統 計 學 (二)

Anderson's Statistics for Business &Economics (14/E)

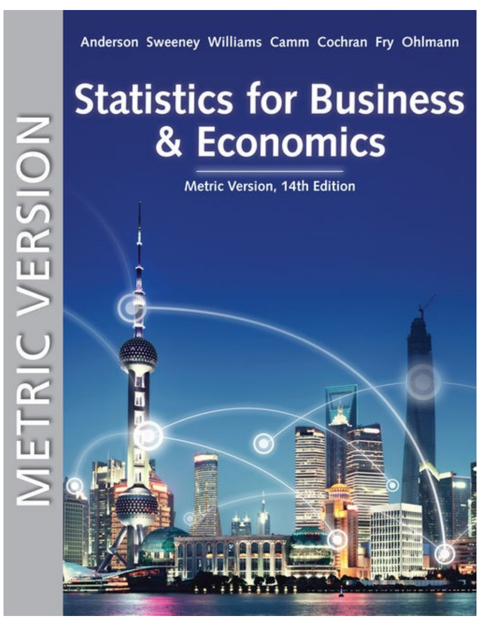
授課教師: 吳漢銘 國立政治大學統計學系

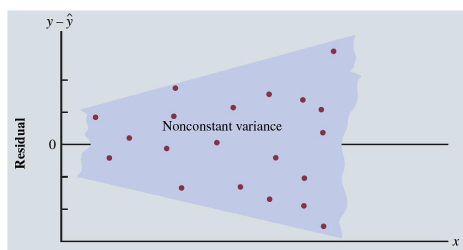
開課單位: 統計系

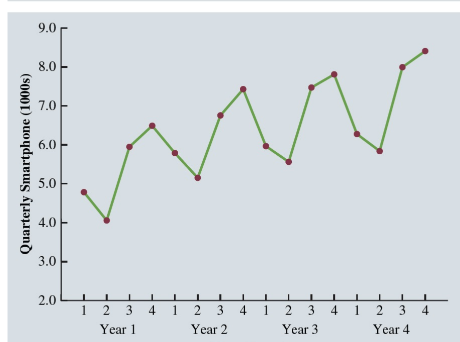
科目代碼: 000360041

教學網站: http://www.hmwu.idv.tw

系級: ________________ 學號: ________________ 姓名: ______________________







111 學年度第 2 學期

目錄

叮嚀

A. 平常就要唸書，做習題。

B. 考過的題目，要主動訂正。

C. 上課以「互相尊重」為最高原則並盡到「告知老師」的義務。

D. 上課可小聲討論、上廁所安靜去回、不鼓勵飲食。(請一定要維護教室整潔)

E. 四不一要: 「上課不聊天，睡覺不趴著，手機不要滑，考試不作弊，要認真。」

統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 10: Inference About Means and Proportions with Two Populations

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: http://www.hmwu.idv.tw

系級: ____________ 學號: ____________ 姓名: ____________

Overview

1. Discuss the statistical inference (____________ and ____________) for two population means (three situations: population standard deviations known, unknown; match samples) and the two population proportions.

2. *Examples*:

   (a) Develop an interval estimate of the difference between the mean starting salary for a population of men and the mean starting salary for a population of women.

   (b) Conduct a hypothesis test to determine whether any difference is present between the proportion of defective parts in a population of parts produced by supplier A and the proportion of defective parts in a population of parts produced by supplier B.

10.1 Inferences About the Difference Between Two Population Means: σ_1 and σ_2 Known

1. __________ denote the mean of population 1 (2), we will focus on inferences about the difference between the means: __________.
2. A simple random sample of __________ units from population 1 (2). The two samples, taken separately and independently, are referred to as ______________ simple random samples.
3. Assume the two population standard deviations, ____________, can be assumed ______________ to collecting the samples.
4. Question: how to compute a ________________ and develop an __________________ of the difference between the two population means when σ_1 and σ_2 are known.

Interval Estimation of $\mu_1 - \mu_2$

1. Example **Greystone Department Stores**, Inc., operates two stores in Buffalo, New York: One is in the inner city and the other is in a suburban shopping center. The regional manager noticed that products that sell well in one store do not always sell well in the other. The manager believes this situation may be attributable to differences in customer demographics at the two locations. Customers may differ in age, education, income, and so on. (對觀察事物提出問題)
2. Suppose the manager asks us to investigate the difference between the ____________ of the customers who shop at the two stores. (針對問題收集資料)
3. Let us define population 1 as all customers who shop at the ________________ and population 2 as all customers who shop at the ________________.
   (a) ______: mean of population 1 (i.e., the mean age of all customers who shop at the inner-city store)
   (b) ______: 5 mean of population 2 (i.e., the mean age of all customers who shop at the suburban store)

4. The difference between the two population means is __________. To estimate $\mu_1 - \mu_2$, we will select a simple random sample of ______ customers from population 1 and a simple random sample of ______ customers from population 2.

5. We then compute the two sample means.

   (a) ______: sample mean age for the simple random sample of n_1 inner-city customers

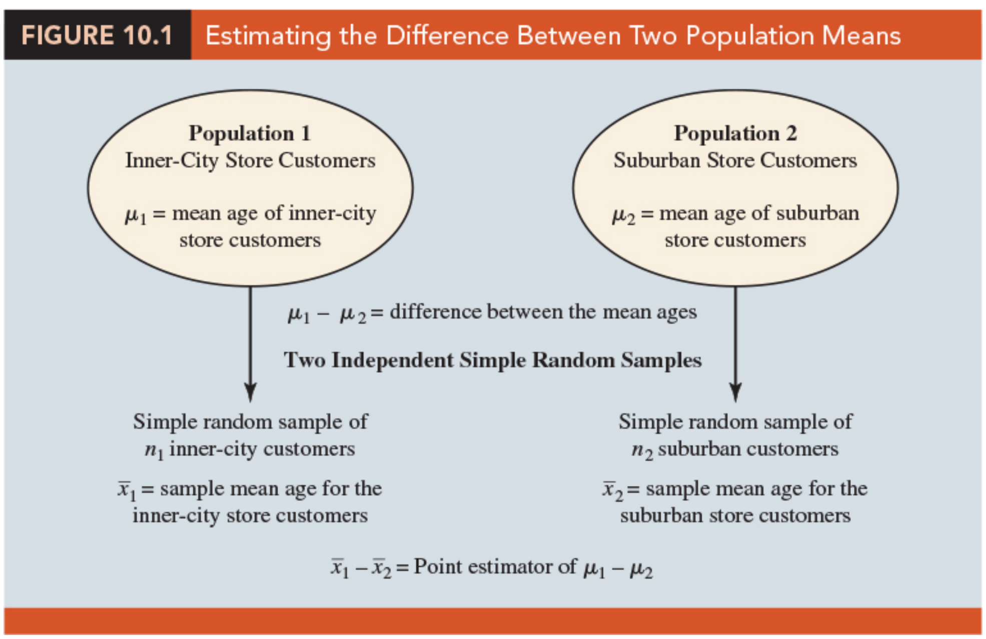
   (b) ______: sample mean age for the simple random sample of n_2 suburban customers

6. The point estimator of the difference between the two ____________________ is the difference between the two ________________.

7. **Point Estimator of the Difference Between Two Population Means**

$$\rule{5cm}{0.4pt} \qquad (10.1)$$

8. (Figure 10.1) the process used to estimate the difference between two population means based on two independent simple random samples.



9. The point estimator $\bar{x}_1 - \bar{x}_2$ has a standard error that describes the ____________ in the sampling distribution of the estimator.

10. **Standard Error of $\bar{x}_1 - \bar{x}_2$** With two independent simple random samples, the standard error of $\bar{x}_1 - \bar{x}_2$ is as follows:

$$\underline{\hspace{6cm}} \qquad (10.2)$$

(証明如下:) (Hint: $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2ab\ Cov(X, Y)$).

11. If both populations have a ___________ distribution, or if the sample sizes are large enough that the _______________________ enables us to conclude that the sampling distributions of $\bar{x}_1$ and $\bar{x}_2$ can be approximated by a normal distribution, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will have a ___________ distribution with mean given by __________. (Denoted by _______________________________)

12. In general, an interval estimate is given by a point estimate ± a margin of error. In the case of estimation of the difference between two population means, an interval estimate will take the following form:

$$\underline{\hspace{8cm}}$$

13. With the sampling distribution of $\bar{x}_1 - \bar{x}_2$ having a normal distribution, we can write the margin of error as follows:

$$\text{Margin of error } = \underline{\hspace{3.5cm}} = \underline{\hspace{4.5cm}} \qquad (10.3)$$

14. **Interval Estimate of the Difference Between Two Population Means: σ_1 and σ_2 Known**

$$\underline{\hspace{7cm}} \qquad (10.4)$$

where $1-\alpha$ is the confidence coefficient.

(公式說明如下:)

✎ Question .... (p485)

Example Greystone example. Based on data from previous customer demographic studies, the two population standard deviations are known with $\sigma_1 = 9$ years and $\sigma_2 = 10$ years. The data collected from the two independent simple random samples of Greystone customers provided the following results.

	Inner City Store	Suburban Store
Sample Size	$n_1 = 36$	$n_2 = 49$
Sample Mean	$\bar{x}_1 = 40$ years	$\bar{x}_2 = 35$ years

Find the margin of error and the 95% confidence interval estimate of the difference between the two population means.

sol:

Hypothesis Tests About $\mu_1 - \mu_2$

1. Let us consider hypothesis tests about the difference between two population means. Using _______ to denote the hypothesized difference between μ_1 and μ_2, the three forms for a hypothesis test are as follows:

Left-tailed test	Right-tailed test	Two-tailed test
H_0 : ________________	H_0 : ________________	H_0 : ________________
H_a : ________________	H_a : ________________	H_a : ________________

2. In many applications, ___________. Using the two-tailed test as an example, when $D_0 = 0$ the null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$.

3. In this case, the null hypothesis is that μ_1 and μ_2 are equal. Rejection of H_0 leads to the conclusion that $H_a : \mu_1 - \mu_2 \neq 0$ is true; that is, μ_1 and μ_2 are not equal.

4. The general steps for conducting hypothesis tests: choose a ______________________, compute the value of the ________________, and find the ___________ to ______________ whether the null hypothesis should be rejected.

5. With two independent simple random samples, we showed that the point estimator $\bar{x}_1 - \bar{x}_2$ has a standard error $\sigma_{\bar{x}_1 - \bar{x}_2}$ given by expression (10.2) and, when the sample sizes are large enough, the distribution of $\bar{x}_1 - \bar{x}_2$ can be described by a ___________ distribution.

6. **Test Statistic for Hypothesis Tests About $\mu_1 - \mu_2$: σ_1 and σ_2 Known**

   (10.5)

   ____

7. We demonstrated a two-tailed hypothesis test about the difference between two population means. Lower tail and upper tail tests can also be considered. These tests use the _____________________ as given in equation (10.5). The procedure for computing the p-value and the rejection rules for these one-tailed tests are the same as those for hypothesis tests involving a single population mean and single population proportion.

Question .... (p486)

As part of a study to evaluate differences in education quality between two training centers, a standardized examination is given to individuals who are trained at the centers. The difference between the mean examination scores is used to assess quality differences between the centers. The population means for the two centers are as follows. μ_1 is the mean examination score for the population of individuals trained at center A, μ_2 is the mean examination score for the population of individuals trained at center B. We begin with the tentative assumption that no difference exists between the training quality provided at the two centers. The standardized examination given previously in a variety of settings always resulted in an examination score standard deviation near 10 points. Thus, we will use this information to assume that the population standard deviations are known with $\sigma_1 = 10$ and $\sigma_2 = 10$. An $\alpha = 0.05$ level of significance is specified for the study. Independent simple random samples of $n_1 = 30$ individuals from training center A and $n_2 = 40$ individuals from training center B are taken. The respective sample means are $\bar{x}_1 = 82$ and $\bar{x}_2 = 78$. Do these data suggest a significant difference between the population means at the two training centers? State the null and alternative hypotheses for this two-tailed test, compute the test statistic, and state the decision rules based on the p-value approach and the critical value approach and make the decision.

sol:

Practical Advice

1. In most applications of the interval estimation and hypothesis testing procedures presented in this section, random samples with ____________ and ____________ are adequate.

2. In cases where either or both sample sizes are less than 30, the ________________ of the populations become important considerations.

3. In general, with smaller sample sizes, it is more important for the analyst to be satisfied that it is reasonable to assume that the distributions of the two populations are at least ________________________.

10.2 Inferences About The Difference Between Two Population Means: σ_1 and σ_2 Unknown

1. Extend the discussion of inferences about the difference between two population means to the case when the two population standard deviations, _______ and _______, are _____________.

2. In this case, we will use the sample standard deviations, _______ and _______, to estimate the unknown population standard deviations.

3. When we use the sample standard deviations, the interval estimation and hypothesis testing procedures will be based on the ________________ rather than the standard normal distribution.

Interval Estimation of $\mu_1 - \mu_2$

1. Let us develop the margin of error and an interval estimate of the difference between these two population means. (Recall) The interval estimate for the case when the

population standard deviations, σ_1 and σ_2, are known.

____

2. With σ_1 and σ_2 unknown, we will use the sample standard deviations s_1 and s_2 to estimate ______________ and replace $z_{\alpha/2}$ with ________.

3. **Interval Estimate of the Difference Between Two Population Means: σ_1 and σ_2 Unknown**

   ______________________________ (10.6)

   where $1 - \alpha$ is the confidence coefficient.

4. In this expression, the use of the t distribution is an __________________, but it provides excellent results and is relatively easy to use. The only difficulty that we encounter in using expression (10.6) is determining the appropriate ______________________ for $t_{\alpha/2}$.

5. Statistical software packages compute the appropriate degrees of freedom automatically. The formula used is as follows:

   Degrees of Freedom: t Distribution With Two Independent Random Samples

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \qquad (10.7)$$

✎ Question .... (p490)

(**Clearwater National Bank example**) Clearwater National Bank is conducting a study designed to identify differences between checking account practices by customers at two of its branch banks. A simple random sample of 28 checking accounts is selected from the Cherry Grove Branch and an in dependent simple random sample of 22 checking accounts is selected from the Beechmont Branch. The current checking account balance is recorded for each of the checking accounts. A summary of the account balances follows:

	Cherry Grove	Beechmont
Sample Size	$n_1 = 28$	$n_2 = 22$
Sample Mean	$\bar{x}_1 = \$1025$	$\bar{x}_2 = \$910$
Sample Standard Deviation	$s_1 = \$150$	$s_2 = \$125$

Clearwater National Bank would like to estimate the difference between the mean checking account balance maintained by the population of Cherry Grove customers and the population of Beechmont customers. Compute a 95% confidence interval estimate of the difference between the population mean checking account balances at the two branch banks.

sol:

Hypothesis Tests About $\mu_1 - \mu_2$

1. (Recall) Letting D_0 denote the hypothesized difference between μ_1 and μ_2, the test statistic used for the case where σ_1 and σ_2 are known is as follows.

   ____

   The test statistic, z, follows the standard normal distribution.

2. When σ_1 and σ_2 are unknown, we use s_1 as an estimator of σ_1 and s_2 as an estimator of σ_2. Substituting these sample standard deviations for σ_1 and σ_2 provides the following test statistic when σ_1 and σ_2 are unknown.

3. **Test Statistic for Hypothesis Tests About $\mu_1 - \mu_2$: σ_1 and σ_2 Unknown**

(10.8)

The degrees of freedom for t are given by equation (10.7).

Question .... (p491)

Consider a new computer software package developed to help systems analysts reduce the time required to design, develop, and implement an information system. To evaluate the benefits of the new software package, a random sample of 24 systems analysts is selected. Each analyst is given specifications for a hypothetical information system. Then 12 of the analysts are instructed to produce the information system by using current technology. The other 12 analysts are trained in the use of the new software package and then instructed to use it to produce the information system. This study involves two populations: a population of systems analysts using the current technology and a population of systems analysts using the new software package. In terms of the time required to complete the information system design project, the population means are as follows. μ_1 is the mean project completion time for systems analysts using the current technology and μ_2 is the mean project completion time for systems analysts using the new software package The researcher in charge of the new software evaluation project hopes to show that the new software package will provide a shorter mean project completion time. Thus, the researcher is looking for evidence to conclude that μ_2 is less than μ_1; in this case, the difference between the two population means, $\mu_1 - \mu_2$, will be greater than zero. Suppose that the 24 analysts complete the study with the results shown in Table 10.1.

TABLE 10.1 Completion Time Data and Summary Statistics for the Software Testing Study

	Current Technology	New Software
	300	274
	280	220
	344	308
	385	336
	372	198
	360	300
	288	315
	321	258
	376	318
	290	310
	301	332
	283	263
Summary Statistics		
Sample size	$n_1 = 12$	$n_2 = 12$
Sample mean	$\bar{x}_1 = 325$ hours	$\bar{x}_2 = 286$ hours
Sample standard deviation	$s_1 = 40$	$s_2 = 44$

Let the level of significance be $\alpha = 0.05$. State the null and the alternative hypothesis, the test statistic, p-value, the rejection rule, make a decision and conclusion.

sol:

TABLE 10.2 Output for the Hypothesis Test on the Difference Between the Current and New Software Technology

	Current	New
Mean	325	286
Variance	1600	1936
Observations	12	12
Hypothesized Mean Difference	0	
Degrees of Freedom	21	
Test Statistic	2.272	
One-Tail *p*-value	0.017	
One-Tail Critical Value	1.717	

(Software Output)

Practical Advice

1. The interval estimation and hypothesis testing procedures presented in this section are __________ and can be used with ________________ sample sizes.
2. In most applications, equal or nearly equal sample sizes such that the total sample size _______________ can be expected to provide very good results even if the populations are not normal.
3. Larger sample sizes are recommended if the distributions of the populations are ________________ or contain __________.
4. Smaller sample sizes should only be used if the analyst is satisfied that the distributions of the populations are at least ______________________.

Notes + Comments

1. How to make inferences about the difference between two population means when σ_1 and σ_2 are equal and unknown (______________)?
2. Based on above assumption, the two sample standard deviations are combined to provide the following pooled sample variance:

____

3. The t test statistic becomes

   and has ______________ degrees of freedom. At this point, the computation of the p-value and the interpretation of the sample results are identical to the procedures discussed earlier in this section.

4. A difficulty with this procedure is that the assumption that the two population standard deviations are equal is usually difficult to __________. Unequal population standard deviations are frequently encountered.

5. Using the pooled procedure may not provide satisfactory results, especially if the sample sizes n_1 and n_2 are __________________.

6. The t procedure that we presented in this section does not require the assumption of equal population standard deviations and can be applied whether the population standard deviations are equal or not. It is a more general procedure and is recommended for most applications.

10.3 Inferences About The Difference Between Two Population Means: Matched Samples

1. Example **Matched**.

   (a) Suppose employees at a manufacturing company can use two different methods to perform a production task. To maximize production output, the company wants to identify the method with the smaller population mean completion time.

   (b) Let _______ denote the population mean completion time for production method 1 and _______ denote the population mean completion time for production method 2.

(c) With no preliminary indication of the preferred production method, we begin by tentatively assuming that the two production methods have the same population mean completion time. Thus, the null hypothesis is ________________.

(d) If this hypothesis is rejected, we can conclude that the population mean completion times differ. In this case, the method providing the smaller mean completion time would be recommended.

(e) The null and alternative hypotheses are written as follows.

____

2. In choosing the sampling procedure that will be used to collect production time data and test the hypotheses, we consider two alternative designs. One is based on ____________________ and the other is based on ________________.

   (a) *Independent sample design*: A simple random sample of workers is selected and each worker in the sample uses method 1. A second independent simple random sample of workers is selected and each worker in this sample uses method 2. The test of the difference between population means is based on the procedures in Section 10.2.

   (b) *Matched sample design*: One simple random sample of workers is selected. Each worker first uses one method and then uses the other method. The order of the two methods is assigned randomly to the workers, with some workers performing method 1 first and others performing method 2 first. Each worker provides ____________________, one value for method 1 and another value for method 2.

3. In the matched sample design the two production methods are tested under similar conditions (i.e., with the same workers); hence this design often leads to a ____________________ than the independent sample design. The primary reason is that in a matched sample design, _______________________ is eliminated because the same workers are used for both production methods.

4. Assuming the analysis of a matched sample design is the method used to test the difference between population means for the two production methods. The key

to the analysis of the matched sample design is to realize that we consider only ________________________.

5. Therefore, we have six data values $(0.6, -0.2, 0.5, 0.3, 0.0,$ and $0.6)$ that will be used to analyze the difference between population means of the two production methods.

6. Let ______ is the mean of the difference in values for the population of workers. With this notation, the null and alternative hypotheses are rewritten as follows.

    ____

7. Assume the population of ____________ has a __________ distribution. This assumption is necessary so that we may use the ______________ for hypothesis testing and interval estimation procedures. Based on this assumption, the following test statistic has a t distribution with ________ degrees of freedom.

8. **Test Statistic for Hypothesis Tests Involving Matched Samples**

    ____

    where

    ____________, and ____________________ (10.9)

Question .... (p498)

(Table 10.3) (Matched Example). A random sample of six workers is used. The data on completion times for the six workers are given in Table 10.3. Note that each worker provides a pair of data values, one for each production method. Also note that the last column contains the difference in completion times d_i for each worker in the sample. Assume that the population of differences has a normal distribution. Test the hypotheses $H_0 : \mu_d = 0$ and $H_a : \mu_d \neq 0$, using $\alpha = 0.05$. Compute the test statistic, the p-value and draw a conclusion. Compute the 95% confidence interval for the difference between the population means of the two production methods. If H_0 is rejected, we can conclude that the population mean completion times differ.

TABLE 10.3 Task Completion Times for a Matched Sample Design

Worker	Completion Time for Method 1 (minutes)	Completion Time for Method 2 (minutes)	Difference in Completion Times (d_i)
1	6.0	5.4	.6
2	5.0	5.2	−.2
3	7.0	6.5	.5
4	6.2	5.9	.3
5	6.0	6.0	.0
6	6.4	5.8	.6

sol:

Area in Upper Tail	0.20	0.10	0.05	0.025	0.01	0.005
t-Value (5 df)	0.920	1.476	2.015	2.571	3.365	4.032

10.4 Inferences About The Difference Between Two Population Proportions

1. Letting ______ denote the proportion for population 1 and ______ denote the proportion for population 2.

2. Consider inferences about the difference between the two population proportions: ________.

3. To make an inference about this difference, we will select two independent random samples consisting of n_1 units from population 1 and n_2 units from population 2.

Interval Estimation of $p_1 - p_2$

1. Example **Tax Preparation Firm**

   A tax preparation firm is interested in comparing the quality of work at two of its regional offices. By randomly selecting samples of tax returns prepared at each office and verifying the sample returns' accuracy, the firm will be able to estimate the proportion of erroneous returns prepared at each office. Of particular interest is the difference between these proportions.

   (a) p_1: proportion of erroneous returns for population 1 (office 1)

   (b) p_2: proportion of erroneous returns for population 2 (office 2)

   (c) ______: sample proportion for a simple random sample from population 1

   (d) ______: sample proportion for a simple random sample from population 2

2. **Point Estimator of the Difference Between Two Population Proportions**

   ________________ (10.10)

3. Thus, the point estimator of the difference between two ____________ proportions is the difference between the __________ proportions of two independent simple random samples.

4. As with other point estimators, the point estimator $\bar{p}_1 - \bar{p}_2$ has a sampling distribution that reflects the possible values of $p_1 - p_2$ if we repeatedly took two independent

random samples. The mean of this sampling distribution is ________ and the standard error of $\bar{p}_1 - \bar{p}_2$ is:

Standard Error of $\bar{p}_1 - \bar{p}_2$

________________________________ . (10.11)

5. If the sample sizes are large enough that ________, ____________, ________, and ____________ are all greater than or equal to _____, the sampling distribution of $\bar{p}_1 - \bar{p}_2$ can be approximated by a __________ distribution.

6. With the sampling distribution of $\bar{p}_1 - \bar{p}_2$ approximated by a normal distribution, we would like to use ____________ as the margin of error.

7. However, $\sigma_{\bar{p}_1 - \bar{p}_2}$ given by equation (10.11) cannot be used directly because the two population proportions, p_1 and p_2, are unknown. Using the sample proportion $\bar{p}_1$ to estimate $\bar{p}_1$ and the sample proportion $\bar{p}_2$ to estimate p_2, the margin of error is:

Margin of error = ____________________________ . (10.12)

8. **Interval Estimate of the Difference Between Two Population Proportions**

   ______________________________________ (10.13)

   where $1 - \alpha$ is the confidence coefficient.

Question .... (p504)

(**Tax Preparation Example**) We find that independent simple random samples from the two offices provide the following information.

Office	1	2
n_i	250	300
Number of returns with errors	35	27

Find a margin of error and interval estimate of the difference between the two population proportions. and 90% confidence interval.

sol:

Hypothesis Tests About $p_1 - p_2$

1. Let us now consider hypothesis tests about no difference between the proportions of two populations. In this case, the three forms for a hypothesis test are as follows:

$$\begin{array}{lll} H_0: p_1 - p_2 \geq 0, & H_0: p_1 - p_2 \leq 0, & H_0: p_1 - p_2 = 0 \\ H_a: p_1 - p_2 < 0 & H_a: p_1 - p_2 > 0 & H_a: p_1 - p_2 \neq 0 \end{array}$$

2. When we assume __________________________, we have $p_1 - p_2 = 0$, which is the same as saying that the population proportions are equal, $p_1 = p_2$.

3. Under the assumption H_0 is true as an equality, the population proportions are equal and _______________. In this case, $\sigma_{\bar{p}_1 - \bar{p}_2}$ becomes **Standard Error of $\bar{p}_1 - \bar{p}_2$ when $p_1 = p_2 = p$**

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \underline{\hspace{5cm}} \tag{10.14}$$

4. With p unknown, we pool, or combine, the point estimators from the two samples ($\bar{p}_1$ and $\bar{p}_2$) to obtain a single point estimator of p as follows:
   Pooled Estimator of p When $p_1 = p_2 = p$

$$\underline{\hspace{4cm}} \tag{10.15}$$

   This pooled estimator of p is a weighted average of $\bar{p}_1$ and $\bar{p}_2$.

5. Substituting $\bar{p}$ for p in equation (10.14), we obtain an estimate of the standard error of $\bar{p}_1 - \bar{p}_2$. This estimate of the standard error is used in the test statistic.

6. The general form of the test statistic for hypothesis tests about the difference between two population proportions is the point estimator divided by the estimate of $\sigma_{\bar{p}_1 - \bar{p}_2}$.

7. **Test Statistic for Hypothesis Tests About** $p_1 - p_2$

(10.16)

This test statistic applies to large sample situations where n_1p_1, $n_1(1-p_1)$, n_2p_2, and $n_2(1-p_2)$ are all greater than or equal to 5.

Question .... (p506)

(**Tax Preparation Firm Example**) Assume that the firm wants to use a hypothesis test to determine whether the error proportions differ between the two offices. A two-tailed test is required. Use $\alpha = 0.10$ as the level of significance. State the null and alternative hypotheses. Compute the test statistic, and the p-value for this two-tailed test. State the decison rule and draw a conclusion.

sol:

😍 EXERCISES

10.1 : 1, 2, 4, 6

10.2 : 9, 10, 13, 14, 15

10.3 : 19, 23, 24

10.4 : 28, 29, 31, 34

SUP : 38, 39, 44

“少花點時間去取悅別人，多花些時間來經營自己。”
“Spend a little more time trying to make something of yourself and a little less time trying to impress people.”

— *早餐俱樂部 (Breakfast Club, 1985)*

統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 11: Inferences About Population Variances

上課時間地點: 二 D56, 商館 260206

授課教師: 吳漢銘 (國立政治大學統計學系副教授)

教學網站: http://www.hmwu.idv.tw

系級: __________ 學號: __________ 姓名: __________

Overview

1. Examine methods of statistical inference involving __________.

2. Example **Production Process of Filling Containers**

   (a) Consider the production process of filling containers with a liquid detergent product. The filling mechanism for the process is adjusted so that the __________ is 450 grams per container.

   (b) Although a mean of 450 grams is desired, the __________ of the filling weights is also critical. That is, even with the filling mechanism properly adjusted for the mean of 450 grams, we cannot expect every container to have exactly 450 grams.

   (c) By selecting a sample of containers, we can compute a __________ for the number of grams placed in a container. This value will serve as an __________ of the variance for the population of containers being filled by the production process.

(d) If the sample variance is modest, the production process will be continued. However, if the sample variance is excessive, ____________ and ____________ may be occurring even though the mean is correct at 450 grams.

(e) In this case, the filling mechanism will be readjusted in an attempt to __________ the filling variance for the containers.

3. In the first section we consider inferences about the variance of a _________ population. Subsequently, we will discuss procedures that can be used to make inferences about the variances of _______ populations.

11.1 Inferences About a Population Variance

1. The sample variance

$$s^2 = \underline{\hspace{4cm}} \qquad (11.1)$$

is the point estimator of the population variance σ^2. In using the sample variance as a basis for making inferences about a population variance, the sampling distribution of the quantity ______________ is helpful.

2. **Sampling Distribution of $(n-1)s^2/\sigma^2$**
Whenever a simple random sample of size n is selected from a __________ population, the sampling distribution of

$$\underline{\hspace{3cm}} \qquad (11.2)$$

is a ____________ distribution with _________ degrees of freedom (denoted by ____________ or ________).

3. 補充說明:

(a) Chi-squared distribution: `https://en.wikipedia.org/wiki/Chi-squared_distribution`

(b) **Theorem**: $X_1, X_2, \cdots, X_n$ are observations of a random sample of size n from the normal distribution $N(\mu, \sigma^2)$. $\bar{X}$ is the sample mean and S^2 is the sample variance. Then

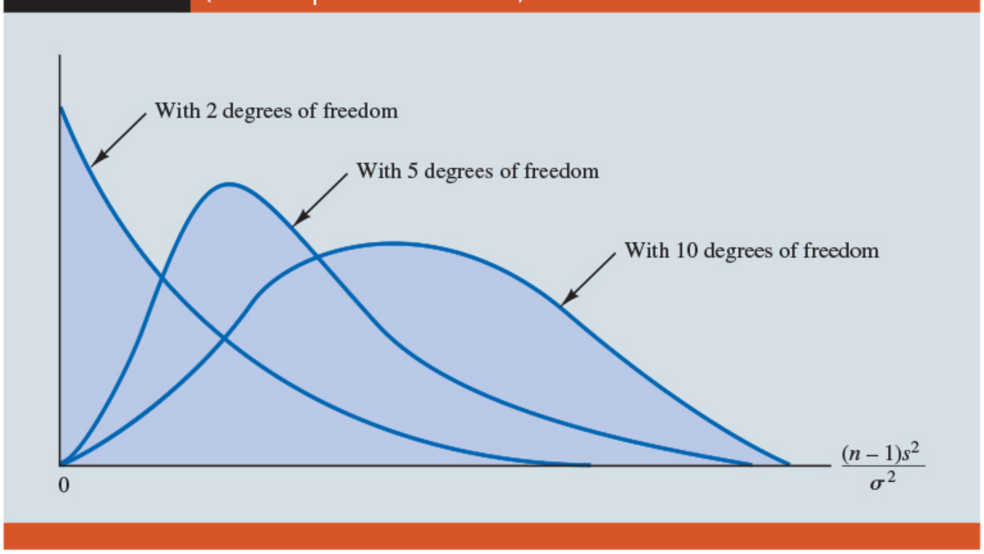
i. $\bar{X}$ and S^2 are independent.

ii. $\dfrac{(n-1)S^2}{\sigma^2} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$

(証明過程) Sampling Distribution of Sample Variance: `https://online.stat.psu.edu/stat414/lesson/26/26.3`

4. (Figure 11.1) shows some possible forms of the sampling distribution of $(n-1)s^2/\sigma^2$. Because the sampling distribution of $(n-1)s^2/\sigma^2$ is known to have a chi-square distribution whenever a simple random sample of size n is selected from a normal population, we can use the chi-square distribution to develop ____________________ and conduct ____________________ about a population variance.

FIGURE 11.1 Examples of the Sampling Distribution of $(n-1)s^2/\sigma^2$ (A Chi-Square Distribution)



Interval Estimation

1. Suppose that we are interested in estimating the population variance for the production filling process. A sample of ___________ containers is taken, and the sample variance for the filling quantities is found to be ______________. However, we know we cannot expect the variance of a sample of 20 containers to provide the

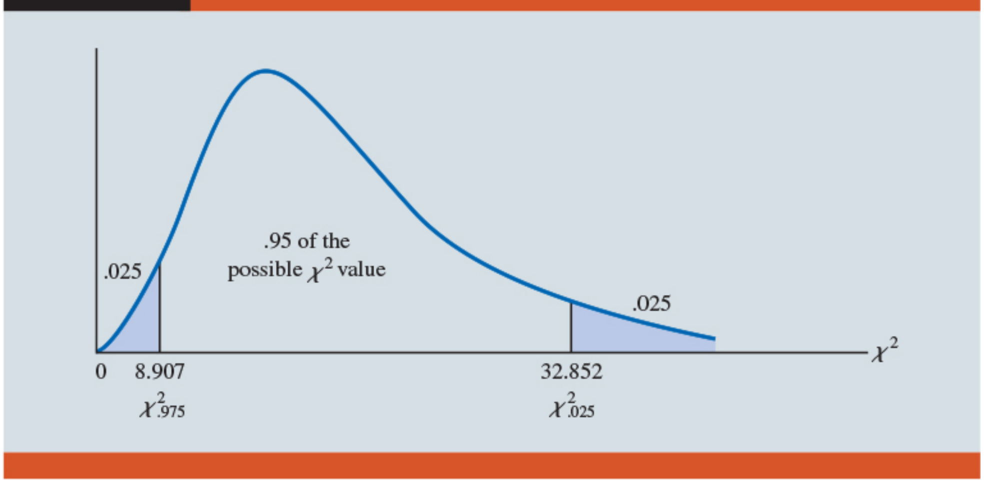
____________ of the variance for the population of containers filled by the production process. Hence, our interest will be in developing an interval estimate for the population variance.

2. (Figure 11.2) We will use the notation ______ a to denote the value for the chi-square distribution that provides an area or probability of ______________ of the χ^2_α value.

   (a) the chi-square distribution with 19 degrees of freedom is shown with ______________ indicating that 2.5% of the chi-square values are to the right of 32.852, and

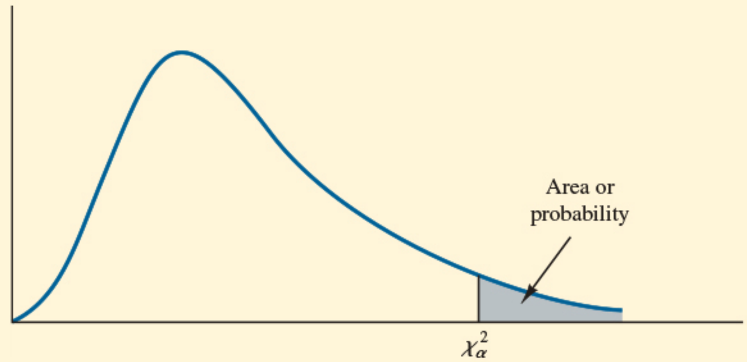
   (b) ______________ indicating that 97.5% of the chi-square values are to the right of 8.907.

FIGURE 11.2 A Chi-Square Distribution with 19 Degrees of Freedom



3. (Table 11.1) Tables of areas or probabilities are readily available for the chi-square distribution. Table 3 of Appendix B provides a more extensive table of chi-square values.

TABLE 11.1 Selected Values from the Chi-Square Distribution Table*



Degrees of Freedom	Area in Upper Tail .99	.975	.95	.90	.10	.05	.025	.01
1	.000	.001	.004	.016	2.706	3.841	5.024	6.635
2	.020	.051	.103	.211	4.605	5.991	7.378	9.210
3	.115	.216	.352	.584	6.251	7.815	9.348	11.345
4	.297	.484	.711	1.064	7.779	9.488	11.143	13.277
5	.554	.831	1.145	1.610	9.236	11.070	12.832	15.086
6	.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

**Note:* A more extensive table is provided as Table 3 of Appendix B.

4. From the graph in Figure 11.2 we see that 0.95, or 95%, of the chi-square values are between ________ and ________. That is, there is a 0.95 probability of obtaining a χ^2 value such that

   ____

5. We stated in expression (11.2) that $(n-1)s^2/\sigma^2$ follows a chi-square distribution; therefore we can substitute $(n-1)s^2/\sigma^2$ for χ^2 and write

   ________________ (11.3)

6. In effect, expression (11.3) provides an interval estimate in that .95, or 95%, of ________________ for $(n-1)s^2/\sigma^2$ will be in the interval $\chi^2_{0.975}$ to $\chi^2_{0.025}$.

7. We now need to do some algebraic manipulations with expression (11.3) to develop an interval estimate for the population variance σ^2. Working with the leftmost inequality in expression (11.3), we have

   $$\chi^2_{0.975} \le \frac{(n-1)s^2}{\sigma^2} \quad \Rightarrow \quad \sigma^2\chi^2_{0.975} \le (n-1)s^2 \quad \Rightarrow \quad \_\_\_\_\_\_\_\_\_\_ \qquad (11.4)$$

8. Performing similar algebraic manipulations with the rightmost inequality in expression (11.3) gives

   ________________ (11.5)

9. The results of expressions (11.4) and (11.5) can be combined to provide a 95% confidence interval estimate for the population variance

   ________________ (11.6)

10. Example **Production Process of Filling Containers** Recall that the sample of 20 containers provided a sample variance of $s^2 = 2.016$. With a sample size of 20, we have 19 degrees of freedom and $\chi^2_{0.975} = 8.907$ and $\chi^2_{0.025} = 32.852$. Using these values in expression (11.6) provides the following interval estimate for the population variance of filling quantities:

    ________________ or $1.166 \le \sigma^2 \le 4.300$

11. Taking the square root of these values provides the following 95% confidence interval for the population standard deviation.

____

12. Thus, we illustrated the process of using the _________________________ to establish ____________________ of a population variance and a population standard deviation.

13. $(1-\alpha)$% **Confidence Interval Estimate of a Population Variance**

______________________________ (11.7)

where the χ^2 values are based on a chi-square distribution with $n-1$ degrees of freedom and where $(1-\alpha)$ is the confidence coefficient.

Hypothesis Testing

1. Using _______ to denote the hypothesized value for the population variance, the three forms for a hypothesis test about a population variance are as follows:

$$\begin{array}{lll} H_0: \sigma^2 \geq \sigma_0^2, & H_0: \sigma^2 \leq \sigma_0^2, & H_0: \sigma^2 = \sigma_0^2 \\ H_0: \sigma^2 < \sigma_0^2, & H_0: \sigma^2 > \sigma_0^2, & H_0: \sigma^2 \neq \sigma_0^2 \end{array}$$

2. These three forms are similar to the three forms used to conduct one-tailed and two-tailed hypothesis tests about ___________________________________.

3. The procedure for conducting a hypothesis test about a population variance uses the hypothesized value for the population variance σ_0^2 and the sample variance s^2 to compute the value of a _______ test statistic.

4. **Test Statistic for Hypothesis Tests About a Population Variance** Assuming that the population has a normal distribution, the test statistic is:

$$\chi^2 = \underline{\qquad\qquad\qquad} \qquad (11.8)$$

where χ^2 has a chi-square distribution with $n-1$ degrees of freedom.

5. After computing the value of the χ^2 test statistic, either the __________ approach or the ______________ approach, may be used to determine whether the null hypothesis can be rejected.

6. Like the t distribution table, the chi-square distribution table does not contain sufficient detail to enable us to determine the p-value exactly. However, we can use the chi-square distribution table to obtain ______________________.

7. (Table 11.2)

TABLE 11.2 Summary of Hypothesis Tests About a Population Variance

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0: \sigma^2 \geq \sigma_0^2$ $H_a: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 \leq \sigma_0^2$ $H_a: \sigma^2 > \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$ $H_a: \sigma^2 \neq \sigma_0^2$
Test Statistic	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$
Rejection Rule: *p*-value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $\chi^2 \leq \chi^2_{(1-\alpha)}$	Reject H_0 if $\chi^2 \geq \chi^2_{\alpha}$	Reject H_0 if $\chi^2 \leq \chi^2_{(1-\alpha/2)}$ or if $\chi^2 \geq \chi^2_{\alpha/2}$

✎ Question .... (p532)

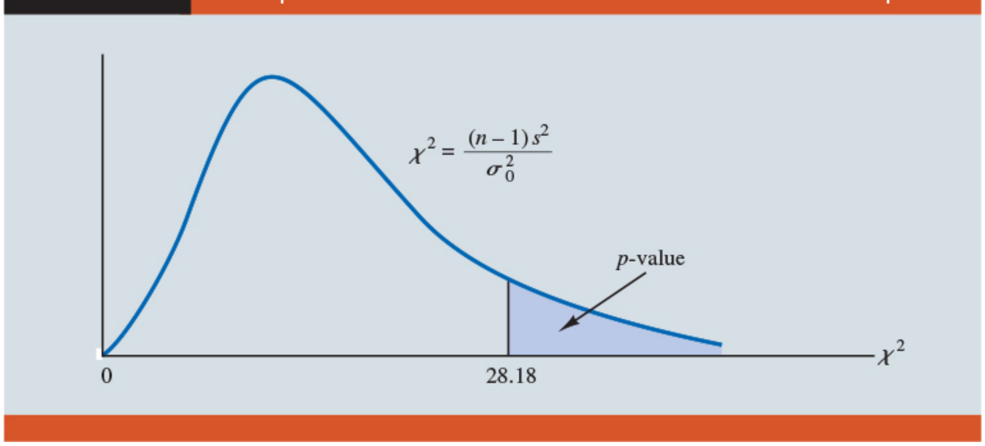
(**The St. Louis Metro Bus Example**) The St. Louis Metro Bus Company wants to promote an image of reliability by encouraging its drivers to maintain consistent schedules. As a standard policy, the company would like arrival times at bus stops to have low variability. In terms of the variance of arrival times, the company standard specifies an arrival time variance of 4 or less when arrival times are measured in minutes. The following hypothesis test is formulated to help the company determine whether the arrival time population variance is excessive.

$$H_0 : \sigma^2 \leq 4, \qquad H_a : \sigma^2 > 4$$

In tentatively assuming H_0 is true, we are assuming that the population variance of arrival times is within the company guideline. We reject H_0 if the sample evidence indicates that the population variance exceeds the guideline. In this case, follow-up steps should be taken to reduce the population variance. We conduct the hypothesis test using a level of significance of $\alpha = 0.05$. Suppose that a random sample of 24 bus arrivals taken at a downtown intersection provides a sample variance of $s^2 = 4.9$. Assuming that the population distribution of arrival times is approximately normal. Conduct a hypothesis testing and draw a conclusion using p-value approach and the critical value approach, separately.

sol:

FIGURE 11.3 Chi-Square Distribution for the St. Louis Metro Bus Example



Area in Upper Tail	0.10	0.05	0.025	0.01
χ^2 Value (23 df)	32.007	35.172	38.076	41.638

✎ Question .... (p533)

Conduct a two-tailed test about a population variance by considering a situation faced by a bureau of motor vehicles. Historically, the variance in test scores for individuals applying for driver's licenses has been $\sigma^2 = 100$. A new examination with new test questions has been developed. Administrators of the bureau of motor vehicles would like the variance in the test scores for the new examination to remain at the historical level. To evaluate the variance in the new examination test scores, the following two-tailed hypothesis test has been proposed.

$$H_0 : \sigma^2 = 100, \quad H_a : \sigma^2 \neq 100$$

Rejection of H_0 will indicate that a change in the variance has occurred and suggest that some questions in the new examination may need revision to make the variance of the new test scores similar to the variance of the old test scores. A sample of 30 applicants for driver's licenses will be given the new version of the examination. We will use a level of significance $\alpha = 0.05$ to conduct the hypothesis test. The sample of 30 examination scores provided a sample variance $s^2 = 162$. Conduct a hypothesis testing and draw a conclusion using p-value approach.

sol:

Area in Upper Tail	0.10	0.05	0.025	0.01
χ^2 Value (29 df)	39.087	42.557	45.722	49.588

11.2 Inferences About Two Population Variances

1. In some statistical applications we may want to compare the variances in product quality resulting from two different production processes, the variances in assembly times for two assembly methods, or the variances in temperatures for two heating devices.

2. In making comparisons about the two population variances, we will be using data collected from ________________________________, one from population 1 and another from population 2. The two sample variances s_1^2 and s_2^2 will be the basis for making inferences about the two population variances σ_1 and σ_2.

3. **Sampling Distribution of s_1^2/s_2^2 When $\sigma_1^2 = \sigma_2^2$**
   Whenever independent simple random samples of sizes n_1 and n_2 are selected from two __________ populations with ________________, the sampling distribution of _________ is an ______ distribution with __________ degrees of freedom for the numerator and __________ degrees of freedom for the denominator; s_1^2 is the sample variance for the random sample of n_1 items from population 1, and s_2^2 is the sample variance for the random sample of n_2 items from population 2.

---

4. 補充說明: The sampling distribution of ratio of variances is given by Prof. R. A. Fisher in 1924. According to Prof. R. A. Fisher, the ratio of two independent chi- square variates when divided by their respective degrees of freedom follows F-distribution as

$$F = \frac{\chi^2_{(n_1-1)}/(n_1-1)}{\chi^2_{(n_2-1)}/(n_2-1)}, \quad \text{Since } \chi^2 = \frac{(n-1)s^2}{\sigma^2}, \text{ therefore}$$

$$F = \frac{((n_1-1)s_1^2/\sigma_1^2)/(n_1-1)}{((n_2-1)s_2^2/\sigma_2^2)/(n_2-1)} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{(n_1-1,n_2-1)}$$

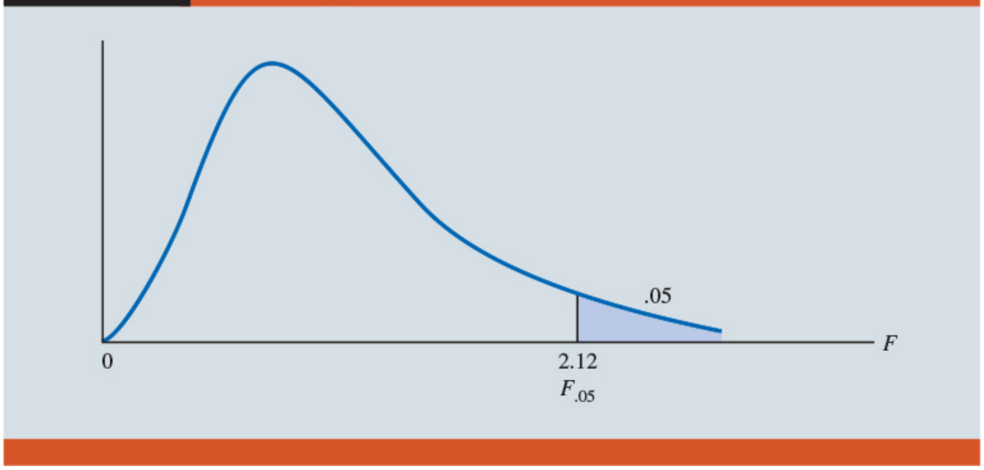
If $\sigma_1^2 = \sigma_2^2$, then

$$F = \frac{s_1^2}{s_2^2} \sim F_{(n_1-1,n_2-1)}$$

Therefore, the sampling distribution of ratio of sample variances follows F-distribution with (n_1-1, n_2-1) degrees of freedom.

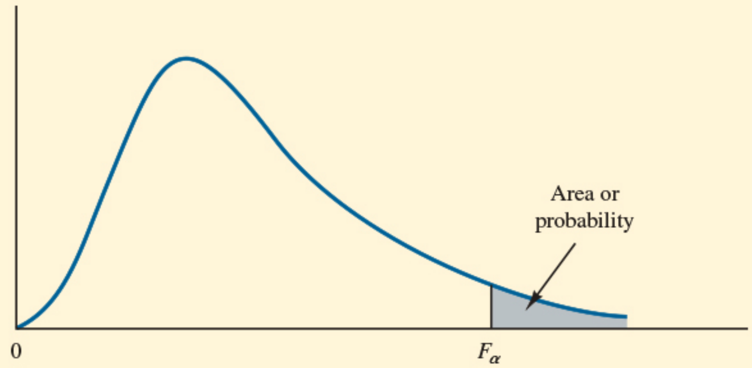
5. (Figure 11.4) The F distribution is ________________ and the F values can ________________. The shape of any particular F distribution depends on its numerator and denominator degrees of freedom.

FIGURE 11.4 *F* Distribution with 20 Degrees of Freedom for the Numerator and 20 Degrees of Freedom for the Denominator



6. (Table 11.3) We will use ______ to denote the value of F that provides an area or probability of a in the ____________ of the distribution. For example, as noted in Figure 11.4, $F_{0.05}(20, 20) = 2.12$ denotes the upper tail area of 0.05 for an F distribution with 20 degrees of freedom for the numerator and 20 degrees of freedom for the denominator.

TABLE 11.3 Selected Values from the *F* Distribution Table*



Denominator Degrees of Freedom	Area in Upper Tail	Numerator Degrees of Freedom 10	15	20	25	30
10	.10	2.32	2.24	2.20	2.17	2.16
	.05	2.98	2.85	2.77	2.73	2.70
	.025	3.72	3.52	3.42	3.35	3.31
	.01	4.85	4.56	4.41	4.31	4.25
15	.10	2.06	1.97	1.92	1.89	1.87
	.05	2.54	2.40	2.33	2.28	2.25
	.025	3.06	2.86	2.76	2.69	2.64
	.01	3.80	3.52	3.37	3.28	3.21
20	.10	1.94	1.84	1.79	1.76	1.74
	.05	2.35	2.20	2.12	2.07	2.04
	.025	2.77	2.57	2.46	2.40	2.35
	.01	3.37	3.09	2.94	2.84	2.78
25	.10	1.87	1.77	1.72	1.68	1.66
	.05	2.24	2.09	2.01	1.96	1.92
	.025	2.61	2.41	2.30	2.23	2.18
	.01	3.13	2.85	2.70	2.60	2.54
30	.10	1.82	1.72	1.67	1.63	1.61
	.05	2.16	2.01	1.93	1.88	1.84
	.025	2.51	2.31	2.20	2.12	2.07
	.01	2.98	2.70	2.55	2.45	2.39

**Note:* A more extensive table is provided as Table 4 of Appendix B.

7. Let us show how the F distribution can be used to conduct a hypothesis test about the variances of two populations. We begin with a test of the equality of two population variances. The hypotheses are stated as follows.

$$\underline{\hspace{3cm}}, \quad \underline{\hspace{3cm}}$$

8. The procedure used to conduct the hypothesis test requires two independent random

samples, one from each population. The two sample variances are then computed. We refer to the population providing the larger sample variance as population 1. Thus, a sample size of ______ and a sample variance of ______ correspond to population 1, and a sample size of ______ and a sample variance of ______ correspond to population 2.

9. **Test Statistic for Hypothesis Tests About Population Variances With** $\sigma_1^2 = \sigma_2^2$
Based on the assumption that both populations have a __________ distribution, the ratio of sample variances provides the following F test statistic:

$$F = \frac{}{} \qquad (11.10)$$

Denoting the population with the larger sample variance as population 1, the test statistic has an F distribution with $n_1 - 1$ degrees of freedom for the numerator and $n_2 - 1$ degrees of freedom for the denominator.

10. Because the F test statistic is constructed with the ________________________ in the numerator, the value of the test statistic will be in the ____________ of the F distribution.

11. (Table 11.4) A summary of hypothesis tests about two population variances.

TABLE 11.4 Summary of Hypothesis Tests About Two Population Variances

	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0: \sigma_1^2 \le \sigma_2^2$ $H_a: \sigma_1^2 > \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_a: \sigma_1^2 \ne \sigma_2^2$ Note: Population 1 has the larger sample variance
Test Statistic	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{s_1^2}{s_2^2}$
Rejection Rule: *p*-value	Reject H_0 if p-value $\le \alpha$	Reject H_0 if p-value $\le \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $F \ge F_\alpha$	Reject H_0 if $F \ge F_{\alpha/2}$

✐ Question .... (p539)

(**Dullus County Schools Example**) Dullus County Schools is renewing its school bus service contract for the coming year and must select one of two bus companies, the Milbank Company or the Gulf Park Company. We will use the variance of the arrival or pickup/delivery times as a primary measure of the quality of the bus service. Low variance values indicate the more consistent and higher-quality service. If the variances of arrival times associated with the two services are equal, Dullus School administrators will select the company offering the better financial terms. However, if the sample data on bus arrival times for the two companies indicate a significant difference between the variances, the administrators may want to give special consideration to the company with the better or lower variance service. The appropriate hypotheses follow.

