

8. Refer to the display obtained by using the paired data consisting of weights (in *lb.*) of 32 cars and their highway fuel consumption amounts (in *mi/gal*). A car weight of 4000 *lb*. to be used for predicting the highway fuel consumption amount

```
The regression equation is
Highway = 50.5 - 0.00587 Weight

Predictor        Coef    SE Coef      T      P
Constant       50.502      2.860  17.66  0.000
Weight     -0.0058685  0.0007859  -7.47  0.000

S = 2.19498   R-Sq = 65.0%   R-Sq(adj) = 63.9%

Predicted Values for New Observations

New
Obs     Fit  SE Fit        95% CI             95% PI
  1  27.028   0.497  (26.013, 28.042)  (22.431, 31.624)

Values of Predictors for New Observations
New
Obs  Weight
  1    4000
```

a) What percentage of the total variation in highway fuel consumption can be explained by the linear correlation between weight and highway fuel consumption?

b) If a car weighs 4000 *lb*., what is the single value that is the best predicted amount of highway fuel consumption? (Assume that there is a linear correlation between weight and highway fuel consumption.)

9. The paired values of the Consumer Price Index (CPI) and the cost of a slice of pizza are shown below

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00

a) Find the explained variation

b) Find the unexplained variation

c) Find the total variation

d) Find the coefficient of determination

e) Find the standard error of estimate s_e

f) Find the predicted cost of a slice of pizza for the year 2001, when the CPI was 187.1.

g) Find a 95% prediction interval estimate of the cost of a slice of pizza when the CPI was 187.1

In each case, there is sufficient evidence to support a claim of a linear correlation so that it is reasonable to use the regression equation when making predictions.

10. Find the best predicted temperature for a recent year in which the concentration (in parts per million) of CO_2 and temperature (in °C) for different years

CO_2	314	317	320	326	331	339	346	354	361	369
Temperature	13.9	14.0	13.9	14.1	14.0	14.3	14.1	14.5	14.5	14.4

a) Find the explained variation
b) Find the unexplained variation
c) Find the total variation
d) Find the coefficient of determination
e) Find the standard error of estimate s_e
f) Find the predicted temperature (in °C) when CO_2 concentration is 370.9 parts per million.
g) Find a 99% prediction interval estimate temperature (in °C) when CO_2 concentration is 370.9 parts per million

In each case, there is sufficient evidence to support a claim of a linear correlation so that it is reasonable to use the regression equation when making predictions.

Find a prediction interval data listed below.

Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00

11. Using: *Cost of a slice of pizza*: \$2.10; 99% *confidence*

12. Using: *Cost of a slice of pizza*: \$2.10; 90% *confidence*

13. Using: *Cost of a slice of pizza*: \$0.50; 95% *confidence*

14. Using: *Cost of a slice of pizza*: \$0.75; 99% *confidence*

Section 4.7 – Goodness-of-Fit

In this section we consider sample data consisting of observed frequency counts arranged in a single row or column (called a one-way frequency table). We will use a hypothesis test for the claim that the observed frequency counts agree with some claimed distribution, so that there is a good fit of the observed data with the claimed distribution.

Definition

A ***goodness-of-fit*** test is used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution.

Notation

O represents the ***observed frequency*** of an outcome.
E represents the ***expected frequency*** of an outcome.
k represents the ***number of different categories*** or outcomes.
n represents the total ***number of trials***.

Requirements

1. The data have been randomly selected.
2. The sample data consist of frequency counts for each of the different categories.
3. For each category, the expected frequency is at least 5. (The expected frequency for a category is the frequency that would occur if the dat a actually have the distribution that is being claimed. There is no requirement that the *observed* frequency for each category must be at least 5.)

Test Statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Critical Values

1. Found in Table A- 4 using $k - 1$ degrees of freedom, where k = number of categories.
2. Goodness-of-fit hypothesis tests are always *right-tailed*.

P-Values

P-values are typically provided by computer software, or a range of *P*-values can be found from Table A-4.