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad H_a : \sigma_1^2 \neq \sigma_2^2.$$

If H_0 can be rejected, the conclusion of unequal service quality is appropriate. We will use a level of significance of $\alpha = 0.10$ to conduct the hypothesis test. A sample of 26 arrival times for the Milbank service provides a sample variance of 48 and a sample of 16 arrival times for the Gulf Park service provides a sample variance of 20. Because the Milbank sample provided the larger sample variance, we will denote Milbank as population 1. Use the p-value approach or the critical value approach to obtain the hypothesis testing conclusion.

sol:

Area in Upper Tail	0.10	0.05	0.025	0.01
F Value ($df_1 = 25$, $df_2 = 15$)	1.89	2.28	2.69	3.28

✎ Question .... (p541)

A one-tailed F test about the variances of two populations by considering a public opinion survey. Samples of 31 men and 41 women will be used to study attitudes about current political issues. The researcher conducting the study wants to test to see whether the sample data indicate that women show a greater variation in attitude on political issues than men. In the form of the one-tailed hypothesis test given previously, women will be denoted as population 1 and men will be denoted as population 2. The hypothesis test will be stated as follows.

$$H_0 : \sigma^2_{women} \leq \sigma^2_{men}, \quad H_a : \sigma^2_{women} > \sigma^2_{men}.$$

A rejection of H_0 gives the researcher the statistical support necessary to conclude that women show a greater variation in attitude on political issues. The survey results provide a sample variance of $s_1^2 = 120$ for women and a sample variance of $s_2^2 = 80$ for men. Use a level of significance $\alpha = 0.05$ to conduct the hypothesis test.

sol:

😍 EXERCISES

11.1 : 2, 3, 5, 9, 10

11.2 : 14, 15, 18, 19

SUP : 26, 29

“永遠不要讓別人的冷漠，影響了你對這世界的熱情。”
“Never allow the indifference of others to affect your passion for this world.”

— *魔女宅急便 (Kiki’s Delivery Service, 1989)*

統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 12: Comparing Multiple Proportions, Test of Independence and Goodness of Fit

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: http://www.hmwu.idv.tw

系級: ____________ 學號: ____________ 姓名: ____________

Overview

1. Consider cases in which the data are __________ by using a test statistic based on the chi-square (______) distribution.
2. In cases in which data are not naturally categorical, we define __________ and consider the observation count in each category. These ______________ are versatile and expand hypothesis testing with the following applications.
   (a) Testing the equality of population proportions for three or more populations. (Chi-Square ________________)
   (b) Testing the independence of two categorical variables. (Chi-square ________________)
   (c) Testing whether a probability distribution for a population follows a specific historical or theoretical probability distribution. (Chi-Square ________________)

12.1 Testing the Equality of Population Proportions for Three or More Populations

1. In this section, we show how the chi-square (χ^2) test statistic can be used to make statistical inferences about the ____________________ for three or more populations.

2. Using the notation

$$\underline{\qquad\quad} = \text{population proportion for population } i, i = 1, 2, 3, \cdots, k.$$

   the hypotheses for the equality of population proportions for $k \geq 3$ populations are as follows:

$$H_0 : \underline{\qquad\qquad\qquad\qquad}, \quad H_a : \text{Not all population proportions are equal}$$

   (a) If the ____________ and the chi-square test computations indicate H_0 cannot be rejected, we cannot detect a difference among the k population proportions.

   (b) However, if the sample data and the chi-square test computations indicate H_0 can be rejected, we have the ________________ to conclude that not all k population proportions are equal; that is, one or more population proportions differ from the other population proportions.

   (c) Further analyses can be done to conclude ____________________ or proportions are significantly different from others.

3. Example Organizations such as J.D. Power and Associates use the proportion of owners likely to repurchase a particular automobile as an indication of customer loyalty for the automobile. An automobile with a greater proportion of owners likely to repurchase is concluded to have greater customer loyalty.

   (a) Suppose that in a particular study we want to compare the customer loyalty for three automobiles: Chevrolet Impala, Ford Fusion, and Honda Accord. The current owners of each of the three automobiles form the three populations for the study. The three population proportions of interest are as follows:

p_1 = proportion likely to repurchase an Impala for the population of Chevrolet Impala owners.

p_2 = proportion likely to repurchase a Fusion for the population of Ford Fusion owners.

p_3 = proportion likely to repurchase an Accord for the population of Honda Accord owners.

(b) The hypotheses are stated as follows:

H_0 : ______________, H_a : Not all population proportions are equal

(c) To conduct this hypothesis test we begin by taking a sample of owners from each of the three populations. Thus we will have a sample of Chevrolet Impala owners, a sample of Ford Fusion owners, and a sample of Honda Accord owners.

(d) Each sample provides ________________ indicating whether the respondents are likely or not likely to repurchase the automobile.

(e) (Table 12.1) The data for samples of 125 Chevrolet Impala owners, 200 Ford Fusion owners, and 175 Honda Accord owners are summarized in Table 12.1.

TABLE 12.1 Sample Results of Likely to Repurchase for Three Populations of Automobile Owners (Observed Frequencies)

		Automobile Owners			
		Chevrolet Impala	**Ford Fusion**	**Honda Accord**	**Total**
Likely to Repurchase	**Yes**	69	120	123	312
	No	56	80	52	188
	Total	125	200	175	500

(f) This table has two rows for the responses Yes and No and three columns, one corresponding to each of the populations. The observed frequencies are summarized in the six cells of the table corresponding to each combination of the likely to repurchase responses and the three populations.

(g) The data in Table 12.1 are the observed frequencies for each of the six cells that represent the six combinations of the likely to ____________ response and the owner population.

(h) If we can determine the expected frequencies under the assumption H_0 is true, we can use the chi-square test statistic to determine whether there is a significant difference between the ____________ and ______________________.

(i) If a ______________________ exists between the observed and expected frequencies, the hypothesis H_0 can be ___________ and there is evidence that not all the population proportions are equal.

4. Expected frequencies for the six cells of the table are based on the following rationale.

   (a) First, we assume that the _________________ of equal population proportions is true.

   (b) Then we note that in the entire sample of 500 owners, a total of 312 owners indicated that they were likely to repurchase their current automobile. Thus, __________________ is the overall sample proportion of owners indicating they are likely to repurchase their current automobile.

   (c) If $H_0 : p_1 = p_2 = p_3$ is true, 0.624 would be the ________________ of the proportion responding likely to repurchase for each of the automobile owner populations.

   (d) So if the assumption of H_0 is true, we would expect 0.624 of the 125 Chevrolet Impala owners, or __________________ owners to indicate they are likely to repurchase the Impala. Using the 0.624 overall sample proportion, we would expect ____________________ of the 200 Ford Fusion owners and ____________________ of the Honda Accord owners to respond that they are likely to repurchase their respective model of automobile.

5. Let us generalize the approach to computing expected frequencies by letting _______ denote the expected frequency for the cell in _________ and ____________ of the table.

6. Note that 312 is the total number of Yes responses (row 1 total), 125 is the total sample size for Chevrolet Impala owners (column 1 total), and 500 is the total sample size. We can show

$$e_{11} = \left(\frac{\text{Row 1 Total}}{\text{Total Sample Size}}\right)(\text{Column 1 Total}) = \underline{\hspace{8cm}}.$$

7. **Expected Frequencies under the Assumption H_0 is True**

$$e_{ij} = \underline{\qquad\qquad\qquad\qquad\qquad\qquad} \tag{12.1}$$

8. (Table 12.2) Use equation (12.1) to verify the other expected frequencies:

TABLE 12.2 Expected Frequencies for Likely to Repurchase for Three Populations of Automobile Owners if H_0 is True

		Automobile Owners			
		Chevrolet Impala	Ford Fusion	Honda Accord	Total
Likely to Repurchase	Yes	78	124.8	109.2	312
	No	47	75.2	65.8	188
	Total	125	200	175	500

9. **Chi-Square Test Statistic**

$$\underline{\qquad\qquad\qquad\qquad\qquad\qquad} \tag{12.2}$$

where

$f_{ij} =$ ________________ for the cell in row i and column j.

$e_{ij} =$ ________________ for the cell in row i and column j.

under the assumption H_0 is true.

10. In a chi-square test involving the equality of k population proportions, the above test statistic has a chi-square distribution with k–1 degrees of freedom (____________) provided the expected frequency is __________ for each cell.

11. (補充說明) Why the test statistic for the chi-square test of homogeneity has a chi-square distribution? See

    (a) The Multinomial Distribution and the Chi-Squared Test for Goodness of Fit: `https://www.stat.berkeley.edu/~stark/SticiGui/Text/chiSquare.htm`,

    (b) 17.1 - Test For Homogeneity: `https://online.stat.psu.edu/stat415/lesson/17/17.1`

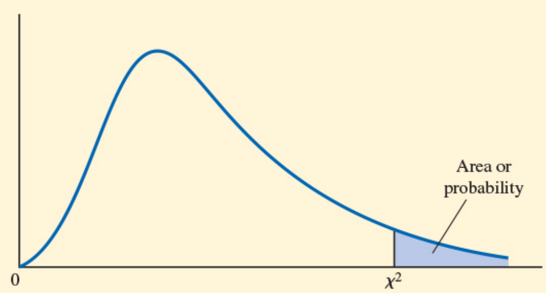
12. (Table 12.3) Computation of the chi-square test statistic:

TABLE 12.3 Computation of the Chi-Square Test Statistic for the Test of Equal Population Proportions

Likely to Repurchase?	Automobile Owner	Observed Frequency f_{ij}	Expected Frequency e_{ij}	Difference $f_{ij} - e_{ij}$	Squared Difference $(f_{ij} - e_{ij})^2$	Squared Difference Divided by Expected Frequency $(f_{ij} - e_{ij})^2/e_{ij}$
Yes	Impala	69	78.0	−9.0	81.00	1.04
Yes	Fusion	120	124.8	−4.8	23.04	.18
Yes	Accord	123	109.2	13.8	190.44	1.74
No	Impala	56	47.0	9.0	81.00	1.72
No	Fusion	80	75.2	4.8	23.04	.31
No	Accord	52	65.8	−13.8	190.44	2.89
	Total	500	500			$\chi^2 = 7.89$

13. (Table 12.4) In order to understand whether or not the value of the test statistic ____________ leads us to reject $H_0 : p_1 = p_2 = p_3$, you will need to understand and refer to values of the chi-square distribution:

TABLE 12.4 Selected Values of the Chi-Square Distribution



Degrees of Freedom	Area in Upper Tail .10	.05	.025	.01	.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.832	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801

(a) Since the expected frequencies shown in Table 12.2 are based on the assumption that $H_0 : p_1 = p_2 = p_3$ is true, observed frequencies, f_{ij}, that are in agreement with expected frequencies, e_{ij}, provide ________________ in equation (12.2). If this is the case, the value of the chi-square test statistic will be relatively small and H_0 ____________.

(b) On the other hand, if the differences between the observed and expected frequencies are large, values of $(f_{ij}-e_{ij})^2$ and the computed value of the test statistic will be ______. In this case, the null hypothesis of equal population proportions __________.

(c) Thus a chi-square test for equal population proportions will always be an __________ with rejection of H_0 occurring when the test statistic is in the upper tail of the chi-square distribution. (Reject H_0 if ________)

(d) We can use the upper tail area of the appropriate chi-square distribution and the _______ approach to determine whether the null hypothesis can be rejected.

14. Example In the automobile brand loyalty study, the three owner populations indicate that the appropriate chi-square distribution has __________ degrees of freedom.

chi-square distribution table					
Area in Upper Tail	0.10	0.05	0.025	0.01	0.005
χ^2 Value (2 *df*)	4.605	5.991	7.378	9.210	10.597

(a) (The p-value approach) We see the upper tail area at ________ is between ______ and ______. Thus, the corresponding upper tail area or _______ must be between ______ and ______. (Software: p-value $= 0.0193$)

(b) With __________, we reject H_0 and conclude that the three population proportions are not all equal and thus there is a difference in brand loyalties among the Chevrolet Impala, Ford Fusion, and Honda Accord owners.

(c) (The critical value approach) With $\alpha = 0.05$ and 2 degrees of freedom, the critical value for the chi-square test statistic is $\chi^2_{0.05,2} = 5.991$. The upper tail rejection region becomes

$$\text{Reject } H_0 \text{ if } \_\_\_\_\_\_\_\_$$

With $7.89 \geq 5.991$, we reject H_0.

(d) Thus, the p-value approach and the critical value approach provide the same hypothesis-testing conclusion.

15. **A Chi-Square Test for the Equality of Population Proportions for $k \geq 3$ Populations**

(1) State the null and alternative hypotheses

H_0 : ____________________, H_a : Not all population proportions are equal

(2) Set the level of significance ______. Select a random sample from each of the populations and record the observed frequencies, _______, in a table with 2 rows and k columns. Assume the null hypothesis is true and compute the expected frequencies, _______.

(3) If ________________, compute the test statistic:

$$\chi^2 =$$

____

(4) Rejection rule (Desicion rule):

i. p-value approach: Reject H_0 if _______________.

ii. Critical value approach: Reject H_0 if _______________.

(5) Make decision.

(6) Draw conclusion with respect to the problem.

A Multiple Comparison Procedure

1. Since the chi-square test indicated that not all population proportions are equal, it is reasonable for us to proceed by attempting to ______________________________ among the population proportions exist.

2. Begin by computing the three sample proportions as follows:

Brand Loyalty	Sample Proportions
Chevrolet Impala	$p_1 = 69/125 = 0.5520$
Ford Fusion	$p_2 = 120/200 = 0.6000$
Honda Accord	$p_3 = 123/175 = 0.7029$

3. For this we will rely on a ____________ procedure that can be used to conduct statistical tests between all pairs of population proportions. In the following, we discuss a multiple comparison procedure known as the ____________.

4. We begin by computing the ____________ between sample proportions for each pair of populations in the study:

   - Chevrolet Impala and Ford Fusion:
     ______ $= |0.5520 - 0.6000| = 0.0480$
   - Chevrolet Impala and Honda Accord:
     ______ $= |0.5520 - 0.7029| = 0.1509$
   - Ford Fusion and Honda Accord:
     ______ $= |0.6000 - 0.7029| = 0.1029$

5. In a second step, we select a ____________ and compute the corresponding ________ for each pairwise comparison using the following expression.

6. **Critical Values for the Marascuilo Pairwise Comparison Procedure for K Population Proportions**:
   For each pairwise comparison compute a critical value as follows:

   $$CV_{ij} = \_\_\_\_\_\_\_\_\_\_\_\_ \qquad (12.3)$$

   where

   χ^2_α = chi-square with a level of significance a and k–1 degrees of freedom

   $\bar{p}_i$ and $\bar{p}_j$ = sample proportions for populations i and j

   n_i and n_j = sample sizes for populations i and j

7. Using the chi-square distribution in Table 12.4, $k-1 = 3-1 = 2$ degrees of freedom, and a 0.05 level of significance, we have $\chi^2_{0.05,2} = 5.991$. Now using the sample proportions $\bar{p}_1 = 0.5520$, $\bar{p}_2 = 0.6000$, and $\bar{p}_3 = 0.7029$, the critical values for the three pairwise comparison tests are as follows:

   (a) Chevrolet Impala and Ford Fusion

   $$CV_{12} = \_\_\_\_\_\_\_\_\_\_\_\_$$

(b) Chevrolet Impala and Honda Accord

$$CV_{13} = \sqrt{5.991}\sqrt{\frac{0.5520(1-0.5520)}{125} + \frac{0.7029(1-0.7029)}{175}} = 0.1379$$

(c) Ford Fusion and Honda Accord

$$CV_{23} = \sqrt{5.991}\sqrt{\frac{0.6000(1-0.6000)}{200} + \frac{0.7029(1-0.7029)}{175}} = 0.1198$$

8. If the absolute value of any pairwise sample proportion difference ____________ exceeds its corresponding critical value, _________, the pairwise difference is ______________ at the 0.05 level of significance and we can conclude that the two corresponding population proportions are different.

9. (Table 12.5) pairwise comparison procedure:

TABLE 12.5 Pairwise Comparison Tests for the Automobile Brand Loyalty Study

Pairwise Comparison	$\|\bar{p}_i - \bar{p}_j\|$	CV_{ij}	Significant if $\|\bar{p}_i - \bar{p}_j\| > CV_{ij}$
Chevrolet Impala vs. Ford Fusion	.0480	.1380	Not significant
Chevrolet Impala vs. Honda Accord	.1509	.1379	Significant
Ford Fusion vs. Honda Accord	.1029	.1198	Not significant

10. The conclusion from the pairwise comparison procedure is that the only significant difference in customer loyalty occurs between the Chevrolet Impala and the Honda Accord. Our sample results indicate that the Honda Accord had a greater population proportion of owners who say they are likely to repurchase the Honda Accord. Thus, we can conclude that the Honda Accord ($\bar{p}_3 = 0.7029$) has a greater customer loyalty than the Chevrolet Impala ($\bar{p}_1 = 0.5520$). The results of the study are inconclusive as to the comparative loyalty of the Ford Fusion.

11. While the Ford Fusion did not show significantly different results when compared to the Chevrolet Impala or Honda Accord, a larger sample may have revealed a significant difference between Ford Fusion and the other two automobiles in terms of customer loyalty.

12. It is not uncommon for a multiple comparison procedure to show significance for some pairwise comparisons and yet not show significance for other pairwise comparisons in the study.

13. (補充說明) Why if $|\bar{p}_i - \bar{p}_j| \geq CV_{ij}$, the pairwise difference is significant?

    - If $X_1 \sim B(n_1, p_1)$, we have $E(X_1) = n_1 p_1$ and $Var(X_1) = n_1 p_1 (1 - p_1)$.
    - $\dfrac{X_1}{n_1} = \bar{p}_1 = \hat{p}_1$.
    - $E(\dfrac{X_1}{n_1}) =$
    - $Var(\dfrac{X_1}{n_1}) =$
    - CLT:
    - Similarly for $X_2 \sim B(n_2, p_2)$.
    - $E(\bar{p}_1 - \bar{p}_2) =$
    - $Var(\bar{p}_1 - \bar{p}_2) =$
    - Under $H_0 : p_1 = p_2$, test statistic: $\bar{p}_1 - \bar{p}_2$

12.2 Test of Independence

1. An important application of a chi-square test involves using sample data to test for the ________________ of two ______________ variables. For this test we take

_______________ from a population and record the observations for two categorical variables.

2. We will summarize the data by counting the number of responses for each combination of a category for variable 1 and a category for variable 2.

3. The null hypothesis for this test is that the two categorical variables are independent. Thus, the test is referred to as a _______________________.

4. Example A beer industry association conducts a survey to determine the preferences of beer drinkers for light, regular, and dark beers.

   (a) A sample of 200 beer drinkers is taken with each person in the sample asked to indicate a preference for one of the three types of beers: light, regular, or dark. At the end of the survey questionnaire, the respondent is asked to provide information on a variety of demographics including gender: male or female.

   (b) A research question of interest to the association is whether preference for the three types of beer is independent of the gender of the beer drinker.

   (c) If the two categorical variables, beer preference and gender, are independent, beer preference does not depend on gender and the preference for light, regular, and dark beer can be expected to be the same for male and female beer drinkers.

   (d) However, if the test conclusion is that the two categorical variables are not independent, we have evidence that beer preference is associated or dependent upon the gender of the beer drinker.

   (e) As a result, we can expect beer preferences to differ for male and female beer drinkers. In this case, a beer manufacturer could use this information to customize its promotions and advertising for the different target markets of male and female beer drinkers.

5. The hypotheses for this test of independence are as follows:

   H_0 : Beer preference is ________________ of gender

   H_a : Beer preference is ____________________ of gender

6. (Table 12.6) Since an objective of the study is to determine if there is difference between the beer preferences for male and female beer drinkers, we consider gender an

________________ and follow the usual practice of making the explanatory variable the ________ variable in the data tabulation table. The beer preference is the ________________ variable and is shown as the ______ variable. The sample results of the 200 beer drinkers in the study are summarized in Table 12.6.

TABLE 12.6 Sample Results for Beer Preferences of Male and Female Beer Drinkers (Observed Frequencies)

		Gender		
		Male	**Female**	**Total**
Beer Preference	**Light**	51	39	90
	Regular	56	21	77
	Dark	25	8	33
	Total	132	68	200

7. For the categorical variable gender, we see 132 of the 200 in the sample were male. This gives us the estimate that ________________, of the beer drinker population is male. Similarly we estimate that ________________, of the beer drinker population is female. Thus male beer drinkers appear to outnumber female beer drinkers approximately 2 to 1.

8. Sample proportions or percentages for the three types of beer are
   (a) Prefer Light Beer ________________
   (b) Prefer Regular Beer $77/200 = 0.385$, or 38.5%
   (c) Prefer Dark Beer $33/200 = 0.165$, or 16.5%

9. Across all beer drinkers in the sample, light beer is preferred most often and dark beer is preferred least often.

10. The computations and formulas used to determine if beer preference and gender are independent are the same as those used for the chi-square test in Section 12.1. Under the assumption that the beer preferences and gender are independent. Thus the expected frequency for row i and column j is given by

$$e_{ij} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \quad (12.4)$$

11. (Table 12.7) expected frequencies

TABLE 12.7 Expected Frequencies If Beer Preference Is Independent of the Gender of the Beer Drinker

		Gender		
		Male	Female	Total
Beer Preference	Light	59.40	30.60	90
	Regular	50.82	26.18	77
	Dark	21.78	11.22	33
	Total	132	68	200

12. The chi-square test statistic.

$$\chi^2 = \qquad\qquad\qquad\qquad \tag{12.5}$$

____

13. With r rows and c columns in the table, the chi-square distribution will have _______________ degrees of freedom provided the expected frequency is ______________ for each cell.

14. (Table 12.8)

TABLE 12.8 Computation of the Chi-Square Test Statistic for the Test of Independence Between Beer Preference and Gender

Beer Preference	Gender	Observed Frequency f_{ij}	Expected Frequency e_{ij}	Difference $f_{ij} - e_{ij}$	Squared Difference $(f_{ij} - e_{ij})^2$	Squared Difference Divided by Expected Frequency $(f_{ij} - e_{ij})^2/e_{ij}$
Light	Male	51	59.40	−8.40	70.56	1.19
Light	Female	39	30.60	8.40	70.56	2.31
Regular	Male	56	50.82	5.18	26.83	.53
Regular	Female	21	26.18	−5.18	26.83	1.02
Dark	Male	25	21.78	3.22	10.37	.48
Dark	Female	8	11.22	−3.22	10.37	.92
	Total	200	200			$\chi^2 = 6.45$

15. The upper tail area of the chi-square distribution with 2 degrees of freedom:

Area in Upper Tail	0.10	0.05	0.025	0.01	0.005
χ^2 Value (2 *df*)	4.605	5.991	7.378	9.210	10.597

16. Thus, we see the upper tail area at ____________ is between ________ and _________, and so the corresponding upper tail area or __________ must be between 0.05 and 0.025. With _______________, we reject H_0 and conclude that beer preference is not independent of the gender of the beer drinker. (Software: p-value = .0398)

17. With $\alpha = 0.05$ and 2 degrees of freedom, the critical value for the chi-square test statistic is $\chi^2_{0.05,2} = 5.991$. The upper tail rejection region becomes

    Reject H_0 if ____________________

    With $6.45 \geq 5.991$, we reject H_0.

18. While we now have evidence that beer preference and gender are not independent, we will need to gain additional insight from the data to assess the nature of the _____________ between these two variables. One way to do this is to compute the probability of the beer preference responses for males and females separately.

Beer Preference	Male	Female
Light		$39/68 = 0.5735$, or 57.35%
Regular	$56/132 = 0.4242$, or 42.42%	$21/68 = 0.3088$, or 30.88%
Dark	$25/132 = 0.1894$, or 18.94%	$8/68 = .1176$, or 11.76%

19. What observations can you make about the association between beer preference and gender in the sample?

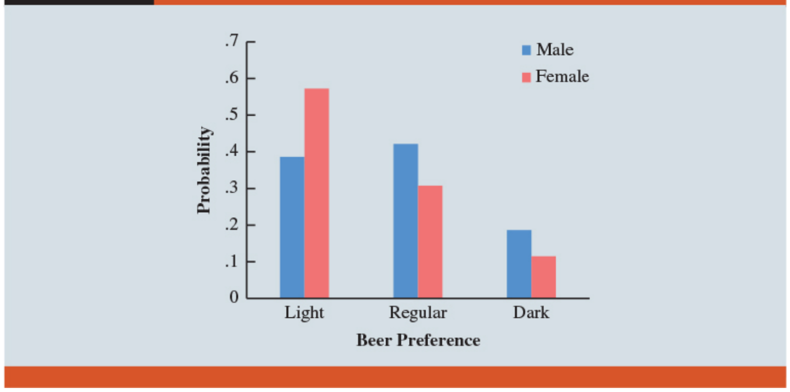
    (a) For female beer drinkers, the highest preference is for light beer at 57.35%.

    (b) For male beer drinkers, regular beer is most frequently preferred at 42.42%.

    (c) While female beer drinkers have a higher preference for light beer than males, male beer drinkers have a higher preference for both regular beer and dark beer.

    (d) (Figure 12.1) Data visualization through bar charts is helpful in gaining insight as to how two categorical variables are associated.

FIGURE 12.1 Bar Chart Comparison of Beer Preference by Gender



20. **Chi-Square Test for Independence of Two Categorical Variables**

    (1) State the null and alternative hypotheses.

    H_0 : The two categorical variables are independent,

    H_a : The two categorical variables are not independent

    (2) Set a level of significance α. Select a random sample from the population and collect data for both variables for every element in the sample. Record the observed frequencies, f_{ij}, in a table with r rows and c columns. The expected frequencies must all be 5 or more for the chi-square test to be valid. Assume the null hypothesis is true and compute the expected frequencies, e_{ij}

    (3) If the expected frequency, e_{ij}, is 5 or more for each cell, compute the test statistic:

    $$\chi^2 = \underline{\qquad\qquad\qquad\qquad}$$

    (4) Rejection rule:

    i. p-value approach: Reject H_0 if p-value $\leq \alpha$.

    ii. Critical value approach: Reject H_0 if $\chi^2 \geq \chi^2_{\alpha,(r-1)(c-1)}$.

    (5) Draw decision and conclusion.

21. Finally, if the null hypothesis of independence is rejected, summarizing the probabilities as shown in the above example will help the analyst determine where the ______________ or ______________ exists for the two categorical variables.

12.3 Goodness of Fit Test

1. In this section we use a chi-square test (goodness of fit tests) to determine whether a population being sampled has a ________________.
   (a) We first consider a population with a historical ________ probability distribution and use a ________ test to determine if new sample data indicate there has been a change in the population distribution compared to the historical distribution.
   (b) We then consider a situation where an assumption is made that a population has a ______ probability distribution and use a goodness of fit test to determine if sample data indicate that the assumption of a normal probability distribution is or is not appropriate.

Multinomial Probability Distribution

1. With a multinomial probability distribution, each element of a population is assigned to one and only one of three or more ________.
2. Wikipedia: Multinomial distribution:
   `https://en.wikipedia.org/wiki/Multinomial_distribution`
3. Example Consider the market share study being conducted by Scott Marketing Research.
   (a) Over the past year, market shares for a certain product have stabilized at 30% for company A, 50% for company B, and 20% for company C. Since each customer is classified as buying from one of these companies, we have a multinomial probability distribution with three possible outcomes.
   (b) The probability for each of the three outcomes is:
      i. p_A = probability a customer purchases the company A product
      ii. p_B = probability a customer purchases the company B product
      iii. p_C = probability a customer purchases the company C product
   (c) Using the historical market shares, we have multinomial probability distribution with $p_A = 0.30$, $p_B = 0.50$, and $p_C = 0.20$.

(d) Company C plans to introduce a "new and improved" product to replace its current entry in the market. Company C has retained Scott Marketing Research to determine whether the new product will alter or change the market shares for the three companies.

(e) Specifically, the Scott Marketing Research study will introduce a sample of customers to the new company C product and then ask the customers to indicate a preference for the company A product, the company B product, or the new company C product.

4. The hypothesis test to determine if the new company C product is likely to change the historical market shares for the three companies.

H_0 : ______________________________

H_a : The population proportions are not $p_A = 0.30, p_B = 0.50,$ and $p_C = 0.20$

5. The null hypothesis is based on the historical multinomial probability distribution for the market shares. If sample results lead to the rejection of H_0, Scott Marketing Research will have evidence to conclude that the introduction of the new company C product will change the market shares.

6. Let us assume that the market research firm has used a consumer panel of 200 customers. Each customer was asked to specify a purchase preference among the three alternatives: company A's product, company B's product, and company C's new product. The 200 responses are summarized:

Observed Frequency

Company A's Product	Company B's Product	Company C's New Product
48	98	54

7. Perform a goodness of fit test that will determine whether the sample of 200 customer purchase preferences is ______________ the null hypothesis.

8. Like other chi-square tests, the goodness of fit test is based on a comparison of observed frequencies with the expected frequencies under the assumption that the null hypothesis is true.

9. The expected frequency for each category is found by multiplying the sample size of 200 by the hypothesized proportion for the category:

Expected Frequency

Company A's Product	Company B's Product	Company C's New Product
$200(0.30) = 60$	$200(0.50) = 100$	$200(0.20) = 40$

10. **Test Statistic for Goodness of Fit**

$$\chi^2 = \underline{\qquad\qquad\qquad} \tag{12.6}$$

where

(a) $f_i =$ ____________ frequency for category i

(b) $e_i =$ ____________ frequency for category i

(c) $k =$ the number of ____________

Note: The test statistic has a chi-square distribution with $k-1$ degrees of freedom provided that the ____________ frequencies are ____________ for all categories.

11. The test for goodness of fit is always a one-tailed test with the rejection occurring in the upper tail of the chi-square distribution:

Reject H_0 if ____________

12. Example (Table 12.9) Let us continue with the Scott Marketing Research example and use the sample data to test the hypothesis that the multinomial population has the market share proportions $p_A = 0.30$, $p_B = 0.50$, and $p_C = 0.20$. We will use an $\alpha = 0.05$ level of significance. We proceed by using the observed and expected frequencies to compute the value of the test statistic.

TABLE 12.9 Computation of the Chi-Square Test Statistic for the Scott Marketing Research Market Share Study

Category	Hypothesized Proportion	Observed Frequency f_i	Expected Frequency e_i	Difference $f_i - e_i$	Squared Difference $(f_i - e_i)^2$	Squared Difference Divided by Expected Frequency $(f_i - e_i)^2/e_i$
Company A	.30	48	60	−12	144	2.40
Company B	.50	98	100	−2	4	.04
Company C	.20	54	40	14	196	4.90
Total		200				$\chi^2 = 7.34$

13. With the expected frequencies all 5 or more, the chi-square test statistic is ____________. We will reject the null hypothesis if the differences between the observed and expected frequencies are large.

14. The test statistic $\chi^2 = 7.34$ is between 5.991 and 7.378. Thus, the corresponding upper tail area or p-value must be between ________________. With ________________, we reject H_0 and conclude that the introduction of the new product by company C will alter the historical market shares. (Software: p-value $= 0.0255$)

Area in Upper Tail	0.10	.05	0.025	0.01	0.005
χ^2 Value (2 df)	4.605	5.991	7.378	9.210	10.597

15. The critical value approach: with $\alpha = 0.05$ and 2 degrees of freedom, the critical value for the test statistic is $\chi^2_{0.05} = 5.991$. The upper tail rejection rule becomes ______________________________. With $7.34 > 5.991$, we reject H_0.

16. Now that we have concluded the introduction of a new company C product will alter the market shares for the three companies, we are interested in knowing more about how the market shares are likely to change.

17. Using the historical market shares and the sample data, we summarize the data as follows:

Company	Historical Market Share (%)	Sample Data Market Share (%)
A	30	____________________
B	50	$98/200 = 0.49$, or 49
C	20	$54/200 = 0.27$, or 27

18. (Figure 12.2) This data visualization process shows that the new product will likely increase the market share for company C. Comparisons for the other two companies indicate that company C's gain in market share will hurt company A more than company B.

FIGURE 12.2 Bar Chart of Market Shares by Company Before and After the New Product for Company C

19. **Multinomial Probability Distribution Goodness of Fit Test**

 (1) State the null and alternative hypotheses.

  i. H_0: The population ____________________ probability distribution with specified probabilities for each of the k categories

  ii. H_a: The population does not follow a multinomial distribution with the specified probabilities for each of the k categories

 (2) Set a level of significance α and select a random sample and record the ____________ frequencies f_i for each category. Assume the null hypothesis is true and determine the ____________ frequency e_i in each category by multiplying the category probability by the sample size.

 (3) If the expected frequency e_i is at least 5 for each category, compute the value of the test statistic.

 (4) Rejection rule: ____________________

i. p-value approach: Reject H_0 if p-value $\leq \alpha$

ii. Critical value approach: Reject H_0 if $\chi^2 \geq \chi^2_{\alpha,k-1}$.

(5) Draw decision and conclusion.

Normal Probability Distribution

1. The goodness of fit test for a ____________ probability distribution is also based on the use of the _______________ distribution.

2. In particular, observed frequencies for several categories of sample data are compared to expected frequencies under the assumption that the _______________ has a normal probability distribution.

3. Because the normal probability distribution is _______________, we must modify the way the _______________ are defined and how the expected frequencies are computed.

4. Example (Table 12.10) Job applicant test data for Chemline, Inc.

TABLE 12.10 Chemline Employee Aptitude Test Scores for 50 Randomly Chosen Job Applicants

71	66	61	65	54	93
60	86	70	70	73	73
55	63	56	62	76	54
82	79	76	68	53	58
85	80	56	61	61	64
65	62	90	69	76	79
77	54	64	74	65	65
61	56	63	80	56	71
79	84				

(a) Chemline hires approximately 400 new employees annually for its four plants located throughout the United States. The personnel director asks whether a normal distribution applies for the population of test scores.

(b) If such a distribution can be used, the distribution would be helpful in evaluating specific test scores; that is, scores in the upper 20%, lower 40%, and so on, could be identified quickly.

(c) Hence, we want to test the null hypothesis that the population of test scores has a normal distribution.

5. Calculations:

$$\bar{x} = \underline{\hspace{5cm}}$$

$$s = \underline{\hspace{8cm}}$$

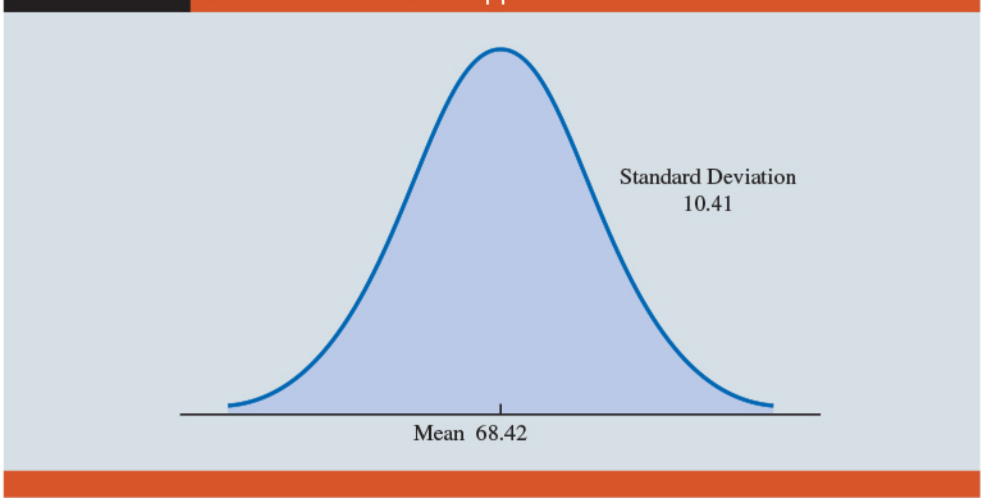
6. Hypotheses about the distribution of the job applicant test scores:

   (a) H_0: The population of test scores has a normal distribution with mean 68.42 and standard deviation 10.41

   (b) H_a: The population of test scores does not have a normal distribution with mean 68.42 and standard deviation 10.41

7. (Figure 12.3) The hypothesized normal distribution:

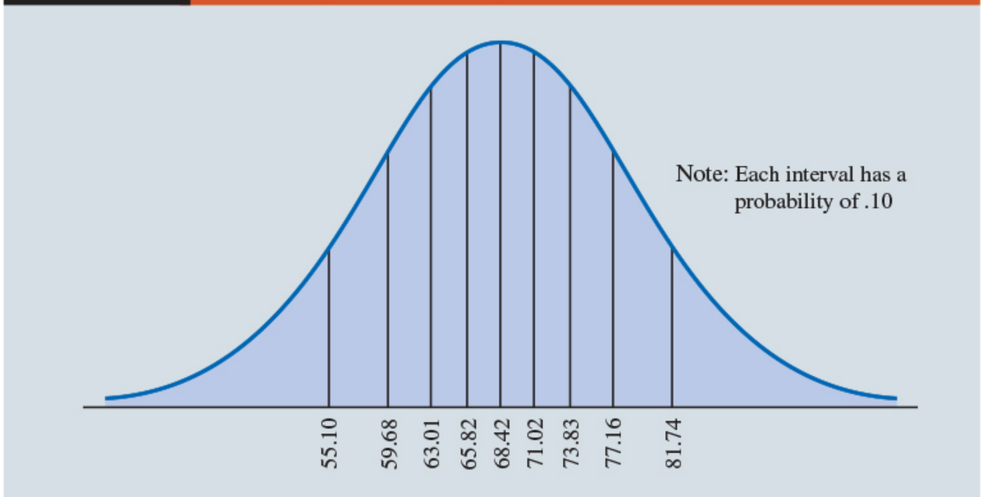
FIGURE 12.3 Hypothesized Normal Distribution of Test Scores for the Chemline Job Applicants



8. (Figure 12.4) Define the categories of test scores such that the expected frequencies will be ______________ for each category. With a sample size of 50, one way of establishing categories is to divide the normal probability distribution into

____

FIGURE 12.4 Normal Distribution for the Chemline Example with 10 Equal-Probability Intervals



9. With a sample size of 50, we would expect ________________ in each interval or category, and the rule of thumb for expected frequencies would be satisfied. Let us look more closely at the procedure for calculating the category boundaries.

   (a) With a continuous probability distribution, establish intervals such that each interval has an expected frequency of _______________.

   (b) First consider the test score cutting off the lowest ________ of the test scores. From the table for the standard normal distribution we find that the z value for this test score is __________. Therefore, the test score of ________________________________ provides this cutoff value for the lowest 10% of the scores.

   (c) For the lowest 20%, we find ______________, and thus ________________________________.

   (d) Working through the normal distribution in that way provides the following test score values:

Percentage	z	Test Score
10%	−1.28	68.42 − 1.28(10.41) = 55.10
20%	−.84	68.42 − .84(10.41) = 59.68
30%	−.52	68.42 − .52(10.41) = 63.01
40%	−.25	68.42 − .25(10.41) = 65.82
50%	.00	68.42 + 0(10.41) = 68.42
60%	+.25	68.42 + .25(10.41) = 71.02
70%	+.52	68.42 + .52(10.41) = 73.83
80%	+.84	68.42 + .84(10.41) = 77.16
90%	+1.28	68.42 + 1.28(10.41) = 81.74

10. (Table 12.11) With the categories or intervals of test scores now defined and with the known expected frequency of five per category, we can return to the sample data of Table 12.10 and determine the observed frequencies for the categories. Doing so provides the results in Table 12.11.

TABLE 12.11 Observed and Expected Frequencies for Chemline Job Applicant Test Scores

Test Score Interval	Observed Frequency f_i	Expected Frequency e_i
Less than 55.10	5	5
55.10 to 59.68	5	5
59.68 to 63.01	9	5
63.01 to 65.82	6	5
65.82 to 68.42	2	5
68.42 to 71.02	5	5
71.02 to 73.83	2	5
73.83 to 77.16	5	5
77.16 to 81.74	5	5
81.74 and over	6	5
Total	50	50

11. (Table 12.12) The value of the test statistic is $\chi^2 = 7.2$.

TABLE 12.12 Computation of the Chi-Square Test Statistic for the Chemline Job Applicant Example

Test Score Interval	Observed Frequency f_i	Expected Frequency e_i	Difference $f_i - e_i$	Squared Difference $(f_i - e_i)^2$	Squared Difference Divided by Expected Frequency $(f_i - e_i)^2/e_i$
Less than 55.10	5	5	0	0	.0
55.10 to 59.68	5	5	0	0	.0
59.68 to 63.01	9	5	4	16	3.2
63.01 to 65.82	6	5	1	1	.2
65.82 to 68.42	2	5	−3	9	1.8
68.42 to 71.02	5	5	0	0	.0
71.02 to 73.83	2	5	−3	9	1.8
73.83 to 77.16	5	5	0	0	.0
77.16 to 81.74	5	5	0	0	.0
81.74 and over	6	5	1	1	.2
Total	50	50			$\chi^2 = 7.2$

12. Using the rule for computing the number of degrees of freedom for the goodness of fit test, we have __________________ degrees of freedom based on $k = 10$ categories and $p = 2$ parameters (mean and standard deviation) estimated from the sample data.

13. Suppose that we test the null hypothesis that the distribution for the test scores is a normal distribution with a 0.10 level of significance.

14. To test this hypothesis, we need to determine the p-value for the test statistic $\chi^2 = 7.2$ by finding the area in the upper tail of a chi-square distribution with 7 degrees of freedom. (Table 12.4) we find that $\chi^2 = 7.2$ provides an area in the upper tail greater than 0.10. Thus, we know that the p-value is greater than 0.10. (Software: p-value $= 0.4084$).

15. With ______________, the hypothesis that the probability distribution for the Chemline job applicant test scores is a normal probability distribution cannot be rejected.

16. **Normal Probability Distribution Goodness of Fit Test**

    (1) State the null and alternative hypotheses.

H_0: The population has a __________ probability distribution.

H_a: The population does ____________________ probability distribution.

(2) Set a level of significance and select a random sample and

(a) Compute the sample mean and sample standard deviation.

(b) Define k intervals of values so that the expected frequency is at ____________ for each interval. Using ___________________________ is a good approach.

(c) Record the ____________ frequency of data values f_i in each interval defined.

(3) Compute the expected number of occurrences e_i for each interval of values. Multiply the ______________ by the ______________ of a normal random variable being in the interval.

(4) Compute the value of the test statistic.

____

(5) Rejection rule:

i. p-value approach: Reject H_0 if p-value $\leq \alpha$

ii. Critical value approach: Reject H_0 if $\chi^2 \geq \chi^2_{\alpha,k-p-1}$

where p is the number of ______________ of the distribution estimated by the sample.

(6) Draw decision and conclusion.

😍 EXERCISES

12.1 : 1, 2, 3, 7

12.2 : 10, 11, 14, 17

12.3 : 19, 23, 25

SUP : 29, 32, 33, 36

“很多時候我們缺的不是機會，而是決心與勇氣。”
“Often times we lack is not the opportunity, but courage and determination.”

— 心靈補手 (Good Will Hunting, 1997)

統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 13: Experimental Design and Analysis of Variance

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: http://www.hmwu.idv.tw

系級: ____________________ 學號: ____________________ 姓名: ____________________

Overview

1. The statistical studies can be classified as either ________________ or ________________.
2. In an experimental statistical study, an experiment is conducted to generate the data.
   (a) An experiment begins with identifying a ____________ of interest.
   (b) Then one or more other variables, thought to be ___________, are identified and ______________, and
   (c) data are collected about how those variables _____________ the variable of interest.
3. In an observational study, data are usually obtained through sample ___________ and not a controlled experiment.
4. Good design principles are still employed, but the ____________ controls associated with an experimental statistical study are often not possible.

5. Example In a study of the relationship between smoking and lung cancer the researcher cannot assign a smoking habit to subjects. The researcher is restricted to simply observing the effects of smoking on people who already smoke and the effects of not smoking on people who do not already smoke.

6. In this chapter we introduce three types of experimental designs: a ______________ ______________ design, a ____________________ design*, and a ____________ experiment*.

7. Analysis of variance (____________) can analyze the results of regression studies involving both experimental and observational data.

13.1 An Introduction to Experimental Design and Analysis of Variance

1. Example Chemitech Inc. developed a new filtration system for municipal water supplies.

   (a) The industrial engineering group is responsible for determining the best assembly method (method A, method B, and method C) for the new filtration system.

   (b) Managers at Chemitech want to determine which assembly method can produce the greatest number of filtration systems per week.

   (c) In the Chemitech experiment, ____________________ is the ________________ variable or __________.

   (d) Because three assembly methods correspond to this factor, we say that three ______________ are associated with this experiment; each treatment corresponds to one of the three assembly methods.

2. (**single-factor experiment**) The Chemitech problem is an example of a ________________ experiment; it involves one ______________ factor (method of assembly).

3. More complex experiments may consist of ___________ factors; some factors may be categorical and others may be quantitative.

4. (**populations**) The three assembly methods or treatments define the three ______________ of interest for the Chemitech experiment. One population is all Chemitech employees who use assembly method A, another is those who use method B, and the third is those who use method C.

5. (**objective**) Note that for each population the _____________ or ____________ variable is the ___________ of filtration systems assembled per week, and the primary statistical objective of the experiment is to determine whether the _________ __________________ produced per week is the same for all three populations (methods).

6. (**experimental units**) Suppose a random sample of three employees is selected from all assembly workers at the Chemitech production facility. In experimental design terminology, the three randomly selected ___________ are the experimental _________.

7. (**completely randomized design**) A ______________________________ requires that each of the three assembly methods or treatments be assigned randomly to one of the experimental units or workers.

   (a) For example, method A might be randomly assigned to the second worker, method B to the first worker, and method C to the third worker.

   (b) Note that this experiment would result in only one measurement or number of units assembled for each treatment.

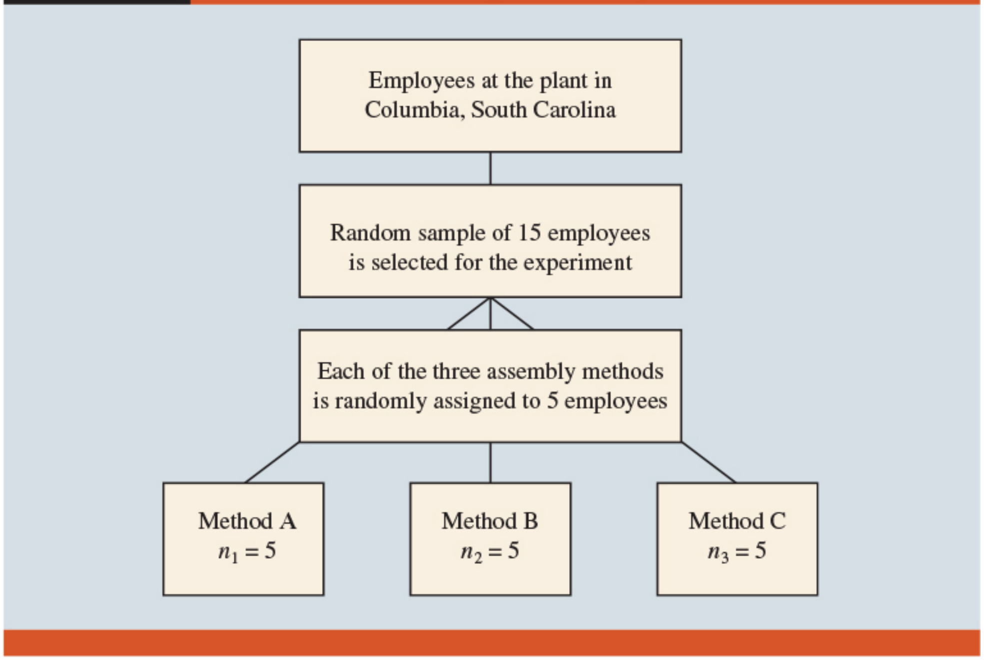
8. (**replicates**) To obtain additional data for each assembly method, we must __________ or ____________ the basic experimental process.

   (a) Suppose, for example, we selected 15 workers and then randomly as signed each of the three treatments to 5 of the workers.

   (b) Because each method of assembly is assigned to 5 workers, we say that ________________ have been obtained.

(Figure 13.1) the completely randomized design for the Chemitech experiment.

FIGURE 13.1 Completely Randomized Design for Evaluating the Chemitech Assembly Method Experiment



Data Collection

1. Once we are satisfied with the experimental design, we proceed by collecting and analyzing the data. In the Chemitech case, the employees would be instructed in how to perform the assembly method assigned to them and then would begin assembling the new filtration systems using that method.

2. (Table 13.1) After this assignment and training, the number of units assembled by each employee during one week is as shown in Table 13.1. The sample means, sample variances, and sample standard deviations for each assembly method are also provided. From these data, ____________ appears to result in higher production rates than either of the other methods.

TABLE 13.1 Number of Units Produced by 15 Workers

	Method		
	A	B	C
	58	58	48
	64	69	57
	55	71	59
	66	64	47
	67	68	49
Sample mean	62	66	52
Sample variance	27.5	26.5	31.0
Sample standard deviation	5.244	5.148	5.568

3. (**question**) is whether the three sample means observed are different enough for us to conclude that the means of the populations corresponding to the three methods of assembly are different.

4. Turn the question to Statistical terms: μ_1, μ_2, μ_3 = mean number of units produced per week using method A, B, C, respectively

5. Although we will never know the actual values of μ_1, μ_2, and μ_3, we want to use the sample means to test the following hypotheses.

$$H_0 : \mu_1 = \mu_2 = \mu_3, \quad H_a : \text{Not all population means are equal}$$

6. The ________________________ is the statistical procedure used to determine whether the observed differences in the three sample means are large enough to reject H_0.

Assumptions for Analysis of Variance

Three assumptions are required to use analysis of variance.

1. For each population, the response variable is ____________________.
   Implication: In the Chemitech experiment, the number of units produced per week (response variable) must be normally distributed for each assembly method.

2. The variance of the ____________________, is the same for all of the populations.

Implication: In the Chemitech experiment, the variance of the number of units produced per week must be the same for each assembly method.

3. The observations must be ______________.
   Implication: In the Chemitech experiment, the number of units produced per week for each employee must be independent of the number of units produced per week for any other employee.

Analysis of Variance: A Conceptual Overview

1. If the means for the three populations are equal, we would expect the three _______________ to be close together.

   (a) The more the sample means ________, the stronger the evidence we have for the conclusion that the population means ________.

   (b) If the ____________ among the sample means is __________ it supports ______; if the variability among the sample means is _________," it supports ______.

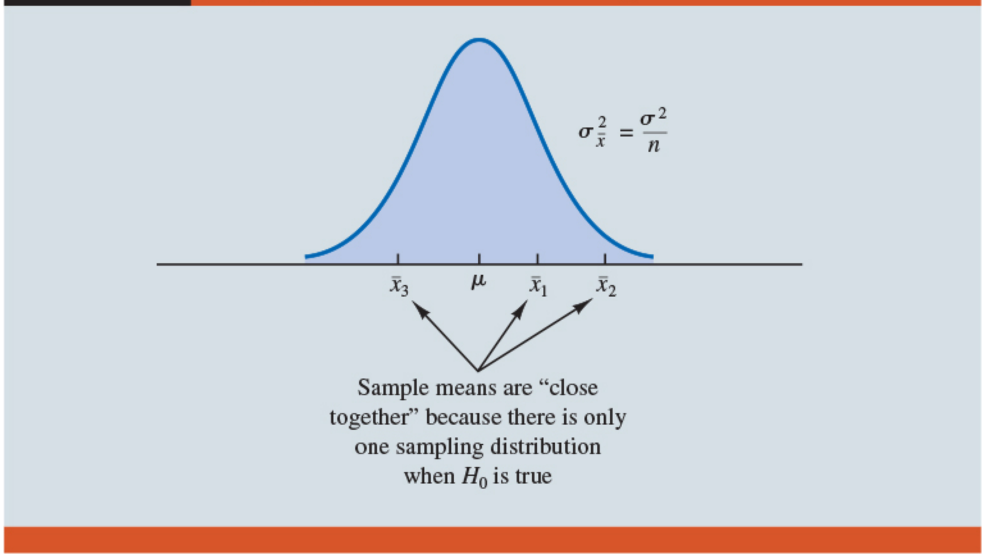
2. If the null hypothesis, $H_0 : \mu_1 = \mu_2 = \mu_3$, is true, we can use the __________________ the sample means to develop an estimate of ______.

   (a) If the assumptions for analysis of variance are satisfied and the null hypothesis is true, each sample will have come from the same __________ distribution with mean _____ and variance ______.

   (b) (Chapter 7) the sampling distribution of the sample mean $\bar{x}$ for a simple random sample of size n from a normal population will be normally distributed with mean _____ and variance ________. (__________________)

   (c) (Figure 13.2) if H_0 is true, we can think of each of the three sample means, $\bar{x}_1 = 62, \bar{x}_2 = 66$, and $\bar{x}_3 = 52$ from Table 13.1, as values drawn at random from the sampling distribution shown in Figure 13.2.

FIGURE 13.2 Sampling Distribution of $\bar{x}$ Given H_0 Is True



3. When the sample sizes are equal, as in the Chemitech experiment, the best estimate of the mean of the sampling distribution of $\bar{x}$ is the __________ or __________________ __________________. In the Chemitech experiment, an estimate of the mean of the sampling distribution of $\bar{x}$ is _________________________. We refer to this estimate as the _______________________.