Expected Frequencies

If all expected frequencies are <u>equal</u>: $E = \frac{n}{k}$

the sum of all observed frequencies divided by the number of categories

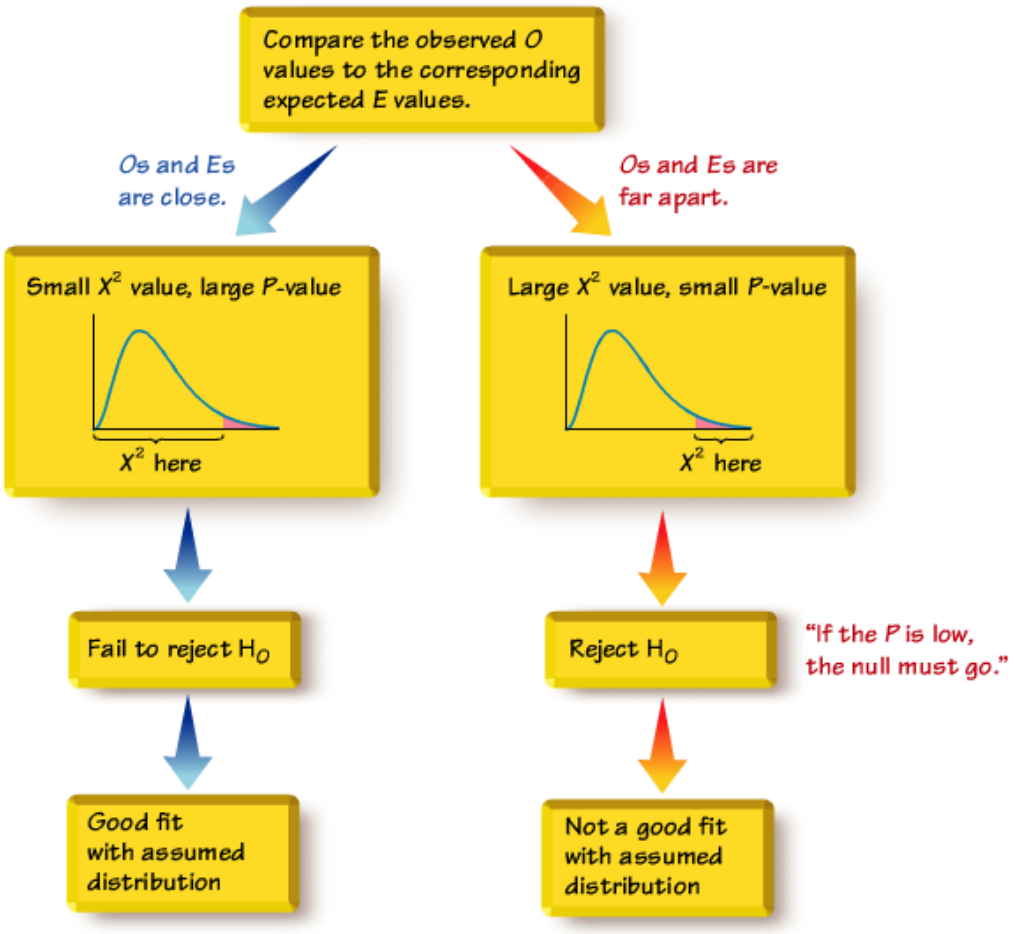
If expected frequencies are <u>not all equal</u>: $E = np$

Each expected frequency is found by multiplying the sum of all observed frequencies by the probability for the category.

Goodness-of-Fit Test

- A close agreement between observed and expected values will lead to a small value of χ^2 and a large *P*-value.
- A large disagreement between observed and expected values will lead to a large value of χ^2 and a small *P*-value.
- A significantly large value of χ^2 will cause a rejection of the null hypothesis of no difference between the observed and the expected.

"If the *P* is low, the null must go."
(If the *P*-value is small, reject the null hypothesis that the distribution is as claimed.)



Example

Last Digit	Frequency
0	7
1	14
2	6
3	10
4	8
5	4
6	5
7	6
8	12
9	8

Data Set 1 in Appendix B includes weights from 40 randomly selected adult males and 40 randomly selected adult females. Those weights were obtained as part of the National Health Examination Survey. When obtaining weights of subjects, it is extremely important to actually weigh individuals instead of asking them to report their weights. By analyzing the last digits of weights, researchers can verify that weights were obtained through actual measurements instead of being reported. When people report weights, they typically round to a whole number, so reported weights tend to have many last digits consisting of 0. In contrast, if people are actually weighed with a scale having precision to the nearest 0.1 pound, the weights tend to have last digits that are uniformly distributed, with 0, 1, 2, … , 9 all occurring with roughly the same frequencies. Table below shows the frequency distribution of the last digits from 80 weights listed in Data Set 1 in Appendix B.