4. An estimate of the variance of the sampling distribution of $\bar{x}$, _______, is provided by the variance of the three sample means.

$$s_{\bar{x}}^2 = \underline{\hspace{8cm}}$$

5. Because _______________, solving for σ^2 gives

$$\sigma^2 = \underline{\hspace{2cm}}$$

Hence,

$$\text{Estimate of } \sigma^2 = n\ (\text{Estimate of } \sigma_{\bar{x}}^2) = \underline{\hspace{4cm}}.$$

6. The result, $ns_{\bar{x}}^2 = 260$, is referred to as the _______________________ estimate of σ^2.

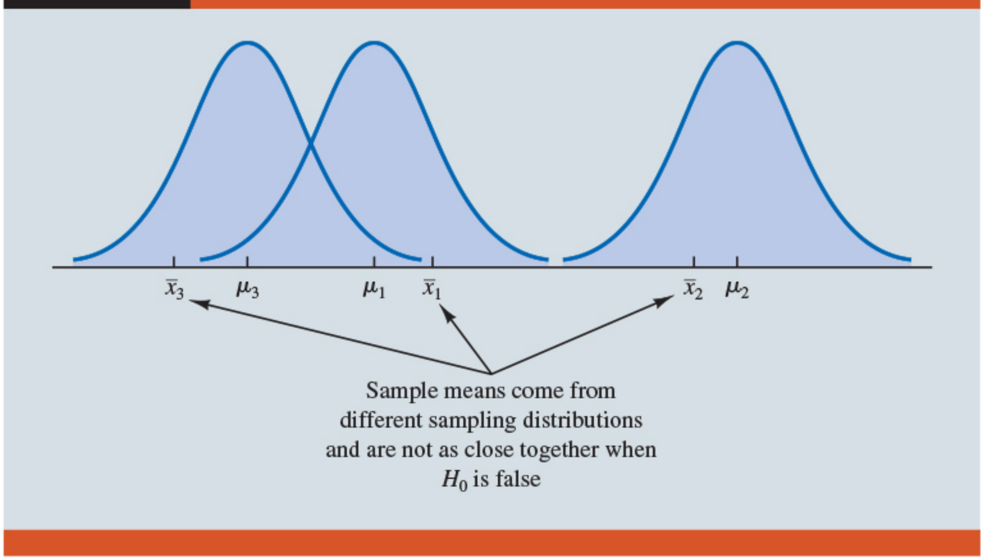
7. The between treatments estimate of σ^2 is based on the assumption that ______________. In this case, each sample comes from the __________ population, and there is only ________ sampling distribution of $\bar{x}$.

8. Illustrate what happens when H_0 is false, suppose the population means all ________.

   (a) Note that because the three samples are from ________ populations with different means, they will result in three ________ sampling distributions.

   (b) (Figure 13.3) The sample means are not as close together as they were when H_0 was true. Thus, $s_{\bar{x}}^2$ will be larger, causing the between treatments estimate of σ^2 to be ________.

FIGURE 13.3 Sampling Distributions of $\bar{x}$ Given H_0 Is False



   (c) In general, when the population means are not equal, the between treatments estimate will ________ the population variance σ^2.

   (d) When a simple random sample is selected from each population, each of the sample variances provides an ________ estimate of σ^2. Hence, we can ________ or ________ the individual estimates of σ^2 into one overall estimate.

   (e) The estimate of σ^2 obtained in this way is called the ________ or ________ estimate of σ^2.

   (f) Because each sample variance provides an estimate of σ^2 based only on the variation within each sample, the within treatments estimate of σ^2 is not affected by whether the population means are equal.

   (g) When the sample sizes are equal, the within treatments estimate of σ^2 can be obtained by computing the ________ of the individual sample variances.

9. Example For the Chemitech experiment we obtain

   Within treatments estimate of σ^2 = ______________________________

   (a) The between treatments estimate of σ^2 (260) is much _________ than the within treatments estimate of σ^2 (28.33).

   (b) The ________ of these two estimates is $260/28.33 = 9.18$.

10. If the null hypothesis is ________,

    (a) The between treatments approach provides a ________ estimate of σ^2.

    (b) The two estimates will be similar and their ratio will be close to _____.

11. If the null hypothesis is ________,

    (a) The between treatments approach________________ σ^2

    (b) the between treatments estimate will be larger than the within treatments estimate, and their ratio will be ________.

12. In the next section we will show how large this ratio must be to reject H_0.

13. **Summary**: The logic behind ANOVA is based on the development of two independent estimates of the common population variance ______.

    (a) One estimate of σ^2 is based on the variability __________ the sample means themselves.

    (b) The other estimate of σ^2 is based on the variability of the data __________ each sample.

    (c) By comparing these two estimates of σ^2, we will be able to determine whether the population means are equal.

補充說明:

ANOVA Table

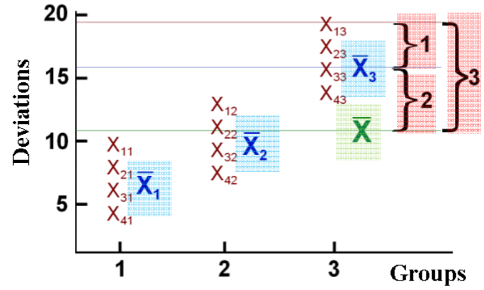
Groups

1	2	... j ...			k
X_{11}	X_{12}	$\cdots$	X_{1j}	$\cdots$	X_{1k}
X_{21}	X_{22}	$\cdots$	X_{2j}	$\cdots$	X_{2k}
			$\cdots$		
X_{i1}	X_{i2}	$\cdots$	X_{ij}	$\cdots$	X_{ik}
$\vdots$					$X_{n_k k}$
	$X_{n_2 2}$	$\cdots$	$\vdots$	$\cdots$	
$X_{n_1 1}$			$X_{n_j j}$		

$$T_j = \sum_{i=1}^{n_j} X_{ij} \qquad \bar{X}_j = \frac{T_j}{n_j}$$

$$T = \sum_{j=1}^{k} T_j \qquad \bar{X} = \frac{T}{N}$$

$$S^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \frac{(X_{ij}-\bar{X})^2}{N-1}$$



$$(X_{ij}-\bar{X}) = (X_{ij}-\bar{X}_j) + (\bar{X}_j-\bar{X})$$

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$X_{ij} = \mu_j + \epsilon_{ij} \qquad i = 1,\cdots,n_j \quad j = 1,\cdots,k$$

$$\epsilon_{ij} \sim N(0,\sigma^2)$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}[(X_{ij}-\bar{X}_j)+(\bar{X}_j-\bar{X})]^2$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{X}_j-\bar{X})^2$$

ANOVA Table

Source	SS	df	MS	F	p
Between	SS_B	$k-1$	MS_B	MS_B/MS_W	< 0.05
Within	SS_W	$N-k$	MS_W		
Total	SS_T	$N-1$			

$$SS_{Total} = SS_{Within} + SS_{Between}$$

$$F = \frac{MS_{Between}}{MS_{Within}}$$

Reject H_0, if $F_{obs} > F_{\{\alpha,k-1,N-k\}}$

13.2 Analysis of Variance and the Completely Randomized Design

1. How analysis of variance can be used to test for the equality of k population means for a ______________ randomized design.

2. The general form of the hypotheses tested is

$$H_0 : \underline{\qquad\qquad\qquad\qquad}, \quad H_a : \text{Not all population means are equal}$$

where μ_j is mean of the jth population.

3. We assume that a simple random sample of size ______ has been selected from each of the k ______________ or ______________.

4. For the resulting sample data, let

   (a) x_{ij}: value of observation i for treatment j, $i = 1, 2, \cdots, n_j$, $j = 1, 2, \cdots, k$
   (b) n_j: number of observations for treatment j.
   (c) $\bar{x}_j$: sample mean for treatment j, ______________.
   (d) s_j^2: sample variance for treatment j, ______________
   (e) s_j: sample standard deviation for treatment j

5. The overall sample mean, denoted ______, is the sum of all the observations divided by the total number of observations:

$$\bar{\bar{x}} = \underline{\qquad\qquad\qquad} \qquad (13.3)$$

where $n_T = n_1 + n_2 + \cdots + n_k \qquad (13.4)$.

6. If the size of each sample is n, $n_T = kn$; the overall sample mean is just the ___________ of the k sample means.

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{kn} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}/n}{k} = \underline{\qquad\qquad} \qquad (13.5)$$

7. The overall sample mean can also be computed as a ________________ of the k sample means.

$$\bar{\bar{x}} = \underline{\qquad\qquad\qquad\qquad\qquad}$$

8. Example Each sample in the Chemitech experiment consists of $n = 5$ observations (Table 13.1), we obtained the following result:

$$\bar{\bar{x}} = \frac{62 + 66 + 52}{3} = 60$$

If the null hypothesis is true ($\mu_1 = \mu_2 = \mu_3 = \mu$), the overall sample mean of 60 is the ________ estimate of the population mean μ.

Between-Treatments Estimate of Population Variance

1. A between treatments estimate of σ^2 when the sample sizes were equal.

   (a) This estimate of σ^2 is called the ______________ due to ______________ and is denoted __________:

   $$MSTR = \frac{}{\rule{5cm}{0.4pt}} \qquad (13.6)$$

   (b) The numerator in equation (13.6) is called the ________________ due to treatments and is denoted _________.

   (c) The denominator, $k-1$, represents the degrees of freedom associated with SSTR.

   (d) **Mean Square Due to Treatments**

   $$MSTR = \frac{SSTR}{k-1} \qquad (13.7)$$

   where

   $$SSTR = \qquad\qquad (13.8)$$

   ____

   (e) If H_0 is true, MSTR provides an ____________ estimate of σ^2. However, if the means of the k populations are not equal, MSTR is not an unbiased estimate of σ^2; in fact, in that case, MSTR should ______________ σ^2.

   (f) If each sample consists of n observations, equation (13.6) can be written as

   $$MSTR = \frac{}{\rule{5cm}{0.4pt}} = \frac{}{\rule{5.5cm}{0.4pt}} = \rule{2cm}{0.4pt}$$

2. Example For the Chemitech data in Table 13.1, we obtain the following results:

$$SSTR = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 5(62-60)^2 + 5(66-60)^2 + 5(52-60)^2 = \_\_\_\_\_\_$$
$$MSTR = \frac{SSTR}{k-1} = \frac{520}{2} = \_\_\_\_\_\_$$

Within-Treatments Estimate of Population Variance

1. A within treatments estimate of σ^2 when the sample sizes were equal.

   (a) This estimate of σ^2 is called the ____________ due to ________ and is denoted ________:

   $$MSE = \frac{\qquad}{\qquad} \qquad (13.9)$$

   (b) The numerator in equation (13.9) is called the ______________ due to error and is denoted ________.

   (c) The denominator of MSE is referred to as the degrees of freedom associated with SSE.

   (d) **Mean Square Due to Error**

   $$MSE = \frac{\qquad}{\qquad} \qquad (13.10)$$
   $$\text{where } SSE = \frac{\qquad}{\qquad} \qquad (13.11)$$

   (e) Note that MSE is based on the variation within each of the treatments; it is not influenced by whether the null hypothesis is true. Thus, MSE ________ provides an __________ estimate of σ^2.

   (f) If each sample has n observations, $n_T = kn$; thus, ______________, and equation (13.9) can be rewritten as

   $$MSE = \frac{\sum_{j=1}^{k}(n-1)s_j^2}{k(n-1)} = \_\_\_\_\_\_$$

   (g) If the sample sizes are the same, MSE is the average of the ________________.

   (h) Note that it is the same result we used in Section 13.1 when we introduced the concept of the within-treatments estimate of σ^2.

2. Example For the Chemitech data in Table 13.1 we obtain the following results.

$$SSE = \sum_{j=1}^{k}(n_j-1)s_j^2 = (5-1)27.5+(5-1)26.5+(5-1)31 = \_\_\_\_\_\_\_$$

$$MSE = \frac{SSE}{n_T-k} = \frac{340}{15-3} = \frac{340}{12} = \_\_\_\_\_\_\_$$

Comparing the Variance Estimates: The F Test

1. If the null hypothesis is ________, MSTR and MSE provide two independent, unbiased estimates of σ^2.
2. (Chapter 11) For ___________ populations, the sampling distribution of the ratio of two independent estimates of σ^2 follows an ______ distribution.
3. Hence, if the null hypothesis is true and the ANOVA assumptions are valid, the sampling distribution of _________________ is an ______ distribution with numerator degrees of freedom equal to ________ and denominator degrees of freedom equal to _________.
   (a) If the null hypothesis is true, the value of $MSTR/MSE$ should appear to have been selected from this ______ distribution.
   (b) If the null hypothesis is false, the value of MSTR/MSE will be ___________ because MSTR overestimates σ^2.
4. Hence, we will reject H_0 if the resulting value of $MSTR/MSE$ appears to be ____________ to have been selected from an F distribution with $k-1$ numerator degrees of freedom and n_T-k denominator degrees of freedom.
5. **Test Statistic for the Equality of K Population Means**

   ________________ (13.12)

6. The test statistic follows an F distribution with $k-1$ degrees of freedom in the numerator and n_T-k degrees of freedom in the denominator. (_________________)
7. Example Let us return to the Chemitech experiment and use a level of significance ____________ to conduct the hypothesis test.

(a) The value of the test statistic is

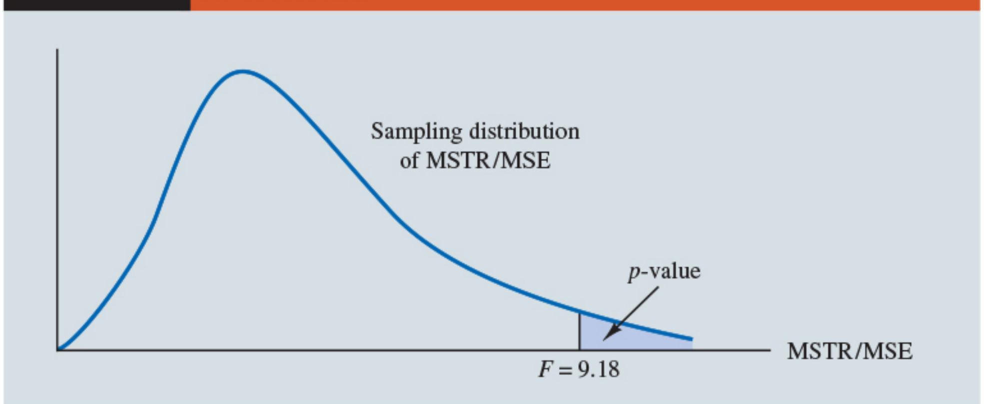
$$F = \frac{MSTR}{MSE} = \underline{\hspace{3cm}}$$

(b) The numerator degrees of freedom is $k-1 = 3-1 = 2$ and the denominator degrees of freedom is $n_T-k = 15-3 = 12$.

(c) Because we will only reject the null hypothesis for large values of the test statistic, the p-value is the ________________ of the F distribution to the right of the test statistic $F = 9.18$.

(d) (Figure 13.4) shows the sampling distribution of $F = MSTR/MSE$, the value of the test statistic, and the upper tail area that is the p-value for the hypothesis test.

FIGURE 13.4 Computation of p-Value Using the Sampling Distribution of MSTR/MSE



(e) (Table 4 of Appendix B) the upper tail of an F distribution with 2 numerator degrees of freedom and 12 denominator degrees of freedom.

Area in Upper Tail	0.10	0.05	0.025	0.01
F Value ($df_1 = 2$, $df_2 = 12$)	2.81	3.89	5.10	6.93

(f) (**the p-value approach**) Because ________________, the area in the upper tail at $F = 9.18$ is less than 0.01. Thus, the p-value is less than 0.01 (Software: p-value $= 0.004$.)

(g) With ________________, H_0 is rejected.

(h) (**conclusion**) The test provides sufficient evidence to conclude that the means of the three populations are not equal. In other words, analysis of variance supports the conclusion that the population mean number of units produced per week for the three assembly methods are not equal.

(i) (**the critical value approach**) With ____________________________, and conclude that the means of the three populations are not equal.

8. **Test for the Equality of K Population Means**

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a: \quad \text{Not all population means are equal}$$

Test Statistic

$$F = \frac{MSTR}{MSE}$$

Rejection Rule

$$p\text{-value approach} \; : \; \text{Reject } H_0 \text{ if } p\text{-value} \leq \alpha$$
$$\text{Critical value approach} \; : \; \text{Reject } H_0 \text{ if } F \geq F_{\alpha.k-1,n_T-k}$$

ANOVA Table

1. (Table 13.2) The results of the preceding calculations can be displayed conveniently in a table referred to as the analysis of variance or ____________ table. The general form of the ANOVA table for a completely randomized design is:

TABLE 13.2 ANOVA Table for a Completely Randomized Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
Treatments	SSTR	$k-1$	$MSTR = \frac{SSTR}{k-1}$	$\frac{MSTR}{MSE}$	
Error	SSE	n_T-k	$MSE = \frac{SSE}{n_T-k}$		
Total	SST	n_T-1			

2. (Table 13.3) JMP/Excel output

TABLE 13.3 Analysis of Variance Table for the Chemitech Experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*	*p*-value
Treatments	520	2	260.00	9.18	.004
Error	340	12	28.33		
Total	860	14			

3. Total sum of squares (SST):

   (a) The sum of squares associated with the source of variation referred to as ____________ is called the total sum of squares (__________).

   (b) $SST =$ __________________, and that the degrees of freedom associated with this _________ sum of squares is the sum of the degrees of freedom associated with the sum of squares due to _______________ and the sum of squares due to _________. (_____________________________.)

   (c) We point out that SST divided by its degrees of freedom $n_T - 1$ is the __________________________ that would be obtained if we treated the entire set of 15 observations as one data set.

   (d) With the entire data set as one sample, the formula for computing the total sum of squares, SST, is

   $$SST = \qquad\qquad\qquad\qquad (13.13)$$

4. ANOVA can be viewed as the process of partitioning the total sum of squares and the degrees of freedom into their corresponding sources: _________________________.

Computer Results for Analysis of Variance

1. (Figure 13.5) JMP/Excel output for the Chemitech experiment:

FIGURE 13.5 Output for the Chemitech Experiment Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	2	520.0	260.00	9.18	.004
Error	12	340.0	28.33		
Total	14	860.0			

Model Summary

S	R-sq	R-sq (adj)
5.32291	60.47%	53.88%

Means

Factor	N	Mean	StDev	95% CI
Method A	5	62.00	5.24	(56.81, 67.19)
Method B	5	66.00	5.15	(60.81, 71.19)
Method C	5	52.00	5.57	(46.81, 57.19)

Pooled StDev = 5.32291

2. The square root of MSE provides the best estimate of the population standard deviation σ. This estimate of σ in Figure 13.5 is Pooled StDev; it is equal to 5.323.

3. A 95% confidence interval estimate of the population mean for Method A.

   (13.15)

   where s is the estimate of the population standard deviation σ. Because the best estimate of σ is provided by the Pooled StDev, we use a value of 5.323 for σ in expression (13.15).

4. The degrees of freedom for the t value is 12, the degrees of freedom associated with the error sum of squares. Hence, with $t_{0.025} = 2.179$ we obtain

$$62 \pm 2.179 \frac{5.323}{\sqrt{5}} = 62 \pm 5.19$$

   Thus, the individual 95% confidence interval for Method A goes from $62 - 5.19 = 56.81$ to $62 + 5.19 = 67.19$.

Testing for the Equality of k Population Means: An Observational Study

1. ANOVA can also be used to test for the equality of three or more population means using data obtained from an ____________________.

2. Example (Table 13.4) National Computer Products, Inc. (NCP) manufactures printers and fax machines at plants located in Atlanta, Dallas, and Seattle. To measure how much employees at these plants know about quality management, a random sample of 6 employees was selected from each plant and the employees selected were given a quality awareness examination. The examination scores for these 18 employees are shown in Table 13.4. Managers want to use these data to test the hypothesis that the mean examination score is the same for all three plants.

TABLE 13.4 Examination Scores for 18 Employees

	Plant 1 Atlanta	Plant 2 Dallas	Plant 3 Seattle
	85	71	59
	75	75	64
	82	73	62
	76	74	69
	71	69	75
	85	82	67
Sample mean	79	74	66
Sample variance	34	20	32
Sample standard deviation	5.83	4.47	5.66

3. Define population 1 as all employees at the Atlanta plant, population 2 as all employees at the Dallas plant, and population 3 as all employees at the Seattle plant. Let ______ mean examination score for population $j, j = 1, 2, 3$

4. Want to use the sample results to test the following hypotheses:

$$H_0: \quad \mu_1 = \mu_2 = \mu_3$$
$$H_a: \quad \text{Not all population means are equal}$$

5. Note that the hypothesis test for the NCP observational study is ____________________ as the hypothesis test for the Chemitech experiment.

6. Even though the same ANOVA methodology is used for the analysis, it is worth noting how the NCP observational statistical study differs from the Chemitech experimental statistical study.

7. The individuals who conducted the NCP study had ______________ over how the plants were assigned to individual employees. That is, the plants were already in operation and a particular employee worked at one of the three plants. All that NCP could do was to select a random sample of 6 employees from each plant and administer the quality awareness examination.

8. To be classified as an ______________________, NCP would have had to be able to randomly select 18 employees and then assign the plants to each employee in a random fashion.

13.3 Multiple Comparison Procedures

1. When we use analysis of variance to test whether the means of k populations are equal, ____________ of the null hypothesis allows us to conclude only that the population means are not all equal.

2. In some cases we will want to go a step further and determine where the differences among means occur.

3. To show how _______________________ procedures can be used to conduct statistical comparisons between pairs of population means.

Fisher's LSD

1. Suppose that analysis of variance provides statistical evidence to __________ the null hypothesis of equal population means. Fisher's __________________________________ procedure can be used to determine where the differences occur.

2. Example (Section 13.1) The Chemitech experiment. Using analysis of variance, we concluded that the mean number of units produced per week are not the same for the three assembly methods. In this case, the followup question is: We believe the assembly methods differ, but where do the differences occur?

3. **Fisher's LSD Procedure**

$$H_0 : \underline{\hspace{3cm}}, \quad H_a : \mu_i \neq \mu_j$$

Test Statistic

$$t = \qquad (13.16)$$

Rejection Rule ____________________

- p-value approach: RejectH_0 if p-value $\leq \alpha$.
- Critical value approach: Reject H_0 if ____________ or ____________.

where the value of $t_{\alpha/2}$ is based on a t distribution with $n_T - k$ degrees of freedom.

✎ Question .... (p616)

For the Chemitech experiment, apply Fisher's LSD Procedure to determine whether there is a significant difference between the means of population 1 (Method A) and population 2 (Method B) at the $\alpha = 0.05$ level of significance.

sol:

Area in Upper Tail	0.20	0.10	0.05	0.025	0.01	0.005
t Value (12 df)	0.873	1.356	1.782	2.179	2.681	3.055

4. Many practitioners find it easier to determine how large the difference between the sample means must be to reject H_0. In this case the test statistic is __________, and the test is conducted by the following procedure.

5. **Fisher's LSD Procedure Based on the Test Statistic $\bar{x}_i - \bar{x}_j$**

$$H_0 : \mu_i = \mu_j, \quad H_a : \mu_i \neq \mu_j$$

Test Statistic

____

Rejection Rule at a Level of Significance α Reject H_0 if $|\bar{x}_i - \bar{x}_j| \geq LSD$ where

$$LSD = \underline{\hspace{6cm}} \qquad (13.17)$$

6. **Confidence Interval Estimate of the Difference Between Two Population Means Using Fisher's LSD Procedure**

$$\underline{\hspace{5cm}} \qquad (13.18)$$

where

$$LSD = \underline{\hspace{6cm}} \qquad (13.19)$$

and $t_{\alpha/2}$ is based on a t distribution with $n_T - k$ degrees of freedom.

7. If the confidence interval in expression (13.18) includes the value ________, we cannot reject the hypothesis that the two population means are equal.

8. However, if the confidence interval does not include the value zero, we conclude that there is a difference between the population means.

✎ Question .... (p617)

For the Chemitech experiment, apply Fisher's LSD Procedure based on the Test Statistic $\bar{x}_i - \bar{x}_j$ to determine whether there is a significant difference (a) between the means of population 1 (Method A) and population 3 (Method C), (b) between the means of population 2 (Method B) and population 3 (Method C) at the $\alpha = 0.05$ level of significance. Find a 95% confidence interval estimate of the difference between the means of populations 1 and 2 and make a conclusion.

sol:

Type I Error Rates

1. ANOVA gave us statistical evidence to reject or not reject the null hypothesis of equal population means.

2. We showed how Fisher's LSD procedure can be used in such cases to determine where the differences occur. Technically, it is referred to as a ______________ or ______________ LSD test because it is employed only if we first find a significant F value by using analysis of variance.

3. To see why this distinction is important in multiple comparison tests, we need to explain the difference between a ____________________ Type I error rate and an ____________________ Type I error rate.

4. In the Chemitech experiment we used Fisher's LSD procedure to make three pairwise comparisons.

Test 1	Test 2	Test 3
$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 = \mu_3$	$H_0 : \mu_2 = \mu_3$
$H_a : \mu_1 \pm \mu_2$	$H_a : \mu_1 \pm \mu_3$	$H_a : \mu_2 \neq \mu_3$

5. In each case, we used a level of significance of $\alpha = 0.05$.

6. Therefore, ______________, if the null hypothesis is true, the probability that we will make a Type I error is $\alpha = 0.05$; hence, the probability that we will not make a Type I error on each test is ______________.

7. In discussing multiple comparison procedures we refer to this probability of a Type I error ($\alpha = 0.05$) as the ______________________________; It indicate the level of significance associated with a ________ pairwise comparison.

8. What is the probability that in making three pairwise comparisons, we will commit a Type I error on ____________ of the three tests?

   (a) To answer this question, note that the probability that we will not make a Type I error on any of the three tests is ________________________.

   (b) The probability of making at least one Type I error is __________________.

   (c) Thus, when we use Fisher's LSD procedure to make all three pairwise comparisons, the Type I error rate associated with this approach is not 0.05, but actually 0.1426; we refer to this error rate as the _________ or ________________ Type I error rate.

9. To avoid confusion, we denote the experimentwise Type I error rate as _______.

10. The experimentwise Type I error rate gets larger for problems with more populations. For example, a problem with five populations has 10 possible pairwise comparisons. If we tested all possible pairwise comparisons by using Fisher's LSD with a comparisonwise error rate of $\alpha = 0.05$, the experimentwise Type I error rate would be ____________________.

11. In such cases, practitioners look to alternatives that provide better control over the experimentwise error rate.

12. One alternative for controlling the overall experimentwise error rate, referred to as the ____________________, involves using a smaller comparisonwise error rate for each test.

13. For example, if we want to test C pairwise comparisons and want the maximum probability of making a Type I error for the overall experiment to be α_{EW}, we simply use a comparisonwise error rate equal to _________.

14. In the Chemitech experiment, if we want to use Fisher's LSD procedure to test all three pairwise comparisons with a maximum experimentwise error rate of ____________, we set the comparisonwise error rate to be ________________.

15. (Recall Chapter 9) For a fixed sample size, any decrease in the probability of making a Type I error will result in an increase in the probability of making a ________ error, which corresponds to accepting the hypothesis that the two population means are equal when in fact they are not equal.

16. As a result, many practitioners are reluctant to perform individual tests with a low comparisonwise Type I error rate because of the increased risk of making a Type II error.

17. Several other procedures, such as ________________ and ________________ __________, have been developed to help in such situations. However, there is considerable controversy in the statistical community as to which procedure is "best." The truth is that no one procedure is best for all types of problems.

13.4 Randomized Block Design*

13.5 Factorial Experiment*

😍 EXERCISES

13.2 : 1, 4, 7, 8, 10

13.3 : 13, 15, 18, 19

SUP : 35, 37

“會讓人後悔的從來都不是失敗，而是當機會出現時你沒有全力以赴。”
“Regrets don’t come from failure, they come from moments you failed to give your best.”

— *墊底辣妹 (Flying Colors, 2015)*

統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 14: Simple Linear Regression

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: http://www.hmwu.idv.tw

系級: ______ 學號: ______ 姓名: ______

Overview

1. Managerial decisions often are based on the relationship between two or more variables.

2. Examples

   (a) After considering the relationship between advertising expenditures and sales, a marketing manager might attempt to ______ sales for a given level of advertising expenditures.

   (b) A public utility might use the relationship between the daily high temperature and the demand for electricity to ______ electricity usage on the basis of next month's anticipated daily high temperatures.

3. Regression analysis can be used to develop ______ showing how the variables are related.

4. In regression terminology, the variable being predicted is called the ______ variable (denoted by ______). The variable or variables being used to predict the value of the dependent variable are called the ______ variables (denoted by ______).

5. **Simple linear regression**: the simplest type of regression analysis involving ________ independent variable and ________ dependent variable in which the relationship between the variables is approximated by a ________________.

6. Regression analysis involving two or more ________________ variables is called ____________ regression analysis.

7. Multiple regression and cases involving ______________ relationships are covered in Chapters 15 and 16.

14.1 Simple Linear Regression Model

1. Example Armand's Pizza Parlors

   (a) Armand's Pizza Parlors is a chain of Italian-food restaurants located in a five-state area. Armand's most successful locations are near college campuses.

   (b) The managers believe that __________________ for these restaurants (denoted by ______) are related positively to the _____________________ population (denoted by ______);

   (c) Restaurants near campuses with a large student population tend to generate more sales than those located near campuses with a small student population.

2. Using regression analysis, we can develop an equation showing how the dependent variable y is related to the independent variable x.

Regression Model and Regression Equation

1. (**population**) In the Armand's Pizza Parlors example, the population consists of all the Armand's restaurants. For every restaurant in the population, there is a value of ______ (student population) and a corresponding value of ______ (quarterly sales).

2. (**regression model**) The ____________ that describes how y is related to x and an ______________ is called the regression model.

3. **Simple Linear Regression Model**

$$\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \qquad (14.1)$$

β_0 and β_1 are referred to as the _______________ of the model, and ϵ (the Greek letter epsilon) is a ____________________ referred to as the ______________.

4. The error term accounts for the __________________ that cannot be explained by the _____________________ between x and y.

5. The population of all Armand's restaurants can also be viewed as a collection of ___________________, one for each distinct value of ______.

   (a) For example, one subpopulation consists of all Armand's restaurants located near college campuses with _________________; another subpopulation consists of all Armand's restaurants located near college campuses with _________________; and so on.

   (b) Each subpopulation has a corresponding __________________________. Thus, a distribution of y values is associated with restaurants located near campuses with 8000 students; a distribution of y values is associated with restaurants located near campuses with 9000 students; and so on.

6. (**regression equation**) Each distribution of y values has its own __________ or __________________. The equation that describes how the expected value of y, denoted $E(y)$, is related to x is called the ______________________.

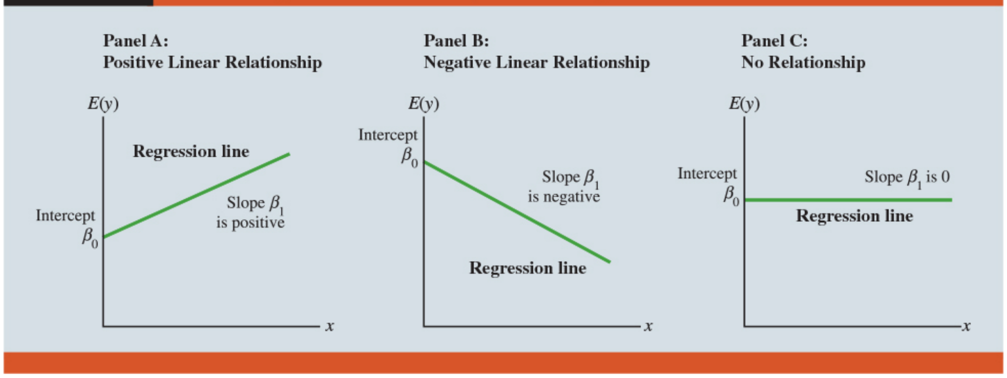
7. **Simple Linear Regression Equation**

$$\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \qquad (14.2)$$

The graph of the simple linear regression equation is a straight line; β_0 is the _______________ of the regression line, β_1 is the __________, and $E(y)$ is the mean or expected value of y for a given value of x.

8. (Figure 14.1) Possible regression lines

FIGURE 14.1 Possible Regression Lines in Simple Linear Regression



Estimated Regression Equation

1. If the values of the population parameters ______ and ______ were known, we could use equation (14.2) to compute the mean value of y for a given value of x.

2. In practice, the parameter values are not known and must be estimated using ______________. Sample statistics (denoted ______ and ______) are computed as estimates of the population parameters β_0 and β_1. Substituting the values of the sample statistics b_0 and b_1 for β_0 and β_1 in the regression equation, we obtain ______________________________.

3. **Estimated Simple Linear Regression Equation**

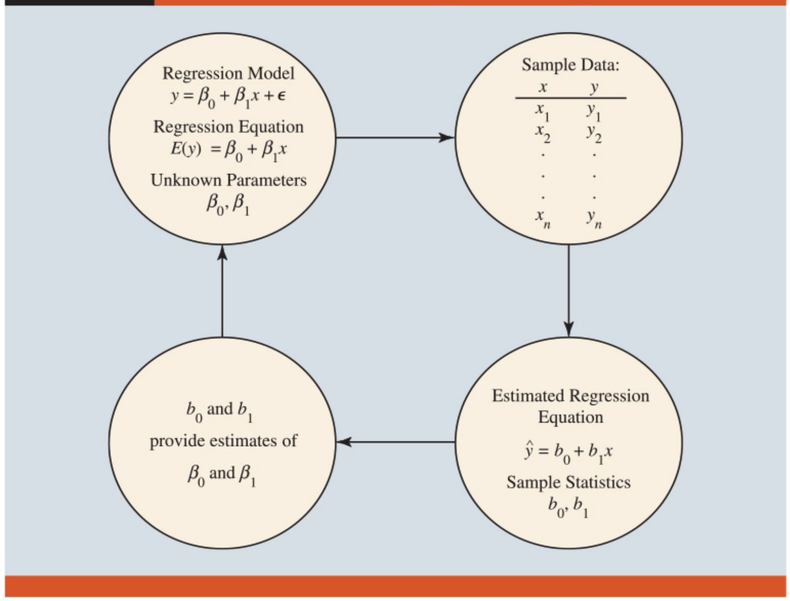
$$\underline{\qquad\qquad\qquad\qquad} \qquad (14.3)$$

4. (**the estimated regression line**) The graph of the estimated simple linear regression equation is called the estimated regression line; b_0 is the y-intercept and b_1 is the slope.

5. In general, _____ is the point estimator of $E(y)$, the mean value of y for a given value of x.

6. (Figure 14.2) A summary of the estimation process for simple linear regression.

FIGURE 14.2 The Estimation Process in Simple Linear Regression



7. Example Armand's Pizza Parlors

   (a) To estimate the mean or expected value of quarterly sales for all restaurants located near campuses with 10,000 students, Armand's would substitute the value of 10,000 for x in equation (14.3).

   (b) In some cases, however, Armand's may be more interested in predicting sales for one particular restaurant.

   (c) For example, suppose Armand's would like to predict quarterly sales for the restaurant they are considering building near Talbot College, a school with 10,000 students. As it turns out, the best predictor of y for a given value of x is also provided by ______.

   (d) Thus, to predict quarterly sales for the restaurant located near Talbot College, Armand's would also substitute the value of 10,000 for x in equation (14.3).

8. The value of $\hat{y}$ provides both a ________________ of $E(y)$ for a given value of x and a ______________ of an individual value of y for a given value of x.

Notes + Comments

1. Regression analysis cannot be interpreted as a procedure for establishing a ___________________ relationship between variables. It can only indicate how or to what extent variables

are ______________ with each other.

2. Any conclusions about cause and effect must be based upon the ______________ of those individuals most knowledgeable about the application.

3. The regression equation in simple linear regression is $E(y) = \beta_0 + \beta_1 x$. More advanced texts in regression analysis often write the regression equation as

____

to emphasize that the regression equation provides the mean value of y for a given value of x.

14.2 Least Squares Method

1. The ________________________ is a procedure for using sample data to find the estimated regression equation.

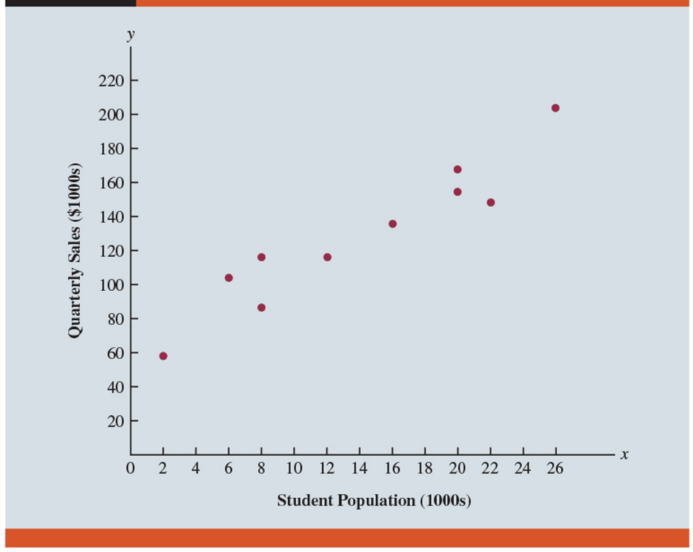
2. (Table 14.1) Example Armand's Pizza Parlor
Suppose data were collected from a sample of 10 Armand's Pizza Parlor restaurants located near college campuses. For the ith observation or restaurant in the sample, x_i is the size of the student population (in thousands) and y_i is the quarterly sales (in thousands of dollars).

TABLE 14.1 Student Population and Quarterly Sales Data for 10 Armand's Pizza Parlors

Restaurant i	Student Population (1000s) x_i	Quarterly Sales ($1000s) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

3. (Figure 14.3) Scatter diagrams for regression analysis are constructed with the independent variable x (student population) on the horizontal axis and the dependent variable y (quarterly sales) on the vertical axis.

FIGURE 14.3 Scatter Diagram of Student Population and Quarterly Sales for Armand's Pizza Parlors



   (a) The ________________ enables us to observe the data graphically and to draw preliminary conclusions about the possible relationship between the variables.

   (b) Quarterly sales appear to be higher at campuses with larger student populations.

   (c) In addition, for these data the relationship between the size of the student population and quarterly sales appears to be approximated by a ________________.

   (d) A ________________ relationship is indicated between x and y.

   (e) We therefore choose the ________________________ model to represent the relationship between quarterly sales and student population.

   (f) Next task is to use the sample data in Table 14.1 to determine the values of b_0 and b_1 in the estimated simple linear regression equation.

4. For the ith restaurant, the estimated regression equation provides

$$\underline{\hspace{3cm}} \qquad (14.4)$$

where

- $\hat{y}_i$: predicted value of quarterly sales (\$1000s) for the ith restaurant
- b_0: the ______________ of the estimated regression line
- b_1: the _________ of the estimated regression line
- x_i: size of the student population (1000s) for the ith restaurant

5. In simple linear regression, each observation __________ consists of two values: one for the independent variable and one for the dependent variable.

6. Every restaurant in the sample will have an observed value of sales y_i and a predicted value of sales $\hat{y}_i$.

7. For the estimated regression line to provide a good fit to the data, we want the differences between the observed sales values and the predicted sales values ______________.

8. (**the least squares method**) The least squares method uses the sample data to provide the values of b_0 and b_1 that ____________ the ____________________ of the _____________ between the observed values of the dependent variable y_i and the predicted values of the dependent variable $\hat{y}_i$.

9. **Least Squares Criterion**

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (14.5)$$

where

- y_i: _________________ of the dependent variable for the ith observation
- $\hat{y}_i$: __________________ of the dependent variable for the ith observation

10. **Slope and Y-Intercept for the Estimated Regression Equation** Differential calculus can be used to show (see Appendix 14.1) that the values of b_0 and b_1 that minimize expression (14.5) can be found by:

$$b_1 = \frac{\qquad\qquad}{\qquad\qquad} = \frac{\qquad\qquad}{\qquad\qquad} \qquad (14.6)$$

$$b_0 = \underline{\qquad\qquad} \qquad (14.7)$$

where

- x_i: value of the independent variable for the ith observation
- y_i: value of the dependent variable for the ith observation
- $\bar{x}$: mean value for the independent variable
- $\bar{y}$: mean value for the dependent variable
- n: total number of observations

補充說明：

✎ Question .... (p660)

Using data in Table 14.2 to calculate the slope and intercept of the estimated regression equation for Armand's Pizza Parlors example.

sol:

$$
\begin{aligned}
\bar{x} &= \frac{\sum x_i}{n} = \frac{140}{10} = 14, \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{1300}{10} = 130 \\
b_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{2840}{568} = 5 \\
b_0 &= \bar{y} - b_1\bar{x} = 130 - 5(14) = 60
\end{aligned}
$$

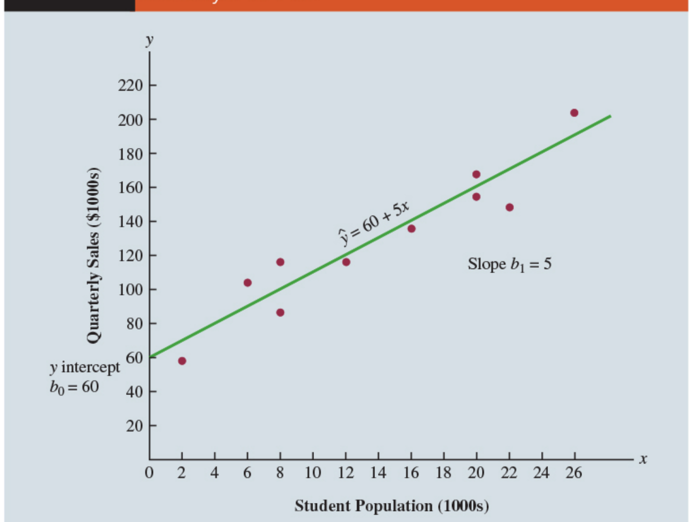
Thus, the estimated regression equation is ______________.

TABLE 14.2 Calculations for the Least Squares Estimated Regression Equation for Armand's Pizza Parlors

Restaurant i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

11. (Figure 14.4) The graph of this equation on the scatter diagram.

FIGURE 14.4 Graph of the Estimated Regression Equation for Armand's Pizza Parlors: $\hat{y} = 60 + 5x$



(a) The slope of the estimated regression equation ________________, implying that as student population increases, sales increase.

(b) We can conclude (based on sales measured in \$1000s and student population in 1000s) that an __________ in the student population of 1000 is associated with an __________ of \$5000 in ________________ sales; that is, quarterly sales are expected to increase by \$5 per student.

12. If we wanted to predict quarterly sales for a restaurant to be located near a campus with 16,000 students, we would compute

$$\hat{y} = \underline{\hspace{4cm}}$$

Hence, we would predict quarterly sales of \$140,000 for this restaurant.

13. This least squares criterion is used to choose the equation that provides the ________.

14. If some other criterion were used, such as minimizing the sum of the ________________ between y_i and $\hat{y}_i$, a different equation would be obtained. In practice, the least squares method is the ______________.

14.3 Coefficient of Determination

1. How well does the estimated regression equation fit the data? The ______________________ provides a measure of the ____________ for the estimated regression equation.

2. (**residual**) For the ith observation, the difference between the observed value of the dependent variable, ____, and the predicted value of the dependent variable, ____, is called the ith residual. The ith residual represents the error in using $\hat{y}_i$ to estimate y_i. Thus, for the ith observation, the residual is ______.

3. (**Sum of Squares Due to Error)** The sum of squares of these residuals or errors is the quantity that is minimized by the least squares method. This quantity, also known as the sum of squares due to error, is denoted by ______:

$$SSE = \underline{\hspace{3cm}} \qquad (14.8)$$

4. (Table 14.3) The value of SSE is a measure of the error in using the estimated regression equation to predict the values of the dependent variable in the sample. $SSE = 1530$ measures the error in using the estimated regression equation $\hat{y} = 60 + 5x$ to predict sales.

TABLE 14.3 Calculation of SSE for Armand's Pizza Parlors

Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales ($1000s)	Predicted Sales $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Squared Error $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
				SSE =	1530

5. Now suppose we are asked to develop an estimate of quarterly sales ____________ knowledge of the size of the student population. Without knowledge of any related variables, we would use the ________________ as an estimate of quarterly sales at any given restaurant.

6. (Table 14.4) (**Total Sum of Squares**) We show the sum of squared deviations obtained by using the ________________________ to predict the value of quarterly sales for each restaurant in the sample. For the ith restaurant in the sample, the difference _________ provides a measure of the error involved in using $\bar{y}$ to predict sales. The corresponding sum of squares, called the total sum of squares, is denoted _________.

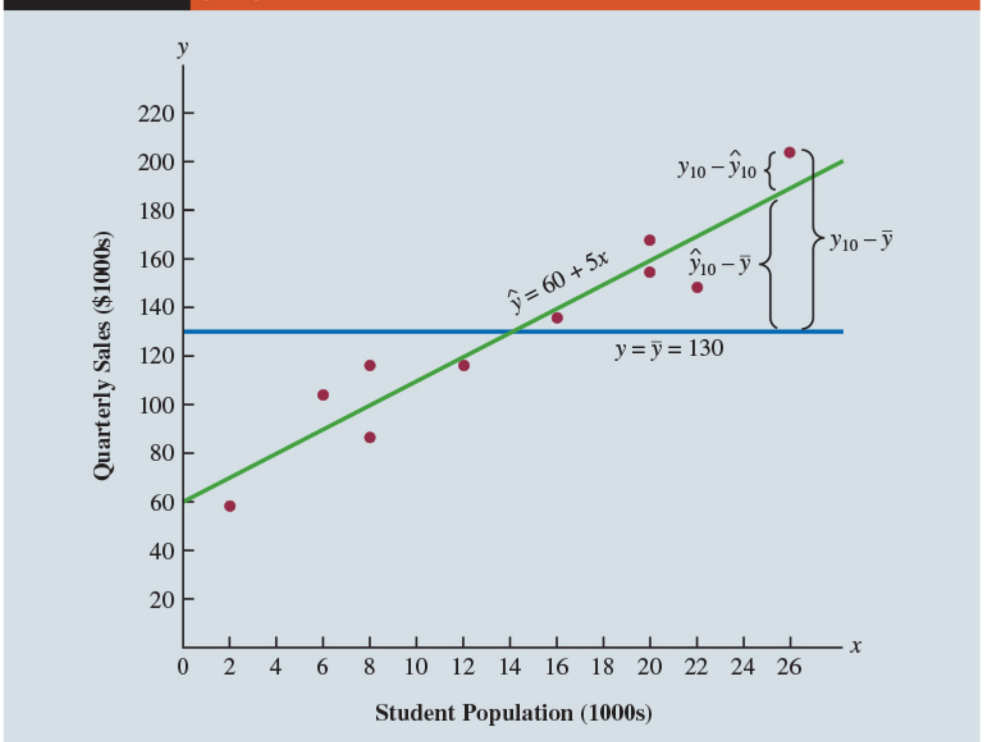
$$SST = \underline{\hspace{3cm}} \qquad (14.9)$$

TABLE 14.4 Computation of the Total Sum of Squares for Armand's Pizza Parlors

Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales ($1000s)	Deviation $y_i - \bar{y}$	Squared Deviation $(y_i - \bar{y})^2$
1	2	58	−72	5184
2	6	105	−25	625
3	8	88	−42	1764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1521
9	22	149	19	361
10	26	202	72	5184
			SST =	15,730

7. (Figure 14.5)

FIGURE 14.5 Deviations About the Estimated Regression Line and the Line $y = \bar{y}$ for Armand's Pizza Parlors



8. We can think of ________ as a measure of how well the observations cluster about

the _________ and SSE as a measure of how well the observations cluster about the _________.

9. (**Sum of Squares Due to Regression**) the sum of squares due to regression, is denoted _________, measures how much the $\hat{y}$ values on the estimated regression line deviate from $\bar{y}$:

$$SSR = \_\_\_\_\_\_\_\_\_\_\_\_\_\_ \qquad (14.10)$$

10. (**Relationship Among SST, SSR, and SSE**) From the preceding discussion, we should expect that SST, SSR, and SSE are related.

$$\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \qquad (14.11)$$

where

- SST: total sum of squares
- SSR: sum of squares due to regression
- SSE: sum of squares due to error

11. SSR can be thought of as the _____________ portion of SST, and SSE can be thought of as the _______________ portion of SST.

12. Example Armand's Pizza Parlors example
we already know that $SSE = 1530$ and $SST = 15,730$; therefore, solving for SSR in equation (14.11), we find that the sum of squares due to regression is

$$SSR = \_\_\_\_\_\_\_\_\_\_\_\_\_\_ = 15,730 - 1530 = 14,200$$

13. How the three sums of squares, SST, SSR, and SSE, can be used to provide a measure of the goodness of fit for the estimated regression equation?

    (a) The estimated regression equation would provide a perfect fit if every value of the dependent variable y_i happened to lie on the estimated regression line.

    (b) In this case, _________ would be zero for each observation, resulting in _____________.

    (c) Because $SST = SSR + SSE$, we see that for a perfect fit SSR must equal SST, and the ratio (________) must equal one.

(d) Poorer fits will result in larger values for SSE. Hence the poorest fit occurs when ____________ and ______________.

14. (**Coefficient of Determination**) The ratio SSR/SST, which will take values between zero and one, is used to evaluate the goodness of fit for the estimated regression equation. This ratio is called the coefficient of determination and is denoted by (______) (Other textbook: ______).

$$r^2 = \underline{\qquad\qquad} \qquad (14.12)$$

15. When we express the coefficient of determination as a percentage, r^2 can be interpreted as the ____________ of the total sum of squares that can be explained by using ______________________________.

16. Example Armand's Pizza Parlors example

(a) The value of the coefficient of determination is

$$r^2 = \frac{SSR}{SST} = \frac{14,200}{15,730} = 0.9027$$

(b) For Armand's Pizza Parlors, we can conclude that 90.27% of the total sum of squares can be explained by using the estimated regression equation $\hat{y} = 60 + 5x$ to predict qua

(c) In other words, ______________________________ can be explained by the linear relationship between the size of the student population and sales. We should be pleased to find such a good fit for the estimated regression equation.

Correlation Coefficient

1. In Chapter 3 we introduced the correlation coefficient as a descriptive measure of the strength of linear association between two variables, x and y. Values of the correlation coefficient are always between ____________.

2. A value of $+1$ indicates that the two variables x and y are ________________ in a __________ linear sense. A value of -1 indicates that x and y are perfectly related in a __________ linear sense, with all data points on a straight line that has a negative slope. Values of the correlation coefficient close to zero indicate that x and y are __________________.

3. (**Sample Correlation Coefficient**) If a regression analysis has already been performed and the coefficient of determination r^2 computed, the sample correlation coefficient can be computed:

$$r_{xy} = \underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad} = \underline{\qquad\qquad\qquad\qquad} \tag{14.13}$$

where b_1 is the slope of the estimated regression equation $\hat{y} = b_0 + b_1 x$

補充說明 : Show that the coefficient of determination of a simple linear regression is the square of the sample correlation coefficient of $(x_1, y_1), \cdots, (x_n, y_n)$.

4. Example Armand's Pizza Parlor example

   the value of the coefficient of determination corresponding to the estimated regression equation $\hat{y} = 60 + 5x$ is 0.9027. Because the slope of the estimated regression equation is positive, equation (14.13) shows that the sample correlation coefficient is $+\sqrt{0.9027} = +0.9501$. (a strong positive linear association exists between x and y.)

5. In the case of a ____________________ between two variables, both the coefficient of determination and the sample correlation coefficient provide measures of the strength of the relationship. The coefficient of determination provides a measure between zero and one, whereas the sample correlation coefficient provides a measure between -1 and $+1$.

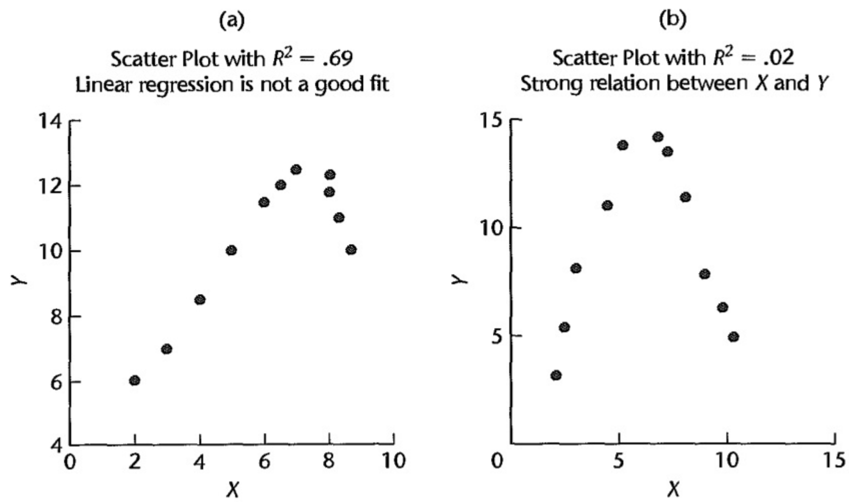
6. Although the sample correlation coefficient is restricted to a linear relationship between two variables, the coefficient of determination can be used for __________ relationships and for relationships that have ____________________________.
Thus, the coefficient of determination provides a wider range of applicability.

(補充) limitations of R^2: three common misunderstandings

Source : Michael H. Kutner et al. (2019), Applied Linear Statistical Models: Applied Linear Regression Models, Mcgraw-Hill Inc., (5th edition)

1. **Misanderstanding 1**: A high R^2 indicates that ________________ can be made. (not necessarily correct)
   (a) (Toluca Company Example) the coefficient of determination was high ($R^2 = 0.82$). Yet the 90 percent prediction interval for the next lot, consisting of 100 units, was wide (332 to 507 hours) and not precise enough to permit management to schedule workers effectively.
   (b) Misunderstanding 1 arises because R^2 measures only a ________________ from SST and provides no information about absolute precision for estimating a mean response or predicting a new observation.
2. **Misunderstanding 2**: A high R^2 indicates that the estimated regression line is a __________. (not necessarily correct)
   (a) (Figure 2.9a) a scatter plot where R^2 is high ($R^2 = 0.69$). Yet a linear regression function would not be a good fit since the regression relation is curvilinear.
3. **Misunderstanding 3**: A R^2 near zero indicates that X and Y are not related. (not necessarily correct).
   (a) (Figure 2.9b) a scatter plot where R^2 between X and Y is $R^2 = 0.02$. Yet X and Y are strongly related; however, the relationship between the two variables is curvilinear.
   (b) Misunderstandings 2 and 3 arise because R^2 measures the degree of ________________ between X and Y, whereas the actual regression relation may be curvilinear.