(For example, the weight of 201.5 lb has a last digit of 5, and this is one of the data values included in Table)

Test the claim that the sample is from a population of weights in which the last digits do not occur with the same frequency. Based on the results, what can we conclude about the procedure used to obtain the weights?

Solution

Requirements are satisfied: randomly selected subjects, frequency counts, expected frequency is 8 (> 5)

Step 1: At least one of the probabilities $p_0, p_1, \ldots, p_9$, is different from the others

Step 2: At least one of the probabilities are the same:

$$p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9$$

Step 3: Null hypothesis contains equality

$H_0 : p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9$

$H_1 :$ At least one probability is different

Step 4: No significance specified, use $\alpha = 0.05$

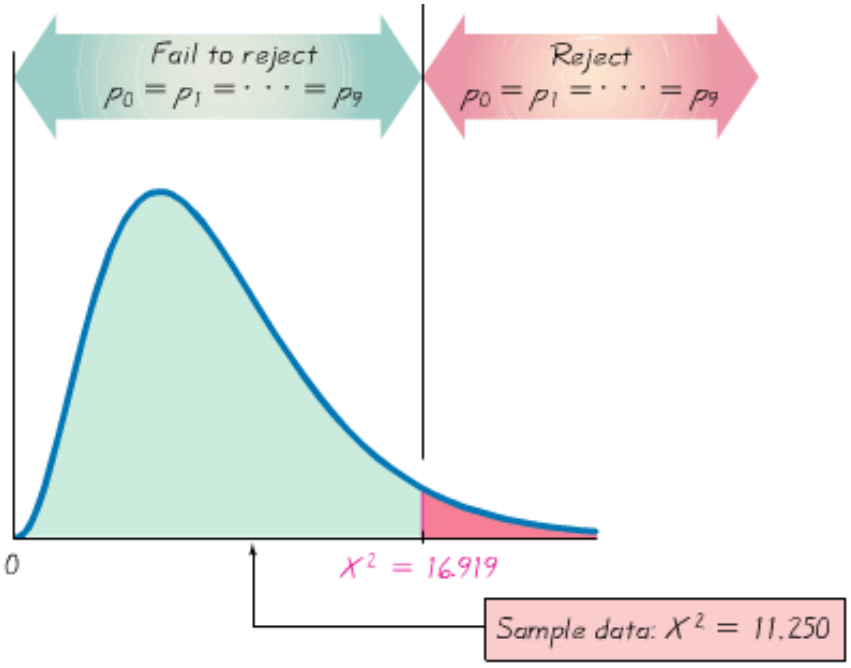
Step 5: Testing whether a uniform distribution so use goodness-of-fit test: χ^2

Step 6: The table (next page) shows the computation of the χ^2 test statistic. The test statistic $\chi^2 = 11.250$, using $\alpha = 0.05$ and $k - 1 = 9$ degrees of freedom, the critical value is $\chi^2 = 16.919$

Step 7: Because the test statistic does not fall in the critical region, there is not sufficient evidence to reject the null hypothesis.

Step 8: There is not sufficient evidence to support the claim that the last digits do not occur with the same relative frequency.

Last Digit	*Observed Frequency O*	*Expected Frequency E*	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	7	8	–1	1	0.125
1	14	8	6	36	4.500
2	6	8	–2	4	0.500
3	10	8	2	4	0.500
4	8	8	0	0	0.000
5	4	8	–4	16	2.000
6	5	8	–3	9	1.125
7	6	8	–2	4	0.500
8	12	8	4	16	2.00
9	8	8	0	0	0.000
				$\chi^2 = \sum \frac{(O-E)^2}{E} = 11.250$	



Conclusion

This goodness-of-fit test suggests that the last digits provide a reasonably good fit with the claimed distribution of equally likely frequencies. Instead of asking the subjects how much they weigh, it appears that their weights were actually measured as they should have been.