FIGURE 2.9
Illustrations of Two Misunderstandings about Coefficient of Determination.



14.4 Model Assumptions

1. In conducting a regression analysis, we begin by making an assumption about the appropriate model for the relationship between the dependent and independent variable(s).

2. For the case of simple linear regression, the assumed regression model is

   ____

3. Then the least squares method is used to develop values for b_0 and b_1, the estimates of the model parameters β_0 and β_1, respectively. The resulting estimated regression equation is

   ____

   Even with a large value of r^2, the estimated regression equation should not be used until further analysis of the appropriateness of the assumed model has been conducted.

4. An important step in determining whether the assumed model is appropriate involves ______________________ of the relationship. The tests of significance in regression analysis are based on the following assumptions about the error term ϵ.

5. **Assumptions About The Error Term ϵ in the Regression Model**

$$y = \beta_0 + \beta_1 x + \epsilon$$

   (a) The error term ϵ is a random variable with a mean or expected value of zero; that is, __________.

   Implication: β_0 and β_1 are constants, thus, for a given value of x, the expected value of y is

   _______________ (14.14)

   As we indicated previously, equation (14.14) is referred to as the regression equation.

   (b) The variance of ϵ, denoted by _____, is the same for all values of x.

   Implication: The variance of y about the regression line equals σ^2 and is the same for _____________.

   (c) The values of ϵ are ____________.

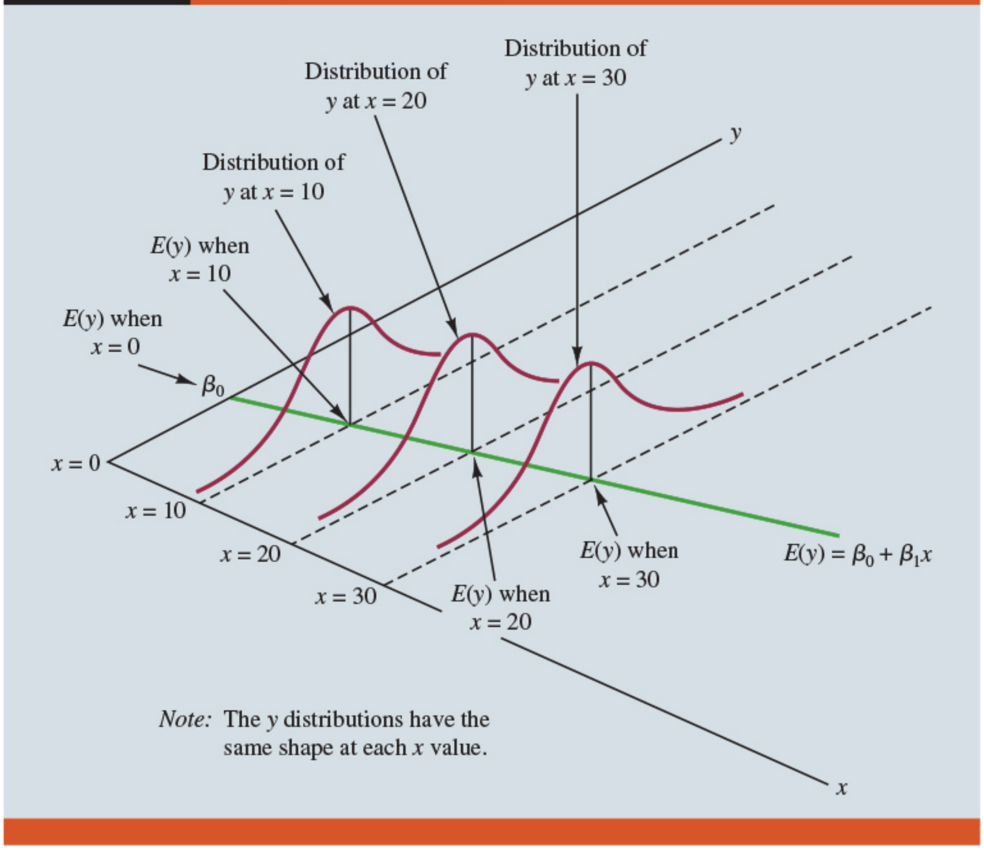
   Implication: The value of ϵ for a particular value of x is not related to the value of ϵ for any other value of x; thus, the value of y for a particular value of x is not related to the value of y for any other value of x.

   (d) The error term ϵ is a __________________ r.v. for all values of x.

   Implication: Because y is a linear function of ϵ, y is also a normally distributed random variable for all values of x.

6. Figure 14.6 illustrates the model assumptions and their implications; note that in this graphical interpretation, the value of _______ changes according to the specific value of x considered. However, regardless of the x value, the probability distribution of ϵ and hence the probability distributions of y are __________ distributed, each with the _____________.

FIGURE 14.6 Assumptions for the Regression Model



7. The specific value of the error ϵ at any particular point depends on whether the actual value of ______ is greater than or less than ________.

8. We assume that a straight line represented by ___________ is the basis for the relationship between the variables.

14.5 Testing for Significance

1. In a simple linear regression equation, the mean or expected value of y is a linear function of x: $E(y) = \beta_0 + \beta_1 x$. If the value of ____________, the mean value of y does not depend on the value of x and hence we would conclude that x and y are ____________________.

2. To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of ____________.

3. Two tests are commonly used. Both require an estimate of ______, the variance of ϵ in the regression model.

Estimate of σ^2

1. From the regression model and its assumptions we can conclude that σ^2, the variance of ϵ, also represents the variance of the y values about the regression line.

2. Thus, ________, the sum of squared residuals, is a measure of the variability of the actual observations about the estimated regression line.

$$SSE = \underline{\hspace{4cm}} = \underline{\hspace{6cm}}$$

3. Statisticians have shown that SSE has ________ degrees of freedom because two parameters (β_0 and β_1) must be estimated to compute SSE.

4. The __________________________ provides the estimate of σ^2; it is SSE divided by its degrees of freedom.

5. **Mean Square Error (Estimate of σ^2)**

$$s^2 = MSE = \underline{\hspace{3cm}} \tag{14.15}$$

6. **Standard Error of the Estimate**

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} \tag{14.16}$$

7. Example Armand's Pizza Parlors example

$$s^2 = MSE = \frac{1530}{8} = 191.25$$

provides an unbiased estimate of σ^2.

$$s = \sqrt{MSE} = \sqrt{191.25} = 13.829.$$

t Test

1. The purpose of the t test is to see whether we can conclude that $\beta_1 \neq 0$. We will use the sample data to test the following hypotheses about the parameter β_1.

   ____

2. If H_0 is rejected, we will conclude that $\beta_1 \neq 0$ and that a ________________________ relationship exists between the two variables.

3. If H_0 cannot be rejected, we will have ______________________ to conclude that a significant relationship exists.

4. The properties of the _______________________ of β_1, the least squares estimator of b_1, provide the basis for the hypothesis test.

5. **Sampling Distribution of b_1**
   - Expected Value: ______________
   - Standard Deviation:

     ____

(証明):

   - Distribution Form: ___________ (14.17)

6. Because we do not know the value of σ, we develop an estimate of σ_{b_1}, denoted s_{b_1}, by estimating σ with s in equation (14.17). Thus, we obtain the following estimate of σ_{b_1}.

7. **Estimated Standard Deviation of b_1**

$$s_{b_1} = \qquad (14.18)$$

____

8. The standard deviation of b_1 is also referred to as the standard error of b_1. Thus, s_{b_1} provides an estimate of the standard error of b_1.

9. The t test for a significant relationship is based on the fact that the test statistic

   ______

   follows a ____ distribution with ______ degrees of freedom. If the null hypothesis is true, then ________ and __________.

10. t **Test for Significance in Simple Linear Regression**

    (a) Hypothesis:

    $$H_0 : \beta_1 = 0, \qquad H_a : \beta_1 \neq 0$$

    (b) Test Statistic: ________ (14.19)

    (c) Rejection Rule:

       i. p-value approach: Reject H_0 if p-value $\leq \alpha$

       ii. Critical value approach: Reject H_0 if __________ or if _________.

    where $t_{\alpha/2}$ is based on a t distribution with $n-2$ degrees of freedom.

Question .... (p678)

Conduct t test of significance for Armand's Pizza Parlors at the $\alpha = 0.01$ level of significance.

sol:

1. Hypothesis: ________________________.
2. Level of significance: __________.
3. Test statistic (under H_0): ____________________.
4. Decision rule

   (a) Reject H_0 if ____________, or

(b) Reject H_0 if ____________________. (Table 2 of Appendix D, upper tail of the t distribution)

5. Decision:

(a) ____.

(b) __________________________.

6. Conclusion: __________________________________.

____
__________________________.

Confidence Interval for β_1

1. The form of a confidence interval for β_1 is as follows:

____

(証明:)

2. The point estimator is _____ and the margin of error is _______.

3. Develop a 99% confidence interval estimate of b_1 for Armand's Pizza Parlors. From Table 2 of Appendix B we find $t_{0.005,8} = 3.355$. Thus, the 99% confidence interval estimate of b_1 is

$$b_1 \pm t_{\alpha/2,n-2} s_{b_1} = \underline{\hspace{4cm}} = 5 \pm 1.95$$

or 3.05 to 6.95.

4. At the $\alpha = 0.01$ level of significance, we can use the 99% confidence interval as an __________ for drawing the hypothesis testing conclusion for the Armand's data.

5. Because 0, the hypothesized value of b_1, is ____________ in the confidence interval (3.05 to 6.95), we can __________ and conclude that a significant statistical relationship exists between the size of the student population and quarterly sales.

6. In general, a confidence interval can be used to test any ________________ about β_1. If the hypothesized value of β_1 is __________ in the confidence interval, do not reject H_0. Otherwise, reject H_0.

F Test

1. Recall: ________________, ________________, ________________.

2. An F test, based on the F probability distribution, can also be used to test for significance in regression. With only __________________, the F test will provide the same conclusion as the t test.

3. But with more than one independent variable, only the F test can be used to test for an ______________ relationship.

4. If the null hypothesis $H_0 : \beta_1 = 0$ is true, the mean square due to regression (__________________), and is denoted _______. In general,

$$MSR = \underline{\hspace{6cm}}$$

5. The regression degrees of freedom is always equal to the __________________ variables in the model. Because we consider only regression models with one independent variable in this chapter, we have __________________.

(証明:)

6. If the null hypothesis ($H_0 : \beta_1 = 0$) is true, ________ and ________ are two independent estimates of σ^2 and the sampling distribution of ____________ follows an F distribution with numerator degrees of freedom equal to one and denominator degrees of freedom equal to $n-2$. Therefore, when $\beta_1 = 0$, the value of MSR/MSE should be close to ______.

7. If the null hypothesis is false ($\beta_1 \neq 0$), MSR will ____________ σ^2 and the value of MSR/MSE will be _________; thus, large values of MSR/MSE lead to the rejection of H_0 and the conclusion that the relationship between x and y is statistically significant.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}, \qquad MSR = \frac{\sum(\hat{y}_i - \bar{y})^2}{1} = \frac{SSR}{1}.$$

$$E(MSE) = \_\_\_\_\_\_, \qquad E(MSR) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ .$$

(証明:)

8. If H_0 is false, MSE still provides an unbiased estimate of σ^2 and MSR overestimates σ^2. If H_0 is true, both MSE and MSR provide unbiased estimates of σ^2; in this case the value of MSR/MSE should be close to 1.

9. F **Test for Significance in Simple Linear Regression**

   (a) Hypothesis: $H_0 : \beta_1 = 0, \qquad H_a : \beta_1 \neq 0$

   (b) Test Statistic: $F = \frac{MSR}{MSE}$ (14.21)

   (c) Rejection Rule:

      i. p-value approach: Reject H_0 if p-value $\leq \alpha$

      ii. Critical value approach: Reject H_0 if ____________

   where F_α is based on an F distribution with 1 degree of freedom in the numerator and $n-2$ degrees of freedom in the denominator.

10. Decision and Conclusion.

✎ Question .... (p680)

Conduct the F test for the Armand's Pizza Parlors example. ($\alpha = 0.01$)

sol:

10. A similar ANOVA table can be used to summarize the results of the F test for significance in regression.

11. (Table 14.5)

TABLE 14.5 General Form of the Anova Table for Simple Linear Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
Regression	SSR	1	$\text{MSR} = \frac{\text{SSR}}{1}$	$F = \frac{\text{MSR}}{\text{MSE}}$	
Error	SSE	$n - 2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$		
Total	SST	$n - 1$			

12. (Table 14.6) ANOVA table with the F test computations performed for Armand's Pizza Parlors.

TABLE 14.6 Anova Table for the Armand's Pizza Parlors Problem

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*	*p*-value
Regression	14,200	1	$\frac{14,200}{1} = 14,200$	$\frac{14,200}{191.25} = 74.25$	.000
Error	1530	8	$\frac{1530}{8} = 191.25$		
Total	15,730	9			

Some Cautions About the Interpretation of Significance Tests

1. Rejecting the null hypothesis $H_0 : \beta_1 = 0$ and concluding that the relationship between x and y is significant does not enable us to conclude that a ________________ relationship is present between x and y.

2. Concluding a cause-and-effect relationship is warranted only if the analyst can provide some type of ______________________ that the relationship is in fact __________.

3. In the Armand's Pizza Parlors example, we can conclude that there is a significant relationship between the size of the student population x and quarterly sales y; moreover, the estimated regression equation $\hat{y} = 60 + 5x$ provides the least squares estimate of the relationship. We cannot, however, conclude that ____________ student population x _______________ in quarterly sales y just because we identified a statistically significant relationship.

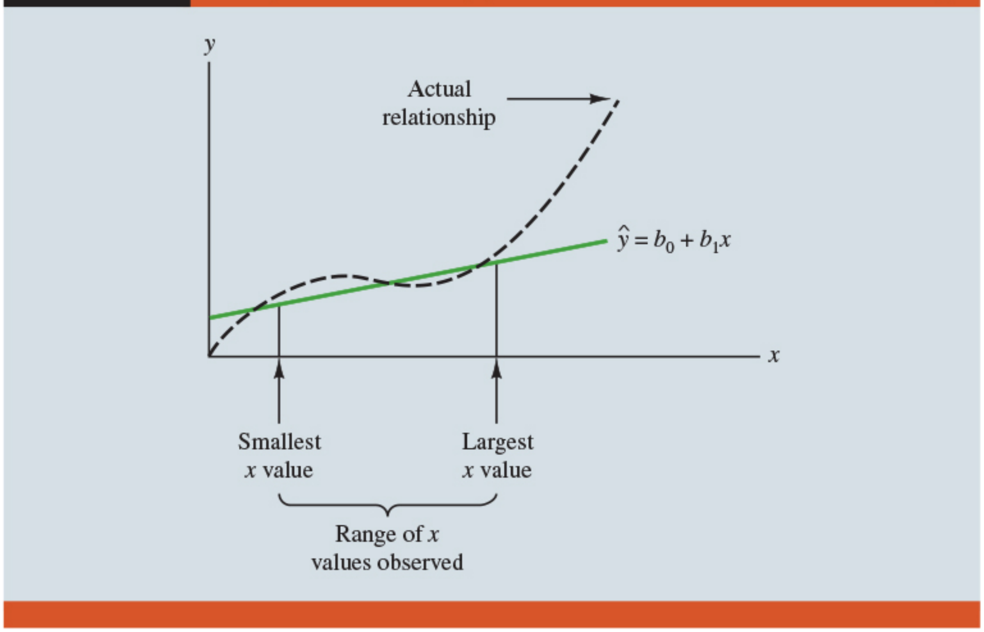
4. Armand's managers felt that increases in the student population were a _____________ of increased quarterly sales. Thus, the result of the significance test enabled them to conclude that a cause-and-effect relationship was present.

5. We can state only that x and y are related and that a linear relationship explains a significant portion of the variability in y over the range of values for x observed in the sample.

6. (Figure 14.7) illustrates this situation. The test for significance calls for the rejection of the null hypothesis $H_0 : \beta_1 = 0$ and leads to the conclusion that x and y are

significantly related, but the figure shows that the actual relationship between x and y is not linear.

FIGURE 14.7 Example of a Linear Approximation of a Nonlinear Relationship



7. Although the linear approximation provided by $\hat{y} = b_0 + b_1 x$ is good over the range of x values observed in the sample, it becomes poor for x values ________________.

8. Given a significant relationship, we should feel confident in using the estimated regression equation for predictions corresponding to x values ________________ of the x values observed in the sample.

9. For Armand's Pizza Parlors, this range corresponds to values of x ________________. Unless other reasons indicate that the model is valid beyond this range, predictions outside the range of the independent variable should be made ________________.

14.6 Using the Estimated Regression Equation for Estimation and Prediction

1. When using the simple linear regression model, we are making an ____________ about the relationship between x and y. We then use the ____________ method to obtain the estimated simple linear regression equation.

2. If a ____________ relationship exists between x and y and the ____________ ____________ shows that the fit is good, the estimated regression equation should be useful for estimation and prediction.

3. Example Armand's Pizza Parlors example

   (a) The estimated regression equation is $\hat{y} = 60 + 5x$. $\hat{y}$ can be used as a point estimator of ________, the mean or expected value of y for a given value of x, and as a predictor of an individual value of _____.

   (b) For example, a point estimate of the mean quarterly sales for all restaurant locations near campuses with $x = 10$ (10,000 students) students is

   $$\hat{y} = \underline{\hspace{4cm}}(\$110,000).$$

   In this case we are using $\hat{y}$ as the ____________ of the mean value of y when $x = 10$.

   (c) For example, to predict quarterly sales for a new restaurant Armand's is considering building near Talbot College, a campus with 10,000 students, we would compute

   $$\hat{y} = \underline{\hspace{4cm}}.$$

   Hence, we would predict quarterly sales of \$110,000 for such a new restaurant. In this case, we are using $\hat{y}$ as the ____________ of y for a new observation when $x = 10$.

4. Notations:

   (a) ______ = the given value of the independent variable x

   (b) ______= the random variable denoting the possible values of the dependent variable y when $x = x^*$

(c) ________ = the mean or expected value of the dependent variable y when $x = x^*$

(d) ______________ = the point estimator of $E(y^*)$ and the predictor of an individual value of y^* when $x = x^*$

5. Example Armand's Pizza Parlors example

   (a) To illustrate the use of this notation, suppose we want to estimate the mean value of quarterly sales for all Armand's restaurants located near a campus with 10,000 students.

   (b) For this case, _________ and ________ denotes the unknown mean value of quarterly sales for all restaurants where $x^* = 10$.

   (c) Thus, the point estimate of $E(y^*)$ is provided by ___________________, or \$110,000.

   (d) But, using this notation, $\hat{y}^* = 110$ is also the __________ of quarterly sales for the new restaurant located near Talbot College, a school with 10,000 students.

Interval Estimation

1. Point estimators and predictors do not provide any information about the __________ associated with the estimate and/or prediction. For that we must develop ___________ intervals and ___________ intervals.

   (a) A confidence interval is an interval estimate of the _______________ for a given value of x.

   (b) A prediction interval is used whenever we want to predict an ___________________ for a new observation corresponding to a given value of x.

2. Although the predictor of y for a given value of x is the same as the point estimator of the mean value of y for a given value of x, the ________________ we obtain for the two cases are different.

3. The margin of error is ________ for a prediction interval.

4. Prediction intervals resemble confidence intervals. However, they differ conceptually. A confidence interval represents an ______________________________ and is an interval that is intended to cover the value of the parameter. A prediction interval is a statement about the value to be taken by a ______________________, the new observation y^*_{new}.

Confidence Interval for the Mean Value of y

1. In general, we cannot expect $\hat{y}^*$ to equal $E(y^*)$ exactly. If we want to make an inference about how close $\hat{y}^*$ is to the true mean value $E(y^*)$, we will have to estimate the variance of $\hat{y}^*$.

2. The formula for estimating the variance of $\hat{y}^*$, denoted by ________, is

$$s^2_{\hat{y}^*} = \underline{\qquad\qquad\qquad\qquad\qquad\qquad} \tag{14.22}$$

where $s^2 =$ __________________________________.

3. The estimate of the standard deviation of $\hat{y}^*$ is given by the square root of equation (14.22).

$$s_{\hat{y}^*} = s\sqrt{\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]} \tag{14.23}$$

4. NOTE:

$$\begin{aligned}
y_i &= \beta_0^* + \beta_1(x_i - \bar{x}) + \epsilon_i, \beta_0^* = \beta_0 + \beta_1\bar{x} \quad \text{(alternative model)} \\
\hat{y}^* &= b_0^* + b_1(x^* - \bar{x}), b_0^* = b_0 + b_1\bar{x} = \bar{y} \\
\hat{y}^* &= \bar{y} + b_1(x^* - \bar{x}) \\
E(\hat{y}^*) &= E(y^*) \\
\sigma^2_{\hat{y}^*} = Var(\hat{y}^*) &= Var(\bar{y} + b_1(x^* - \bar{x})) \\
&= Var(\bar{y}) + Var(b_1(x^* - \bar{x})) \\
&= \frac{\sigma^2}{n} + (x^* - \bar{x})^2\frac{\sigma^2}{\sum(x_i - \bar{x})^2} \\
&= \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right].
\end{aligned}$$

5. Example Armand's Pizza Parlors

   $s = 13.829$. With $x^* = 10$, $\bar{x} = 14$, and $\sum(x_i - \bar{x})^2 = 568$, we can use equation (14.23) to obtain

   $$s_{\hat{y}^*} = \underline{\hspace{8cm}}$$

6. Theorem:

   $$\frac{\hat{y}^* - E(y^*)}{s_{\hat{y}^*}} \sim t_{(n-2)}$$

7. **Confidence Interval for $E(y^*)$**

   $$\underline{\hspace{3cm}} \qquad (14.24)$$

   where the confidence coefficient is $1-\alpha$ and $t_{\alpha/2}$ is based on the t distribution with $(n\text{–}2)$ degrees of freedom.

8. Example Armand's Pizza Parlors

   (a) Develop a 95% confidence interval of the mean quarterly sales for all Armand's restaurants located near campuses with 10,000 students.

   (b) We have ________________. Thus, with ____________ and a margin of error of ____________________________, the 95% confidence interval estimate is 110 ± 11.415.

   (c) In dollars, the 95% confidence interval for the mean quarterly sales of all restaurants near campuses with 10,000 students is \$110,000 ± \$11,415. Therefore, the 95% confidence interval for the ______________________ when the student population is 10,000 is ______________________.

9. Note that the estimated standard deviation of $\hat{y}^*$ given by equation (14.23) is smallest when _____________.

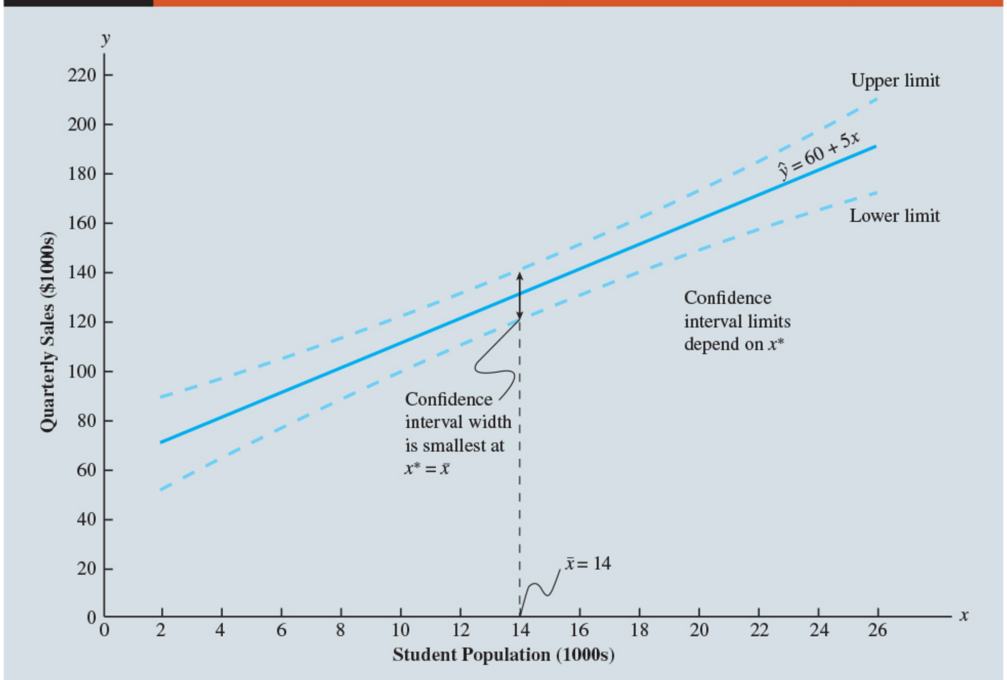
10. In this case the estimated standard deviation of $\hat{y}^*$ becomes

    $$s_{\hat{y}^*} = s\sqrt{\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]} = \underline{\hspace{2cm}}$$

    This result implies that we can make the best or most precise estimate of the mean value of y whenever $x^* = \bar{x}$.

11. (Figure 14.8) the further x^* is from $\bar{x}$, the larger $x^*-\bar{x}$ becomes. As a result, the confidence interval for the mean value of y will become wider as x^* deviates more from $\bar{x}$.

FIGURE 14.8 Confidence Intervals for the Mean Sales y at Given Values of Student Population x



Prediction Interval for an Individual Value of y

1. The predictor of y^*, the value of y corresponding to the given x^*, is $\hat{y}^* = \beta_0 + \beta_1 x^*$.

2. For the new restaurant located near Talbot College, $x^* = 10$ and the prediction of quarterly sales is $\hat{y}^* = 60 + 5(10) = 110$, or $110,000. Note that the prediction of quarterly sales for the new Armand's restaurant near Talbot College is the ________ as the point estimate of the mean sales for all Armand's restaurants located near campuses with 10,000 students.

3. Determine the variance associated with using $\hat{y}^*$ as a predictor of y when $x = x^*$. This variance is made up of the sum of the following two components.

   (a) The (estimated) variance of the y^* values about the mean $E(y^*)$: ______.

   (b) The (estimated) variance associated with using $\hat{y}^*$ to estimate $E(y^*)$: ______.

4. The formula for estimating the variance corresponding to the prediction of the value of y when $x = x^*$, denoted s^2_{pred}, is

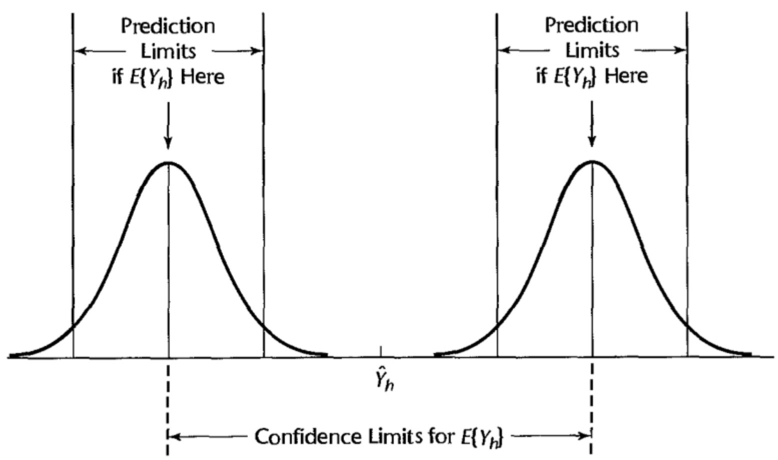
$$s^2_{pred} = \underline{\hspace{3cm}}$$
$$= \underline{\hspace{7cm}}$$
$$= \underline{\hspace{7cm}} \quad (14.25)$$

5. Theorem:

$$\frac{\hat{y}^* - y^*_{new}}{s_{pred}} \sim t_{(n-2)}$$

$$\begin{aligned} \sigma^2_{pred} &= Var(\hat{y}^* - y^*_{new}) \\ &= Var(\hat{y}^*) + Var(y^*_{new}) \\ &= Var(\hat{y}^*) + \sigma^2 \end{aligned}$$

FIGURE 2.5 **Prediction of $Y_{h(new)}$ when Parameters Unknown.**



6. (Armand's Pizza Parlors) the estimated standard deviation corresponding to the prediction of quarterly sales for a new restaurant located near Talbot College, a campus with 10,000 students, is computed as follows.

$$s_{pred} = \underline{\hspace{12cm}}$$

7. **Prediction Interval For y^***

$$\underline{\hspace{4cm}} \quad (14.27)$$

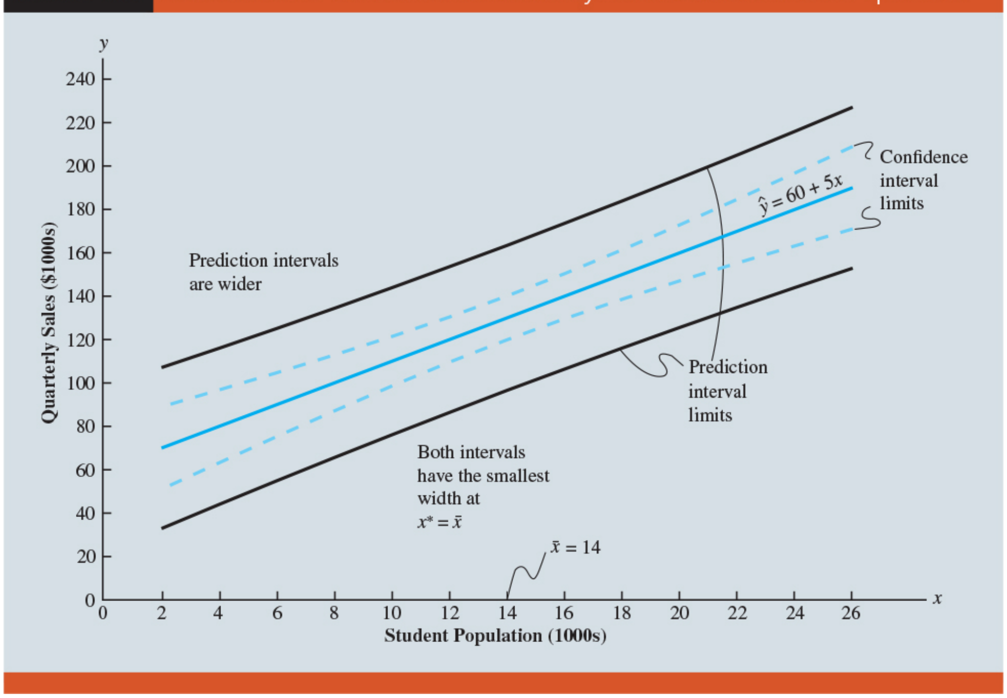
where the confidence coefficient is $1-\alpha$ and $t_{\alpha/2}$ is based on the t distribution with n–2 degrees of freedom.

8. (Armand's Pizza Parlors) The 95% prediction interval for quarterly sales for the new Armand's restaurant located near Talbot College, with $\hat{y}^* = 110$ and a margin of error of ______________________________, the 95% prediction interval is 110 ± 33.875 ($76,125 to $143,875).

9. Note that the prediction interval for the new restaurant located near Talbot College, a campus with 10,000 students, is wider than the confidence interval for the mean quarterly sales of all restaurants located near campuses with 10,000 students. The difference reflects the fact that we are able to estimate the mean value of y _______________ than we can predict an individual value of y.

10. (Figure 14.9) Confidence intervals and prediction intervals are both more precise when the value of the independent variable x^* is closer to $\bar{x}$.

FIGURE 14.9 Confidence and Prediction Intervals for Sales y at Given Values of Student Population x



14.7 Computer Solution

FIGURE 14.10 Output for the Armand's Pizza Parlors Problem

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	14200.0	14200.0	74.25	.000
Error	8	1530.0	191.2		
Total	9	15730.0			

← ANOVA table

Model Summary

S	R-sq	R-sq(adj)
13.8293	90.27%	89.06%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	60.00	9.23	6.50	.000
Population	5.000	.580	8.62	.000

Regression Equation

Sales = 60.00 + 5.000 Population ← Estimated Regression Equation

Prediction for Sales

Variable	Setting
Population	10

Fit	SE Fit	95% CI	95% PI
110	4.95099	(98.5830, 121.417)	(76.1275, 143.873)

← Interval Estimates

14.8 Residual Analysis: Validating Model Assumptions

1. Residual for observation i: the difference between the observed value of the dependent variable (y_i) and the predicted value of the dependent variable ($\hat{y}_i$), __________.

2. An analysis of the corresponding residuals will help determine whether the assumptions made about the regression model are appropriate.

3. (Table 14.7)

TABLE 14.7 Residuals for Armand's Pizza Parlors

Student Population x_i	Sales y_i	Predicted Sales $\hat{y}_i = 60 + 5x_i$	Residuals $y_i - \hat{y}_i$
2	58	70	−12
6	105	90	15
8	88	100	−12
8	118	100	18
12	117	120	−3
16	137	140	−3
20	157	160	−3
20	169	160	9
22	149	170	−21
26	202	190	12

4. Example Armand's Pizza Parlors

(a) A simple linear regression model was assumed.

$$y = \beta_0 + \beta_1 x + \epsilon \qquad (14.29)$$

This model indicates that we assumed quarterly sales (y) to be a linear function of the size of the student population (x) plus an error term ϵ. In Section 14.4 we made the following assumptions about the error term ϵ.

1. ______________.
2. The variance of ϵ is the same for all values of x. _________________.
3. The values of ϵ are _________________.
4. The error term ϵ has a ________________________.

(b) These assumptions provide the theoretical basis for the __________ and the ___________ used to determine whether the relationship between x and y is significant, and for the ______________________________________ estimates presented in Section 14.6.

5. If the assumptions about the error term ϵ appear _________________, the hypothesis tests about the significance of the regression relationship and the interval estimation results _____________________.

6. Much of residual analysis is based on an examination of graphical plots:

   (a) A plot of the ____________ against values of the independent variable ______.

   (b) A plot of ____________ against the __________________ of the dependent variable y

   (c) A ______________________ plot.

   (d) A ____________________ plot.

Residual Plot Against x

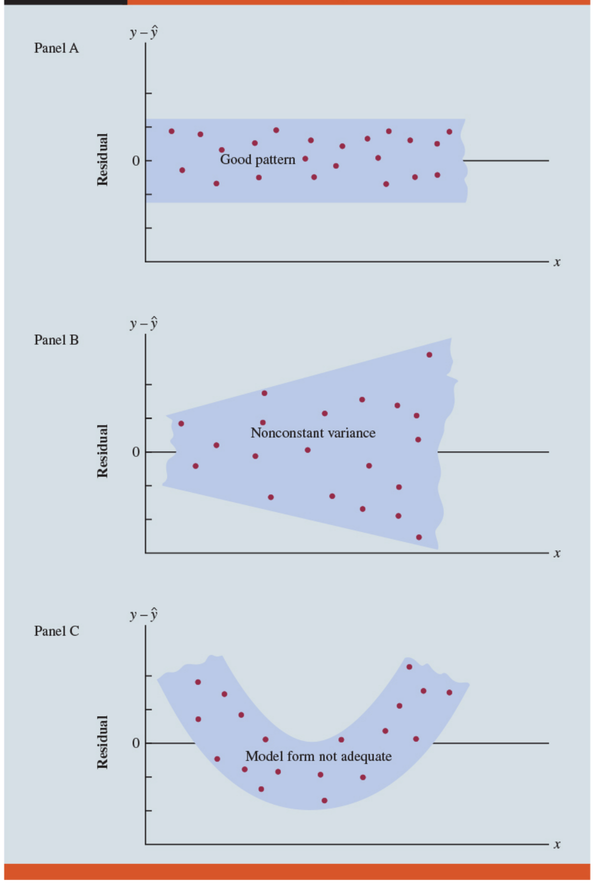
1. (Figure 14.12)

   (a) Panel A: If the assumption that the _______________ is the same for all values of x and the assumed regression model is an adequate representation of the relationship between the variables, the residual plot should give an overall impression of a __________________________.

   (b) Panel B: if the _______________ is not the same for all values of x—for example, if variability about the regression line is greater for larger values of x.

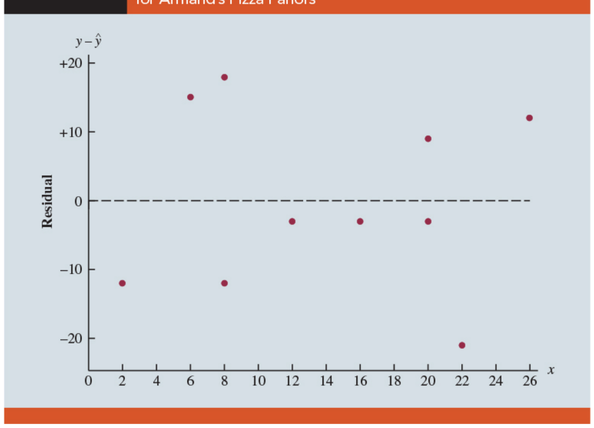
   (c) Panel C: we would conclude that the assumed regression model is not an adequate representation of the relationship between the variables. A _____________ regression model or ____________ regression model should be considered.

FIGURE 14.12 Residual Plots from Three Regression Studies



2. (Figure 14.11) Example Armand's Pizza Parlors:

FIGURE 14.11 Plot of the Residuals Against the Independent Variable x for Armand's Pizza Parlors



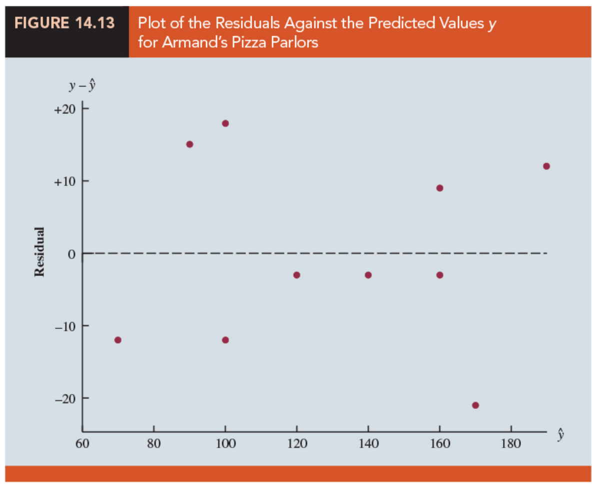
The residual plot does not provide evidence that the assumptions made for Armand's regression model should be challenged. At this point, we are confident in the conclusion that Armand's simple linear regression model is ________.

3. Experience and good judgment are always factors in the effective interpretation of residual plots.

Residual Plot Against $\hat{y}$

1. Another residual plot represents the predicted value of the dependent variable $\hat{y}$ on the horizontal axis and the residual values on the vertical axis.

2. (Figure 14.13) With the Armand's data from Table 14.7,



Note that the pattern of this residual plot is the same as the pattern of the residual plot against the independent variable x.

3. For ____________________ analysis, the residual plot against $\hat{y}$ is more widely used because of the presence of more than one independent variable.

Standardized Residuals

1. A random variable is standardized by subtracting its mean and dividing the result by its standard deviation.

2. With the least squares method, the mean of the residuals is ________. Thus, simply dividing each residual by its ____________________ provides the standardized residual.

3. **Standard Deviation of the ith Residual**

$$s_{y_i-\hat{y}_i} = \underline{\hspace{4cm}} \qquad (14.30)$$

$s_{y_i-\hat{y}_i}$ = the standard deviation of residual i

s = the standard error of the estimate

h_i = $\underline{\hspace{6cm}}$ (14.31)

4. **Standardized Residual for Observation i**

$$\underline{\hspace{3cm}} \qquad (14.32)$$

5. (Table 14.8) the standardized residuals for Armand's Pizza Parlors.

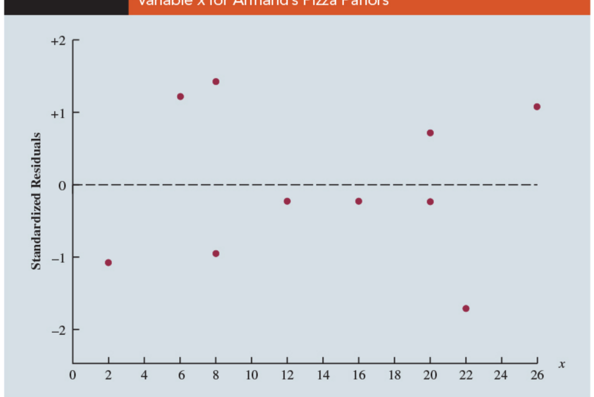
TABLE 14.8 Computation of Standardized Residuals for Armand's Pizza Parlors

Restaurant i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$	h_i	$s_{y_i-\hat{y}_i}$	$y_i - \hat{y}_i$	Standardized Residual
1	2	−12	144	.2535	.3535	11.1193	−12	−1.0792
2	6	−8	64	.1127	.2127	12.2709	15	1.2224
3	8	−6	36	.0634	.1634	12.6493	−12	−.9487
4	8	−6	36	.0634	.1634	12.6493	18	1.4230
5	12	−2	4	.0070	.1070	13.0682	−3	−.2296
6	16	2	4	.0070	.1070	13.0682	−3	−.2296
7	20	6	36	.0634	.1634	12.6493	−3	−.2372
8	20	6	36	.0634	.1634	12.6493	9	.7115
9	22	8	64	.1127	.2127	12.2709	−21	−1.7114
10	26	12	144	.2535	.3535	11.1193	12	1.0792
		Total	568					

Note: The values of the residuals were computed in Table 14.7.

6. (Figure 14.14)

FIGURE 14.14 Plot of the Standardized Residuals Against the Independent Variable x for Armand's Pizza Parlors



7. The standardized residual plot can provide insight about the assumption that the error term ϵ has a ____________________. If this assumption is satisfied, the distribution of the standardized residuals should appear to come from a __________________ probability distribution.

8. Thus, when looking at a standardized residual plot, we should expect to see approximately ________ of the standardized residuals between ______________.

9. We see in Figure 14.14 that for the Armand's example all standardized residuals are between -2 and $+2$. Therefore, on the basis of the standardized residuals, this plot gives us no reason to question the assumption that ϵ has a normal distribution.

Normal Probability Plot

1. Another approach for determining the validity of the assumption that the error term has a normal distribution is the normal probability plot.

2. To show how a normal probability plot is developed, we introduce the concept of ________________.

   (a) Suppose 10 values are selected randomly from a normal probability distribution with a mean of zero and a standard deviation of one, and that the sampling process is repeated over and over with the values in each sample of 10 ______________________________.

   (b) The random variable representing the smallest value obtained in repeated sampling is called the ____________________.

   (c) Statisticians show that for samples of size 10 from a standard normal probability distribution, the expected value of the first-order statistic is -1.55. This expected value is called a ________________.

   NOTE: Compute the expected values of order statistics for a random sample from a standard normal distribution: `evNormOrdStats {EnvStats}`

   https://search.r-project.org/CRAN/refmans/EnvStats/html/evNormOrdStats.html

   (d) (Table 14.9) For the case with a sample of size $n = 10$, there are 10 order statistics and 10 normal.

TABLE 14.9

Normal Scores For $n = 10$

Order Statistic	Normal Score
1	−1.55
2	−1.00
3	−.65
4	−.37
5	−.12
6	.12
7	.37
8	.65
9	1.00
10	1.55

TABLE 14.10

Normal Scores and Ordered Standardized Residuals for Armand's Pizza Parlors

Normal Scores	Ordered Standardized Residuals
−1.55	−1.7114
−1.00	−1.0792
−.65	−.9487
−.37	−.2372
−.12	−.2296
.12	−.2296
.37	.7115
.65	1.0792
1.00	1.2224
1.55	1.4230

```
> data.frame(p, qnorm(p))
            p   qnorm.p.
1  0.00000000       -Inf
2  0.09090909 -1.3351777
3  0.18181818 -0.9084579
4  0.27272727 -0.6045853
5  0.36363636 -0.3487557
6  0.45454545 -0.1141853
7  0.54545455  0.1141853
8  0.63636364  0.3487557
9  0.72727273  0.6045853
10 0.81818182  0.9084579
11 0.90909091  1.3351777
12 1.00000000        Inf
```

(e) Let us now show how the 10 normal scores can be used to determine whether the standardized residuals for Armand's Pizza Parlors appear to come from a standard normal probability distribution.

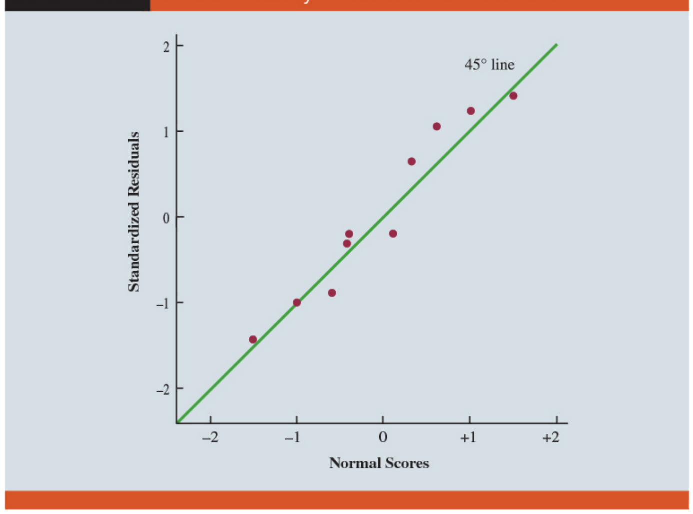
(f) (Table 14.10) The 10 normal scores and the ordered standardized residuals are shown together in Table 14.10. If the normality assumption is satisfied, the smallest standardized residual should be close to the smallest normal score, the next smallest standardized residual should be close to the next smallest normal score, and so on.

(g) A normal probability plot: a plot with the ________________ on the horizontal axis and the corresponding ________________________ on the vertical axis.

(h) If the standardized residuals are approximately normally distributed, the plotted points should cluster closely around a ________________ passing through the _________.

3. (Figure14.15) the normal probability plot for the Armand's Pizza Parlors example: conclude that the assumption of the error term having a normal probability distribution is reasonable.

FIGURE 14.15 Normal Probability Plot for Armand's Pizza Parlors



4. Any substantial curvature in the normal probability plot is evidence that the residuals have not come from a normal distribution.

14.9 Residual Analysis: Outliers and Influential Observations

Detecting Outliers

1. (Figure 14.16) is a scatter diagram for a data set that contains an __________, a data point (observation) that does not fit the trend shown by the remaining data.

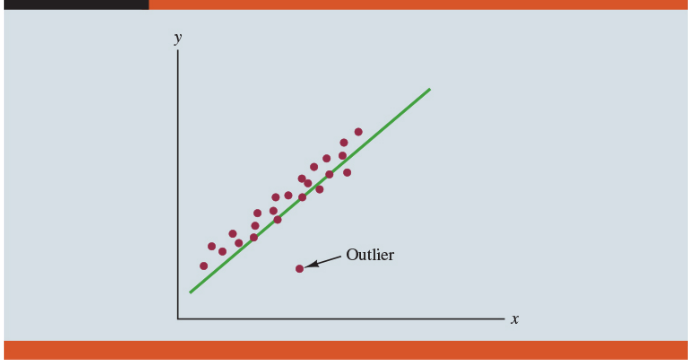


FIGURE 14.16 Data Set with an Outlier

2. Outliers represent observations that are suspect and warrant careful examination. They may represent ____________ data; if so, the data should be ____________.

3. They may signal a violation of model assumptions; if so, ________________ should be considered.

4. Finally, they may simply be ___________ values that occurred by chance. In this case, they should be retained.

5. (Table 14.11) The process of detecting outliers: Except for observation 4 ($x_4 = 3$, $y_4 = 75$), a pattern suggesting a negative linear relationship is apparent. Indeed, given the pattern of the rest of the data, we would expect y_4 to be much smaller and hence would identify the corresponding observation as an outlier.

6. For the case of simple linear regression, one can often detect outliers by simply examining the _________________.

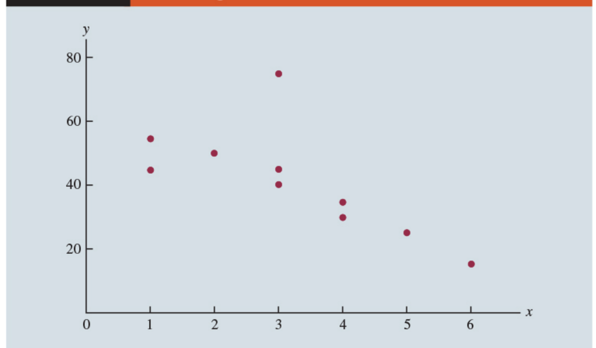
7. The _______________________ can also be used to identify outliers. If an observation deviates greatly from the pattern of the rest of the data, the corresponding standardized residual will be large in absolute value.

TABLE 14.11

Data Set Illustrating the Effect of an Outlier

x_i	y_i
1	45
1	55
2	50
3	75
3	40
3	45
4	30
4	35
5	25
6	15

FIGURE 14.17 Scatter Diagram for Outlier Data Set



8. (Figure 14.18) the output from a regression analysis. The highlighted portion of the output shows that the standardized residual for observation 4 is 2.67. With normally distributed errors, standardized residuals should be outside the range of −2 to +2 approximately 5% of the time.

FIGURE 14.18 Output for Regression Analysis of the Outlier Data Set

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1268.2	1268.2	7.90	.023
Error	8	1284.3	160.5		
Total	9	2552.5			

Model Summary

S	R-sq	R-sq(adj)
12.6704	49.68%	43.39%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	64.96	9.26	7.02	.000
x	−7.33	2.6	−2.81	.023

Regression Equation

y = 64.96 - 7.33 x

Observation	Predicted y	Residuals	Standard Residuals
1	57.6271	−12.6271	−1.0570
2	57.6271	−2.6271	−.2199
3	50.2966	−.2966	−.0248
4	42.9661	32.0339	2.6816
5	42.9661	−2.9661	−.2483
6	42.9661	2.0339	.1703
7	35.6356	−5.6356	−.4718
8	35.6356	−.6356	−.0532
9	28.3051	−3.3051	−.2767
10	20.9746	−5.9746	−.5001

9. In deciding how to handle an outlier, we should first check to see whether it is a ____________________. Perhaps an __________ was made in initially recording the data or in entering the data into the computer file.

10. (Figure 14.19) For example, suppose that in checking the data for the outlier in Table 14.11, we find an error; the correct value for observation 4 is $x_4 = 3, y_4 = 30$. Figure 14.19 is a portion of the output obtained after correction of the value of y_4. We see that using the incorrect data value substantially affected the goodness of fit. With the correct data, the value of __________ increased from 49.68% to 83.8% and the value of _______ decreased from 64.96 to 59.24. The __________ of the line changed from −7.33 to −6.95.

FIGURE 14.19 Output for the Revised Outlier Data Set

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1139.66	1139.66	41.38	.000
Error	8	220.34	27.54		
Total	9	1360.00			

Model Summary

S	R-sq	R-sq(adj)
5.24808	83.80%	81.77%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	59.24	3.83	15.45	.000
x	−6.95	1.08	−6.43	.000

Regression Equation

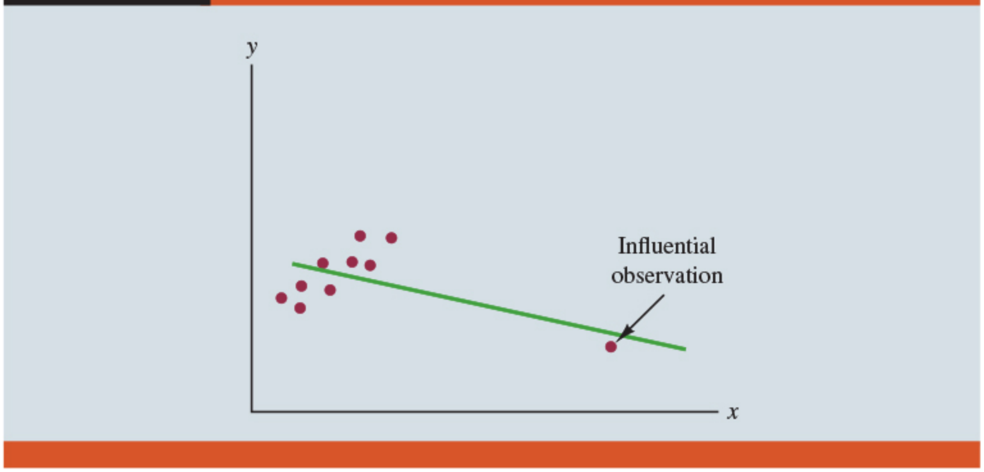
y = 59.24 − 6.95 x

11. The identification of the outlier enabled us to correct the data error and improve the regression results.

Detecting Influential Observations

1. (Figure 14.20) shows an example of an influential observation in simple linear regression.

FIGURE 14.20 Data Set with an Influential Observation



The estimated regression line has a negative slope. However, if the influential observation were dropped from the data set, the slope of the estimated regression line would change from negative to positive and the y-intercept would be smaller. Clearly, this one observation is much more influential in determining the estimated regression line than any of the others.

2. Influential observations can be identified from a ________________ when only one independent variable is present.

3. An influential observation may be an __________ (an observation with a y value that deviates substantially from the trend), it may correspond to an x value far away from its mean (e.g., see Figure 14.20), or it may be caused by a combination of the two (a somewhat off-trend y value and a somewhat extreme x value).

4. The presence of the influential observation in Figure 14.20, if valid, would suggest trying to obtain data on intermediate values of x to understand better the relationship between x and y.

5. Observations with ________________ for the independent variables are called high ________________. The influential observation in Figure 14.20 is a point with high leverage.

6. The leverage of an observation is determined by how far the values of the independent variables are from their ______________.