Example

Table below lists the numbers of games played in the baseball World Series. That table also includes the expected proportions for the numbers of games in a World Series, assuming that in each series, both teams have about the same chance of winning. Use a 0.05 significance level to test the claim that the actual numbers of games fit the distribution indicated by the probabilities.

Games Playe4d	4	5	6	7
Actual World Series Contests	19	21	22	37
Expected Proportion	$\frac{2}{16}$	$\frac{4}{16}$	$\frac{5}{16}$	$\frac{5}{16}$

Solution

Step 1: The original claim: $p_4 = \frac{2}{16}, \quad p_5 = \frac{4}{16}, \quad p_6 = \frac{5}{16}, \quad p_7 = \frac{5}{16}$

Step 2: If the original claim is false, then at least one of the proportions does not have the value as claimed.

Step 3: null hypothesis contains equality

$H_0 : \; p_4 = \frac{2}{16}, \quad p_5 = \frac{4}{16}, \quad p_6 = \frac{5}{16}, \quad p_7 = \frac{5}{16}$

$H_1 :$ At least one probability is not equal to the given claimed value.

Step 4: The significance level is $\alpha = 0.05$

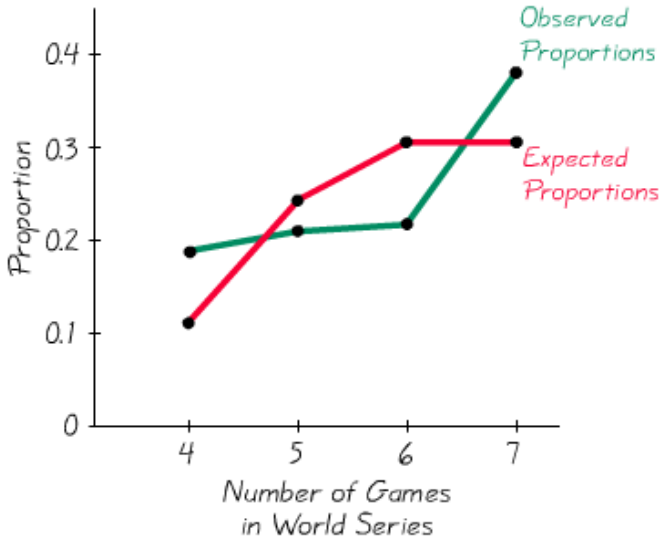
Step 5: Testing whether the distribution of numbers of games in World Series contests is as claimed, use goodness-of-fit test: χ^2

Last Digit	*Observed Frequency O*	*Expected Frequency E*	$O - E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
4	19	$99 \cdot \frac{4}{16} = 24.7500$	6.6250	43.8906	3.5467
5	21	$99 \cdot \frac{2}{16} = 12.3750$	–3.750	14.0625	0.5682
6	22	$99 \cdot \frac{5}{16} = 30.9375$	–8.9375	79.8789	2.5819
7	37	$99 \cdot \frac{5}{16} = 30.9375$	6.0625	36.7539	1.1880
				$\chi^2 = \sum \frac{(O-E)^2}{E} = 7.885$	

Step 6: Table above shows the computation of the χ^2 test statistic. The test statistic $\chi^2 = 7.885$, using $\alpha = 0.05$ and $k - 1 = 3$ degrees of freedom, the *P*-value is 0.048

Step 7: The *P*-value of 0.048 is less than the significance level of 0.05, so there is sufficient evidence to reject the null hypothesis. Also the test statistic of $\chi^2 = 7.885$ is in critical region bounded by the critical value of 7.185.

Step 8: There is sufficient evidence to warrant rejection of the claim that actual numbers of games in World Series contests fit the distribution indicated by the expected proportions.



Conclusion

This goodness-of-fit test suggests that the numbers of games in World Series contest do not fit the distribution expected from probability calculations

Exercises *Section* 4.7 – Goodness-of-Fit

1. A poll typically involves the selection of random digits to be used for telephone numbers. The New York Times states that "within each (telephone) exchange, random digits were added to form a complete telephone number, thus permitting access to listed and unlisted numbers. "When such digits are randomly generated, what is the distribution of those digits? Given such randomly generated digits, what is a test for "goodness-off-fit"?