7. For the single-independent-variable case, the leverage of the ith observation, denoted h_i, can be computed by using equation (14.33).

$$h_i = \underline{\hspace{4cm}}$$

Definition and properties of leverages:
`https://online.stat.psu.edu/stat501/lesson/11/11.2`

8. From the formula, it is clear that the farther x_i is from its mean $\bar{x}$, the higher the leverage of observation i.

9. (Figure 14.21) a scatter diagram for the data set in Table 14.12, it is clear that observation 7 ($x = 70, y = 100$) is an observation with an extreme value of x. Hence, we would expect it to be identified as a point with high leverage:

$$h_7 = \underline{\hspace{4cm}} = \underline{\hspace{6cm}}$$

10. For the case of simple linear regression, observations have high leverage if $h_i > 6/n$ or 0.99, whichever is smaller.

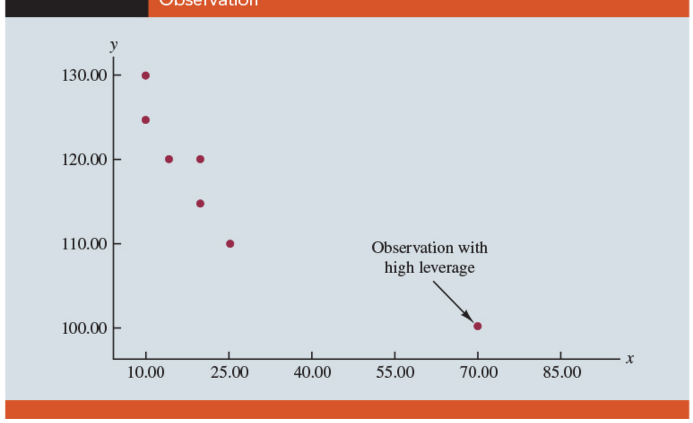
11. For the data set in Table 14.12, $6/n = 6/7 = 0.86$. Because $h_7 = 0.94 > 0.86$, we will identify observation 7 as an observation whose x value gives it large influence.

12. Influential observations that are caused by an interaction of large residuals and high leverage can be difficult to detect. Diagnostic procedures are available that take both into account in determining when an observation is influential. One such measure, called ____________________, will be discussed in Chapter 15.

TABLE 14.12
Data Set with a High Leverage Observation

x_i	y_i
10	125
10	130
15	120
20	115
20	120
25	110
70	100

FIGURE 14.21 Scatter Diagram for the Data Set with a High Leverage Observation



😍 EXERCISES

14.2 : 1, 5, 6

14.3 : 15, 19, 20

14.5 : 23, 26, 27, 30

14.6 : 32, 36, 37

14.7 : 40, 41

14.8 : 45, 47

14.9 : 50, 52

SUP : 59, 67

"不要畏懼失敗，你應該要擔心沒有機會嘗試，但你有的是機會嘗試!"
"Don't fear failure. Be afraid of not having the chance, you have the chance!"
— *汽車總動員 3: 閃電再起 (Cars 3, 2017)*

統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 15: Multiple Regression

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: http://www.hmwu.idv.tw

系級: ____________ 學號: ____________ 姓名: ____________

15.1 Multiple Regression Model

1. (Recall) that the variable being predicted or explained is called the ____________ variable and the variable being used to predict or explain the dependent variable is called the ____________ variable.

2. Multiple regression analysis is the study of how a dependent variable y is related to ________________ variables. In the general case, we will use ____ to denote the number of independent variables.

3. The concepts of a regression model and a regression equation introduced in the preceding chapter are ____________ in the multiple regression case.

4. **Multiple regression model**: The equation that describes how the dependent variable y is related to the independent variables $x_1, x_2, \cdots, x_p$ and an error term is called the multiple regression model.

   ________________________________ (15.1)

5. In the multiple regression model, $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ are the ____________ and the error term ϵ is a ______________. y is a linear function of $x_1, x_2, \cdots, x_p$ plus the error term ϵ.

6. The error term accounts for the ____________ in y that ______________ by the linear effect of the p independent variables.

7. (**Multiple regression equation**):The equation that describes how the mean value of y is related to $x_1, x_2, \cdots, x_p$ is called the multiple regression equation.

$$\rule{6cm}{0.4pt} \qquad (15.2)$$

under the assumption that the mean or expected value of ϵ is zero.

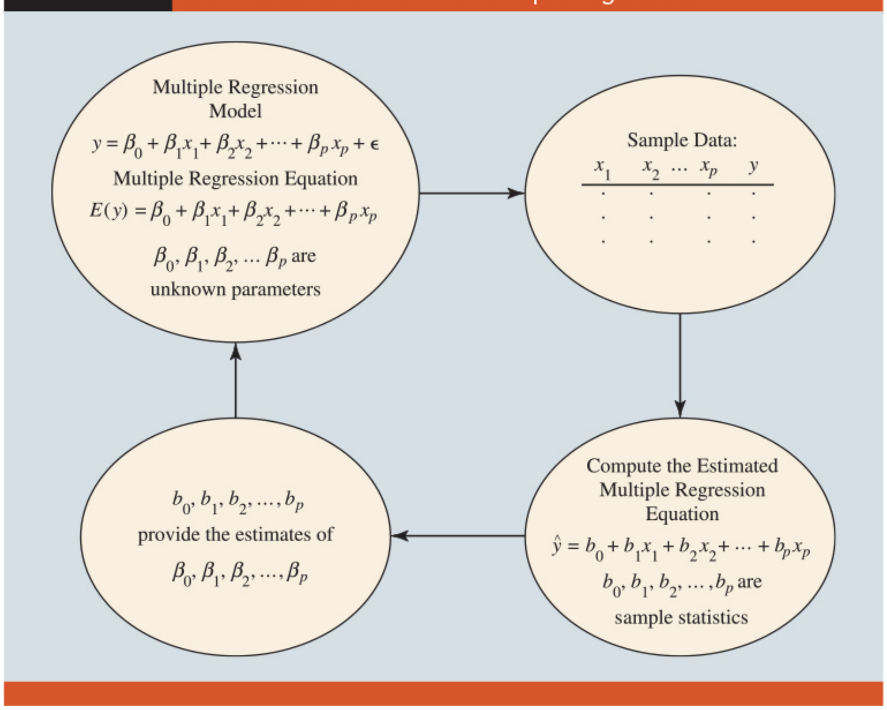
8. **The estimated multiple regression equation**:

$$\rule{5.5cm}{0.4pt} \qquad (15.3)$$

where $b_0, b_1, b_2, \cdots, b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ and $\hat{y}$ is the predicted value of the dependent variable

9. (Figure 15.1)

FIGURE 15.1 The Estimation Process for Multiple Regression



15.2 Least Squares Method

1. The least squares method is used to develop the estimated multiple regression equation:

   ______________________ (15.4)

   where y_i is observed value of the dependent variable for the ith observation, $\hat{y}_i$ is predicted value of the dependent variable for the ith observation

2. In multiple regression, however, the presentation of the formulas for the regression coefficients $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ involves the use of ____________________ and is beyond the scope of this text.

3. Therefore, in presenting multiple regression, we focus on how statistical software can be used to obtain the estimated regression equation and other information. The emphasis will be on how to ______________ the computer output rather than on how to make the multiple regression computations.

An Example: Butler Trucking Company

1. The Butler Trucking Company, an independent trucking company in southern California.

2. A major portion of Butler's business involves deliveries throughout its local area. To develop better work schedules, the managers want to predict the total daily travel time for their drivers.

   (a) Initially the managers believed that the total daily travel time would be closely related to the number of miles traveled in making the daily deliveries.

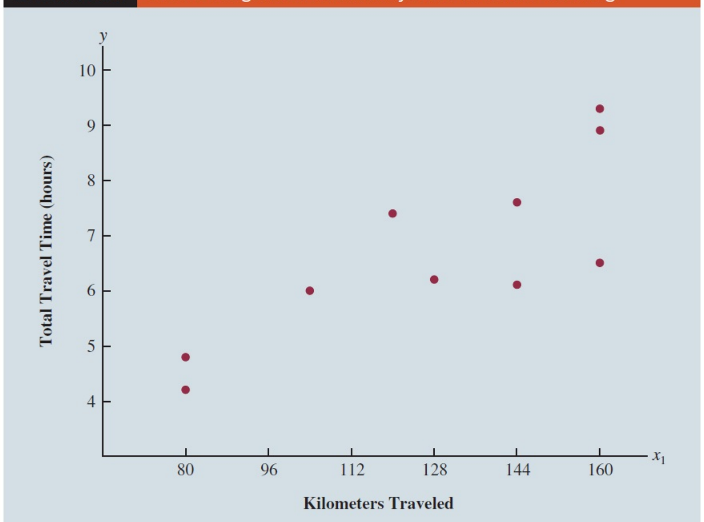
   (b) (Table 15.1)(Figure 15.2) A simple random sample of 10 driving assignments provided the data shown in Table 15.1 and the scatter diagram shown in Figure 15.2.

TABLE 15.1 Preliminary Data for Butler Trucking

Driving Assignment	x_1 = Kilometers Traveled	y = Travel Time (hours)
1	160	9.3
2	80	4.8
3	160	8.9
4	160	6.5
5	80	4.2
6	128	6.2
7	120	7.4
8	104	6.0
9	144	7.6
10	144	6.1

Source: PC Magazine website, April, 2015. (https://www.pcmag.com/reviews/monitors)

FIGURE 15.2 Scatter Diagram of Preliminary Data for Butler Trucking



(c) After reviewing this scatter diagram, the managers hypothesized that the simple linear regression model $y = \beta_0 + \beta_1 x_1 + \epsilon$ could be used to describe the relationship between the total travel time (y) and the number of miles traveled (x_1).

(d) (Figure 15.3) we show statistical software output from applying simple linear regression to the data in Table 15.1. The estimated regression equation is ____________________

i. At the 0.05 level of significance, the F value of __________ and its corresponding p-value of __________ indicate that the relationship is significant; that is, we can reject $H_0 : \beta_1 = 0$ because the p-value is less than $\alpha = 0.05$.

ii. Note that the same conclusion is obtained from the t value of __________ and its associated p-value of __________.

iii. Thus, we can conclude that the relationship between the total travel time and the number of miles traveled is ______________; longer travel times are associated with more miles traveled.

FIGURE 15.3 Output for Butler Trucking with One Independent Variable

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	15.871	15.8713	15.81	.004
Error	8	8.029	1.0036		
Total	9	23.900			

Model Summary

S	R-sq	R-sq (adj)
1.00179	66.41%	62.21%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	1.27	1.40	.91	.390
Kilometers	.0424	.0107	3.98	.004

Regression Equation

Time = 1.27 + .0424 Kilometers

iv. With a coefficient of determination (expressed as a percentage) of ________________, we see that ___________________________ in travel time can be explained by the linear effect of the number of miles traveled.

3. (Table 15.2) The managers might want to consider adding a second independent variable (number of deliveries) to explain some of the remaining variability in the dependent variable.

TABLE 15.2 Data for Butler Trucking with Kilometers Traveled (x_1) and Number of Deliveries (x_2) as the Independent Variables

Driving Assignment	x_1 = Kilometers Traveled	x_2 = Number of Deliveries	y = Travel Time (hours)
1	160	4	9.3
2	80	3	4.8
3	160	4	8.9
4	160	2	6.5
5	80	2	4.2
6	128	2	6.2
7	120	3	7.4
8	104	4	6.0
9	144	3	7.6
10	144	2	6.1

4. (Figure 15.4) Computer output with both miles traveled (x_1) and number of deliveries (x_2) as independent variables is shown in Figure 15.4. The estimated regression equation is

$$\hat{y} = \underline{\hspace{5cm}} \qquad (15.6)$$

FIGURE 15.4 Output for Butler Trucking with Two Independent Variables

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	21.6006	10.8003	32.88	.000
Error	7	2.2994	.3285		
Total	9	23.900			

Model Summary

S	R-sq	R-sq (adj)
.573142	90.38%	87.63%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	−.869	.952	−.91	.392
Kilometers	.03821	.00618	6.18	.000
Deliveries	.923	.221	4.18	.004

Regression Equation

Time = −.869 + .03821 Kilometers + 0.923 Deliveries

Note on Interpretation of Coefficients

1. One observation can be made at this point about the relationship between the estimated regression equation with only the miles traveled as an independent variable and the equation that includes the ______________ as a second independent variable.

2. The value of ______ is not the same in both cases. In simple linear regression, we interpret β_1 as an estimate of the change in y for a ______________ in the independent variable.

3. In multiple regression analysis, we interpret each regression coefficient as follows: b_i represents an estimate of the ______________ corresponding to a ______________ ________ when all other independent variables are ______________.

4. Butler Trucking example

   (a) $\beta_1 = 0.06113$, an estimate of the expected increase in travel time corresponding to an increase of one mile in the distance traveled when the number of deliveries is held constant is 0.06113 hours.

   (b) $\beta_2 = 0.923$, an estimate of the expected increase in travel time corresponding to an increase of one delivery when the number of miles traveled is held constant is 0.923 hours.

15.3 Multiple Coefficient of Determination

1. In simple linear regression, we showed that the total sum of squares can be partitioned into two components: the sum of squares due to regression and the sum of squares due to error. The same procedure applies to the sum of squares in multiple regression.

   ______________________ (15.7)

   where

   SST: total sum of squares = ______________.

   SSR: sum of squares due to regression = ______________.

   SSE: sum of squares due to error = ______________.

2. Example Butler Trucking problem (Figure 15.4) $SST = 23.900$, $SSR = 21.6006$, and $SSE = 2.2994$.

3. With only one independent variable (number of miles traveled), the output in Figure 15.3 shows that $SST = 23.900$, $SSR = 15.871$, and $SSE = 8.029$. The value of SST is the same in both cases because it does not depend on $\hat{y}$, but SSR increases and SSE decreases when a second independent variable (number of deliveries) is added.

4. The multiple coefficient of determination, denoted R^2, measures the goodness of fit for the estimated multiple regression equation.

   $$\underline{\hspace{6em}} \qquad (15.8)$$

5. The multiple coefficient of determination can be interpreted as the ____________ ____________ in the dependent variable that can be explained by the estimated multiple regression equation.

6. Hence, when multiplied by 100, it can be interpreted as the percentage of the variability in y that can be explained ____________________.

7. Example In the two-independent-variable Butler Trucking example, with $SSR = 21.6006$ and $SST = 23.900$, we have $R^2 = 21.6006/23.900 = 0.9038$.

8. Therefore, 90.38% of the variability in travel time y is explained by the estimated multiple regression equation with miles traveled and number of deliveries as the independent variables.

9. (Figure 15.3) the R-sq value for the estimated regression equation with only one independent variable, number of miles traveled (x_1), is 66.41%. Thus, the percentage of the variability in travel times that is explained by the estimated regression equation increases from ____________ when number of deliveries is added as a second independent variable.

10. In general, R^2 always increases as independent variables are added to the model.

11. Many analysts prefer adjusting R^2 for the number of independent variables to avoid ____________ the impact of adding an independent variable on the amount of variability exlained by the estimated regression equation.

12. With n denoting the number of observations and p denoting the number of independent variables, the adjusted multiple coefficient of determination is computed as follows:

    $$\underline{\hspace{12em}} \qquad (15.9)$$

13. Example With $n = 10$ and $p = 2$, we have

    $$R^2 = 1 - (1 - 0.9038)\frac{10 - 1}{10 - 2 - 1}$$

14. Thus, after adjusting for the two independent variables, we have an adjusted multiple coefficient of determination of 0.8763. This value (expressed as a percentage) is provided in the output in Figure 15.4 as ____________________.

15. If the value of R^2 is small and the model contains a large number of independent variables, the adjusted coefficient of determination can take a _______________; in such cases, statistical software usually sets the adjusted coefficient of determination to _______.

15.4 Model Assumptions

1. The multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad (15.10)$$

2. The assumptions about the _____________ in the multiple regression model:

   (1) The error term ϵ is a random variable with mean or expected value of zero; that is, ___________.

   Implication: For given values of $x_1, x_2, \cdots, x_p$, the expected, or average, value of y is given by

   $$E(y) = \underline{\hspace{8cm}} \qquad (15.11)$$

   Equation (15.11) is the __________________________. $E(y)$ represents the average of all possible values of y that might occur for the given values of $x_1, x_2, \cdots, x_p$.

   (2) The variance of ϵ is denoted by σ^2 and is the same for all values of the independent variables $x_1, x_2, \cdots, x_p$; that is, _____________.

   Implication: The variance of y about the regression line equals ______ and is the same for all values of $x_1, x_2, \cdots, x_p$.

(3) The values of ϵ are ________________.

Implication: The value of ϵ for a particular set of values for the independent variables is not related to the value of ϵ for any other set of values.

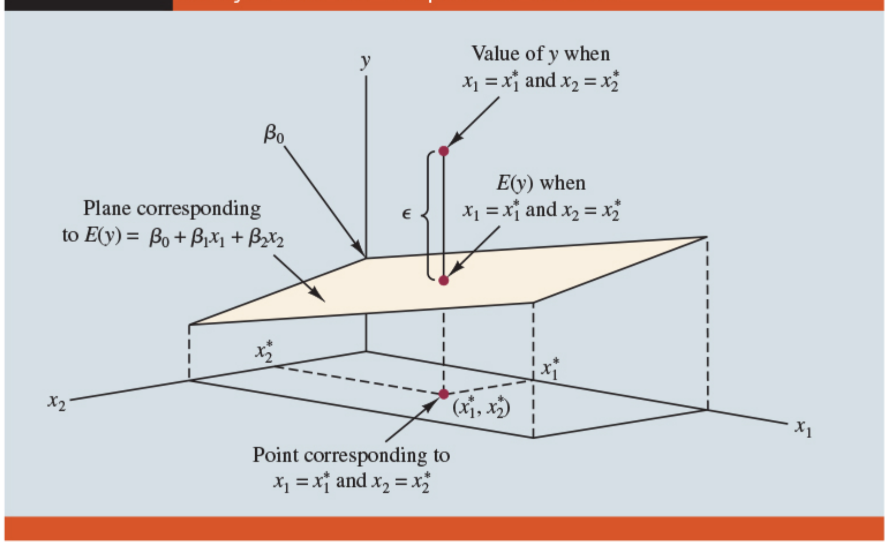
(4) The error term ϵ is a ________________________ random variable reflecting the deviation between the ______ value and the _________ given by $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.

Implication: Because $\beta_0, \beta_1, \cdots, \beta_p$ are _____________ for the given values of $x_1, x_2, \cdots, x_p$, the dependent variable y is also a _____________ distributed random variable.

3. (Figure 15.5) Consider the following two-independent-variable multiple regression equation.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

FIGURE 15.5 Graph of the Regression Equation for Multiple Regression Analysis with Two Independent Variables



4. Note that the value of ϵ shown is the ______________ between the actual y value and the expected value of y, $E(y)$, when $x_1 = x_1^*$ and $x_2 = x_2^*$.

5. In regression analysis, the term response variable is often used in place of the term ______________________. Furthermore, since the multiple regression equation generates a plane or surface, its graph is called a ____________________.

15.5 Testing for Significance

1. In simple linear regression, both ________ and an _________ provide the same conclusion; that is, if the null hypothesis is rejected, we conclude that _________.

2. In multiple regression, the t test and the F test have different purposes.

   (a) The F test is used to determine whether a significant relationship exists between the dependent variable and the set of ______ the independent variables; we will refer to the F test as the test for ___________________.

   (b) If the F test shows an overall significance, the ________ is used to determine whether each of the individual independent variables is significant. A separate t test is conducted for each of the independent variables in the model; we refer to each of these t tests as a test for ______________________.

3. In the material that follows, we will explain the F test and the t test and apply each to the Butler Trucking Company example.

F Test

1. The hypotheses for the F test involve the parameters of the multiple regression model.

$$\begin{aligned} H_0 &: \underline{\hspace{6cm}} \\ H_a &: \text{One or more of the parameters are not equal to zero} \end{aligned}$$

2. If H_0 is rejected, the test gives us ___________________________ to conclude that one or more of the parameters are not equal to zero and that the ___________________ between y and the set of independent variables $x_1, x_2, \cdots, x_p$ is ____________.

3. However, if H_0 cannot be rejected, we do not have ___________________ to conclude that a significant relationship is present.

4. (Review)(Chapter 14)

   (a) A mean square is a ________________ divided by its corresponding degrees of freedom.

(b) In the multiple regression case, the total sum of squares (SST) has ________ degrees of freedom, the sum of squares due to regression (SSR) has ______ degrees of freedom, and the sum of squares due to error (SSE) has __________ degrees of freedom.

(c) Hence, the mean square due to regression (MSR) is __________ and the mean square due to error (MSE) is ________________.

(d) MSE provides an unbiased estimate of _______, the variance of the error term ϵ.

(e) If ___________________________ is true, _________ also provides an unbiased estimate of σ^2, and the value of MSR/MSE should be close to _____.

(f) However, if H_0 is false, MSR _______________ σ^2 and the value of MSR/MSE becomes _________.

5. To determine how large the value of _______________ must be to reject H_0, we make use of the fact that if ____________ and the ______________ about the multiple regression model are ________, the sampling distribution of MSR/MSE is an ______ distribution with _____ degrees of freedom in the numerator and __________ in the denominator.

6. **F test for overall significance**

(a) Hypothesis:

$$
\begin{aligned}
H_0 &: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\
H_a &: \text{One or more of the parameters are not equal to zero}
\end{aligned}
$$

(b) Test statistic:

______________ (15.14)

(c) Rejection rule:

i. p-value approach: Reject H_0 if ______________.

ii. Critical value approach: Reject H_0 if ________________

TABLE 15.3 ANOVA Table for a Multiple Regression Model with p Independent Variables

Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SSR	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	
Total	SST	$n-1$		

7. Example Butler Trucking Company

(a) Hypotheses:

$$H_0 \;:\; \beta_1 = \beta_2 = 0$$
$$H_a \;:\; \beta_1 \text{ and/or } \beta_2 \text{ is not equal to zero}$$

(b) (Figure 15.6)

FIGURE 15.6 Output for Butler Trucking with Two Independent Variables, Kilometers Traveled (x_1) and Number of Deliveries (x_2)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	21.6006	10.8003	32.88	.000
Error	7	2.2994	.3285		
Total	9	23.900			

Model Summary

S	R-sq	R-sq (adj)
.573142	90.38%	87.63%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	−.869	.952	−.91	.392
Kilometers	.03821	.00618	6.18	.000
Deliveries	.923	.221	4.18	.004

Regression Equation

Time = −.869 + .03821 Kilometers + .923 Deliveries

(c) $MSR = 10.8003$ and $MSE = 0.3285$, $F =$ __________. Using $\alpha = 0.01$, ________________. With $F = 32.88 > 9.55$, we reject $H_0 : \beta_1 = \beta_2 = 0$.

(d) Using $\alpha = 0.01$, the p-value $= 0.000$ indicates that we can reject $H_0 : \beta_1 = \beta_2 = 0$ because the p-value is less than $\alpha = 0.01$.

(e) Conclude that a ____________________ is present between travel time y and the two independent variables, miles traveled and number of deliveries.

t Test

1. If the F test shows that the multiple regression relationship is significant, a t test can be conducted to determine the significance of each of the ____________ parameters.

2. **The t test for individual significance**

   (a) Hypothesis: For any parameter β_i

   $$\begin{aligned} H_0 &: \underline{\hspace{3cm}} \\ H_a &: \beta_i \neq 0 \end{aligned}$$

   (b) Test statistic:

   $$\underline{\hspace{3cm}} \qquad (15.15)$$

   (c) Rejection rule:

   i. p-value approach: Reject H_0 if p-value $\leq \alpha$.

   ii. Critical value approach: Reject H_0 if _______________ or if ________________.

3. In the test statistic, s_{b_i} is the estimate of the standard deviation of b_i. The value of s_{b_i} will be provided by the computer software package.

   補充:

   The multiple regression model

   $$\underline{\hspace{8cm}},$$

   or

   $$\underline{\hspace{10cm}}.$$

   (a) In the matrix notation:

   $$\underline{\hspace{3cm}} \quad \text{or} \quad \mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}.$$

(b) Design matrix $\mathbf{X}$:

$$\mathbf{X} =$$

(c) Use Least-squares to fit a regression line to the data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, \cdots, x_{i,p-1}\}$

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i-\mathbf{x}_i\boldsymbol{\beta})^2. \\ \frac{\partial Q}{\partial \boldsymbol{\beta}} &= -2\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \\ \Rightarrow & \quad (\mathbf{X}'\mathbf{X})\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \\ \Rightarrow & \quad \hat{\boldsymbol{\beta}} = \mathbf{b} = \underline{\hspace{3cm}} \end{aligned}$$

(d) Variance of the sampling distribution of $b_i, i = 1, 2, \cdots, p$.

$$Var(b_i) = \frac{\sigma^2}{(n-1)S_{x_i}^2(1-R_i^2)},$$

where $S_{x_i}^2$ is the sample variance of variable x_i and R_i^2 is R-square of the regression of x_i on the rest of the explanatory variables of the models (including the constant term). Note that the variance should be conditional on the observed values of the explanatory variables.

4. Example Butler Trucking Company

(a) (Figure 15.6) that shows the output for the t-ratio calculations:

$$b_1 = 0.06113, b_2 = 0.923, s_{b_1} = 0.00989, s_{b_2} = 0.221$$

(b) The test statistic for the hypotheses involving parameters β_1 and β_2:

$$t = 0.06113/0.00989 = 6.18, \quad t = 0.923/0.221 = 4.18$$

(c) Using $\alpha = 0.01$, the p-values of ________ and ________ in the output indicate that we can reject $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. Hence, both parameters are statistically significant.

(d) Alternatively, ________________. With $6.18 > 3.499$, we reject $H_0 : \beta_1 = 0$. Similarly, with $4.18 > 3.499$, we reject $H_0 : \beta_2 = 0$.

Multicollinearity

1. We use the term ______________________ in regression analysis to refer to any variable being used to predict or explain the value of the dependent variable.

2. The term does not mean, however, that the independent variables ______________ are independent in any statistical sense. On the contrary, most independent variables in a multiple regression problem are _____________ to some degree with one another.

3. Example Butler Trucking Example

   (a) Butler Trucking example involves the two independent variables x_1 (miles traveled) and x_2 (number of deliveries), we could treat the miles traveled as the dependent variable and the number of deliveries as the independent variable to determine whether those two variables are themselves related.

   (b) Compute the sample correlation coefficient $r(x_1, x_2) = 0.16$ and find that some degree of linear association between the two independent variables.

4. In multiple regression analysis, ___________________ refers to the correlation among the independent variables.

5. Example Modified Butler Trucking Example, the potential problems of multicollinearity.

   (a) Consider a modification of the Butler Trucking example. Instead of x_2 being the number of deliveries, let x_2 denote the number of gallons of gasoline consumed. Clearly, x_1 (the miles traveled) and x_2 are related; that is, we know that the number of gallons of gasoline used depends on the number of miles traveled.

   (b) We would conclude logically that x_1 and x_2 are highly correlated independent variables.

   (c) Assume that we obtain the equation $\hat{y} = b_0 + b_1x_1 + b_2x_2$ and find that the F test shows the relationship to be significant. Then suppose we conduct a t test on β_1 to determine whether $\beta_1 \neq 0$, and we cannot reject $H_0 : \beta_1 = 0$. Does this result mean that travel time is not related to miles traveled? Not necessarily.

(d) What it probably means is that with ________________________, x_1 does not make a significant contribution to determining the value of y.

(e) This interpretation makes sense in our example; if we know the amount of gasoline consumed (x_2), we do not gain much additional information useful in predicting y by knowing the miles traveled (x_1).

(f) Similarly, a t test might lead us to conclude $\beta_2 = 0$ on the grounds that, with x_1 in the model, knowledge of the amount of gasoline consumed does not add much.

6. To summarize, in __________ for the significance of individual parameters, the difficulty caused by multicollinearity is that it is possible to conclude that _________ of the individual parameters is significantly different from zero when an __________ on the __________ multiple regression equation indicates a significant relationship.

7. Statisticians have developed several _________ for determining whether multicollinearity is high enough to cause problems.

8. According to the rule of thumb test, multicollinearity is a potential problem if the absolute value of the _____________________________ exceeds _______ for any two of the independent variables.

9. The other types of tests are more advanced and beyond the scope of this text. If possible, every attempt should be made to avoid including independent variables that are highly correlated.

10. When multicollinearity is severe,

    (a) it is not possible to determine the separate effect of any particular independent variable on the dependent variable.

    (b) we can have difficulty interpreting the results of t tests on the individual parameters.

    (c) Least squares estimates may have the wrong sign.

11. 補充:

    (a) Multicollinearity in Regression Analysis: Problems, Detection, and Solutions https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

(b) Multicollinearity in Regression: Why it is a problem? How to check and fix it

https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea

(c) Eight Ways to Detect Multicollinearity

https://www.theanalysisfactor.com/eight-ways-to-detect-multicollinearity/

(d) Multicollinearity (Wikipedia)

https://en.wikipedia.org/wiki/Multicollinearity

15.6 Using the Estimated Regression Equation for Estimation and Prediction

1. The procedures for estimating the mean value of y and predicting an individual value of y in multiple regression are similar to those in regression analysis involving one independent variable.

2. We substitute the given values of $x_1, x_2, \cdots, x_p$ into the estimated regression equation and use the corresponding value of $\hat{y}$ as the ________________.

3. Example Butler Trucking example

   (a) We want to use the estimated regression equation involving x_1 (miles traveled) and x_2 (number of deliveries) to develop two interval estimates:

      i. A ____________________ of the mean travel time for all trucks that travel 100 miles and make two deliveries.

      ii. A ____________________ of the travel time for one specific truck that travels 100 miles and makes two deliveries

   (b) Using the estimated regression equation $\hat{y} = -0.869 + 0.06113x_1 + 0.923x_2$ with $x_1 = 100$ and $x_2 = 2$, we obtain

$$\hat{y} = \underline{\hspace{6cm}}$$

Hence, the point estimate of travel time in both cases is approximately seven hours.

(c) To develop interval estimates for the mean value of y and for an individual value of y, we use a procedure similar to that for regression analysis involving one independent variable. The formulas required are beyond the scope of the text, but statistical ____________ for multiple regression analysis will often provide confidence intervals once the values of $x_1, x_2, \cdots, x_p$ are specified by the user.

(d) (Table 15.4)

TABLE 15.4 The 95% Confidence and Prediction Intervals for Butler Trucking

Value of x_1	Value of x_2	95% Confidence Interval Lower Limit	95% Confidence Interval Upper Limit	95% Prediction Interval Lower Limit	95% Prediction Interval Upper Limit
160	4	8.135	9.742	7.363	10.514
80	3	4.127	5.789	3.369	6.548
160	4	8.135	9.742	7.363	10.514
160	2	6.258	7.925	5.500	8.683
80	2	3.146	4.924	2.414	5.656
128	2	5.232	6.505	4.372	7.366
120	3	6.037	6.936	5.059	7.915
104	4	5.960	7.637	5.205	8.392
144	3	6.917	7.891	5.964	8.844
144	2	5.776	7.184	4.953	8.007
120	4	6.669	8.152	5.865	8.955

(e) Note that the interval estimate for an individual value of y is ______________ the interval estimate for the expected value of y. This difference simply reflects the fact that for given values of x_1 and x_2 we can estimate the mean travel time for all trucks with ________________ than we can predict the travel time for one specific truck.

15.7 Categorical Independent Variables

(a) Thus far, the examples we have considered involved ______________ independent variables such as student population, distance traveled, and number of deliveries.

(b) In many situations, however, we must work with ______________ independent variables such as gender (male, female), method of payment (cash, credit card, check), and so on.

An Example: Johnson Filtration, Inc.

(a) (Background) Johnson Filtration, Inc., provides maintenance service for water-filtration systems throughout southern Florida. Customers contact Johnson with requests for maintenance service on their water-filtration systems. To estimate the service time and the service cost, Johnson′ s managers want to predict the repair time necessary for each maintenance request.

(b) (Dependent variable/Independent variables) Hence, repair time in hours is the dependent variable. Repair time is believed to be related to two factors, the number of months since the last maintenance service and the type of repair problem (mechanical or electrical).

(c) (Data)(Table 15.5)

TABLE 15.5 Data for the Johnson Filtration Example

Service Call	Months Since Last Service	Type of Repair	Repair Time in Hours
1	2	Electrical	2.9
2	6	Mechanical	3.0
3	8	Electrical	4.8
4	3	Mechanical	1.8
5	2	Electrical	2.9
6	7	Electrical	4.9
7	9	Mechanical	4.2
8	8	Mechanical	4.8
9	4	Electrical	4.4
10	6	Electrical	4.5

(d) (SLR) Let y denote the repair time in hours and x_1 denote the number of months since the last maintenance service. The regression model that uses only x_1 to predict y is $y = \beta_0 + \beta_1 x_1 + \epsilon$

(e) (Figure 15.7)

FIGURE 15.7 Output for Johnson Filtration with Months Since Last Service (x_1) as the Independent Variable

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5.596	5.5960	9.17	.016
Error	8	4.880	.6100		
Total	9	10.476			

Model Summary

S	R-sq	R-sq (adj)
.781022	53.42%	47.59%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	2.147	.605	3.55	.008
Months Since Last Service	.304	.100	3.03	.016

Regression Equation

Repair Time (hours) = 2.147 + .304 Months Since Last Service

i. The estimated regression equation is ______________.

ii. At the 0.05 level of significance, the p-value of ________ for the t (or F) test indicates that the number of months since the last service is significantly related to repair time.

iii. R-sq = ________ indicates that x_1 alone explains ________ of the ________ in repair time.

4. To incorporate the type of repair into the regression model, we define

$$x_2 = \begin{cases} \_\_\_\_\_, & \text{if the type of repair is mechanical} \\ \_\_\_\_\_, & \text{if the type of repair is electrical} \end{cases}$$

5. In regression analysis x_2 is called a ________ or ______________.

6. Using this dummy variable, we can write the multiple regression model as

$$y = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

7. (Table 15.6) Data for the Johnson Filtration Example with Type of Repair Indicated by a Dummy Variable ($x_2 = 0$ for Mechanical; $x_2 = 1$ for Electrical)

TABLE 15.6 Data for the Johnson Filtration Example with Type of Repair Indicated by a Dummy Variable (x_2 = 0 for Mechanical; x_2 = 1 for Electrical)

Customer	Months Since Last Service (x_1)	Type of Repair (x_2)	Repair Time in Hours (y)
1	2	1	2.9
2	6	0	3.0
3	8	1	4.8
4	3	0	1.8
5	2	1	2.9
6	7	1	4.9
7	9	0	4.2
8	8	0	4.8
9	4	1	4.4
10	6	1	4.5

8. (Figure 15.7) Output for Johnson Filtration with Months Since Last Service (x_1) as the Independent Variable

FIGURE 15.8 Output for Johnson Filtration with Months Since Last Service (x_1) and Type of Repair (x_2) as the Independent Variables

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	9.0009	4.50046	21.36	.001
Error	7	1.4751	.21073		
Total	9	10.4760			

Model Summary

S	R-sq	R-sq (adj)
.459048	85.92%	81.90%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	.930	.467	1.99	.087
Months Since Last Service	.3876	.0626	6.20	.000
Type of Repair	1.263	.314	4.02	.005

Regression Equation

Repair Time (hours) = .930 + .3876 Months Since Last Service + 1.263 Type of Repair

(a) The estimated multiple regression equation is

$$\underline{\hspace{6cm}} \qquad (15.17)$$

(b) At the 0.05 level of significance, the p-value of ________ associated with the F test (____________) indicates that the regression relationship is significant.

(c) The t test shows that both months since last service (p-value = __________) and type of repair (p-value = __________) are statistically significant.

(d) In addition, R-Sq = __________ and R-Sq (adj) = __________ indicate that the estimated regression equation does a good job of explaining the variability in repair times.

(e) Thus, equation (15.17) should prove helpful in predicting the repair time necessary for the various service calls.

Interpreting the Parameters

1. The multiple regression equation for the Johnson Filtration example is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (15.18)$$

2. Consider the case when $x_2 = 0$ (mechanical repair). Using ____________________ to denote the mean or expected value of repair time given a mechanical repair, we have

$$E(y|\text{mechanical}) = \underline{\qquad\qquad\qquad} = \underline{\qquad\qquad} \qquad (15.19)$$

3. Similarly, for an electrical repair ($x_2 = 1$), we have

$$E(y|\text{electrical}) = \underline{\qquad\qquad\qquad} = \underline{\qquad\qquad\qquad} \qquad (15.20)$$

4. Comparing equations (15.19) and (15.20), we see that the mean repair time is a linear function of _______ for both mechanical and electrical repairs. The slope of both equations is _______, but the _______________ differs.

5. The y-intercept is _______ in equation (15.19) for mechanical repairs and _____________ in equation (15.20) for electrical repairs.

6. The interpretation of β_2 is that it indicates the ______________ between the _____________________ for an electrical repair and the mean repair time for a mechanical repair.

(a) If ___________, the mean repair time for an electrical repair will be ________________ that for a mechanical repair;

(b) if ________, the mean repair time for an electrical repair will be ________ that for a mechanical repair.

(c) if ________, there is ________ in the mean repair time between electrical and mechanical repairs and the type of repair is ________ to the repair time.

7. Using the estimated multiple regression equation $\hat{y} = 0.93 + 0.3876x_1 + 1.263x_2$, we see that 0.93 is the estimate of β_0 and 1.263 is the estimate of β_2.

8. Thus, when $x_2 = 0$ (mechanical repair)

$$\hat{y} = 0.93 + 0.3876x_1 \qquad (15.21)$$

and when $x_2 = 1$ (electrical repair)

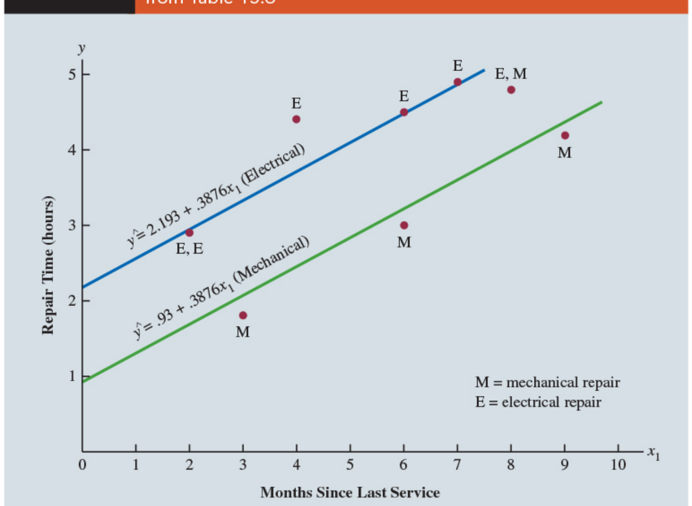
$$\hat{y} = 0.93 + 0.3876x_1 + 1.263(1) = 2.193 + 0.3876x_1 \qquad (15.22)$$

9. In effect, the use of a dummy variable for type of repair provides ________ ________ that can be used to predict the repair time, one corresponding to mechanical repairs and one corresponding to electrical repairs.

10. In addition, with $\beta_2 = 1.263$, we learn that, on average, electrical repairs require ________ than mechanical repairs.

11. (Figure 15.9) Scatter Diagram for the Johnson Filtration Repair Data

FIGURE 15.9 Scatter Diagram for the Johnson Filtration Repair Data from Table 15.6



More Complex Categorical Variables

1. If a categorical variable has k levels, $k-1$ dummy variables are required, with each dummy variable being coded as ________.

2. Example Suppose a manufacturer of copy machines organized the sales territories for a particular state into three regions: A, B, and C. The managers want to use regression analysis to help predict the number of copiers sold per week.

3. With the number of units sold as the dependent variable, they are considering several independent variables (the number of sales personnel, advertising expenditures, and so on).

4. Suppose the managers believe sales region is also an important factor in predicting the number of copiers sold. Because sales region is a categorical variable with three levels, A, B and C, we will need __________ dummy variables to represent the sales region. Each variable can be coded 0 or 1:

$$x_1 = \begin{cases} 1, & \text{if sales region B} \\ 0, & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{if sales region C} \\ 0, & \text{otherwise} \end{cases}$$

5. We have the following values of x_1 and x_2:

Region	x_1	x_2
A	0	0
B	1	0
C	0	1

6. Observations corresponding to region A would be coded ______________; observations corresponding to region B would be coded ______________; and observations corresponding to region C would be coded ______________.

7. The regression equation relating the expected value of the number of units sold, $E(y)$, to the dummy variables would be written as

$$E(y) = \underline{\hspace{4cm}}$$

8. To help us interpret the parameters β_0, β_1, and β_2, consider the following three variations of the regression equation.

$$\begin{aligned} E(y|\text{region A}) &= \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0 \\ E(y|\text{region B}) &= \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1 \\ E(y|\text{region C}) &= \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2 \end{aligned}$$

   (a) Thus, β_0 is the mean or expected value of sales for ____________;

   (b) β_1 is the ____________ between the mean number of units sold in ____________ and the mean number of units sold in ____________;

   (c) and β_2 is the ____________ between the mean number of units sold in ____________ and the mean number of units sold in ____________.

9. Two dummy variables were required because sales region is a categorical variable with three levels.

10. The assignment was ____________. For example, we could have chosen $x_1 = 1, x_2 = 0$ to indicate region A, $x_1 = 0, x_2 = 0$ to indicate region B, and $x_1 = 0, x_2 = 1$ to indicate region C.

Region	x_1	x_2
A	1	0
B	0	0
C	0	1

In that case, β_1 would have been interpreted as the mean difference between regions A and B and β_2 as the mean difference between regions C and B.

11. The important point to remember is that when a categorical variable has k levels, $k-1$ dummy variables are required in the multiple regression analysis. Thus, if the sales region example had a fourth region, labeled D, three dummy variables would be necessary. For example, the three dummy variables can be coded as follows.

$$x_1 = \begin{cases} 1, & \text{if sales region B} \\ 0, & \text{otherwise} \end{cases} \quad x_2 = \begin{cases} 1, & \text{if sales region C} \\ 0, & \text{otherwise} \end{cases} \quad x_3 = \begin{cases} 1, & \text{if sales region D} \\ 0, & \text{otherwise} \end{cases}$$

15.8 Residual Analysis

1. **Standardized Residual for Observation** i

$$\frac{\qquad}{\qquad} \qquad (15.23)$$

where $s_{y_i-\hat{y}_i}$ is the standard deviation of residual i.

2. **Standard Deviation of Residual** i

$$s_{y_i-\hat{y}_i} = \underline{\qquad\qquad} \qquad (15.24)$$

where s is the standard error of the estimate and h_i is the ________ of observation i. (________, ______________)

3. (Chapter 14) the leverage of an observation is determined by how far the values of the ________________ are from their _______.

4. The computation of h_i, $s_{y_i-\hat{y}_i}$, and hence the standardized residual for observation i in multiple regression analysis is too complex to be done by hand. However, the standardized residuals can be easily obtained as part of the output from statistical software.

5. Example Butler Trucking example

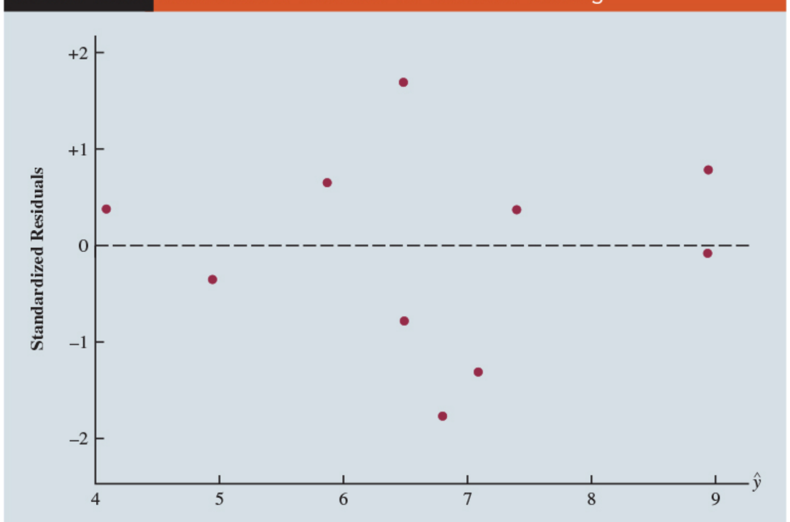
   (a) (Table 15.7) the estimated regression equation $\hat{y} = -0.869 + 0.03821x_1 + 0.923x_2$.

TABLE 15.7 Residuals and Standardized Residuals for the Butler Trucking Regression Analysis

Kilometers Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Predicted Value ($\hat{y}$)	Residual ($y - \hat{y}$)	Standardized Residual
160	4	9.3	8.93846	.361541	.78344
80	3	4.8	4.95830	−.158304	−.34962
160	4	8.9	8.93846	−.038460	−.08334
160	2	6.5	7.09161	−.591609	−1.30929
80	2	4.2	4.03488	.165121	.38167
128	2	6.2	5.86892	.331083	.65431
120	3	7.4	6.48667	.913331	1.68917
104	4	6.0	6.79875	−.798749	−1.77372
144	3	7.6	7.40369	.196311	.36703
144	2	6.1	6.48026	−.380263	−.77639

(b) (Figure 15.10) This standardized residual plot does not indicate any unusual abnormalities. All the standardized residuals are between ______________; hence, we have no reason to question the assumption that the error term ϵ is normally distributed. We conclude that the model assumptions are ______________.

FIGURE 15.10 Standardized Residual Plot for Butler Trucking



(c) (Recall Section 14.8) A ________________________ also can be used to determine whether the distribution of ϵ appears to be normal. The same procedure is appropriate for multiple regression.

Detecting Outliers

1. An outlier is an observation that is ____________ in comparison with the other data. An outlier does not fit the ____________ of the other data.

2. (Chapter 14) An observation is classified as an outlier if the value of its ________________________ is less than −2 or greater than +2.

3. (Table 15.7) Applying this rule to the standardized residuals for the Butler Trucking example, We do not detect any outliers in the data set.

4. In general, the presence of one or more outliers in a data set tends to increase ______, the standard error of the estimate, and hence increase __________, the standard deviation of residual i.

5. Because $s_{y_i-\hat{y}_i}$ appears in the denominator of the formula for the standardized residual (15.23), the size of the standardized residual will ____________ as s ____________. As a result, even though a residual may be unusually large, the large denominator in expression (15.23) may cause the standardized residual rule to fail to identify the observation as being an outlier.

6. We can circumvent this difficulty by using a form of the standardized residuals called ______________________________.

Studentized Deleted Residuals and Outliers

1. Suppose the ith observation is deleted from the data set and a new estimated regression equation is developed with the remaining $n-1$ observations.

2. Let ________ denote the standard error of the estimate based on the data set with the ________ observation deleted. If we compute the standard deviation of residual i using $s_{(i)}$ instead of s, and then compute the standardized residual for observation i using the ________________ value, the resulting standardized residual is called a ______________________________.

3. If the ith observation is an outlier, $s_{(i)}$ will be ________ than s. The absolute value of the ith studentized deleted residual therefore will be ______________ the absolute value of the standardized residual.

4. Studentized deleted residuals may detect outliers that standardized residuals do not detect.

5. The t distribution can be used to determine whether the studentized deleted residuals indicate the presence of outliers.

   (a) If we delete the ith observation, the number of observations in the reduced data set is $n-1$; in this case the error sum of squares has ________ degrees of freedom.

   (b) Example For the Butler Trucking example with $n = 10$ and $p = 2$, the degrees of freedom for the error sum of squares with the ith observation deleted is $9-2-1 = 6$. At $\alpha = 0.05$ level of significance, the t distribution shows that with six degrees of freedom, ________.

   (c) If the value of the ith studentized deleted residual is ________ or ________, we can conclude that the ith observation is an outlier.

   (d) (Table 15.8) Butler Trucking example, outliers are not present in the data set.

TABLE 15.8 Studentized Deleted Residuals for Butler Trucking

Kilometers Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Standardized Residual	Studentized Deleted Residual
160	4	9.3	.78344	.75939
80	3	4.8	−.34962	−.32654
160	4	8.9	−.08334	−.07720
160	2	6.5	−1.30929	−1.39494
80	2	4.2	.38167	.35709
128	2	6.2	.65431	.62519
120	3	7.4	1.68917	2.03187
104	4	6.0	−1.77372	−2.21314
144	3	7.6	.36703	.34312
144	2	6.1	−.77639	−.75190

Influential Observations

1. (Section 14.9) we discussed how the leverage of an observation can be used to identify observations for which the value of the ________ variable may have a strong

influence on the regression results.

2. The leverage of an observation, denoted h_i, measures how far the values of the independent variables are from their mean values.

3. We use the rule of thumb ________________ to identify influential observations.

4. Example Butler Trucking example ($n = 10, p = 2$)

   (a) The critical value for leverage is $3(2 + 1)/10 = 0.9$.

   (b) (Table 15.9) Because h_i does not exceed 0.9, we do not detect influential observations in the data set.

TABLE 15.9 Leverage and Cook's Distance Measures for Butler Trucking

Kilometers Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Leverage (h_i)	Cook's D (D_i)
160	4	9.3	.351704	.110994
80	3	4.8	.375863	.024536
160	4	8.9	.351704	.001256
160	2	6.5	.378451	.347923
80	2	4.2	.430220	.036663
128	2	6.2	.220557	.040381
120	3	7.4	.110009	.117562
104	4	6.0	.382657	.650029
144	3	7.6	.129098	.006656
144	2	6.1	.269737	.074217

Using Cook′ s Distance Measure to Identify

1. A problem that can arise in using leverage to identify influential observations is that an observation can be identified as having ________________ and not necessarily be influential in terms of the resulting ______________________________.

   (a) (Table 15.10) Because the leverage for the eighth observation is ______________ (the critical leverage value), this observation is identified as influential.

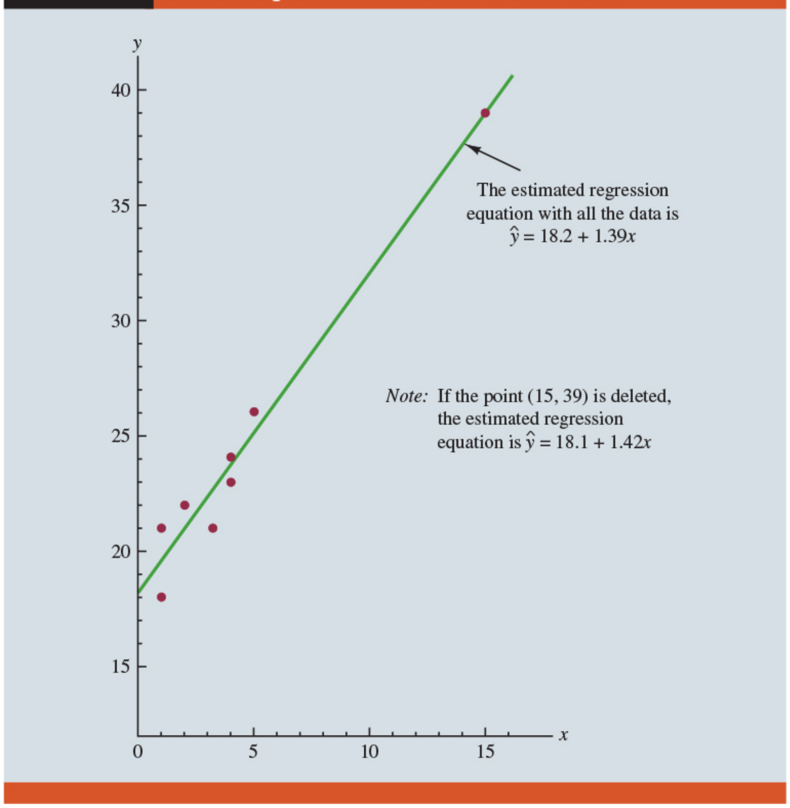
TABLE 15.10

Data Set Illustrating Potential Problem Using the Leverage Criterion

x_i	y_i	Leverage h_i
1	18	.204170
1	21	.204170
2	22	.164205
3	21	.138141
4	23	.125977
4	24	.125977
5	26	.127715
15	39	.909644

(b) (Figure 15.11) the estimated regression equation: $\hat{y} = 18.2 + 1.39x$

FIGURE 15.11 Scatter Diagram for the Data Set in Table 15.10



(c) Delete the observation $x = 15, y = 39$ from the data set and fit a new estimated regression equation to the remaining seven observations; the new estimated regression equation is $\hat{y} = 18.1 + 1.42x$

(d) We note that the y-intercept and slope of the new estimated regression equation are very close to the values obtained using all the data.

(e) Although the leverage criterion identified the eighth observation as influential, this observation clearly had little influence on the results obtained. Thus, in some situations using only leverage to identify influential observations can lead to wrong conclusions.

2. **Cook′ s distance measure** uses both the leverage of observation i, h_i, and the residual for observation i, $(y_i - \hat{y}_i)$, to determine whether the observation is influential.

$$D_i = \underline{\hspace{6cm}}$$

(a) The value of Cook′ s distance measure will be large and indicate an influential observation if the residual or the leverage is large.

(b) As a rule of thumb, values of ______ indicate that the ith observation is influential and should be studied further.

(c) Example (Table 15.9) Cook′ s distance measure for the Butler Trucking problem. Observation 8 with $D_i = 0.650029 < 1$, we should not be concerned about the presence of influential observations in the Butler Trucking data set.

15.9 Logistic Regression

1. In many regression applications, the dependent variable may only assume ______________.

2. Example A bank might want to develop an estimated regression equation for predicting whether a person will be approved for a credit card. The dependent variable can be coded as ______ if the bank ________ the request for a credit card and ______ if the bank ______ the request for a credit card.