2. When generating random digits, we can test the generated digits for goodness-of-fit with the distribution in which all of the digits are equally likely. What does an exceptionally large value of the χ^2 test statistic suggest about the goodness-of-fit? What does an exceptionally small value of the χ^2 test statistic (such as 0.002) suggest about the goodness-of-fit?

3. You purchased a slot machine, and tested it by playing it 1197 times. There are 10 different categories of outcome, including no win, win jackpot, win with three bells, and so on. When testing the claim the observed outcomes agree with the expected frequencies, the author obtained a test statistic of $\chi^2 = 8.185$. Use a 0.05 significance level to test the claim that the actual outcomes agree with the expected frequencies. Does the slot machine appear to be functioning as expected? Conduct the hypothesis test and the test statistic, critical value and/or *P*-value, and state the conclusion.

4. Do "*A*" students tend to sit in a particular part of the classroom? The author recorded the locations of the students who received grades *A*, with these results: 17 sat in the front, 9 sat in the middle, and 5 sat in the back of the classroom. When testing the assumption that the "*A*" students are distributed evenly throughout the room, the author obtained the test statistic of $\chi^2 = 7.226$. If using a 0.05 significance level, is there sufficient evidence to support the claim that the "*A*" students are not evenly distributed throughout the classroom? If so, does that mean you can increase your likelihood of getting an *A* by sitting in the front of the room? Conduct the hypothesis test and the test statistic, critical value and/or *P*-value, and state the conclusion.

5. Randomly selected nonfat occupational injuries and illnesses are categorized according to the day of the week that they first occurred, and the results are listed below. Use a 0.05 significance level to test the claim that such injuries and illness occur with equal frequency on the different days of the week. Conduct the hypothesis test and the test statistic, critical value and/or *P*-value, and state the conclusion.

Day	*Mon*	*Tues*	*Wed*	*Thurs*	*Fri*
Number	23	23	21	21	19

6. Records of randomly selected births were obtained and categorized according to the day of the week that they occurred. Because babies are unfamiliar with our schedule of weekdays, a reasonable claim is that occur on the different days with equal frequency. Use a 0.01 significance level to test that

claim. Can you provide an explanation for the result? Conduct the hypothesis test and the test statistic, critical value and/or *P*-value, and state the conclusion.

Day	*Sun*	*Mon*	*Tues*	*Wed*	*Thurs*	*Fri*	*Sat*
Number of births	77	110	124	122	120	123	97

7. The table below lists the frequency of wins for different post positions in the Kentucky Derby horse race. A post position of 1 is closest to the inside rail, so that horse has the shortest distance to run. (Because the number of horses varies from year to year, only the first ten post positions are included.) Use a 0.05 significance level to test the claim that the likelihood of winning is the same for the different post positions. Based on the result, should bettor consider the post position of a horse racing in the Kentucky Derby? Conduct the hypothesis test and the test statistic, critical value and/or *P*-value, and state the conclusion.

Post Position	1	2	3	4	5	6	7	8	9	10
Wins	19	14	11	14	14	7	8	11	5	11

8. The table below lists the cases of violent crimes are randomly selected and categorized by month. Use a 0.01 significance level to test the claim that the rate of violent crime is the same for each month. Can you explain the result? Conduct the hypothesis test and the test statistic, critical value and/or *P*-value, and state the conclusion.

Month	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Number	786	704	835	826	900	868	920	901	856	862	783	797

9. The table below lists the results of the Advanced Placement Biology class conducted genetics experiments with fruit flies. Use a 0.05 significance level to test the claim that the observed frequencies agree with the proportions that were expected according to principles of genetics Conduct the hypothesis test and the test statistic, critical value and/or *P*-value, and state the conclusion.

Characteristic	***Red eye / normal wing***	***Sepia eye / normal wing***	***Red eye / vestigial wing***	***Sepia eye / vestigial wing***
Frequency	59	15	2	4

10. The table below lists the claims that its M&M plain candies are distributed with the following color percentages: 16% green, 20% orange, 14% yellow, 24% blue, 13% red, and 13% brown. Use a 0.05 significance level to test the claim that the color distribution is as claimed.