3. Using ____________ we can estimate the _________ that the bank will approve the request for a credit card given a particular set of values for the chosen independent variables.

4. Example Simmons Stores. Let us consider an application of logistic regression involving a direct mail promotion being used by Simmons Stores.

(a) Simmons owns and operates a national chain of women's apparel stores. Five thousand copies of an expensive four-color sales catalog have been printed, and each catalog includes a coupon that provides a $50 discount on purchases of $200 or more. The catalogs are expensive and Simmons would like to send them to only those customers who have a high probability of using the coupon.

(b) Management believes that annual spending at Simmons Stores and whether a customer has a Simmons credit card are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon.

(c) Simmons conducted a pilot study using a random sample of 50 Simmons credit card customers and 50 other customers who do not have a Simmons credit card. Simmons sent the catalog to each of the 100 customers selected. At the end of a test period, Simmons noted whether each customer had used her or his coupon.

(d) (Table 15.11) The amount each customer spent last year at Simmons is shown in thousands of dollars and the credit card information has been coded as 1 if the customer has a Simmons credit card and 0 if not. In the Coupon column, a 1 is recorded if the sampled customer used the coupon and 0 if not.

TABLE 15.11 Partial Sample Data for the Simmons Stores Example

Customer	Annual Spending ($1000)	Simmons Card	Coupon
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

(e) We might think of building a ________________ model using the data in Table 15.11 to help Simmons estimate whether a catalog recipient will use the coupon. We would use Annual Spending ($1000) and Simmons Card as independent variables and Coupon as the dependent variable.

5. Because the dependent variable may only assume the values of 0 or 1, however, the ______________________________ model is not applicable. This example shows the type of situation for which logistic regression was developed.

Logistic Regression Equation

1. In multiple regression analysis, the mean or expected value of y is referred to as the multiple regression equation.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (15.26)$$

2. (**Logistic Regression Equation**) In logistic regression, statistical theory as well as practice has shown that the relationship between $E(y)$ and $x_1, x_2, \cdots, x_p$ is better described by the following nonlinear equation.

$$E(y) = \underline{\hspace{6cm}} \qquad (15.27)$$

3. If the two values of the dependent variable y are coded as 0 or 1, the value of $E(y)$ in equation (15.27) provides the ________________________ given a particular set of values for the independent variables $x_1, x_2, \cdots, x_p$.

4. Because of the interpretation of $E(y)$ as a probability, the logistic regression equation is often written:

$$E(y) = \underline{\hspace{6cm}} \qquad (15.28)$$

5. Example Suppose the model involves only one independent variable x and the values of the model parameters are $\beta_0 = -7$ and $\beta_1 = 3$. The logistic regression equation corresponding to these parameter values is

$$E(y) = \underline{\hspace{9cm}} \qquad (15.29)$$

   (a) (Figure 15.12) shows a graph of equation (15.29). Note that the graph is _____________. The value of $E(y)$ ranges from __________.

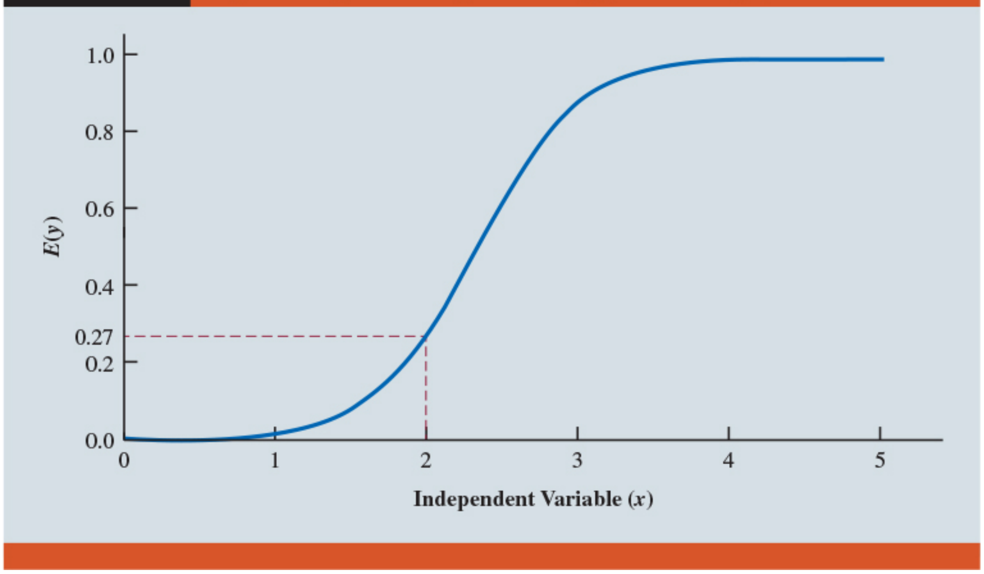


FIGURE 15.12 Logistic Regression Equation for $\beta_0 = -7$ and $\beta_1 = 3$

(b) For example, when $x = 2$, $E(y)$ is approximately 0.27. Also note that the value of $E(y)$ gradually approaches ______ as the value of x becomes __________ and the value of $E(y)$ approaches ______ as the value of x becomes ___________.

(c) For example, when $x = 2$, $E(y) = 0.269$. Note also that the values of $E(y)$, representing _______________, increase fairly rapidly as x ________________________. The fact that the values of $E(y)$ range from 0 to 1 and that the curve is S-shaped makes equation (15.29) ideally suited to model the probability the dependent variable is equal to 1.

Estimating the Logistic Regression Equation

1. The ___________________ of the logistic regression equation makes the method of computing estimates more complex and beyond the scope of this text. We use statistical _____________ to provide the estimates.

2. The estimated logistic regression equation is

$$\hat{y} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \qquad (15.30)$$

3. Here, $\hat{y}$ provides an ____ given a particular set of values for the independent variables.

4. Example Simmons Stores

(a) The variables are defined:

$$y = \begin{cases} \_\_\_\_\_\_ & \text{if the customer did not use the coupon} \\ \_\_\_\_\_\_ & \text{if the customer used the coupon} \end{cases}$$

$$x_1 = \text{ annual spending at Simmons Stores (\$1000s)}$$

$$x_2 = \begin{cases} \_\_\_\_\_\_ & \text{if the customer does not have a Simmons credit card} \\ \_\_\_\_\_\_ & \text{if the customer has a Simmons credit card} \end{cases}$$

(b) Thus, we choose a logistic regression equation with two independent variables.

$$E(y) = \frac{}{\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_} \qquad (15.31)$$

Using the sample data (see Table 15.11), we used statistical software to compute estimates of the model parameters b_0, b_1, and b_2.

(c) (Figure 15.13)

FIGURE 15.13 Logistic Regression Output for the Simmons Stores Example

Significance Tests

Term	Degrees of Freedom	χ^2	p-Value
Whole Model	2	13.63	.0011
Spending	1	7.56	.0060
Card	1	6.41	.0013

Parameter Estimates

Term	Estimate	Standard Error
Intercept	−2.146	.577
Spending	.342	.129
Card	1.099	.44

Odds Ratios

Term	Odds Ratio	Lower 95%	Upper 95%
Spending	1.4073	1.0936	1.8109
Card	3.0000	1.2550	7.1730

(d) We see that $\beta_0 = -2.146$, $\beta_1 = 0.342$, and $\beta_2 = 1.099$. Thus, the estimated logistic regression equation is

$$\hat{y} = \frac{e^{-2.146+0.342x_1+1.099x_2}}{1+e^{-2.146+0.342x_1+1.099x_2}} \qquad (15.32)$$

(e) An estimate the probability of using the coupon for customers who spend \$2000 annually and do not have a Simmons credit card is approximately 0.19 ($x_1 = 2$ and $x_2 = 0$):

$$\hat{y} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

(f) An estimate the probability of using the coupon for customers who spent \$2000 last year and have a Simmons credit card is approximately 0.41. ($x_1 = 2$ and $x_2 = 1$):

$$\hat{y} = \frac{e^{-2.146+0.342(2)+1.099(1)}}{1+e^{-2.146+0.342(2)+1.099(1)}} = \frac{e^{-0.363}}{1+e^{-0.363}} = 0.4102$$

(g) It appears that the probability of using the coupon is ____________ for customers with a Simmons credit card.

(h) Before reaching any conclusions, however, we need to assess the statistical ________________.

Testing for Significance

1. Testing for significance in logistic regression is similar to testing for significance in multiple regression.

2. First we conduct a test for ______________. For the Simmons Stores example, the hypotheses for the test of overall significance follow:

   H_0 : ____________

   H_a : One or both of the parameters is not equal to zero

   (a) The test for overall significance is based upon the value of a ________ statistic. If the null hypothesis is true, the sampling distribution of χ^2 follows a chi-square distribution with degrees of freedom equal to the ____________________ in the model.

   (b) (Figure 15.13) The calculations of χ^2 is __________________. The value of χ^2 and its corresponding p-value in the Whole Model row of the Significance Tests table is 13.63 and its p-value is 0.0011. Thus, at any level of significance $\alpha \geq 0.0011$, we would reject the null hypothesis and conclude that the overall model is significant.

(c) NOTE:

i. Logistic Regression: https://online.stat.psu.edu/stat462/node/207/

ii. Logistic regression (Wikipedia): https://en.wikipedia.org/wiki/Logistic_regression

3. If the χ^2 test shows an overall significance, another __________ can be used to determine whether each of the ____________ independent variables is making a significant contribution to the overall model.

(a) For the independent variables x_i, the hypotheses are

____

(b) If the null hypothesis is true, the sampling distribution of χ^2 follows a chi-square distribution with one degree of freedom.

(c) (Figure 15.13) The Spending and Card rows of the Significance Tests table of Figure 15.13 contain the values of χ^2 and their corresponding p-values test for the estimated coefficients. Suppose we use $\alpha = 0.05$ to test for the significance of the independent variables in the Simmons model.

(d) For the independent variable Spending (x_1) the χ^2 value is ________ and the corresponding p-value is __________. Thus, at the 0.05 level of significance we can reject $H_0 : \beta_1 = 0$.

(e) In a similar fashion we can also reject $H_0 : \beta_2 = 0$ because the p-value corresponding to Card's ____________ is __________. Hence, at the 0.05 level of significance, both independent variables are statistically significant.

Managerial Use

1. We described how to develop the estimated logistic regression equation and how to test it for significance.

2. Example For Simmons Stores, we already computed $P(y = 1|x_1 = 2, x_2 = 1) = 0.4102$ and $P(y = 1|x_1 = 2, x_2 = 0) = 0.1881$. These probabilities indicate that for customers with annual spending of \$2000 the presence of a Simmons credit card ________________________ of using the coupon.

3. (Table 15.12) The estimated probabilities for values of annual spending ranging from \$1000 to \$7000 for both customers who have a Simmons credit card and customers who do not have a Simmons credit card.

TABLE 15.12 Estimated Probabilities for Simmons Stores

		Annual Spending						
		\$1000	\$2000	\$3000	\$4000	\$5000	\$6000	\$7000
Credit Card	Yes	.3307	.4102	.4948	.5796	.6599	.7320	.7936
	No	.1414	.1881	.2460	.3148	.3927	.4765	.5617

4. How can Simmons use this information to better target customers for the new promotion? Suppose Simmons wants to send the promotional catalog only to customers who have a ________________ probability of using the coupon. Using the estimated probabilities in Table 15.12, Simmons promotion strategy would be:

   (a) Customers who have a Simmons credit card: Send the catalog to every customer who spent a\$2000 or more last year.

   (b) Customers who do not have a Simmons credit card: Send the catalog to every customer who spent _________ or more last year.

5. The probability of using the coupon for customers who do not have a Simmons credit card but spend \$5000 annually is _________. Thus, Simmons may want to consider revising this strategy by including those customers who ______________ a credit card, as long as they spent _________ or more last year.

Interpreting the Logistic Regression Equation

1. With logistic regression, it is difficult to interpret the relation between the independent variables and the ______________________ directly because the logistic regression equation is ____________.

2. The relationship can be interpreted indirectly using a concept called the _____________(勝算比).

3. The ________(勝算) in favor of an event occurring is defined as the probability the event ____________ divided by the probability the event ______________. In logistic regression the event of interest is always ________.

4. Given a particular set of values for the independent variables, the odds in favor of $y = 1$ can be calculated as follows:

$$odds = \frac{\qquad\qquad}{\qquad\qquad} = \frac{\qquad\qquad}{\qquad\qquad} \tag{15.33}$$

5. The odds ratio is the odds that $y = 1$ given that one of the independent variables has been increased by ______________ divided by the odds that $y = 1$ given ____________ in the values for the independent variables __________.

   (a) **Odds Ratio**

   $$\text{Odds Ratio} = \frac{\qquad}{\qquad} \tag{15.34}$$

   (b) For example, suppose we want to compare the odds of using the coupon for customers who spend \$2000 annually and have a Simmons credit card ($x_1 = 2$ and $x_2 = 1$) to the odds of using the coupon for customers who spend \$2000 annually and do not have a Simmons credit card ($x_1 = 2$ and $x_2 = 0$).

   (c) We are interested in interpreting the effect of a one-unit increase in the independent variable x_2. In this case

   $$odd_{s1} = \frac{\qquad\qquad}{\qquad\qquad}$$

   and

   $$odd_{s0} = \frac{\qquad\qquad}{\qquad\qquad}$$

   (d) Previously we showed that an estimate of the probability that $y = 1$ given $x_1 = 2$ and $x_2 = 1$ is 0.4102, and an estimate of the probability that $y = 1$ given $x_1 = 2$ and $x_2 = 0$ is 0.1881. Thus,

   $$\text{estimate of } odd_{s1} = \frac{0.4102}{1 - 0.4102} = 0.6956$$

   and

   $$\text{estimate of } odd_{s2} = \frac{0.1881}{1 - 0.1881} = 0.2318$$

   The estimated odds ratio is

   $$\text{estimated odds ratio} = \frac{0.6956}{0.2318} = 3.00$$

(e) Thus, we can conclude that the ________________ in favor of using the coupon for customers who spent \$2000 last year and have a Simmons credit card are ____________________ the estimated odds in favor of using the coupon for customers who spent \$2000 last year and do not have a Simmons credit card.

6. The odds ratio measures the impact on the odds of a one-unit increase in ___________ of the independent variables.

7. The odds ratio for each independent variable is computed while holding all the other independent variables ___________. But it does not matter what constant values are used for the other independent variables. For instance, if we computed the odds ratio for the Simmons credit card variable (x_2) using \$3000, instead of \$2000, as the value for the annual spending variable (x_1), we would still obtain the _____________ for the estimated odds ratio (3.00). Thus, we can conclude that the estimated odds of using the coupon for customers who have a Simmons credit card are 3 times greater than the estimated odds of using the coupon for customers who do not have a Simmons credit card.

8. (Figure 15.13) the estimated odds ratios for each of the independent variables. The estimated odds ratio for Spending (x_1) is _________ and the estimated odds ratio for Card (x_2) is _________.

9. Let us now consider the interpretation of the estimated odds ratio for the continuous independent variable x_1. The value of 1.4073 in the Odds Ratio column of the output tells us that the ________________ in favor of using the coupon for customers who spent \$3000 last year is ________________________ the estimated odds in favor of using the coupon for customers who spent \$2000 last year.

10. A unique relationship exists between the ____________ for a variable and its corresponding ____________________. For each independent variable in a logistic regression equation it can be shown that

____

(a) To illustrate this relationship, consider the independent variable x_1 in the Simmons example. The estimated odds ratio for x_1 is

$$\text{Estimated odds ratio } = e^{b_1} = e^{0.342} = 1.407$$

Similarly, the estimated odds ratio for x_2 is

$$\text{Estimated odds ratio } = e^{b_2} = e^{1.099} = 3.000$$

(b) 補充:

$$\begin{aligned}
\hat{p} &= \frac{e^{b_0+b_1x_1}}{1+e^{b_0+b_1x_1}}, \quad 1-\hat{p} = \frac{1}{1+e^{b_0+b_1x_1}} \\
\ln(\hat{p}) - \ln(1-\hat{p}) &= \ln(e^{b_0+b_1x_1}) - \ln(1+e^{b_0+b_1x_1}) - \ln(1) + \ln(1+e^{b_0+b_1x_1}) \\
\ln(\frac{\hat{p}}{1-\hat{p}}) &= b_0 + b_1x_1 \\
\frac{\hat{p}}{1-\hat{p}} &= e^{b_0+b_1x_1} \\
\frac{\hat{p}}{1-\hat{p}}|_{x_1=0} &= e^{b_0}, \qquad \frac{\hat{p}}{1-\hat{p}}|_{x_1=1} = e^{b_0+b_1} \\
&\Rightarrow \text{odds ratio}|_{x=1/x=0} = e^{b_1}
\end{aligned}$$

11. The odds ratio for an independent variable represents the ____________________ for a ___________ change in the independent variable holding all the other independent variables ___________.

(a) Suppose that we want to consider the effect of a change of more than one unit, say c units. For instance, suppose in the Simmons example that we want to compare the odds of using the coupon for customers who spend \$5000 annually ($x_1 = 5$) to the odds of using the coupon for customers who spend \$2000 annually ($x_1 = 2$). In this case $c = 5-2 = 3$ and the corresponding estimated odds ratio is

____

(b) This result indicates that the estimated odds of using the coupon for customers who spend \$5000 annually is ____________ greater than the estimated odds of using the coupon for customers who spend \$2000 annually.

(c) In other words, the estimated odds ratio for an increase of \$3000 in annual spending is 2.79.

(d) In general, the odds ratio enables us to compare the odds for two different events. If the value of the odds ratio is ______, the odds for both events are the same. Thus, if the independent variable we are considering (such as Simmons credit card status) has a ________________ on the probability of the event occurring, the corresponding odds ratio will be ________________.

12. (Figure 15.13) Most statistical software packages provide a confidence interval for the odds ratio. The Odds Ratio table in Figure 15.13 provides a 95% confidence interval for each of the odds ratios.

(a) For example, the point estimate of the odds ratio for x_1 is 1.4073 and the 95% confidence interval is __________________. Because the confidence interval does not contain the value of ______, we can conclude that x_1 has a _____________ relationship with the estimated odds ratio.

(b) Similarly, the 95% confidence interval for the odds ratio for x_2 is __________________. Because this interval does not contain the value of 1, we can also conclude that x_2 has a significant relationship with the odds ratio.

Logit Transformation

1. It can be shown that

____

2. This equation shows that the natural logarithm of the odds in favor of $y = 1$ is a linear function of the independent variables. This linear function is called the ________. We will use the notation __________________ to denote the logit.

3. Logit

$$g(x_1, x_2, \cdots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (15.35)$$

4. Substituting $g(x_1, x_2, \cdots, x_p)$ for $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ in equation (15.27), we can write the logistic regression equation as

$$E(y) = \frac{\qquad}{\qquad} \qquad (15.36)$$

5. Once we estimate the parameters in the logistic regression equation, we can compute an estimate of the logit. Using $\hat{g}(x_1, x_2, \cdots, x_p)$ to denote the estimated logit, we obtain

$$\text{Estimated Logit } \underline{\hspace{8cm}} \tag{15.37}$$

6. Thus, in terms of the estimated logit, the estimated regression equation is

$$\hat{y} = \frac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}{1+e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}} = \frac{e^{\hat{g}(x_1,x_2,\cdots,x_p)}}{1+e^{\hat{g}(x_1,x_2,\cdots,x_p)}} \tag{15.38}$$

7. For the Simmons Stores example, the estimated logit is

$$\underline{\hspace{7cm}}$$

   and the estimated regression equation is

$$\hat{y} = \underline{\hspace{3cm}} = \underline{\hspace{5cm}}$$

   Thus, because of the unique relationship between the estimated logit and the estimated logistic regression equation, we can compute the estimated probabilities for Simmons Stores by dividing $e^{\hat{g}(x_1,x_2)}$ by $1 + e^{\hat{g}(x_1,x_2)}$.

😍 EXERCISES

15.2 : 1, 5, 6

15.3 : 11, 14, 15

15.5 : 19, 23, 24

15.6 : 27, 29

15.7 : 32, 34, 35

15.8 : 40, 41

15.9 : 44, 46, 48

SUP : 51, 55.

“你無法改變別人的長相，但我們可以改變我們看人的方式。”
“You can not change someone’s looks, but we can change the way we look.”
— *奇蹟男孩 (Wonder, 2017)*

統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 17: Time Series Analysis and Forecasting

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: `http://www.hmwu.idv.tw`

系級: ____________________ 學號: ____________________ 姓名: ____________________

17.1 Time Series Patterns

1. **time series**: A ______________ is a sequence of observations on a variable measured at successive points in time or over successive periods of time.

2. The measurements may be taken every hour, day, week, month, or year, or at any other __________________. (this textbook limits the discussion to time series in which the values of the series are recorded at equal intervals)

3. The ___________ of the data is an important factor in understanding how the time series has behaved in the ________. If such behavior can be expected to continue in the __________, we can use the past pattern to guide us in selecting an appropriate ______________ method.

4. A __________________ is a graphical presentation of the relationship between time and the time series variable; ________ is on the horizontal axis and the time series __________ are shown on the vertical axis. A time series plot is useful to identify the underlying pattern in the data.

5. Some of the common types of data patterns that can be identified when examining a time series plot: horizontal pattern, trend pattern, seasonal pattern, trend and seasonal pattern, and cyclical pattern.

Horizontal Pattern

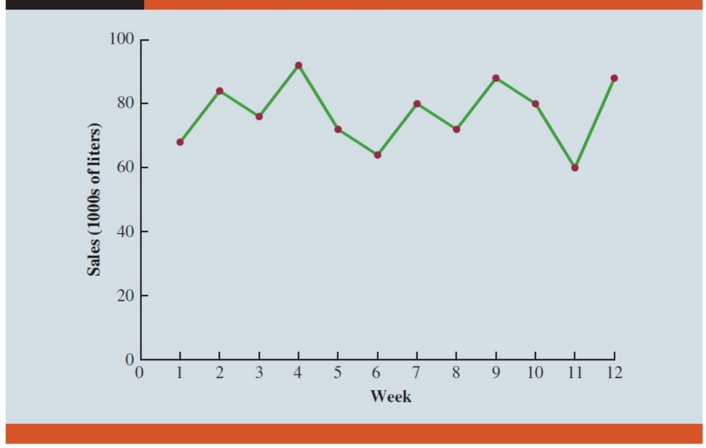
1. A horizontal pattern exists when the data ____________ around a _________________.

2. Example (Table 17.1) (Figure 17.1) These data show the number of gallons of gasoline sold by a gasoline distributor in Bennington, Vermont, over the past 12 weeks.

TABLE 17.1
Gasoline Sales Time Series

Week	Sales (1000s of gallons)
1	17
2	21
3	19
4	23
5	18
6	16
7	20
8	18
9	22
10	20
11	15
12	22

FIGURE 17.1 Gasoline Sales Time Series Plot



The average value or mean for this time series is 19.25 gallons (1000s) per week. Although ____________________ is present, we would say that these data follow a horizontal pattern.

3. The term _____________ time series is used to denote a time series whose statistical properties are _____________________.

4. In particular this means that

   (a) The process generating the data has a ________________.

   (b) The variability of the time series is ___________ over time.

5. A time series plot for a stationary time series will always exhibit a ___________________. But simply observing a horizontal pattern is not sufficient evidence to conclude that the time series is stationary.

6. More advanced texts on forecasting discuss procedures for determining if a time series is stationary and provide methods for transforming a time series that is not stationary into a stationary series.

7. Changes in business conditions can often result in a time series that has a horizontal pattern __________ to a new level.

   (a) Example For instance, suppose the gasoline distributor signs a contract with the Vermont State Police to provide gasoline for state police cars located in southern Vermont. With this new contract, the distributor expects to see a major increase in weekly sales starting in week 13.

   (b) (Table 17.2) The number of gallons of gasoline sold for the original time series and for the 10 weeks after signing the new contract.

TABLE 17.2

Gasoline Sales Time Series After Obtaining the Contract with the Vermont State Police

Week	Sales (1000s of liters)
1	68
2	84
3	76
4	92
5	72
6	64
7	80
8	72
9	88
10	80
11	60
12	88
13	124
14	136
15	124
16	132
17	112
18	128
19	120
20	116
21	136
22	132

FIGURE 17.2 Gasoline Sales Time Series Plot After Obtaining the Contract with the Vermont State Police

   (c) (Figure 17.2) Note the increased level of the time series beginning in week 13. This change in the level of the time series makes it more __________ to choose an appropriate forecasting method.

8. Selecting a forecasting method that adapts well to ____________________ of a time series is an important consideration in many practical applications.

Trend Pattern

1. Although time series data generally exhibit random fluctuations, a time series may also show gradual ______________________ to relatively higher or lower values over a __________ period of time.

2. If a time series plot exhibits this type of behavior, we say that a ________________ exists.

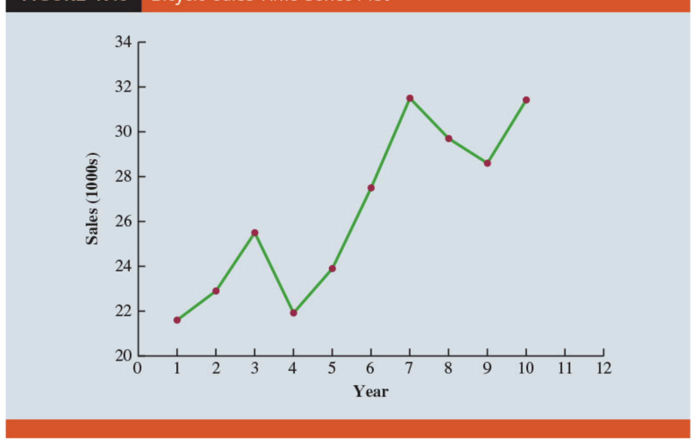
3. A trend is usually the result of ___________________ such as population increases or decreases, changing demographic characteristics of the population, technology, and/or consumer preferences.

4. Example (Table 17.3) (Figure 17.3) Consider the time series of bicycle sales for a particular manufacturer over the past 10 years.

TABLE 17.3
Bicycle Sales Time Series

Year	Sales (1000s)
1	21.6
2	22.9
3	25.5
4	21.9
5	23.9
6	27.5
7	31.5
8	29.7
9	28.6
10	31.4

FIGURE 17.3 Bicycle Sales Time Series Plot



Visual inspection of the time series plot shows some up and down movement over the past 10 years, but the time series also seems to have a __________________________ or ________________. The trend for the bicycle sales time series appears to be _________ and increasing over time.

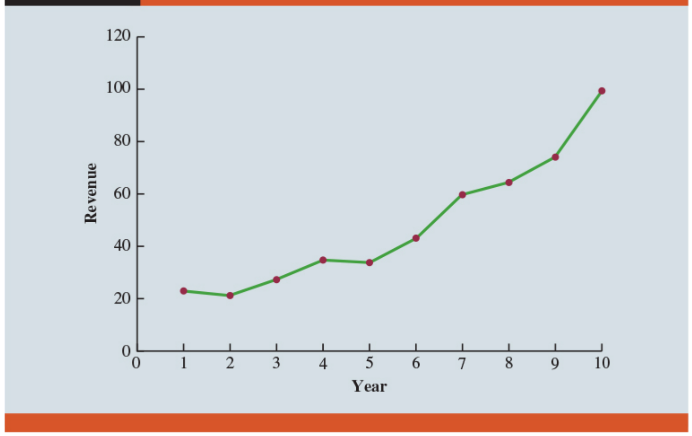
5. Example (Table 17.4) (Figure 17.4) The data show the sales for a cholesterol drug since the company won FDA approval for it 10 years ago.

TABLE 17.4

Cholesterol Revenue Time Series ($Millions)

Year	Revenue
1	23.1
2	21.3
3	27.4
4	34.6
5	33.8
6	43.2
7	59.5
8	64.4
9	74.2
10	99.3

FIGURE 17.4 Cholesterol Revenue Times Series Plot ($Millions)



The time series increases in a nonlinear fashion; that is, the ________________ of revenue does not increase by a constant amount from one year to the next. In fact, the revenue appears to be growing in an ______________ fashion.

6. Exponential relationships such as this are appropriate when the percentage change from one period to the next is relatively ___________.

Seasonal Pattern

1. The trend of a time series can be identified by analyzing multiyear movements in ________________. Seasonal patterns are recognized by seeing the ________________________ over successive periods of time.

2. Example For example, a manufacturer of swimming pools expects low sales activity in the fall and winter months, with peak sales in the spring and summer months. Manufacturers of snow removal equipment and heavy clothing, however, expect just the opposite yearly pattern.

3. The pattern for a time series plot that exhibits a repeating pattern over a one-year period due to seasonal influences is called a ___________ pattern.

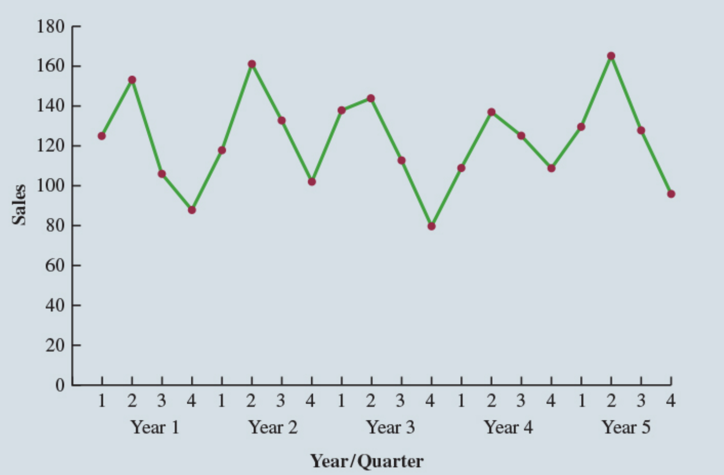
4. Example Daily traffic volume shows within-the-day "seasonal" behavior, with peak levels occurring during rush hours, moderate flow during the rest of the day and early evening, and light flow from midnight to early morning.

5. Example (Table 17.5) (Figure 17.5) As an example of a seasonal pattern, consider the number of umbrellas sold at a clothing store over the past five years.

TABLE 17.5 Umbrella Sales Time Series

Year	Quarter	Sales
1	1	125
	2	153
	3	106
	4	88
2	1	118
	2	161
	3	133
	4	102
3	1	138
	2	144
	3	113
	4	80
4	1	109
	2	137
	3	125
	4	109
5	1	130
	2	165
	3	128
	4	96

FIGURE 17.5 Umbrella Sales Time Series Plot



The time series plot does not indicate any ______________ in sales. The data follow a __________ pattern. But closer inspection of the time series plot reveals a ______________ in the data. That is, the first and third quarters have moderate sales, the second quarter has the highest sales, and the fourth quarter tends to have the lowest sales volume. Thus, we would conclude that a ________________ pattern is present.

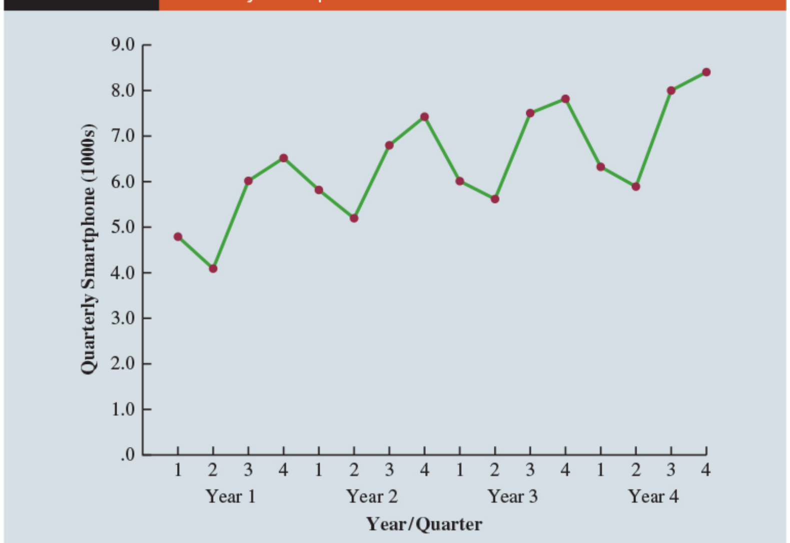
Trend and Seasonal Pattern

1. Some time series include a combination of a trend and seasonal pattern.
2. Example (Table 17.6) (Figure 17.6) The smartphone sales for a particular manufacturer over the past four years.

TABLE 17.6 Quarterly Smartphone Sales Time Series

Year	Quarter	Sales (1000s)
1	1	4.8
	2	4.1
	3	6.0
	4	6.5
2	1	5.8
	2	5.2
	3	6.8
	4	7.4
3	1	6.0
	2	5.6
	3	7.5
	4	7.8
4	1	6.3
	2	5.9
	3	8.0
	4	8.4

FIGURE 17.6 Quarterly Smartphone Sales Time Series Plot



3. Clearly, an increasing trend is present.

4. But, Figure 17.6 also indicates that sales are lowest in the second quarter of each year and increase in quarters 3 and 4. Thus, we conclude that a seasonal pattern also exists for smartphone sales.

5. In such cases we need to use a forecasting method that has the capability to deal with both ______________________.

Cyclical Pattern

1. A __________ pattern exists if the time series plot shows an alternating sequence of points below and above the ____________ lasting more than one year.

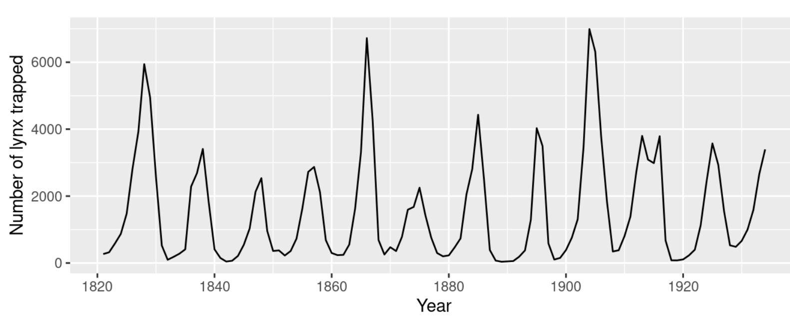
2. Often, the cyclical component of a time series is due to _________________________.

3. Example For example, periods of moderate inflation followed by periods of rapid inflation can lead to time series that alternate __________________ a generally increasing trend line (e.g., a time series for housing costs).

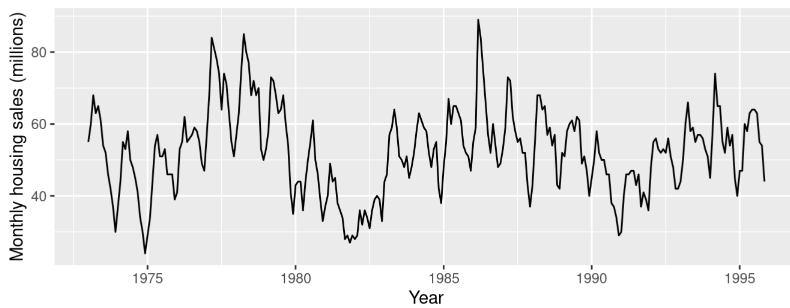
4. A cyclical pattern repeats with some ___________________________. Cyclical patterns differ from seasonal patterns in that cyclical patterns occur over multiple years, whereas seasonal patterns occur _________________.

5. More Example `https://robjhyndman.com/hyndsight/cyclicts/`

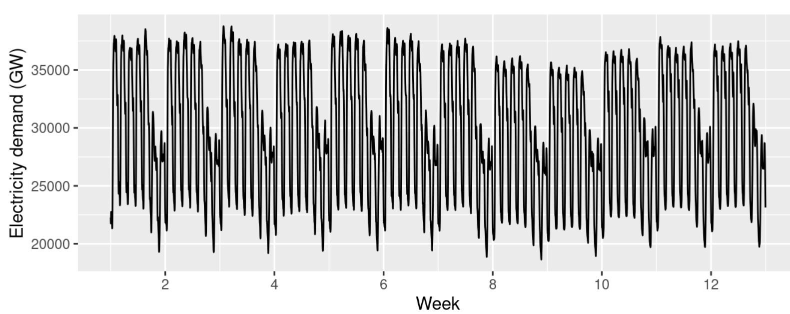
   (a) The plot shows the famous Canadian lynx (山貓) data –the number of lynx trapped each year in the McKenzie (麥肯錫) river district of northwest Canada (1821-1934). These show clear aperiodic (非週期性的) population cycles of approximately 10 years. The cycles are not of fixed length –some last 8 or 9 years and others last longer than 10 years.



(b) The plot shows the monthly sales of new one-family houses sold in the USA (1973-1995). There is strong seasonality within each year, as well as some strong cyclic behaviour with period about $6-10$ years.



(c) The plot shows half-hourly electricity demand in England and Wales from Monday 5 June 2000 to Sunday 27 August 2000. Here there are two types of seasonality –a _________ pattern and a ___________ pattern. If we collected data over a few years, we would also see there is an ___________ pattern. If we collected data over a few decades, we may even see a longer cyclic pattern.



6. Business cycles are extremely difficult, if not impossible, to forecast. As a result, cyclical effects are often combined with long-term trend effects and referred to as ______________________.

Selecting a Forecasting Method

1. The underlying pattern in the time series is an important factor in selecting a forecasting method. Thus, a ____________________ should be one of the first things developed when trying to determine which forecasting method to use.

2. The next two sections illustrate methods that can be used in situations where the underlying pattern is horizontal; in other words, no trend or seasonal effects are present. We then consider methods appropriate when trend and/or seasonality are present in the data.

17.2 Forecast Accuracy

1. The simplest of all the forecasting methods (a ________________________): an approach that uses the ______________ week's sales volume as the forecast for the next week.

2. (Table 17.7) The distributor sold 68 thousand gallons of gasoline in week 1; this value is used as the forecast for week 2. Next, we use 84, the actual value of sales in week 2, as the forecast for week 3, and so on.

TABLE 17.7 Computing Forecasts and Measures of Forecast Accuracy Using the Most Recent Value as the Forecast for the Next Period

Week	Time Series Value	Forecast	Forecast Error	Absolute Value of Forecast Error	Squared Forecast Error	Percentage Error	Absolute Value of Percentage Error
1	68						
2	84	68	16	16	256	19.05	19.05
3	76	84	−8	8	64	−10.53	10.53
4	92	76	16	16	256	17.39	17.39
5	72	92	−20	20	400	−27.78	27.78
6	64	72	−8	8	64	−12.50	12.50
7	80	64	16	16	256	20.00	20.00
8	72	80	−8	8	64	−11.11	11.11
9	88	72	16	16	256	18.18	18.18
10	80	88	−8	8	64	−10.00	10.00
11	60	80	−20	20	400	−33.33	33.33
12	88	60	28	28	784	31.82	31.82
		Totals	20	164	2864	1.19	211.69

3. The key concept associated with measuring forecast accuracy is __________________, defined as

____

(a) Example Because the distributor actually sold 84 thousand gallons of gasoline in week 2 and the forecast, using the sales volume in week 1, was 68 thousand gallons, the forecast error in week 2 is

$$\text{Forecast Error in week 2} = \underline{\hspace{3cm}}$$

(b) The fact that the forecast error is positive indicates that in week 2 the forecasting method ________________ the actual value of sales. Next, we use 84, the actual value of sales in week 2, as the forecast for week 3. Since the actual value of sales in week 3 is 76, the forecast error for week 3 is $76-84 = -8$. In this case, the negative forecast error indicates that in week 3 the forecast ________________ the actual value.

(c) A simple measure of forecast accuracy is the ________________________________. Table 17.7 shows that the sum of the forecast errors for the gasoline sales time series is 20; thus, the mean or average forecast error is ______________.

(d) Because positive and negative forecast errors tend to ________ one another, the mean error is likely to be small; thus, the mean error is not a very useful measure of forecast accuracy.

4. The ____________________, denoted ________, is the average of the absolute values of the forecast errors.

   (a) MAE is a measure of forecast accuracy that avoids the problem of positive and negative forecast errors offsetting one another.

   (b) (Table 17.7) the sum of the absolute values of the forecast errors is 164:

$$\text{MAE} = \text{average of the absolute value of forecast errors} = \underline{\hspace{3cm}}$$

5. Another measure that avoids the problem of positive and negative forecast errors offsetting each other is obtained by computing the average of the __________ forecast errors (____________________, is denoted ________.)

$$\text{MSE} = \text{average of the sum of squared forecast errors} = \underline{\hspace{3.5cm}}$$

6. The size of MAE and MSE depends upon the ________________. As a result, it is difficult to make comparisons for different time intervals, such as comparing

a method of forecasting monthly gasoline sales to a method of forecasting weekly sales, or to make comparisons across different time series.

7. The ______________________________, denoted __________, is a percentage error corresponding to the __________ of 84 in week 2 is computed by dividing the ______________ in week 2 by the ______________ in week 2 and multiplying the result by _______.

   (a) For week 2 the percentage error is computed as follows:

   $$\text{Percentage error for week 2} = \frac{}{} = 16/84 \times (100) = 19.05\%$$

   Thus, the forecast error for week 2 is 19.05% of the observed value in week 2.

   (b) The sum of the absolute values of the percentage errors is 211.69:

   $$\begin{aligned}\text{MAPE} &= \text{average of the absolute value of percentage forecast errors} \\ &= \underline{}\end{aligned}$$

8. Summarizing, using the naive (most recent observation) forecasting method, we obtained the following measures of forecast accuracy:

$$\text{MAE} = 3.73, \quad \text{MSE} = 16.27, \quad \text{MAPE} = 19.24\%$$

9. These measures of forecast accuracy simply measure how well the forecasting method is able to ________________________ of the time series.

10. Suppose we want to forecast sales for a ___________________, such as week 13. In this case the forecast for week 13 is 88, the actual value of the time series in week 12. Is this an accurate estimate of sales for week 13? Unfortunately, there is no way to address the issue of ___________ associated with forecasts for ___________________. But, if we select a forecasting method that works well for the historical data, and we think that the historical pattern will continue into the future, we should obtain results that will ultimately be shown to be good.

11. (Table 17.8) Suppose we use the ______________________________ available as the forecast for the next period. We begin by developing a forecast for week 2. Since there is only one historical value available prior to week 2, the forecast for week 2

is just the time series value in week 1; thus, the forecast for week 2 is 84 thousand gallons of gasoline. To compute the forecast for week 3, we take the average of the sales values in weeks 1 and 2. Thus,

$$\text{Forecast for week 3} = \underline{\hspace{3cm}}$$

TABLE 17.8 Computing Forecasts and Measures of Forecast Accuracy Using the Average of All the Historical Data as the Forecast for the Next Period

Week	Time Series Value	Forecast	Forecast Error	Absolute Value of Forecast Error	Squared Forecast Error	Percentage Error	Absolute Value of Percentage Error
1	68						
2	84	68.00	16.00	16.00	256.00	19.05	19.05
3	76	76.00	.00	.00	.00	.00	.00
4	92	76.00	16.00	16.00	256.00	17.39	17.39
5	72	80.00	−8.00	8.00	64.00	−11.11	11.11
6	64	78.40	−14.40	14.40	207.36	−22.50	22.50
7	80	76.00	4.00	4.00	16.00	5.00	5.00
8	72	76.57	−4.57	4.57	20.90	−6.35	6.35
9	88	76.00	12.00	12.00	144.00	13.64	13.64
10	80	77.33	2.67	2.67	7.11	3.33	3.33
11	60	77.60	−17.60	17.60	309.76	−29.33	29.33
12	88	76.00	12.00	12.00	144.00	13.64	13.64
		Totals	18.10	107.24	1425.13	2.76	141.34

12. Comparing the values of MAE, MSE, and MAPE for each method:

	Naive Method	Average of Past Values
MAE	14.91	9.75
MSE	260.36	129.56
MAPE	19.24%	12.85%

13. For every measure, the average of past values provides ____________ forecasts than using the most recent observation as the forecast for the next period.

14. In general, if the underlying time series is __________, the average of all the historical data will always provide the best results.

    (a) (Recall Table 17.2) But suppose that the underlying time series is not stationary. Note the ____________ in week 13 for the resulting time series. When a shift to a new level like this occurs, it takes a long time for the forecasting method that uses the average of all the historical data to adjust to the new level of the time series.

(b) In this case, the simple naive method adjusts very rapidly to the change in level because it uses the most recent observation available as the forecast.

(c) Measures of forecast accuracy are important factors in comparing different forecasting methods, but we have to be careful not to rely upon them too heavily.

(d) Good judgment and knowledge about business conditions that might affect the forecast also have to be carefully considered when selecting a method. And ________________________ is not the only consideration, especially if the time series is likely to change in the future.

17.3 Moving Averages and Exponential Smoothing

1. Three forecasting methods that are appropriate for a time series with a horizontal pattern: __________ averages, ___________ moving averages, and _____________ smoothing.
2. The objective of each of these methods is to smooth out the ____________________ in the time series, they are referred to as ____________ methods.
3. These methods are easy to use and generally provide a high level of ___________ for short-range ___________, such as a forecast for the next time period.

Moving Averages

1. (**Moving Average Forecast of Order** k) The moving averages method uses the average of the most recent k data values in the time series as the forecast for the next period:

$$F_{t+1} = \frac{\sum(\text{most recent } k \text{ data values})}{k} = \underline{\hspace{6cm}} \qquad (17.1)$$

where F_{t+1} is the forecast of the times series for period $t+1$ and Y_t is the actual value of the time series in period t.

2. The average will change, or move, as new observations become available.

   (a) To use moving averages to forecast a time series, we must first select the ________, or number of time series values, to be included in the moving average.

   (b) If only the ______________ values of the time series are considered relevant, a small value of k is preferred.

   (c) If _____________ values are considered relevant, then a larger value of k is better.

   (d) A time series with a horizontal pattern can shift to a new level over time. A moving average will adapt to the new level of the series and resume providing good forecasts in k periods.

   (e) Thus, a smaller value of k will ______________ in a time series more quickly. But larger values of k will be more effective in ________________ the random fluctuations over time.

3. Example (Recall Table 17.1 and Figure 17.1) the gasoline sales data

   (a) The time series plot in Figure 17.1 indicates that the gasoline sales time series has a ____________________. Thus, the smoothing methods of this section are applicable.

   (b) Use a three-week moving average ($k = 3$), the forecast of sales in week 4 using the average of the time series values in weeks 1–3:

   $$F_4 = \text{average of weeks 1–3} = \underline{\hspace{5cm}}$$

   Thus, the moving average forecast of sales in week 4 is 76 or 76,000 liters of gasoline.

   (c) The actual value observed in week 4 is 92, the ________________ in week 4 is $92 - 76 = 16$.

   (d) (Table 17.9) The forecast of sales in week 5 by averaging the time series values in weeks 2–4.

   $$F_5 = \text{average of weeks 2–4} = \frac{84 + 76 + 92}{3} = 84$$

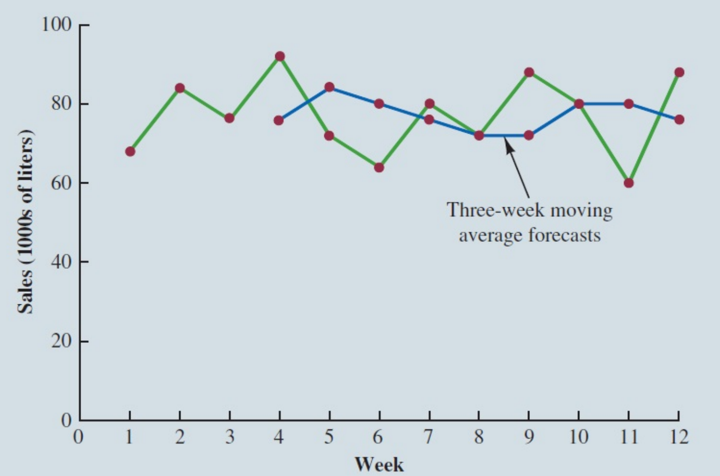
Hence, the forecast of sales in week 5 is 84 and the error associated with this forecast is $72 - 84 = -12$.

TABLE 17.9 Summary of Three-Week Moving Average Calculations

Week	Time Series Value	Forecast	Forecast Error	Absolute Value of Forecast Error	Squared Forecast Error	Percentage Error	Absolute Value of Percentage Error
1	68						
2	84						
3	76						
4	92	76	16	16	256	17.39	17.39
5	72	84	−12	12	144	−16.67	16.67
6	64	80	−16	16	256	−25.00	25.00
7	80	76	4	4	16	5.00	5.00
8	72	72	0	0	0	.00	.00
9	88	72	16	16	256	18.18	18.18
10	80	80	0	0	0	.00	.00
11	60	80	−20	20	400	−33.33	33.33
12	88	76	12	12	144	13.64	13.64
		Totals	0	96	1472	−20.79	129.21

(e) (Figure 17.7) Note how the graph of the moving average forecasts has tended to _______________ the random fluctuations in the time series.

FIGURE 17.7 Gasoline Sales Time Series Plot and Three-Week Moving Average Forecasts



(f) To forecast sales in week 13, the next time period in the future, we simply compute the average of the time series values in weeks 10, 11, and 12.

$$F_{13} = \text{average of weeks 10–12} = \frac{80 + 60 + 88}{3} = 76$$

(g) **Forecast Accuracy** Using the three-week moving average calculations in Table 17.9, the values for these three measures of forecast accuracy (MAE, MSE, and MAPE) are

$$\begin{aligned}\text{MAE} &= \frac{96}{9} = 10.67 \quad \text{(mean absolute error)}\\ \text{MSE} &= \frac{1472}{9} = 163.56 \quad \text{(mean squared error)}\\ \text{MAPE} &= \frac{129.21}{9} = 14.36\% \quad \text{(mean absolute percentage error)}\end{aligned}$$

(h) (Recall Section 17.2) Using the most recent observation as the forecast for the next week (a moving average of order $k = 1$) resulted in values of MAE = 14.91, MSE = 260.36, and MAPE = 19.24%. Thus, in each case the three-week moving average approach provided ________________ forecasts than simply using the most recent observation as the forecast.

4. To determine if a moving average with a different order k can provide more accurate forecasts, we recommend using ________________ to determine the value of k that minimizes MSE.

5. For the gasoline sales time series, it can be shown that the minimum value of MSE corresponds to a moving average of order ________________________. If we are willing to assume that the order of the moving average that is best for the historical data will also be best for future values of the time series, the most accurate moving average forecasts of gasoline sales can be obtained using a moving average of order $k = 6$.

Weighted Moving Averages

1. In the moving averages method, each observation in the moving average calculation receives the ______________.

2. One variation, known as weighted moving averages, involves selecting a _________________ for each data value and then computing a weighted average of the most recent k values as the forecast.

3. In most cases, the ______________ observation receives the ______________, and the weight decreases for older data values.

4. A moving average forecast of order $k = 3$ is just a special case of the weighted moving averages method in which each weight is equal to 1/3. Note that for the weighted moving average method the sum of the weights is equal to ______.

5. Example We assign a weight of ________ to the most recent observation, a weight of ________ to the second most recent observation, and a weight of ________ to the third most recent observation. Using this weighted average, our forecast for week 4 is:

   Forecast for week 4 = ______________________________

6. To use the weighted moving averages method, we must first select the number of data values to be included in the weighted moving average and then choose weights for each of the data values. In general, if we believe that the ______________ is a better predictor of the future than the distant past, ________________ should be given to the more recent observations. However, when the time series is highly variable, selecting approximately ________________ for the data values may be best.

7. **Forecast Accuracy** To determine whether one particular combination of number of data values and weights provides a more accurate forecast than another combination, we recommend using _________ as the measure of forecast accuracy. That is, if we assume that the combination that is best for the ________ will also be best for the __________, we would use the combination of number of data values and weights that minimizes MSE for the historical time series to forecast the next value in the time series.

Exponential Smoothing

1. Exponential smoothing also uses a weighted average of past time series values as a forecast; it is a special case of the weighted moving averages method in which we select __________________—the weight for the ______________ observation.

2. The weights for the other data values are computed automatically and become smaller as the observations move farther into the past.

3. **Exponential Smoothing Forecast**

$$F_{t+1} = \underline{\hspace{4cm}} \qquad (17.2)$$

where

F_{t+1}: forecast of the time series for period $(t+1)$

Y_t: actual value of the time series in period t

F_t: forecast of the time series for period t

α: ______________________ $(0 \leq \alpha \leq 1)$

4. Equation (17.2) shows that the forecast for period $t+1$ is a weighted average of the actual value in period t and the forecast for period t.

5. The weight given to the actual value in period t is the smoothing constant ______ and the weight given to the forecast in period t is _________.

6. Let us illustrate by working with a time series involving only three periods of data: Y_1, Y_2, and Y_3.

   (a) To initiate the calculations, we let F_1 equal the actual value of the time series in period 1; that is, $F_1 = Y_1$. Hence, the forecast for period 2 is

   $$F_2 = \underline{\hspace{4cm}} = \underline{\hspace{5cm}}$$

   We see that the exponential smoothing forecast for period 2 is equal to the actual value of the time series in period ______.

   (b) The forecast for period 3 is

   $$F_3 = \underline{\hspace{7cm}}$$

   (c) Finally, substituting this expression for F3 in the expression for F4, we obtain

   $$\begin{aligned} F_4 &= \alpha Y_3 + (1-\alpha)F_3 \\ &= \alpha Y_3 + (1-\alpha)[\alpha Y_2 + (1-\alpha)Y_1] \\ &= \underline{\hspace{6cm}} \end{aligned}$$

(d) We now see that F_4 is a weighted average of the first three time series values. The sum of the coefficients, or weights, for Y_1, Y_2, and Y_3 equals 1.

(e) A similar argument can be made to show that, in general, any forecast F_{t+1} is a weighted average of all the previous time series values.

✏ Question .... (p876)

Use exponential smoothing approach with a smoothing parameter $\alpha = 0.2$ to obtain F_2, F_3, F_4 and F_{13} for the gasoline sales time series in Table 17.1 and Figure 17.1. Start the calculations, set the exponential smoothing forecast for period 2 equal to the actual value of the time series in period 1.

sol:

TABLE 17.10 Summary of the Exponential Smoothing Forecasts and Forecast Errors for the Gasoline Sales Time Series with Smoothing Constant $\alpha = .2$

Week	Time Series Value	Forecast	Forecast Error	Squared Forecast Error
1	68			
2	84	68.00	16.00	256.00
3	76	71.20	4.80	23.04
4	92	72.16	19.84	393.63
5	72	76.13	−4.13	17.06
6	64	75.30	−11.30	127.69
7	80	73.04	6.96	48.44
8	72	74.43	−2.43	5.90
9	88	73.95	14.05	197.40
10	80	76.76	3.24	10.50
11	60	77.41	−17.41	303.11
12	88	73.92	14.08	198.25
		Totals	43.70	1581.02

FIGURE 17.8 Actual and Forecast Gasoline Sales Time Series with Smoothing Constant $\alpha = .2$

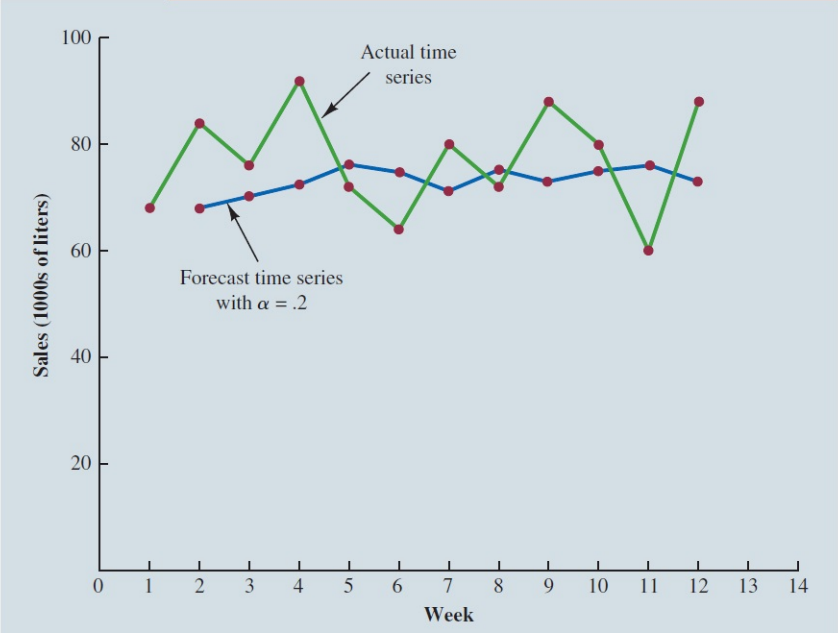


TABLE 17.11 Summary of the Exponential Smoothing Forecasts and Forecast Errors for the Gasoline Sales Time Series with Smoothing Constant $\alpha = .3$

Week	Time Series Value	Forecast	Forecast Error	Squared Forecast Error
1	68			
2	84	68.00	16.00	256.00
3	76	72.80	3.20	10.24
4	92	73.76	18.24	332.70
5	72	79.23	−7.23	52.27
6	64	77.06	−13.06	170.56
7	80	73.14	6.86	47.06
8	72	75.20	−3.20	10.24
9	88	74.24	13.76	189.34
10	80	78.37	1.63	2.66
11	60	78.86	−18.86	355.70
12	88	73.20	14.80	219.04
		Totals	32.14	1645.81

1. **Forecast Accuracy** (Table 17.10)(Figure 17.8)(Table 17.11) The criterion we will use to determine a desirable value for the smoothing constant a is the same as the criterion we proposed for determining the order or number of periods of data to include in the moving averages calculation. That is, we choose the value of a that ____________________.

2. The exponential smoothing results with $\alpha = 0.2$: the value of the sum of squared forecast errors is 98.80; hence ________________________. The exponential smoothing results with $\alpha = 0.3$: the value of the sum of squared forecast errors is 102.83; hence ________________________.

3. Thus, we would be inclined to prefer the original smoothing constant of $\alpha = 0.2$. Using a ________________ calculation with other values of α, we can find a "good value for the smoothing constant.

17.4 Trend Projection

1. We present two forecasting methods in this section that are appropriate for time series exhibiting a ________________.

(a) First, we show how ______________________ can be used to forecast a time series with a linear trend.

(b) Next we show how the ______________ capability of regression analysis can also be used to forecast time series with a ____________ or ____________ trend.

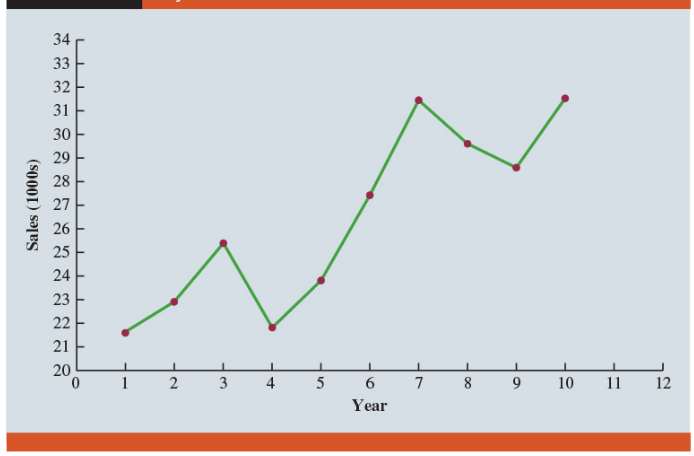
Linear Trend Regression

1. (Table 17.12) (Figure 17.9) the bicycle sales time series: the linear trend line provides a reasonable approximation of the long-run movement in the series.

TABLE 17.12
Bicycle Sales Time Series

Year	Sales (1000s)
1	21.6
2	22.9
3	25.5
4	21.9
5	23.9
6	27.5
7	31.5
8	29.7
9	28.6
10	31.4

FIGURE 17.9 Bicycle Sales Time Series Plot



2. The estimated regression equation describing a ______________ relationship between an independent variable x and a dependent variable y is written as

____

where $\hat{y}$ is the estimated or predicted value of y.

3. To emphasize the fact that in forecasting the independent variable is time, we will replace _____ with _____ and _____ with _____ to emphasize that we are estimating the trend for a time series.

4. **Linear Trend Equation**

______________ (17.4)

where

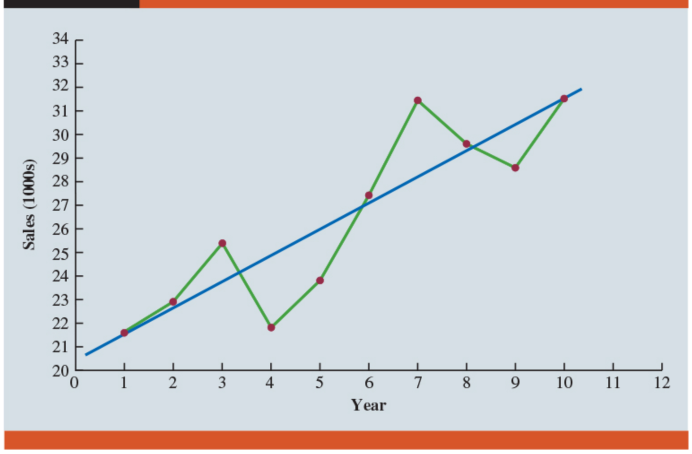
T_t = linear trend forecast in period t

b_0 = intercept of the linear trend line

b_1 = slope of the linear trend line

t = time period, $t = 1$ ($t = n$) corresponding to the first time (most recent) series observation

FIGURE 17.10 Trend Represented by a Linear Function for the Bicycle Sales Time Series



5. **Computing the Slope and Intercept for a Linear Trend**

$$b_1 = \frac{\qquad\qquad}{\qquad\qquad} = \frac{\qquad\qquad}{\qquad\qquad} \qquad (17.5)$$

$$b_0 = \underline{\qquad\qquad} \qquad (17.6)$$

where

Y_t = value of the time series in period t

n = number of time periods (number of observations)

$\bar{Y}$ = average value of the time series

$\bar{t}$ = average value of t

6. Example the bicycle sales time series

TABLE 17.13 Summary of Linear Trend Calculations for the Bicycle Sales Time Series

	t	Y_t	$t-\bar{t}$	$Y_t-\bar{Y}$	$(t-\bar{t})(Y_t-\bar{Y})$	$(t-\bar{t})^2$
	1	21.6	−4.5	−4.85	21.825	20.25
	2	22.9	−3.5	−3.55	12.425	12.25
	3	25.5	−2.5	−.95	2.375	6.25
	4	21.9	−1.5	−4.55	6.825	2.25
	5	23.9	−.5	−2.55	1.275	.25
	6	27.5	.5	1.05	.525	.25
	7	31.5	1.5	5.05	7.575	2.25
	8	29.7	2.5	3.25	8.125	6.25
	9	28.6	3.5	2.15	7.525	12.25
	10	31.4	4.5	4.95	22.275	20.25
Totals	55	264.5			90.750	82.50

(a) The linear trend equation is

$$T_t = \underline{\hspace{3cm}}$$

(b) The slope of 1.1 indicates that over the past 10 years the firm experienced an ____________ of about ____________ per year.

(c) If we assume that the past 10-year trend in sales is a good indicator of the future, this trend equation can be used to develop forecasts for future time periods. For example, the next year's trend projection or forecast, T_{11}.

$$T_{11} = \underline{\hspace{5cm}}$$

Thus, using trend projection, we would forecast sales of 32,500 bicycles next year.

7. (Table 17.14) Use MSE to compute the accuracy associated with the trend projection forecasting method.

TABLE 17.14 Summary of the Linear Trend Forecasts and Forecast Errors for the Bicycle Sales Time Series

Year	Sales (1000s) Y_t	Forecast T_t	Forecast Error	Squared Forecast Error
1	21.6	21.5	.1	.01
2	22.9	22.6	.3	.09
3	25.5	23.7	1.8	3.24
4	21.9	24.8	−2.9	8.41
5	23.9	25.9	−2.0	4.00
6	27.5	27.0	.5	.25
7	31.5	28.1	3.4	11.56
8	29.7	29.2	.5	.25
9	28.6	30.3	−1.7	2.89
10	31.4	31.4	.0	.00
			Total	30.70

$$MSE = \underline{\hspace{6cm}}$$

8. Because ______________ in forecasting is the same as the standard regression analysis procedure applied to time-series data, we can use statistical software to perform the calculations.

9. (Figure 17.11) the value of MSE in the ANOVA table is

$$MSE = \frac{\text{Sum of Squares Due to Error}}{\text{Degrees of Freedom}} = \underline{\hspace{3cm}}$$

FIGURE 17.11 Regression Output for the Bicycle Sales Time Series

Regression Analysis: Sales Versus Year

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	99.82	99.825	26.01	.001
Error	8	30.70	3.837		
Total	9	130.52			

Model Summary

S	R-sq
1.95895	76.48%

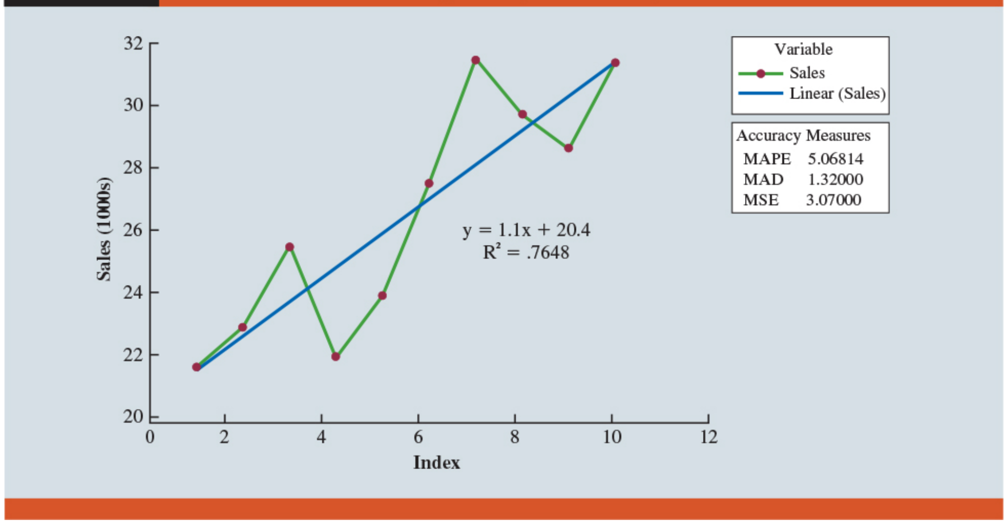
Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	20.40	1.34	15.24	.000
Year	1.10	.216	5.10	.001

Regression Equation

Sales = 20.40 + 1.100 Year

FIGURE 17.12 Time Series Linear Trend Analysis Output for the Bicycle Sales Time Series



10. This value of MSE __________ from the value of MSE that we computed previously because the sum of squared errors is divided by ______ instead of _______; thus, MSE in the regression output is not the ______________________________________.

11. **NOTE**: Most forecasting packages, however, compute MSE by taking the average of the squared errors. Thus, when using time series packages to develop a trend equation, the value of MSE that is reported may differ slightly from the value you would obtain using a general regression approach.

Nonlinear Trend Regression

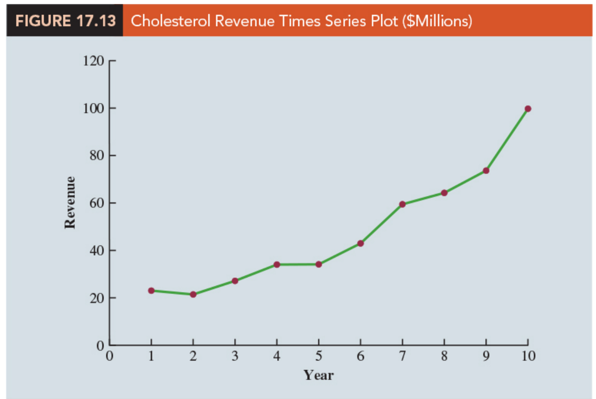
1. Example (Table 17.15) (Figure 17.13) Consider the annual revenue in millions of dollars for a cholesterol drug for the first 10 years of sales.

TABLE 17.15

Cholesterol Revenue Time Series ($ Millions)

Year (t)	Revenue ($ millions)
1	23.1
2	21.3
3	27.4
4	34.6
5	33.8
6	43.2
7	59.5
8	64.4
9	74.2
10	99.3

FIGURE 17.13 Cholesterol Revenue Times Series Plot ($Millions)



2. The time series plot indicates an ____________ or ________ trend. A curvilinear function appears to be needed to model the long-term trend.

3. **Quadratic Trend Equation** A variety of nonlinear functions can be used to develop an estimate of the trend for the cholesterol time series. For instance, consider the following quadratic trend equation:

   ____________ (17.7)

4. (Figure 17.14) a portion of the multiple regression output for the quadratic trend model;

FIGURE 17.14 Quadratic Trend Regression Output for the Cholesterol Revenue Time Series

Regression Analysis: Revenue Versus Year, YearSq

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5770.13	2885.06	182.52	.000
Error	7	110.65	15.81		
Total	9	5880.78			

Model Summary

S	R-sq
3.97578	98.12%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	24.18	4.68	5.17	.001
Year	−2.11	1.95	−1.08	.317
YearSq	.922	.173	5.33	.001

Regression Equation

Revenue = 24.18 − 2.11 Year + .922 YearSq

The estimated regression equation is

$$\text{Revenue (\$millions)} = 24.18 - 2.11\text{Year} + 0.922\text{YearSq}$$

5. **Exponential Trend Equation**

   __________ (17.8)

6. Suppose $b_0 = 16.71$, and $b_1 = 0.1697$, T_t is not increasing by a constant amount as in the case of the linear trend model but by a ______________.

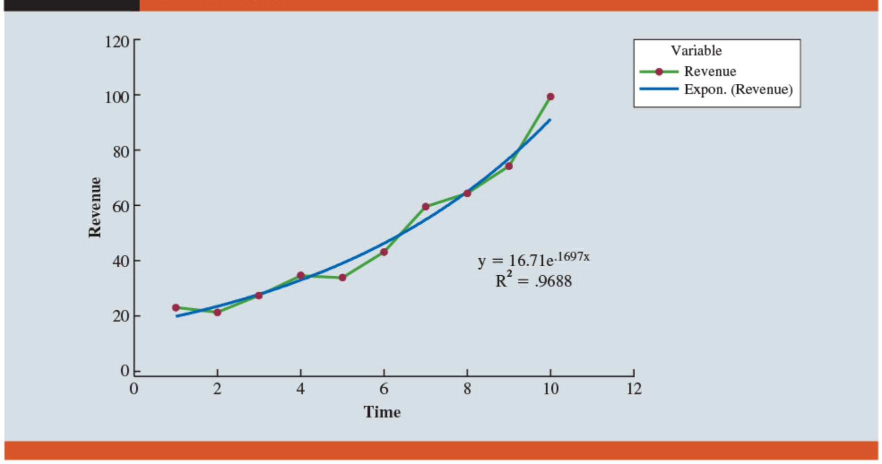
7. In this exponential trend model, multiplicative factor is ________________, so the constant percentage increase from time period to time period is _________.

8. Many statistical software packages have the capability to compute an exponential trend equation directly. Some software packages only provide linear trend, but by applying a natural *log* transformation to both sides of the equality in equation (17.8) we can apply the equivalent linear form:

____

(Figure 17.15)

FIGURE 17.15 Time Series Exponential Growth Trend Analysis Output for the Cholesterol Sales Time Series



17.5 Seasonality and Trend

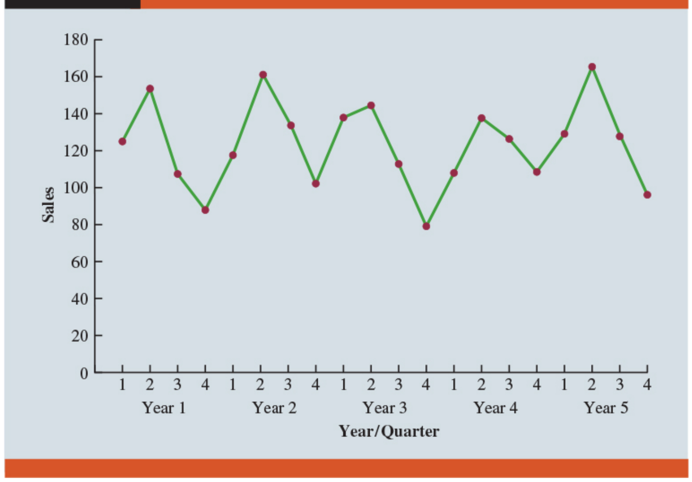
Seasonality Without Trend

1. Example (Table 17.16)(Figure 17.16) Consider the number of umbrellas sold at a clothing store over the past five years.

TABLE 17.16
Umbrella Sales Time Series

Year	Quarter	Sales
1	1	125
	2	153
	3	106
	4	88
2	1	118
	2	161
	3	133
	4	102
3	1	138
	2	144
	3	113
	4	80
4	1	109
	2	137
	3	125
	4	109
5	1	130
	2	165
	3	128
	4	96

FIGURE 17.16 Umbrella Sales Time Series Plot



2. The time series plot does not indicate any ________ trend in sales. The first and third quarters have moderate sales, the second quarter has the highest sales, and the fourth quarter tends to be the lowest quarter in terms of sales volume. Thus, we would conclude that a ________ pattern is present.

3. Just like using ________ to deal with an independent variable in a standard regression analysis, we can use the same approach to model a time series with a seasonal pattern by treating the season as a ________.

4. Recall that when a categorical variable has k levels, ________ dummy variables are required. Thus, to model the ________ in the umbrella time series we need 4−1 = 3 dummy variables:

$$Qtr1 = \begin{cases} 1 & \text{if Quarter 1} \\ 0 & \text{otherwise} \end{cases}, Qtr2 = \begin{cases} 1 & \text{if Quarter 2} \\ 0 & \text{otherwise} \end{cases}, Qtr3 = \begin{cases} 1 & \text{if Quarter 3} \\ 0 & \text{otherwise} \end{cases}$$

5. Using $\hat{Y}$ to denote the estimated or forecasted value of sales, the general form of the estimated regression equation relating the number of umbrellas sold to the quarter

the sales take place:

---

6. (Table 17.17) the umbrella sales time series with the coded values of the dummy variables.

TABLE 17.17 Umbrella Sales Time Series with Dummy Variables

Year	Quarter	Qtr1	Qtr2	Qtr3	Sales
1	1	1	0	0	125
	2	0	1	0	153
	3	0	0	1	106
	4	0	0	0	88
2	1	1	0	0	118
	2	0	1	0	161
	3	0	0	1	133
	4	0	0	0	102
3	1	1	0	0	138
	2	0	1	0	144
	3	0	0	1	113
	4	0	0	0	80
4	1	1	0	0	109
	2	0	1	0	137
	3	0	0	1	125
	4	0	0	0	109
5	1	1	0	0	130
	2	0	1	0	165
	3	0	0	1	128
	4	0	0	0	96

7. (Figure 17.17) the computer output: the estimated multiple regression equation obtained is

$$Sales = 95.00 + 29.00\ Qtr1 + 57.00\ Qtr2 + 26.00\ Qtr3$$

We can use this equation to forecast quarterly sales for next year.

Quarter 1: $Sales = 95.0 + 29.0(1) + 57.0(0) + 26.0(0) = 124.$

Quarter 2: $Sales = 95.0 + 29.0(0) + 57.0(1) + 26.0(0) = 152.$

Quarter 3: $Sales = 95.0 + 29.0(0) + 57.0(0) + 26.0(1) = 121.$

Quarter 4: $Sales = 95.0 + 29.0(0) + 57.0(1) + 26.0(0) = 95.$

FIGURE 17.17 Regression Output for the Umbrella Sales Time Series

Term	Coef	SE Coef	T-Value	P-Value
Constant	95.00	5.06	18.76	.000
Qtr1	29.00	7.16	4.05	.001
Qtr2	57.00	7.16	7.96	.000
Qtr3	26.00	7.16	3.63	.002

Regression Equation

Sales = 95.00 + 29.00 Qtr1 + 57.00 Qtr2 + 26.00 Qtr3

8. The regression output shown in Figure 17.17 provides additional information that can be used to assess the ____________ of the forecast and determine the ______________ of the results.

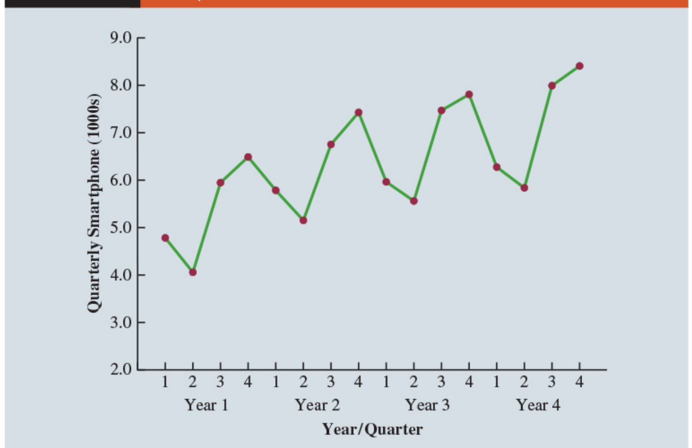
Seasonality and Trend

1. Example (Table 17.18) (Figure 17.18) The quarterly smartphone sales.

TABLE 17.18 Smartphone Sales Time Series

Year	Quarter	Sales (1000s)
1	1	4.8
	2	4.1
	3	6.0
	4	6.5
2	1	5.8
	2	5.2
	3	6.8
	4	7.4
3	1	6.0
	2	5.6
	3	7.5
	4	7.8
4	1	6.3
	2	5.9
	3	8.0
	4	8.4

FIGURE 17.18 Smartphone Sales Time Series Plot



2. The sales are lowest in the second quarter of each year and increase in quarters 3 and 4. Thus, we conclude that a __________________ exists for smartphone sales.

3. But the time series also has an ____________________ that will need to be accounted for in order to develop accurate forecasts of quarterly sales.

4. This is easily handled by combining the _________________________ for seasonality with the time series ____________________ for hand-ling linear trend.

5. The general form of the estimated multiple regression equation for modeling both the quarterly seasonal effects and the linear trend in the smartphone time series:

____

where

$\hat{Y}_t$ = estimate or forecast of sales in time period t

$Qtr1 = 1$ $(Qtr2 = 1)(Qtr3 = 1)$ if time period t corresponds to the first (second) (third) quarter of the year; 0 otherwise.

6. (Table 17.19) revised smartphone sales time series that includes the coded values of the dummy variables and the time period t.

TABLE 17.19 Smartphone Sales Time Series with Dummy Variables and Time Period

Year	Quarter	Qtr1	Qtr2	Qtr3	Period	Sales (1000s)
1	1	1	0	0	1	4.8
	2	0	1	0	2	4.1
	3	0	0	1	3	6.0
	4	0	0	0	4	6.5
2	1	1	0	0	5	5.8
	2	0	1	0	6	5.2
	3	0	0	1	7	6.8
	4	0	0	0	8	7.4
3	1	1	0	0	9	6.0
	2	0	1	0	10	5.6
	3	0	0	1	11	7.5
	4	0	0	0	12	7.8
4	1	1	0	0	13	6.3
	2	0	1	0	14	5.9
	3	0	0	1	15	8.0
	4	0	0	0	16	8.4

7. (Figure 17.19) The estimated multiple regression equation is

$$Sales(1000s) = 6.069 - 1.363\ Qtr1 - 2.034\ Qtr2 - 0.304 Qtr3 + 0.1456\ t \qquad (17.9)$$

8. Forecast for Time Period 17 (Quarter 1 in Year 5):

$$Sales(1000s) = 6.069 + 1.363(1) + 2.034(0) + 0.304(0) + 0.1456(17) = 7.18$$

Thus, accounting for the seasonal effects and the linear trend in smartphone sales, the estimates of quarterly sales in year 5 are 7180, 6660, 8530, and 8980.

9. The dummy variables in the estimated multiple regression equation actually provide ______________________ regression equations, one for each quarter. If time period t corresponds to quarter 1, the estimate of quarterly sales is

$$\begin{aligned} Quarter1 : Sales &= 6.069 - 1.363(1) - 2.034(0) - 0.304(0) + 0.1456(t) \\ &= 4.71 + 0.1456t \\ Quarter2 : Sales &= 4.04 + 0.1456t \\ Quarter3 : Sales &= 5.77 + 0.1456t \\ Quarter4 : Sales &= 6.07 + 0.1456t \end{aligned}$$

10. The _________ of the trend line for each quarterly forecast equation is __________, indicating a __________ in sales of about 146 sets per quarter.

11. The intercept for the Quarter 1 equation is 4.71 and the intercept for Quarter 4 equation is 6.07. Thus, sales in Quarter 1 are ______________________ or ____________ ____________ in Quarter 4.

12. The estimated regression coefficient for $Qtr1$ in equation (17.9) provides an estimate of the difference in sales between Quarter ______ and Quarter ______.

13. Similar interpretations can be provided for __________, the estimated regression coefficient for dummy variable $Qtr2$, and __________, the estimated regression coefficient for dummy variable $Qtr3$.

Models Based on Monthly Data

1. For monthly data, season is a categorical variable with 12 levels and thus ______________ dummy variables are required.

$$Month1 = \begin{cases} 1 & \text{if January} \\ 0 & \text{otherwise} \end{cases}, \ldots, \; Month11 = \begin{cases} 1 & \text{if November} \\ 0 & \text{otherwise} \end{cases}$$

2. Other than this change, the multiple regression approach for handling seasonality remains ____________.

17.6 Time Series Decomposition*

☺ **SUPPLEMENTARY EXERCISES**: 41, 44, 47

😍 EXERCISES

17.2 : 1, 4

17.3 : 5, 9, 11, 14

17.4 : 17, 20, 22, 26

17.5 : 28, 30, 33

SUP : 41, 44, 47

“有時候壞事是注定要發生，而我們卻無能為力。那我們何必擔心呢?”
“Look, sometimes bad things happen —and there’ s nothing you can do about it. So why worry?”

— *獅子王 (The Lion King, 2019)*

統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 18: Nonparametric Methods

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: http://www.hmwu.idv.tw

系級: ____________________ 學號: ____________________ 姓名: ____________________

Overview

1. The statistical methods for inference presented previously are ________________________.
2. The parametric methods begin with an ________________ about the probability distribution of the ________________ which is often that the population has a ____________ distribution.
3. Based upon this assumption, statisticians are able to derive the ________________________ that can be used to make ______________ about one or more parameters of the population, such as the population mean or the population standard deviation.
   (a) (Recall Chapter 9) An inference about a population mean that was based on an assumption that the population had a normal probability distribution with unknown parameters μ and σ.
   (b) Using the sample standard deviation s to estimate the population standard deviation σ.
   (c) The test statistic for making an inference about the population mean was shown to have a t distribution.

(d) The t distribution was used to compute confidence intervals and conduct hypothesis tests about the mean of a normally distributed population.

4. In this chapter we present ________________ methods which can be used to make inferences about a population without requiring an assumption about the specific form of the population's probability distribution.

   (a) (First section) how the binomial distribution uses two categories of data to make an inference about a ____________________.

   (b) (Next three sections) how ______________ data are used in nonparametric tests about two or more populations.

   (c) (Final section) use rank-ordered data to compute the _________________ for two variables.

5. For this reason, these nonparametric methods are also called _________________________.

6. The computations used in the nonparametric methods are generally done with _________________. Whenever the data are quantitative, we will transform the data into categorical data in order to conduct the nonparametric test.

18.1 Sign Test

Hypothesis Test About a Population Median

1. The ___________ provides a nonparametric procedure for testing a hypothesis about the value of a ____________________.

2. If we consider a population where _______________ is exactly equal to the median, the median is the measure of __________________ that divides the population so that ________ of the values are greater than the median and ________ of the values are less than the median.

3. Whenever a population distribution is __________, the median is often preferred over the mean as the best measure of central location for the population.

4. Example The weekly sales of Cape May Potato Chips by the Lawler Grocery Store chain.

   (a) Lawler's management made the decision to carry the new potato chip product based on the manufacture's estimate that the ______________ should be \$450 per week on a per store basis.

   (b) (Table 18.1) After carrying the product for three-months, Lawler's management requested the following hypothesis test about the population median weekly sales:

$$H_0 : \text{Median} = 450$$
$$H_a : \text{Median} \neq 450$$

TABLE 18.1 One-Week Sales of Cape May Potato Chips at 10 Lawler Grocery Stores

Store Number	Weekly Sales (\$)	Store Number	Weekly Sales (\$)
56	485	63	474
19	562	39	662
36	415	84	380
128	860	102	515
12	426	44	721

   (c) (Table 18.2) In conducting the sign test, we compare each sample observation to the ____________________ of the population median.

      i. If the observation is greater than the hypothesized value, we record a plus sign ________

      ii. If the observation is less than the hypothesized value, we record a minus sign ________

      iii. If an observation is exactly equal to the hypothesized value, the observation is _____________ from the sample and the analysis proceeds with the smaller sample size, using only the observations where a plus sign or a minus sign has been recorded.

TABLE 18.2 Lawler Sample Data for the Sign Test About the Population Median Weekly Sales

Store Number	Weekly Sales ($)	Sign	Store Number	Weekly Sales ($)	Sign
56	485	+	63	474	+
19	562	+	39	662	+
36	415	−	84	380	−
128	860	+	102	515	+
12	426	−	44	721	+

(d) Note that there are 7 plus signs and 3 minus signs.

5. The assigning of the plus signs and minus signs has made the situation a ____________ ____________ application. The sample size __________ is the number of trials. There are two outcomes possible per trial, a ________ sign or a __________ sign, and the trials are independent. Let _____ denote the probability of a plus sign.

6. If the population median is 450, p would equal ________ as there should be 50% plus signs and 50% minus signs in the population. Thus, in terms of the binomial probability p, the sign test hypotheses about the population median are converted to the following hypotheses about the binomial probability p.

$$\begin{aligned} H_0 &: \text{Median} = 450 \\ H_a &: \text{Median} \neq 450 \end{aligned} \quad \Rightarrow \quad \underline{\hspace{4cm}}$$

(a) If H_0 cannot be rejected, we cannot conclude that p is different from 0.50 and thus we cannot conclude that the population median is different from 450.

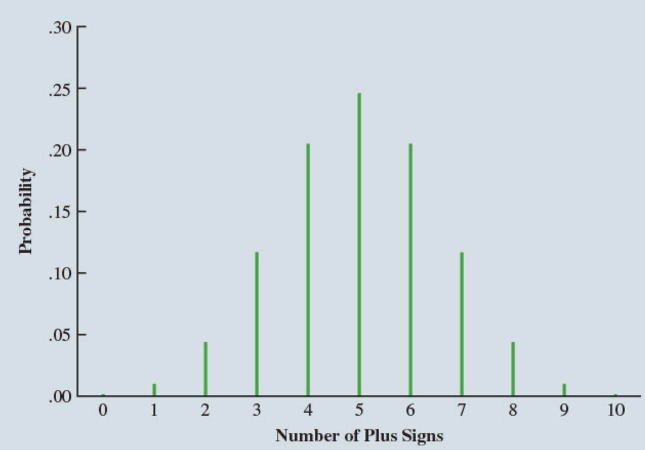
(b) If H_0 is rejected, we can conclude that p is not equal to 0.50 and thus the population median is not equal to 450.

7. (Table 5 in Appendix B)(Table 18.3)(Figure 18.1) With $n = 10$ stores or trials and $p = 0.50$, obtain the binomial probabilities for the number of plus signs under the assumption H_0 is true. (________________________)

TABLE 18.3 Binomial Probabilities with $n = 10$ and $p = .50$

Number of Plus Signs	Probability
0	.0010
1	.0098
2	.0439
3	.1172
4	.2051
5	.2461
6	.2051
7	.1172
8	.0439
9	.0098
10	.0010

FIGURE 18.1 Binomial Sampling Distribution for the Number of Plus Signs When $n = 10$ and $p = .50$



(a) Use a 0.10 level of significance for the test.

(b) Since the observed number of plus signs for the sample data, 7, is in the upper tail of the binomial distribution, we compute the probability of obtaining 7 or more plus signs

______________________________.

(c) Since we are using a two-tailed hypothesis test, this upper tail probability is doubled to obtain the ______________________.

(d) With __________, we cannot reject H_0. In terms of the binomial probability p, we cannot reject $H_0 : p = 0.50$, and thus we cannot reject the hypothesis that the population median is \$450.

8. The one-tailed sign tests about a population median:

   (a) Formulated the hypotheses as an ________________:

   $$H_0 : \text{Median } \le 450$$
   $$H_a : \text{Median } > 450$$

   (b) The corresponding p-value is equal to the binomial probability that the number of plus signs is __________________________ found in the sample.

   (c) This one-tailed p-value: ____.

   (d) If the example were converted to a lower tail test, the p-value would have been the probability of obtaining 7 or fewer plus signs.

   (e) The binomial probabilities provided in Table 5 of Appendix B can be used to compute the p-value when the sample size is ____________.

   (f) With larger sample sizes, we rely on the _________________________________ of the binomial distribution to compute the p-value; this makes the computations quicker and easier.

Use the Normal Distribution to Approximate the Binomial Probability

1. Example One year ago the median price of a new home was \$236,000. However, a current downturn in the economy has real estate firms using sample data on recent home sales to determine if the population median price of a new home is less today than it was a year ago.

   (a) The hypothesis test about the population median price of a new home is as follows:

   $$H_0 : \underline{\qquad\qquad\qquad}$$
   $$H_a : \underline{\qquad\qquad\qquad}$$

   (b) We will use a 0.05 level of significance to conduct this test. A random sample of ______ recent new home sales found ______ homes sold for more than \$236,000, ______ homes sold for less than \$236,000, and _______ home sold for \$236,000.

(c) After deleting the home that sold for the hypothesized median price of \$236,000, the sign test continues with 22 plus signs, 38 minus signs, and a sample of ________.

(d) The null hypothesis that the population median is greater than or equal to \$236,000 is expressed by the binomial distribution hypothesis ________.

(e) If H_0 were true as an equality, we would expect ________ homes to have a plus sign.

(f) The sample result showing 22 plus signs is in the lower tail of the binomial distribution. Thus, the p-value is the probability of ________ when $p = 0.50$.

(g) While it is possible to compute the exact binomial probabilities for $0, 1, 2, \cdots$ to 22 and sum these probabilities, we will use the normal distribution approximation of the binomial distribution to make this computation easier.

2. **Normal approximation of the sampling distribution of the number of plus signs when** $H_0 : p = 0.50$: For this approximation (________), the mean and standard deviation of the normal distribution are:

$$\text{Mean} : \mu = \underline{\qquad\qquad} \tag{18.1}$$

$$\text{Standard deviation} : \sigma = \underline{\qquad\qquad\qquad} \tag{18.2}$$

3. With $n = 60$ homes and $p = 0.50$, the sampling distribution of the number of plus signs can be approximated by a normal distribution with

$$\mu = 0.50n = \underline{\qquad\qquad} \qquad \sigma = \sqrt{0.25n} = \underline{\qquad\qquad}$$

4. The binomial probability distribution is discrete and the normal probability distribution is continuous. To account for this, the binomial probability of 22 is computed by the normal probability interval ________. The 0.5 added to and subtracted from 22 is called the ________ factor.

5. Thus, to compute the p-value for 22 or fewer plus signs we use the normal distribution with $\mu = 30$ and $\sigma = 3.873$ to compute the probability that the normal random variable, X, has a value less than or equal to 22.5.

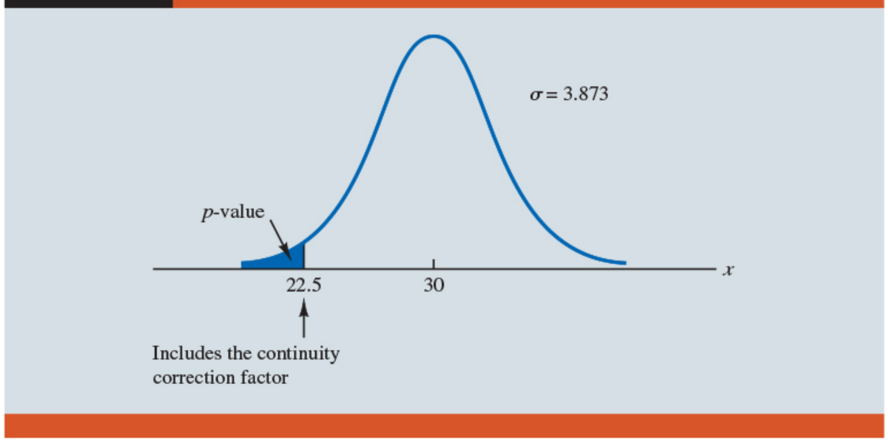
____

6. (Figure 18.2) Using this normal distribution, we compute the p-value as follows:

   p-value = ____

7. With $0.0262 < 0.05$, we __________ the null hypothesis and conclude that the median price of a new home is _____________ the \$236,000 median price a year ago.

FIGURE 18.2 Normal Distribution Approximation of the p-Value for the Sign Test About the Median Price of New Homes



Hypothesis Test with Matched Samples

1. (Recall Chapter 10) Using ____________________ and assuming that the differences between the pairs of matched observations were _____________ distributed, the ______ distribution was used to make an inference about the difference between the means of the two populations.

2. Use the nonparametric sign test to analyze ___________________ data. the sign test enables us to analyze categorical as well as quantitative data and requires no assumption about the distribution of the differences.

3. This type of matched-sample design occurs in ___________________ when a sample of n potential customers is asked to compare two brands of a product such as coffee, soft drinks, or detergents. Without obtaining a quantitative measure of each individual's preference for the brands, each individual is asked to state a brand preference.

4. Example Sun Coast Farms produces an orange juice product called Citrus Valley. The primary competition for Citrus Valley comes from the producer of an orange juice known as Tropical Orange. In a consumer preference comparison of the two brands, 14 individuals were given unmarked samples of the two orange juice products. The brand each individual tasted first was selected randomly.

   (a) If the individual selected Citrus Valley as the more preferred, a ________ was recorded.

   (b) If the individual selected Tropical Orange as the more preferred, a ________ was recorded.

   (c) If the individual was unable to express a difference in preference for the two products, ________ was recorded.

5. (Table 18.4) Deleting the two individuals who could not express a preference for either brand, the data have been converted to a sign test with ________ signs and ________ signs for the ________ individuals who could express a preference for one of the two brands.

TABLE 18.4 Preference Data for the Sun Coast Farms Taste Test

Individual	Preference	Sign	Individual	Preference	Sign
1	Tropical Orange	−	8	Tropical Orange	−
2	Tropical Orange	−	9	Tropical Orange	−
3	Citrus Valley	+	10	No Preference	
4	Tropical Orange	−	11	Tropical Orange	−
5	Tropical Orange	−	12	Citrus Valley	+
6	No Preference		13	Tropical Orange	−
7	Tropical Orange	−	14	Tropical Orange	−

6. Letting ____ indicate the proportion of the population of customers who prefer Citrus Valley orange juice, we want to test the hypotheses that there is no difference between the preferences for the two brands as follows:

$$H_0 : \_\_\_\_\_\_\_\_$$
$$H_a : \_\_\_\_\_\_\_\_$$

7. If H_0 cannot be rejected, we cannot conclude that there is a difference in preference for the two brands. However, if H_0 can be rejected, we can conclude that the consumer preferences differ for the two brands.

8. (Table 18.5) We will conduct the sign test ($\alpha = 0.05$). The sampling distribution for the number of plus signs is a __________ distribution with $p = 0.50$ and $n = 12$. (____________________)

TABLE 18.5 Binomial Probabilities with $n = 12$ and $p = .50$

Number of Plus Signs	Probability
0	.0002
1	.0029
2	.0161
3	.0537
4	.1208
5	.1934
6	.2256
7	.1934
8	.1208
9	.0537
10	.0161
11	.0029
12	.0002

9. Under the assumption H_0 is true, we would expect __________________ plus signs. With only two plus signs in the sample, the results are in the __________ of the binomial distribution.

10. To compute the p-value for this two-tailed test, we first compute the probability of 2 or fewer plus signs and then ________ this value. Using the binomial probabilities of 0, 1, and 2 shown in Table 18.5, the p-value is

$$p\text{-value} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

11. We reject H_0. The taste test provides evidence that consumer preference ________________ for the two brands of orange juice. We would advise Sun Coast Farms of this result and conclude that the competitor's Tropical Orange product is the more preferred. Sun Coast Farms can then pursue a strategy to address this issue.

12. Similar to other uses of the sign test, one-tailed tests may be used depending upon the application.

13. As the sample size becomes large, the __________________________ of the binomial distribution will ease the computation.

14. While the Sun Coast Farms sign test for matched samples used categorical preference data, the sign test for matched samples can be used with ____________ data as well.
    (a) This would be particularly helpful if the ________________ are ____________ distributed and are ________.
    (b) In this case a positive difference is assigned a plus sign, a negative difference is assigned a negative sign, and a zero difference is removed from the sample.
    (c) The sign test computations proceed as before.

18.2 Wilcoxon Signed-Rank Test

1. (Recall Chapter 10) The parametric test for the _______________ experiment requires quantitative data and the assumption that the ___________ between the paired observations are normally distributed. The _____________ can then be used to make an inference about the difference between the means of the two populations.
2. The ___________________ test is a nonparametric procedure for analyzing data from a ________________________. The test uses ________________ but does not require the assumption that the differences between the paired observations are normally distributed.
3. It only requires the assumption that the ___________ between the paired observations have a ___________ distribution.
4. This occurs whenever the ________ of the two populations are the same and the focus is on determining if there is a difference between the _________ of the two populations.

5. Example Production Task Completion Times: Consider a manufacturing firm that is attempting to determine whether two production methods differ in terms of task completion time.

   (a) (Table 18.6) Using a matched-samples experimental design, 11 randomly selected workers completed the production task two times, once using method A and once using method B. The production method that the worker used first was randomly selected.

TABLE 18.6 Production Task Completion Times (Minutes)

Worker	Method A	Method B	Difference
1	10.2	9.5	.7
2	9.6	9.8	−.2
3	9.2	8.8	.4
4	10.6	10.1	.5
5	9.9	10.3	−.4
6	10.2	9.3	.9
7	10.6	10.5	.1
8	10.0	10.0	.0
9	11.2	10.6	.6
10	10.7	10.2	.5
11	10.6	9.8	.8

   (b) A ______ difference indicates that method A required more time; a ______ difference indicates that method B required more time.

   (c) Do the data indicate that the two production methods differ significantly in terms of completion times? If we assume that the differences have a ______ distribution but not necessarily a normal distribution, the Wilcoxon signed-rank test applies.

   (d) In particular, we will use the Wilcoxon signed-rank test for the difference between the ______ completion times for the two production methods.

$$H_0 : \text{\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_}$$

$$H_a : \text{Median for method A} - \text{Median for method B} \neq 0$$

   (e) If H_0 cannot be rejected, we will not be able to conclude that the median completion times are different. However, if H_0 is rejected, we will conclude that the median completion times are different.

6. The Wilcoxon signed-rank test steps ($\alpha = 0.05$):

(a) Discard the difference of ________ for worker 8 and then compute the ________________ ________________ for the remaining 10 workers.

(b) Rank these absolute differences from ________________. The first (second) smallest absolute difference of 0.1 (0.2) for worker 7 (2) is assigned the rank of 1 (2). This ranking of absolute differences continues with the largest absolute difference of 0.9 for worker 6 being assigned the rank of 10. The ________ absolute differences of 0.4 (0.5) for workers 3 and 5 (4 and 10) are assigned the ________________ of 3.5 (5.5).

(c) (Table 18.7) Each rank is given the ________ of the original difference for the worker.

TABLE 18.7 Ranking the Absolute Differences and the Signed Ranks for the Production Task Completion Times

Worker	Difference	Absolute Difference	Rank	Signed Ranks Negative	Signed Ranks Positive
1	.7	.7	8		8
2	−.2	.2	2	−2	
3	.4	.4	3.5		3.5
4	.5	.5	5.5		5.5
5	−.4	.4	3.5	−3.5	
6	.9	.9	10		10
7	.1	.1	1		1
8	.0				
9	.6	.6	7		7
10	.5	.5	5.5		5.5
11	.8	.8	9		9
				Sum of Positive Signed Ranks T^+ =	49.5

(d) Let ________ denote the sum of the positive signed ranks ($T^+ = 49.5$). We will use T^+ as the Wilcoxon signed-rank test statistic.

(e) **Sampling Distribution of T^+ for the Wilcoxon Signed-Rank Test**: If the medians of the two populations are equal and the number of matched pairs is 10 or more, the sampling distribution of T^+ can be approximated by a ________________:

$$\text{Mean} : \mu_{T^+} = \underline{\hspace{3cm}} \quad (18.3)$$

$$\text{Standard deviation}: \sigma_{T^+} = \underline{\hspace{6cm}} \quad (18.4)$$

Distribution Form: Approximately normal for ________.

7. Example Production Task Completion Times:

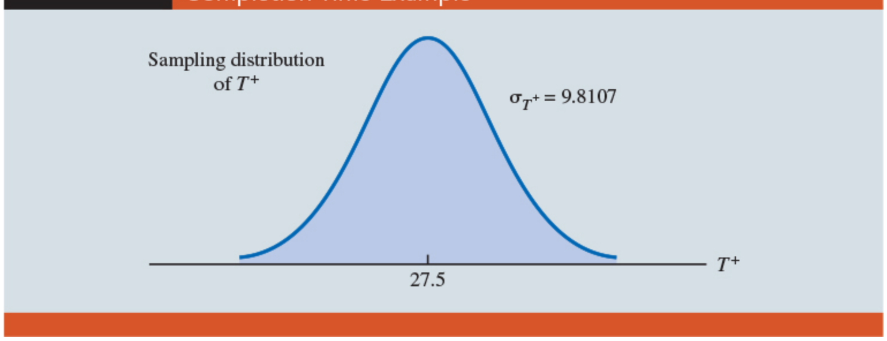
(a) After discarding the observation of a zero difference for worker 8, the analysis continues with the $n = 10$ matched pairs.

$$\mu_{T^+} = \frac{n(n+1)}{4} = \underline{\hspace{4cm}} = 27.5$$

$$\sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \underline{\hspace{9cm}} = 9.8107$$

(b) (Figure 18.3) The sampling distribution of the T^+ test statistic.

FIGURE 18.3 Sampling Distribution of T^+ for the Production Task Completion Time Example



(c) Compute the two-tailed p-value for the hypothesis that the median completion times for the two production methods are equal. Since the test statistic $T^+ = 49.5$ is in the ____________ of the sampling distribution, we begin by computing the upper tail probability ______________.

(d) Since the sum of the positive ranks T^+ is discrete and the normal distribution is continuous, we will obtain the best approximation by including the ____________________ factor. Thus, the discrete probability of ____________ is approximated by the normal probability interval, __________, and the probability that $T^+ \geq 49.5$ is approximated by:

$$P(T^+ \geq 49.5) = \underline{\hspace{9cm}}$$

(e) Using the standard normal distribution table and $z = 2.19$, we see that the two-tailed p-value = ______________________. With the p-value ≤ 0.05, we reject H_0 and conclude that the median completion times for the two production methods are not equal.

(f) With T^+ being in the upper tail of the sampling distribution, we see that method A led to the longer completion times. We would expect management to conclude that method B is the faster or better production method.

8. **One-tailed Wilcoxon signed-rank tests** are possible. For example, if initially we had been looking for statistical evidence to conclude method A had the larger median completion time than method B:

$$H_0 \;:\; \underline{\hspace{8cm}}$$
$$H_a \;:\; \text{Median for method A} - \text{Median for method B} > 0$$

9. (Recall Section 18.1) the sign test could be used for both a hypothesis test about a population median and a hypothesis test with matched samples.

10. The Wilcoxon signed-rank test can also be used for a nonparametric test about a ____________________. This test makes no assumption about the population distribution other than that it is _____________.

11. If this symmetric assumption is appropriate, the Wilcoxon signed-rank test is the preferred nonparametric test for a population median. However, if the population is __________, the sign test is preferred.

12. With the Wilcoxon signed-rank test, the differences between the ______________ and the ____________________ of the population median are used instead of the differences between the matched-pair observations.

13. NOTES+COMMENTS:

(a) The Wilcoxon signed-rank test for a population median is based on the assumption that the population is symmetric. With this assumption, the population __________ is equal to the population _________. Thus, the Wilcoxon signed-rank test can also be used as a test about the _______________________________.

(b) There are several variations of the Wilcoxon signed-rank test that generally provide similar but not identical results. The test we use in section 18.2 is based on a ________________________ (which is much easier to calculate).

(c) JMP uses the exact Wilcoxon signed-rank test when $n \leq 20$ and a Student's t approximation when $n > 20$.

18.3 Mann-Whitney-Wilcoxon (MWW) Test

1. (Recall Chapter 10) A hypothesis test (t-test) about the difference between the means of two populations using two independent samples:

   (a) This parametric test required _______________ data and the assumption that both populations had a ___________ distribution.

   (b) If population standard deviations σ_1 and σ_2 were unknown, the sample standard deviations s_1 and s_2 provided estimates of σ_1 and σ_2.

   (c) The t distribution was used to make an ____________ about the difference between the means of the two populations.

2. We present a nonparametric test for the difference between two populations based on two independent samples. It can be used with either ___________ data or _______________ data and it does not require the assumption that the populations have a normal distribution.

3. Versions of the test were developed jointly by Mann and Whitney and also by Wilcoxon. As a result, the test has been referred to as the ______________________ and the __________________________. The tests are equivalent and both versions provide the same conclusion. We will refer to this nonparametric test as the ___________________________________ test (e.g., a two-tailed test):

   H_0 : The two populations are ____________

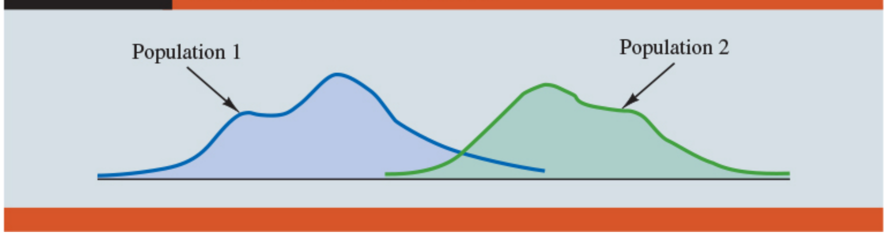
   H_a : The two populations are not identical

(i.e., either population may provide the smaller or larger values.)

4. If H_0 is rejected, we are using the test to conclude that the populations are not identical and that population 1 tends to provide either ________________ values than population 2.

5. (Figure 18.4) A situation where population 1 tends to provide smaller values than population 2. (Note that it is not necessary that all values from population 1 be less than all values from population 2.)

FIGURE 18.4 Two Populations Are Not Identical with Population 1 Tending to Provide the Smaller Values



6. First illustrate the MWW test using ______________ with ________________. Later, we will introduce a ______________ approximation based on the __________ distribution that will simplify the calculations required by the MWW test.

7. Example Consider the on-the-job performance ratings for employees at a Showtime Cinemas 20-screen multiplex movie theater.

   (a) During an employee performance review, the theater manager rated all 35 employees from best (rating 1) to worst (rating 35) in the theater's annual report. Knowing that the part-time employees were primarily college and high school students, the district manager asked if there was evidence of a significant difference in performance for college students compared to high school students.

   (b) In terms of the population of college students and the population of high school students who could be considered for employment at the theater, the hypotheses were:

H_0 : College and high school student populations are identical in terms of performance

H_a : College and high school student populations are not identical in terms of performance

(c) (Table 18.8) The theater manager's overall performance rating based on all 35 employees was recorded for each of these employees.

TABLE 18.8 Performance Ratings for a Sample of College Students and a Sample of High School Students Working at Showtime Cinemas

College Student	Manager's Performance Rating	High School Student	Manager's Performance Rating
1	15	1	18
2	3	2	20
3	23	3	32
4	8	4	9
		5	25

(d) (Table 18.9)(**The combined-sample ranks**) Use a 0.05 level of significance for this test and ________ the combined samples ________________.

TABLE 18.9 Ranks for the Nine Students in the Showtime Cinemas Combined Samples

College Student	Manager's Performance Rating	Rank	High School Student	Manager's Performance Rating	Rank
1	15	4	1	18	5
2	3	1	2	20	6
3	23	7	3	32	9
4	8	2	4	9	3
	Sum of Ranks	14	5	25	8
				Sum of Ranks	31

(e) **Sum the ranks** for each sample as shown in Table 18.9. The sum of ranks for the first sample will be the test statistic W for the MWW test: $W = 4 + 1 + 7 + 2 = 14$.

(f) We will always follow the procedure of using the sum of the ranks for ____________ as the ________________.

8. Why the sum of the ranks will help us select between the two hypotheses: H_0: The two populations are identical and H_a: The two populations are not identical.

(a) Letting C denote a college student and H denote a high school student, suppose the ranks of the nine students had the following order with the four college students having the four lowest ranks.

Rank	1	2	3	4	5	6	7	8	9
Student	C	C	C	C	H	H	H	H	H

(b) Notice that this permutation or ordering separates the two samples, with the college students all having a ____________ than the high school students.

(c) This is a strong indication that the two populations are ____________. The sum of ranks for the college students in this case is ____________________.

(d) Now consider a ranking where the four college students have the four highest ranks.

Rank	1	2	3	4	5	6	7	8	9
Student	H	H	H	H	H	C	C	C	C

This is another strong indication that the two populations are not identical. The sum of ranks for the college students in this case is ____________________.

(e) Thus, we see that the ______________ for the college students must be between 10 and 30. Values of __________ imply that college students have lower ranks than the high school students, whereas values of __________ imply that college students have higher ranks than the high school students.

(f) Either of these extremes would signal the two populations are not identical. However, if the two populations are identical, we would expect a ______ in the ordering of the C's and H's so that the sum of ranks W is closer to the ________ of the two extremes, or nearer to ______________.

9. (Figure 18.5)(Table 18.10) Making the assumption that the two populations are identical, we used a computer program to compute __________________ for the nine students. For each ordering, we computed the _______________ for the college students. This provided the probability distribution showing the exact sampling distribution of W.

FIGURE 18.5 Exact Sampling Distribution of the Sum of the Ranks for the Sample of College Students

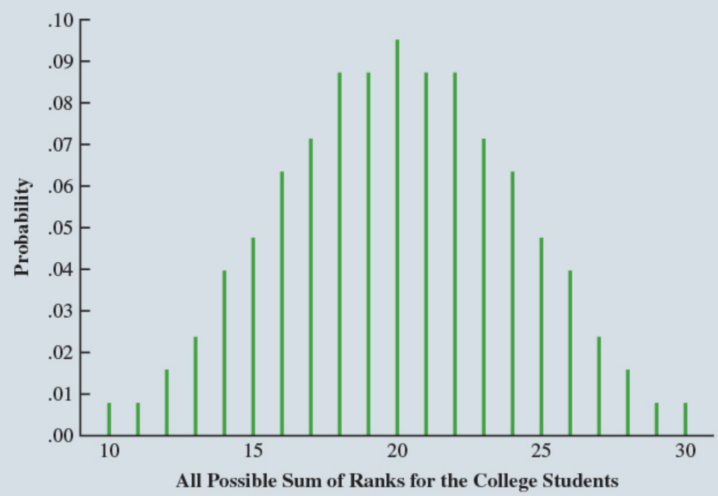


TABLE 18.10 Probabilities for the Exact Sampling Distribution of the Sum of the Ranks for the Sample of College Students

W	Probability	W	Probability
10	.0079	20	.0952
11	.0079	21	.0873
12	.0159	22	.0873
13	.0238	23	.0714
14	.0397	24	.0635
15	.0476	25	.0476
16	.0635	26	.0397
17	.0714	27	.0238
18	.0873	28	.0159
19	.0873	29	.0079
		30	.0079

10. Use the sampling distribution of W in Figure 18.5 to compute the p-value for the test. Table 18.9 shows that the sum of ranks for the four college student is __________. Because this value of W is in the __________ of the sampling distribution, we begin by computing the lower tail probability __________:

$$\begin{aligned} P(W \leq 14) &= P(10) + P(11) + P(12) + P(13) + P(14) \\ &= 0.0079 + 0.0079 + 0.0159 + 0.0238 + 0.0397 = 0.0952 \end{aligned}$$

11. The two-tailed p-value __________. With $\alpha = 0.05$ as the level of significance and p-value > 0.05, the MWW test conclusion is that we cannot reject

the null hypothesis that the populations of college and high school students are identical.

12. Use the same combined-sample ranking procedure and use the ___________ distribution approximation of W to compute the p-value and draw the conclusion.

13. Example Third National Bank.

(a) The bank manager is monitoring the balances maintained in checking accounts at two branch banks and is wondering if the populations of account balances at the two branch banks are identical.

(b) (Table 18.11) Two independent samples of checking accounts are taken with sample sizes $n_1 = 12$ at branch 1 and $n_2 = 10$

TABLE 18.11 Account Balances for Two Branches of Third National Bank

Branch 1		Branch 2	
Account	Balance ($)	Account	Balance ($)
1	1095	1	885
2	955	2	850
3	1200	3	915
4	1195	4	950
5	925	5	800
6	950	6	750
7	805	7	865
8	945	8	1000
9	875	9	1050
10	1055	10	935
11	1025		
12	975		

(c) (Table 18.12) The first step in the MWW test is to rank the combined data from the lowest to highest values. In that case of the two or more values are the same, the tied values are assigned the average rank of their positions in the combined data set.

TABLE 18.12 Assigned Ranks for the Combined Account Balance Samples

Branch	Account	Balance	Rank
2	6	750	1
2	5	800	2
1	7	805	3
2	2	850	4
2	7	865	5
1	9	875	6
2	1	885	7
2	3	915	8
1	5	925	9
2	10	935	10
1	8	945	11
1	6	950	12.5
2	4	950	12.5
1	2	955	14
1	12	975	15
2	8	1000	16
1	11	1025	17
2	9	1050	18
1	10	1055	19
1	1	1095	20
1	4	1195	21
1	3	1200	22

(d) (Table 18.13) The next step is to sum the ranks for each sample: 169.5 for sample 1 and 83.5 for sample 2 are shown. Thus, we have $W = 169.5$. When both samples sizes are ____________, a normal approximation of the sampling distribution of W can be used.

TABLE 18.13 Combined Ranking of the Data in the Two Samples from Third National Bank

	Branch 1			Branch 2	
Account	Balance ($)	Rank	Account	Balance ($)	Rank
1	1095	20	1	885	7
2	955	14	2	850	4
3	1200	22	3	915	8
4	1195	21	4	950	12.5
5	925	9	5	800	2
6	950	12.5	6	750	1
7	805	3	7	865	5
8	945	11	8	1000	16
9	875	6	9	1050	18
10	1055	19	10	935	10
11	1025	17		Sum of Ranks	83.5
12	975	15			
	Sum of Ranks	169.5			

14. Under the assumption that the null hypothesis is true and the populations are identical, the sampling distribution of the test statistic W is:

$$\text{Mean} : \underline{\qquad\qquad\qquad\qquad\qquad} \tag{18.5}$$

Standard deviation : ________________________ (18.6)

Distribution form: Approximately normal provided _______________.

15. Since the test statistic W is discrete and the normal distribution is continuous, we will again use the ________________ factor for the normal distribution approximation.

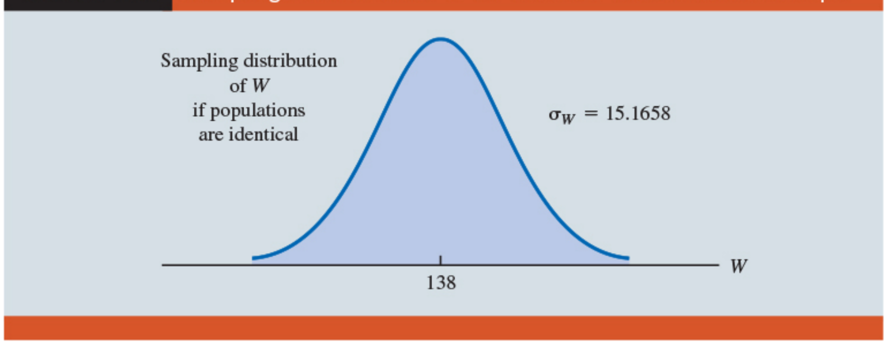
16. Example Third National Bank.

    (a) (Figure 18.6) Given the sample sizes ________________, equations (18.5) and (18.6) provide the following mean and standard deviation for the sampling distribution:

$$\text{Mean} : \mu_W = (1/2)(12)(12 + 10 + 1) = 138$$

$$\text{Standard deviation} : \sigma_W = \sqrt{(1/12)(12)(10)(12 + 10 + 1)} = 15.1658$$

FIGURE 18.6 Sampling Distribution of W for the Third National Bank Example



    (b) With $W = 169.5$ in the __________ of the sampling distribution, we have the following p-value calculation:

$$P(W \geq 169.5) = $$

____

    (c) Using the standard normal random variable and $z = 2.04$, the two-tailed p-value __________________. With p-value ≤ 0.05, _________ and conclude that the two populations of account balances are not identical. The upper tail value for test statistic W indicates that the population of account balances at branch 1 tends to be _______.

17. Some applications of the MWW test make it appropriate to assume that the two populations have ________________ and if the populations differ, it is only by a ________ in the location of the distributions.

18. If the two populations have the ______________, the hypothesis test may be stated in terms of the difference between the two ____________________. Any difference between the medians can be interpreted as the shift in location of one population compared to the other. In this case, the three forms of the MWW test about the medians ($M_i, i = 1, 2$) of the two populations are as follows:

Two-Tailed Test	Lower Tail Test	Upper Tail Test
H_0: $M_1 - M_2 = 0$	H_0: $M_1 - M_2 \geq 0$	H_0: $M_1 - M_2 \leq 0$
H_a: $M_1 - M_2 \neq 0$	H_a: $M_1 - M_2 < 0$	H_a: $M_1 - M_2 > 0$

18.4 Kruskal-Wallis Test

1. (Recall Chapter 13, ANOVA) We considered a parametric test for three or more populations when we used __________________ and assumed that the populations had normal distributions with the same standard deviations. Based on an independent random sample from each population, we used the ________________ to test for differences among the __________________.

2. The nonparametric ____________________ is based on the analysis of independent random samples from each of _____ populations. This procedure can be used with either __________ data or ______________ data and does not require the assumption that the populations have normal distributions:

$$H_0 : \text{All populations are } \_\_\_\_\_\_\_\_\_\_\_$$
$$H_a : \text{Not all populations are identical}$$

3. If H_0 is rejected, we will conclude that there is a difference among the populations with one or more populations tending to provide __________________ values compared to the other populations.

4. Example Performance Evaluation Ratings for 20 Williams Employees

(a) (Table 18.14) Williams Manufacturing Company hires employees for its management staff from three different colleges. Recently, the company's personnel director began reviewing the annual performance reports for the management staff in an attempt to determine whether there are differences in the performance ratings among the managers who graduated from the three colleges. The performance rating shown for each manager is recorded on a scale from 0 to 100, with 100 being the highest possible rating.

TABLE 18.14 Performance Evaluation Ratings for 20 Williams Employees

College A	College B	College C
25	60	50
70	20	70
60	30	60
85	15	80
95	40	90
90	35	70
80		75

(b) Suppose we want to test whether the three populations of managers are identical in terms of ________________. We will use a 0.05 level of significance for the test.

(c) (Table 18.15) The first step in the Kruskal-Wallis procedure is to __________ ________________ from lowest to highest values. Note that we assigned the average ranks to tied performance ratings of 60, 70, 80, and 90.

TABLE 18.15 Combined Rankings for the Three Samples

College A	Rank	College B	Rank	College C	Rank
25	3	60	9	50	7
70	12	20	2	70	12
60	9	30	4	60	9
85	17	15	1	80	15.5
95	20	40	6	90	18.5
90	18.5	35	5	70	12
80	15.5	Sum of Ranks	27	75	14
Sum of Ranks	95			Sum of Ranks	88

5. **The Kruskal-Wallis test statistic**:

(18.7)

____

where

- k = the number of populations
- n_i =the number of observations in sample i
- $n_T = \sum_{i=1}^{k} n_i$ = the total number of observations in all samples
- R_i = the sum of the ranks for sample i

(a) Kruskal and Wallis were able to show that, under the null hypothesis assumption of identical populations, the sampling distribution of H can be approximated by a ____________ distribution with _________ degrees of freedom.

(b) This approximation is generally acceptable if the ______________ for each of the k populations are all __________________________.

(c) The null hypothesis of identical populations will be rejected if the test statistic H is large. As a result, the Kruskal-Wallis test is _________ expressed as an ____________ test.

6. Example Performance Evaluation Ratings for 20 Williams Employees

(a) The value of the Kruskal-Wallis test statistic:

$$H = \frac{12}{20(21)}\left[\frac{(95)^2}{7} + \frac{(27)^2}{6} + \frac{(88)^2}{7}\right] - 3(20+1) = 8.92$$

(b) We find ________________ has an area of 0.025 in the upper tail of the chi-square distribution and ______________ has an area of 0.01 in the upper tail of the chi-square distribution.

(c) With $H = 8.92$ between 7.378 and 9.21, we can conclude that the p-value is between 0.025 and 0.01. Because p-value $\leq \alpha = 0.05$, we reject H_0 and conclude that the three populations are not all the same. The three populations of performance ratings are not identical and differ significantly depending upon the college.

(d) Because the sum of the ranks is relatively low for the sample of managers who graduated from ___________, it would be reasonable for the company to either reduce its recruiting from college B, or at least evaluate the college B graduates more thoroughly before making a hiring decision.

7. In some applications of the Kruskal-Wallis test, it may be appropriate to make the assumption that the populations have ________________ and if they differ, it is only by a ________________ for one or more of the populations.

8. If the k populations are assumed to have the same shape, the hypothesis test can be stated in terms of the ____________________. In this case, the hypotheses for the Kruskal-Wallis test would be written as follows:

$$H_0 : M_1 = M_2 = \cdots = M_k$$
$$H_a : \text{Not all Medians are equal}$$

9. NOTES+COMMENTS: The example in this section used quantitative data on employee performance ratings to conduct the Kruskal- Wallis test. This test could also have been used if the data were the ________________ of the 20 employees in terms of performance. In this case, the test would use the ordinal data directly. The step of converting the quantitative data into rank-ordered data would not be necessary.

18.5 Rank Correlation

1. (Recall Chapter 3) The Pearson product moment correlation coefficient is a measure of the ________________ between two variables using quantitative data.

2. The Spearman rank-correlation coefficient has been developed for a correlation measure of association between two variables when ________________________ are available:

   ____________________ (18.8)(   )

   where

   n =the number of observations in the sample

   x_i = the rank of observation i with respect to the first variable

y_i = the rank of observation i with respect to the second variable

____

3. The Spearman rank-correlation coefficient ranges from ______________ and its interpretation is similar to the Pearson product moment correlation coefficient for quantitative data.

4. A rank-correlation coefficient near ______________ indicates a strong ______________ ______________ association between the ranks for the two variables.

5. Example Sales Potential and Actual Two-Year Sales Data

   (a) A company wants to determine whether individuals who had a greater potential at the time of employment turn out to have higher sales records. To investigate, the personnel director reviewed the original job interview reports, academic records, and letters of recommendation for 10 current members of the sales force.

   (b) After the review, the director ranked the 10 individuals in terms of their potential for success at the time of employment and assigned the individual who had the most potential the rank of 1.

   (c) (Table 18.16) Data were then collected on the actual sales for each individual during their first two years of employment. On the basis of the actual sales records, a second ranking of the 10 individuals based on sales performance was obtained.

TABLE 18.16 Sales Potential and Actual Two-Year Sales Data

Salesperson	Ranking of Potential	Two-Year Sales (units)	Ranking According to Two-Year Sales
A	2	400	1
B	4	360	3
C	7	300	5
D	1	295	6
E	6	280	7
F	3	350	4
G	10	200	10
H	9	260	8
I	8	220	9
J	5	385	2

(d) (Table 18.17) Computation of the Spearman Rank-Correlation Coefficient for Sales Potential and Sales Performance

TABLE 18.17 Computation of the Spearman Rank-Correlation Coefficient for Sales Potential and Sales Performance

Salesperson	x_i = Ranking of Potential	y_i = Ranking of Sales Performance	$d_i = x_i - y_i$	d_i^2
A	2	1	1	1
B	4	3	1	1
C	7	5	2	4
D	1	6	−5	25
E	6	7	−1	1
F	3	4	−1	1
G	10	10	0	0
H	9	8	1	1
I	8	9	−1	1
J	5	2	3	9
				$\Sigma d_i^2 = 44$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2+1)} = 1 - \frac{6(44)}{10(100-1)} = .733$$

(e) $r_s = 0.733$ indicates a ________________ between the ranks based on potential and the ranks based on sales performance. Individuals who ranked higher in potential at the time of employment tended to rank higher in two-year sales performance.

6. Use the sample rank correlation r_s to make an inference about the population rank correlation coefficient ρ_s:

$$H_0 : \_\_\_\_\_\_ \qquad H_a : \_\_\_\_\_\_$$

7. (**S ampling distribution of** r_s) Under the assumption that the null hypothesis is true and the population rank-correlation coefficient is 0, the following sampling distribution of r_s can be used to conduct the test.

Mean : ________ (18.9)

Standard deviation : ____________ (18.10)

Distribution form: Approximately normal provided ________

8. Example Sales Potential and Actual Two-Year Sales Data

(a) The sample rank-correlation coefficient for sales potential and sales performance is ____________. Using equation (18.9), we have __________, and using equation (18.10), we have _________________________.

(b) With the sampling distribution of r_s approximated by a normal distribution, the standard normal random variable z becomes the test statistic with

____

(c) Using the standard normal probability table and $z = 2.20$, we find the two-tailed p-value ______________________. With a 0.05 level of significance, p-value $\leq \alpha$. Thus, we ________ the null hypothesis that the population rank-correlation coefficient is zero.

(d) The test result shows that there is a ____________ rank correlation between potential at the time of employment and actual sales performance.

😍 EXERCISES

18.1 : 1, 3, 6, 9

18.2 : 12, 15, 17

18.3 : 18, 21

18.4 : 26, 29

18.5 : 32, 35

SUP : 39, 41, 45

“對一個不滿意的人生你只有兩種選擇，強迫自己接受，或說服自己改變。”
“You can only do one of two things to an unsatisfying life: force yourself to accept it, or convince yourself to change.”

— 媽的多重宇宙 (Everything Everywhere All at Once, 2022)

考試科目： 統計學 (一)					開課班別： 統計學整合開課	命題教授: 吳漢銘
考試日期：3 月 17 日 (四) 16:10-17:30	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 3頁，印刷份數: 56 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (範圍: §10∼§11)	O	×	×	×	×	

注意事項: (1) 答案卷請寫上姓名及學號。(2) 請按題號順序書寫。(3) 每一題號需置於答案卷最左邊。(4) 中英文作答皆可。(5) 可用鉛筆。(6) 需要計算過程。(7) 同時交回答案卷、題目卷、計算紙。(8) 總分共 120 分。

1. (10%) 試舉生活上一個應用「Hypothesis Testing for the Difference Between Two Population Means」的情境範例，同時說明，資料中的變數需滿足什麼條件才能合適地應用此統計方法。

2. (20%) Consider the following data for two independent random samples taken from two normal populations.

Sample 1:	10, 7, 13, 7, 9, 8
Sample 2:	8, 7, 8, 4, 6, 9

   (a) Compute the two sample means.
   (b) Compute the two sample standard deviations.
   (c) What is the point estimate of the difference between the two population means?
   (d) What is the 90% confidence interval estimate of the difference between the two population means?

3. (20%) **Gender Differences in Raise or Promotion Expectations**. The Adecco Workplace Insights Survey sampled men and women workers and asked if they expected to get a raise or promotion this year. Suppose the survey sampled 200 men and 200 women. If 104 of the men replied Yes and 74 of the women replied Yes, are the results statistically significant in that you can conclude a greater proportion of men are expecting to get a raise or a promotion this year?

   (a) State the hypothesis test in terms of the population proportion of men and the population proportion of women.
   (b) What is the sample proportion for men? For women?
   (c) Use a .01 level of significance. What is the p-value and what is your conclusion?

4. (10%) A sample of 16 items provides a sample standard deviation of 9.5. Test the following hypotheses using $\alpha = 0.05$. What is your conclusion? Use both the p-value approach and the critical value approach.

$$H_0 : \sigma^2 \geq 50, \qquad H_a : \sigma^2 < 50$$

5. (20%) **Price Comparison of Smoothie Blenders**. A personal fitness company produces both a deluxe and a standard model of a smoothie blender for home use. Selling prices obtained from a sample of retail outlets follow.

	Model Price ($)						
Retail Outlet	1	2	3	4	5	6	7
Standard	39	39	45	38	40	39	35
Deluxe	27	28	35	30	30	34	29

(a) The manufacturer's suggested retail prices for the two models show a $10 price differential. Use a 0.05 level of significance and test that the mean difference between the prices of the two models is $10.

(b) What is the 95% confidence interval for the difference between the mean prices of the two models?

6. (20%) **Smartphone Battery Life**. Battery life is an important issue for many smartphone owners. Public health studies have examined ″low-battery anxiety″ and acute anxiety called nomophobia that results when a smartphone user′ s phone battery charge runs low and then dies. Battery life between charges for the Samsung Galaxy S9 averages 31 hours when the primary use is talk time and 10 hours when the primary use is Internet applications. Because the mean hours for talk time usage is greater than the mean hours for Internet usage, the question was raised as to whether the variance in hours of usage is also greater when the primary use is talk time. Sample data showing battery life between charges for the two applications follows.

Primary Use: Talking: `35.8, 22.2, 24.0, 32.6, 18.5, 42.5, 28.0, 23.8, 30.0, 22.8, 20.3, 35.5`

Primary Use: Internet: `14.0, 12.5, 16.4, 11.9, 9.9, 3.1, 5.4, 11.0, 15.2, 4.0, 4.7`

(a) Formulate hypotheses about the two population variances that can be used to determine if the population variance in battery life is greater for the talk time application.

(b) What are the standard deviations of battery life for the two samples?

(c) Conduct the hypothesis test and compute the p-value. Using a 0.05 level of significance, what is your conclusion?

機率表

Upper tail probability.

i	p	z	t_6	t_9	t_{11}	χ^2_{15}	χ^2_{16}	$F_{11,10}$	$F_{10,11}$
1	0.200	0.8416	0.9057	0.8834	0.8755	19.3107	20.4651	1.7235	1.6940
2	0.100	1.2816	1.4398	1.3830	1.3634	22.3071	23.5418	2.3018	2.2482
3	0.050	1.6449	1.9432	1.8331	1.7959	24.9958	26.2962	2.9430	2.8536
4	0.025	1.9600	2.4469	2.2622	2.2010	27.4884	28.8454	3.6649	3.5257
5	0.010	2.3263	3.1427	2.8214	2.7181	30.5779	31.9999	4.7715	4.5393
6	0.005	2.5758	3.7074	3.2498	3.1058	32.8013	34.2672	5.7462	5.4183

公式

- $df = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$

國立政治大學 110 學年度第 2 學期 線上小考 (2) 考試命題紙

考試科目： 統計學 (一)	開課班別： 統計學整合開課	命題教授: 吳漢銘

考試日期：5 月 12 日 (四) 16:10-17:30	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共2頁，印刷份數: 56 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (範圍: §14)	O	×	×	×	×	

注意事項: (1) 答案卷請寫上姓名及學號。(2) 請按題號順序書寫。(3) 每一題號需置於答案卷最左邊。(4) 中英文作答皆可。(5) 可用鉛筆。(6) 需要計算過程。(7) 上傳答案卷: `http://www.hmwu.idv.tw`，點選【作業考試上傳區】，login: `stat110`，password: `xxxx`。(8) 答題卷拍照請務必清晰、需滿版平整，周遭不可有雜物入鏡。答題卷檔名: 學號-姓名-頁碼.jpg 或學號-姓名-頁碼.pdf。(9) 總分共 120 分。

1. (20%) 名詞解釋 (不能只列出公式，需解釋其意思): (a) Coefficient of determination (判定係數)，(b) Residuals (殘差)，(c) Normal probability plot (常態機率圖)(Hint: 用途為何?)。(d) Leverage of observation i and the high leverage points (第 i 個觀察值的槓桿及高度槓桿點)。

2. (30%) The simple linear regression (SLR).

   (a) (5%) What is a SLR model? (Explain all the symbols and characters in the model.)

   (b) (10%) In the SLR model, what assumptions are made about the error term?

   (c) (15%) Derive the estimated regression coefficients (b_0, b_1) using least square method.

3. (10%) When applying a SLR analysis to a given dataset, $\{y_i, x_i\}_{i=1}^n$, let $\hat{y}^*$ be the point estimate of $E(y^*)$. The formula for estimating the variance of $\hat{y}^*$, denoted by $s^2_{\hat{y}^*}$, is

$$s^2_{\hat{y}^*} = s^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \qquad (14.22)$$

   Also, let $\hat{y}^*$ be the predictor of an individual value of y^* when $x = x^*$. The formula for estimating the variance corresponding to the prediction of the value of y when $x = x^*$, denoted s^2_{pred}, is

$$s^2_{pred} = s^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \qquad (14.25)$$

   Why does Equation (14.25) have an extra term s^2 when compared to Equation (14.22)?

4. (20%) How do you validate the assumptions of a SLR model? (請依課本第 14 章內容回答。)

5. (20%) When you apply SLR analysis to a given data set, how can you detect outliers and influential observations? (請依課本第 14 章內容回答。)

國立政治大學 110 學年度第 2 學期 線上小考 (2) 考試命題紙

考試科目： 統計學 (一)　　開課班別： 統計學整合開課　　命題教授: 吳漢銘

考試日期：5 月 12 日 (四) 16:10-17:30	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共2頁，印刷份數: 56 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (範圍: §14)	O	×	×	×	×	

6. (20%) **Estimating Setup Time**. Sherry is a production manager for a small manufacturing shop and is interested in developing a predictive model to estimate the time to produce an order of a given size - that is, the total time to produce a certain quantity of the product. She has collected data on the total time to produce 30 different orders of various quantities in the file Setup.

```
Quantity: 105, 125, 135, ...
Total Time: 173, 188, 222, ...
```

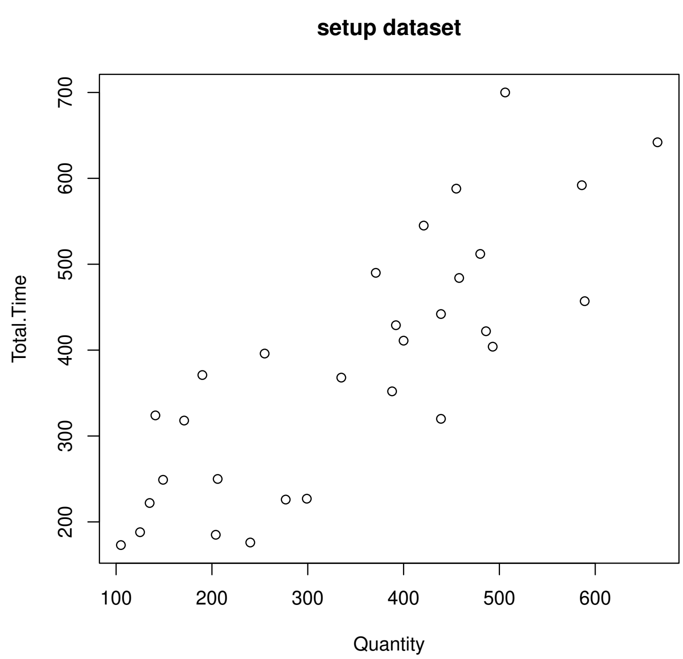
The scatter diagram with quantity as the independent variable and the report of a SLR model analysis for this data are given below.

(a) (5%) Develop the estimated regression equation. Interpret the intercept and slope.

(b) (5%) Test for a significant relationship (t-test, F-test). Use $\alpha = 0.05$. (The interpretation is required.)

(c) (5%) Interpret the coefficient of determination.

(d) (5%) Did the estimated regression equation provide a good fit?




```
Call:
lm(formula = Total.Time ~ Quantity, data = setup)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.5419    38.4893   3.054  0.00491 **
Quantity      0.7631     0.1015   7.521 3.42e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 85.59 on 28 degrees of freedom
Multiple R-squared:  0.6689,    Adjusted R-squared:  0.6571
F-statistic: 56.57 on 1 and 28 DF,  p-value: 3.42e-08

Analysis of Variance Table

Response: Total.Time
          Df Sum Sq Mean Sq F value   Pr(>F)
Quantity   1 414398  414398   56.57 3.42e-08 ***
Residuals 28 205111    7325
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


注意： 1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。

2、考試時不得有交談、窺視、夾帶、抄襲、傳遞、代考或其它作弊等舞弊行為，考畢務必交卷，不得攜卷出場，違者依考場規則議處。

國立政治大學 110 學年度第 2 學期 線上小考 (3) 考試命題紙

考試科目： 統計學 (一)　　開課班別： 統計學整合開課　　命題教授: 吳漢銘

考試日期：6 月 02 日 (四) 16:10-17:30	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共4頁，印刷份數: 56 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (範圍: §15)	O	×	×	×	×	

注意事項: (1) 答案卷請寫上科目、系級、學號及姓名。(2) 請按題號順序書寫。(3) 每一題號需置於答案卷最左邊。(4) 中英文作答皆可。(5) 建議用深色原子筆。(6) 需要計算過程。(7) 上傳答案卷: `http://www.hmwu.idv.tw`，點選【作業考試上傳區】，login: `stat110`，password: `xxxx`。(8) 答題卷拍照請務必清晰、需滿版平整，周遭不可有雜物入鏡。答題卷檔名: 學號-姓名-exam3.pdf。(9) 總分共 120 分。(10) 請複寫下列宣誓詞至答案卷上。

> 本人姓名 重視榮譽，以認真負責的態度，參與於本次線上遠距考試，恪遵各項考試規則，無任何不法或舞弊情事，如違誓言，願受校方最嚴厲之處罰，謹誓。

1. (15%) 名詞解釋 (不能只列出公式，需解釋其意思): (a) Logistic regression (邏輯斯迴歸)。(b) Odds (勝算)。(c) Odds ratio (勝算比)。(不能僅寫「勝算比是兩個勝算的比值」)。

2. (15%) Suppose the equation for the multiple linear regression (MLR) model is expressed by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

   (a) (5%) What is the estimated multiple regression equation? (請依課本所使用的符號作答)

   (b) (10%) How do you interpret each regression coefficient? (Except for the intercept.)

3. (20%) Apply a MLR to a given dataset, $\{y_i, x_{i1}, x_{i2}, \cdots, x_{ip}\}_{i=1}^n$,

   (a) (5%) What is the definition of the multiple coefficient of determination (多重判定係數)? (記做 R^2)

   (b) (5%) How to interpret R^2?

   (c) (10%) The formula for the adjusted R^2 is $R_a^2 = 1 - (1 - R^2)\dfrac{n-1}{n-p-1}$。What is the purpose of adjusting for R^2?

4. (15%) Multicollinearity (共線性)

   (a) (5%) What does the multicollinearity mean in the multiple regression analysis?

   (b) (10%) What are the effects when the multicollinearity is severe? (請依課本內容作答)

6. (15%) Data for two variables, x and y, follow.

i	1	2	3	4	5	mean	sum square
x_i	22	24	26	28	40	$\bar{x} = 28$	$\sum x_i^2 = 4120$
y_i	12	21	31	35	70	$\bar{y} = 33.8$	$\sum y_i^2 = 7671$

(a) (0%) Develop the estimated regression equation for these data.

(b) (5%) Compute the studentized deleted residual for the 4th observation ($i = 4$). At the 0.05 level of significance, can this observation be classified as an outlier? Explain.

(c) (5%) Compute the leverage value for the 4th observation ($i = 4$). Do there appear to be an influential observation? Explain.

(d) (5%) Compute Cook's distance measure for the 4th observation ($i = 4$). Is this observation influential? Explain.

(公式見最後一頁)

```
> x <- c(22, 24, 26, 28, 40)
> y <- c(12, 21, 31, 35, 70)
> mylm <- lm(y~x)
> summary(mylm)

Call:
lm(formula = y ~ x)

Residuals:
    1     2     3     4     5 
-3.14 -0.36  3.42  1.20 -1.12 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -53.2800     5.7864  -9.208 0.002709 ** 
x             3.1100     0.2016  15.428 0.000592 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.851 on 3 degrees of freedom
Multiple R-squared:  0.9876,	Adjusted R-squared:  0.9834 
F-statistic:   238 on 1 and 3 DF,  p-value: 0.0005915

> anova(mylm)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)    
x          1 1934.42 1934.42  238.03 0.0005915 ***
Residuals  3   24.38    8.13                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> # remove i=4
> x.4 <- c(22, 24, 26, 40)
> y.4 <- c(12, 21, 31, 70)
> mylm.4 <- lm(y.4~x.4)
> summary(mylm.4)

Call:
lm(formula = y.4 ~ x.4)

Residuals:
    1     2     3     4 
-2.84 -0.06  3.72 -0.82 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -53.5800     6.8614  -7.809  0.01601 * 
x.4           3.1100     0.2376  13.090  0.00579 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.36 on 2 degrees of freedom
Multiple R-squared:  0.9885,	Adjusted R-squared:  0.9827 
F-statistic: 171.3 on 1 and 2 DF,  p-value: 0.005786

> anova(mylm.4)
Analysis of Variance Table

Response: y.4
          Df  Sum Sq Mean Sq F value   Pr(>F)   
x.4        1 1934.42 1934.42  171.34 0.005786 **
Residuals  2   22.58   11.29                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

考試日期：6 月 02 日 (四) 16:10-17:30	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共4頁，印刷份數: 56 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (範圍: §15)	O	×	×	×	×	

6. (40%) **College Retention**. Over the past few years the percentage of students who leave Lakeland College at the end of the first year has increased. Last year Lakeland started a voluntary one-week orientation program to help first-year students adjust to campus life. If Lakeland is able to show that the orientation program has a positive effect on retention, they will consider making the program a requirement for all first-year students. Lakeland's administration also suspects that students with lower GPAs have a higher probability of leaving Lakeland at the end of the first year. In order to investigate the relation of these variables to retention, Lakeland selected a random sample of 100 students from last year's entering class. The data are contained in the data set named Lakeland; a portion of the data follows.

Student:	1	2	3	⋯	98	99	100
GAP:	3.78	2.38	1.30	⋯	2.57	1.70	3.85
Program:	1	0	0	⋯	1	1	1
Return:	1	1	0	⋯	1	1	1

The dependent variable was coded as $y = 1$ if the student returned to Lakeland for the sophomore year and $y = 0$ if not. The two independent variables are:

$$x_1 = \text{GPA at the end of the first semester}$$

$$x_2 = \begin{cases} 0 & \text{if the student did not attend the orientation program} \\ 1 & \text{if the student attended the orientation program} \end{cases}$$

(a) (5%) Write the logistic regression equation relating x_1 and x_2 to y.

(b) (5%) What is the interpretation of $E(y)$ when $x_2 = 0$?

(c) (5%) Use both independent variables and statistical software to compute the estimated logit.

(d) (5%) Conduct a test for overall significance using $\alpha = 0.05$.

(e) (5%) Use $\alpha = 0.05$ to determine whether each of the independent variables is significant.

(f) (5%) Use the estimated logit computed in part (c) to estimate the probability that students with a 2.5 grade point average who did not attend the orientation program will return to Lakeland for their sophomore year. What is the estimated probability for students with a 2.5 grade point average who attended the orientation program?

(g) (5%) What is the estimated odds ratio for the orientation program? Interpret it.

(h) (5%) Would you recommend making the orientation program a required activity? Why or why not?

```
glm(formula = Return ~ GPA + Program, family = binomial, data = Lakeland)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9610  -0.4828   0.2848   0.5980   1.8154

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.8926     1.7472  -3.945 7.98e-05 ***
GPA           2.5388     0.6729   3.773 0.000161 ***
Program1      1.5608     0.5631   2.772 0.005579 **
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 128.207  on 99  degrees of freedom
Residual deviance:  80.338  on 97  degrees of freedom
AIC: 86.338

Number of Fisher Scoring iterations: 5

>
> anova(model, test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: Return
Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                       99    128.207
GPA      1   40.008        98     88.199 2.53e-10 ***
Program  1    7.862        97     80.338 0.005049 **
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
```

公式:

1. $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$
2. $\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}}$, $s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i}$
3. $D_i = \frac{(y_i - \hat{y}_i)^2}{(p+1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right]$

<table>
<tr><td colspan="8">國立政治大學 110 學年度第 2 學期 期中考 考試命題紙
考試科目： 統計學 (二) 開課班別： 統計學整合開課 命題教授: 吳漢銘</td></tr>
<tr><td>考試日期：4 月 12 日 (二) 13:10-14:40</td><td colspan="5">※准帶項目打「O」，否則打「×」</td><td rowspan="3">1. 需加發計算紙或答案紙請在試題內封袋備註。
2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：</td></tr>
<tr><td>本試題共 3頁，印刷份數: 56 份</td><td>計算機</td><td>課本</td><td>筆記</td><td>字典</td><td>手機平板筆電</td></tr>
<tr><td>備註：注意事項要看!! (範圍: §10~§13)</td><td>O</td><td>×</td><td>×</td><td>×</td><td>×</td></tr>
</table>

注意事項: (1) 答案卷請寫上姓名及學號。(2) 請按題號順序書寫。(3) 每一題號需置於答案卷最左邊。(4) 中英文作答皆可。(5) 可用鉛筆。(6) 需要計算過程。(7) 同時交回答案卷、題目卷、計算紙。(8) 總分共 120 分。(9) 下列假設檢定問題，一律需寫出虛無假設、檢定統計量、決策法則及結論。(10) 小數請計算至小數點以下 4 位。

1. (25%) **統計名詞解釋**:
   (a) 配對樣本 (Matched samples)
   (b) 卡方適合度檢定 (Chi-square goodness of fit test)
   (c) 變異數分析表格 (ANOVA table)
   (d) 均方誤差 (Mean square error, Mean square due to error)
   (e) 多重比較 (multiple comparison)(ps. 於變異數分析中，拒絕虛無假設後所要進行的程序)

2. (15%) 採用變異數分析 (ANOVA) 於多個母體平均數的比較所要求的三個前置假設 (assumptions) 為何?

3. (10%) Let $X_1, X_2, \cdots, X_n$ are observations of a random sample of size n from the normal distribution $N(\mu, \sigma^2)$. $\bar{X}$ is the sample mean and S^2 is the sample variance. Derive the $(1-\alpha)\%$ confidence interval for population variance. (不能只有寫信賴區間的公式)

4. (20%) **Weekly Demand at Whole Foods Market**. The manager at a Whole Foods Market is responsible for managing store inventory. The mathematical models that she uses to determine how much inventory to stock rely on product demand being normally distributed. In particular, the weekly demand of sriracha chili kale chips at a Whole Foods Market store is believed to be normally distributed. Use a goodness of fit test and the following data to test this assumption. Use $\alpha = 0.10$

   18 20 22 27 22 25 22 27 25 24 26 23 20 24 26
   27 25 19 21 25 26 25 31 29 25 25 28 26 28 24

考試日期：4 月 12 日 (二) 13:10-14:40	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 3頁，印刷份數: 56 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (範圍: §10~§13)	O	×	×	×	×	

5. (20%) **Company Reputation and Management Quality Survey**. The Wall Street Journal Annual Corporate Perceptions Study surveyed readers and asked how they rated the quality of management and the reputation of the company for more than 250 world-wide corporations. Both the quality of management and the reputation of the company were rated on a categorical scale of excellent, good, and fair categorical. Assume the sample data for 200 respondents below applies to this study.

Reputation of Company	Quality of Management	Number
Excellent	Excellent	40
Excellent	Good	35
Excellent	Fair	25
Good	Excellent	25
Good	Good	35
Good	Fair	10
Fair	Excellent	5
Fair	Good	10
Fair	Fair	15

(a) Use a 0.05 level of significance and test for independence of the quality of management and the reputation of the company. What is the p-value and what is your conclusion?

(b) If there is a dependence or association between the two ratings, discuss and use probabilities to justify your answer.

6. (30%) The following data are from a completely randomized design. ($\alpha = 0.05$).

Treatment	Values
1	63, 47, 54, 40
2	82, 72, 88, 66, 77
3	69, 58, 62

(a) Use analysis of variance to test for a significant difference among the means of the three treatments. (需列出虛無假設及完整 ANOVA Table)

(b) Use Fisher's LSD procedure to determine which means are different.

考試科目： 統計學 (二)	開課班別： 統計學整合開課					命題教授: 吳漢銘
考試日期：4 月 12 日 (二) 13:10-14:40	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 3頁，印刷份數: 56 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (範圍: §10~§13)	O	×	×	×	×	

機率表

Upper tail probability.

i	p	z	t_6	t_9	t_{11}	χ^2_4	χ^2_5	$F_{2,9}$	$F_{3,9}$
1	0.200	0.8416	0.9057	0.8834	0.8755	5.9886	7.2893	1.9349	1.9007
2	0.100	1.2816	1.4398	1.3830	1.3634	7.7794	9.2364	3.0065	2.8129
3	0.050	1.6449	1.9432	1.8331	1.7959	9.4877	11.0705	4.2565	3.8625
4	0.025	1.9600	2.4469	2.2622	2.2010	11.1433	12.8325	5.7147	5.0781
5	0.010	2.3263	3.1427	2.8214	2.7181	13.2767	15.0863	8.0215	6.9919
6	0.005	2.5758	3.7074	3.2498	3.1058	14.8603	16.7496	10.1067	8.7171

Standard Normal Distribution

Percentage	z
10%	-1.28
20%	-0.84
30%	-0.52
40%	-0.25
50%	0.00
60%	0.25
70%	0.52
80%	0.84
90%	1.28

國立政治大學 110 學年度第 2 學期 期末考 考試命題紙

考試科目： 統計學 (二)	開課班別： 統計學整合開課				命題教授: 吳漢銘
考試日期：6 月 14 日 (二) 13:10-14:40	※准帶項目打「O」，否則打「×」				1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共5頁，印刷份數: 56 份	計算機	課本	筆記	字典	手機平板筆電
備註：注意事項要看!! (範圍: §14~§18)	O	×	×	×	×

注意事項: (1) 答案卷請寫上科目、系級、學號及姓名。(2) 請按題號順序書寫。(3) 每一題號需置於答案卷最左邊。(4) 中英文作答皆可。(5) 建議用深色原子筆。(6) 需要計算過程。(7) 上傳答案卷: `http://www.hmwu.idv.tw`，點選【作業考試上傳區】，login: `stat110`，password: `xxxx`。(8) Adobe Scan APP 答題卷拍照請務必清晰、需滿版平整，周遭不可有雜物入鏡。答題卷檔名: 學號-姓名-FinalExam.pdf。(9) 總分共 120 分。(10) 小數請計算至小數點以下 4 位。(11) 請複寫下列宣誓詞至答案卷上。

> 本人姓名 重視榮譽，以認真負責的態度，參與於本次線上遠距考試，恪遵各項考試規則，無任何不法或舞弊情事，如違誓言，願受校方最嚴厲之處罰，謹誓。

1. (20%) **統計名詞解釋**:

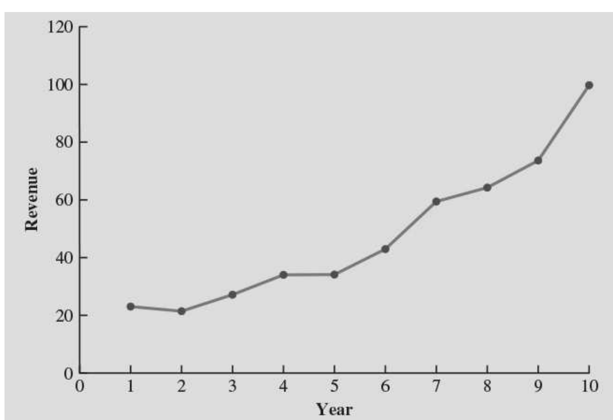
   (a) Studentized deleted residuals (學生化刪除殘差、t 化去點殘差)

   (b) Moving averages (移動平均)

   (c) Nonparametric methods (無母數方法)

   (d) Residual analysis (殘差分析)

2. (20%) 若觀察到一公司近 10 年的營運收入 (Revenue) 之時間序列圖如下:



假設有一位分析師想依此資料建立時間序列預測模型，他考量三個預測模型: (a)(5%) linear trend regression, (b)(5%) quadratic trend regression, (c)(10%) exponential trend regression。請你依此資料，幫他寫出上述三個模型的迴歸方程式。(需解釋方程式中所使用符號的意思，例如，t 為時間 (年)，$t = 1, 2, \cdots, 10$。)

考試日期：6 月 14 日 (二) 13:10-14:40	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共5頁，印刷份數: 56 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (範圍: §14~§18)	O	×	×	×	×	

3. (25%) 某家公司過去一年的月營收如下:

240 350 230 260 280 320 220 310 240 310 240 230

(a) (5%) 請徒手畫出此資料的時間序列圖。此資料是屬於何種時間序列樣態 (pattern)?

(b) (6%) 計算 3-month 的移動平均 (下表中的 `MA3`)，exponential smoothing ($\alpha = 0.2$) 預測模型 (下表中的 `ES0.2`) 以及各自的三種預測誤差指標。(計算空格的部份 (1)~(3) 即可，要有計算過程) (表中的 $|P.Error|$ 為 absolute value of percentage error.)

(c) (4%) 計算兩方法的 MSE，若依此指標，哪個方法較好?

(d) (10%) 各以兩模型預測次年一月的營收。

t	Revenue	MA3	$\lvert Error\rvert$	$(Error)^2$	$\lvert P.Error\rvert$	ES0.2	$\lvert Error\rvert$	$(Error)^2$	$\lvert P.Error\rvert$
1	240								
2	350					240.00	110.00	12100.00	31.43
3	230	273.33	43.33	1877.78	18.84	262.00	32.00	1024.00	13.91
4	260	280.00	20.00	400.00	7.69	255.60	4.40	19.36	1.69
5	280	256.67	23.33	544.44	8.33	256.48	23.52	553.19	8.40
6	320	(1)	33.33	1111.11	(2)	(3)	58.82	3459.32	18.38
7	220	273.33	53.33	2844.44	24.24	272.95	52.95	2803.41	24.07
8	310	283.33	26.67	711.11	8.60	262.36	47.64	2269.78	15.37
9	240	256.67	16.67	277.78	6.94	271.89	31.89	1016.73	13.29
10	310	286.67	23.33	544.44	7.53	265.51	44.49	1979.45	14.35
11	240	263.33	23.33	544.44	9.72	274.41	34.41	1183.85	14.34
12	230	260.00	30.00	900.00	13.04	267.53	37.53	1408.18	16.32
Total	3230	2720.00	293.33	9755.56	115.36	2889.90	477.64	27817.28	171.54

4. (20%) 小明是富二代，爸爸開了兩家球鞋製造工廠 (A、B)，小明想了解這兩家工廠生產的球鞋產量是否有差異，於是在某些固定時段，從兩家工廠各自隨機收集了以下資料 (仟雙為單位)。請你幫他選擇合適的無母數統計方法做檢定 (顯著水準 (α) 定為 0.05)。(請寫出無母數檢定方法名稱、虛無假設、擇一假設、檢定統計量值、以 normal distribution approximation 算出 p 值、決策法則及結論。所需公式及機率值見最後頁)

no	1	2	3	4	5	6	7
A	68	99	42	31	54	25	49
B	105	35	34	87	73	41	57

5. (25%) **Risk of a Stroke**. A 10-year study conducted by the American Heart Association provided data on how age, blood pressure, and smoking relate to the risk of strokes. Assume that the following data are from a portion of this study. Risk is interpreted as the probability (times 100) that the patient will have a stroke over the next 10-year period. For the smoking variable, define a dummy variable with 1 indicating a smoker and 0 indicating a nonsmoker. (R 報表在最後頁)

Risk	Age	Pressure	Smoker	Code of Smoker
12	57	152	No	0
24	67	163	No	0
13	58	155	No	0
		...		
37	59	207	Yes	1

(a) (5%) Develop an estimated regression equation that relates risk of a stroke to the person's age, blood pressure, whether the person is a smoker and and an interaction term between smoking status and blood pressure. (要說明所使用符號的意思。)

(b) (5%) Is smoking a significant factor in the risk of a stroke? Explain. Use $\alpha = 0.05$.

(c) (5%) What are the MSR and MSE for this analysis?

(d) (5%) What are the values R^2 and adjusted R^2. How to interpret these two values?

(e) (5%) What is the probability of a stroke over the next 10 years for Art Speen, a 68-year-old smoker who has blood pressure of 175? What action might the physician recommend for this patient?

6. (10%) 本課程教科書的第 15.9 節 Logistic Regression (第 777 頁) 中，有一段敘述是有關於勝算比 (Odds ratio) 與迴歸係數 (β_i) 的關係:「For each independent variable in a logistic regression equation it can be shown that Odds ratio $= e^{\beta_i}$」。請証明簡單羅吉斯迴歸 (y is binary, X is a single predictor) 中，

$$\text{Odds ratio} = e^{\beta_1}.$$

(提示: 以 odds of success 的表達法，寫出羅吉斯迴歸方程式，再依 odds ratio 的定義可導出上式)

```
> summary(Stroke.lm)

Call:
lm(formula = Risk ~ Age + Pressure + Smoker, data = Stroke)

Residuals:
     Min       1Q   Median       3Q      Max
-13.1064  -1.5715   0.4225   3.4855   8.5561

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -91.75950   15.22276  -6.028 1.76e-05 ***
Age           1.07674    0.16596   6.488 7.49e-06 ***
Pressure      0.25181    0.04523   5.568 4.24e-05 ***
SmokerYes     8.73987    3.00082   2.912   0.0102 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.757 on 16 degrees of freedom
Multiple R-squared:  0.8735,    Adjusted R-squared:  0.8498
F-statistic: 36.82 on 3 and 16 DF,  p-value: 2.064e-07

> anova(Stroke.lm)
Analysis of Variance Table

Response: Risk
          Df  Sum Sq Mean Sq F value    Pr(>F)
Age        1 1771.98 1771.98 53.4726 1.743e-06 ***
Pressure   1 1607.66 1607.66 48.5138 3.185e-06 ***
Smoker     1  281.10  281.10  8.4826   0.01017 *
Residuals 16  530.21   33.14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

考試日期：6 月 14 日 (二) 13:10-14:40	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共5頁，印刷份數: 56 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (範圍: §14∼§18)	O	×	×	×	×	

機率表

Upper tail probability.

i	p	z	t_6	t_9	t_{11}	χ^2_4	χ^2_5	$F_{2,9}$	$F_{3,9}$
1	0.200	0.8416	0.9057	0.8834	0.8755	5.9886	7.2893	1.9349	1.9007
2	0.100	1.2816	1.4398	1.3830	1.3634	7.7794	9.2364	3.0065	2.8129
3	0.050	1.6449	1.9432	1.8331	1.7959	9.4877	11.0705	4.2565	3.8625
4	0.025	1.9600	2.4469	2.2622	2.2010	11.1433	12.8325	5.7147	5.0781
5	0.010	2.3263	3.1427	2.8214	2.7181	13.2767	15.0863	8.0215	6.9919
6	0.005	2.5758	3.7074	3.2498	3.1058	14.8603	16.7496	10.1067	8.7171

R **機率分佈相關指令:** `d, q, p, r`

- `pnorm(-0.83) = 0.2033, pnorm(-0.77) = 0.2206, pnorm(-0.70) = 0.2420, pnorm(-0.64) = 0.2611, pnorm(-0.57) = 0.2843.`

公式

- Mean: $\mu = np = 0.5n$, Standard deviation: $\sigma = \sqrt{np(1-p)} = \sqrt{0.25n}$
- Mean: $\mu_{T^+} = \frac{n(n+1)}{4}$, Standard deviation: $\sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$, Distribution Form: Approximately normal for $n \geq 10$.
- Mean: $\mu_W = (1/2)n_1(n_1+n_2+1)$, Standard deviation: $\sigma_W = \sqrt{(1/12)n_1n_2(n_1+n_2+1)}$ Distribution form: Approximately normal provided $n_1 \geq 7$ and $n_2 \geq 7$.
- $h_i = \dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$
- $\dfrac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}}$, $s_{y_i-\hat{y}_i} = s\sqrt{1-h_i}$
- $D_i = \dfrac{(y_i - \hat{y}_i)^2}{(p+1)s^2}\left[\dfrac{h_i}{(1-h_i)^2}\right]$

“不管你再怎麼努力，還是有人會忽略你的付出；就為了自己而奮鬥吧。”
“Your efforts will always be neglected no matter how hard you try; so fight for yourself.”
— *緊急迫降 (Emergency Declaration, 2022)*