Contributor name: Jasleen Kaur Sondhi

Book Proposed: Introductory Statistics by Sheldon M. Ross

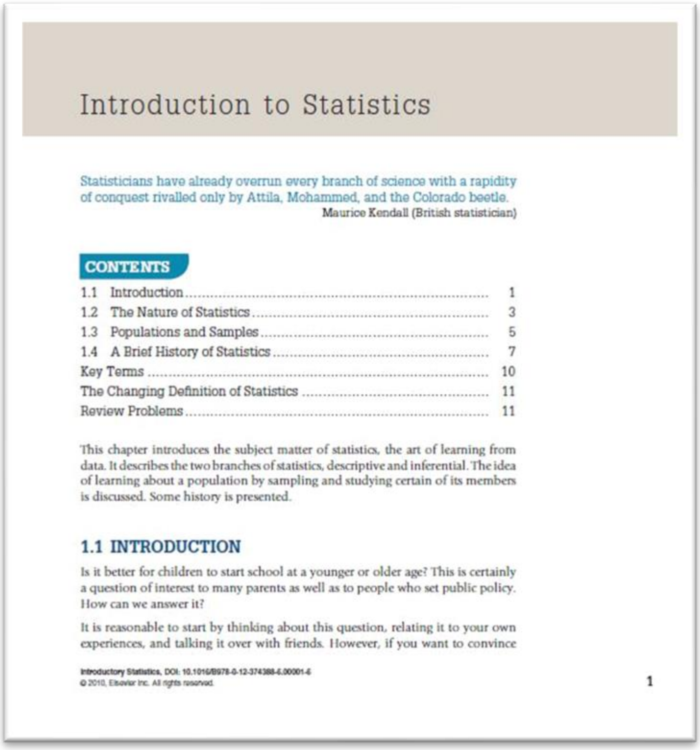
Total Chapters: 15

Total Examples: 194

Codable Examples: 183

Chapter 1: Introduction to Statistics

This Chapter is Non-Codable (Reason: The chapter contains no example problems and mostly contains definitions.)



Introduction to Statistics

Statisticians have already overrun every branch of science with a rapidity of conquest rivalled only by Attila, Mohammed, and the Colorado beetle.

Maurice Kendall (British statistician)

CONTENTS

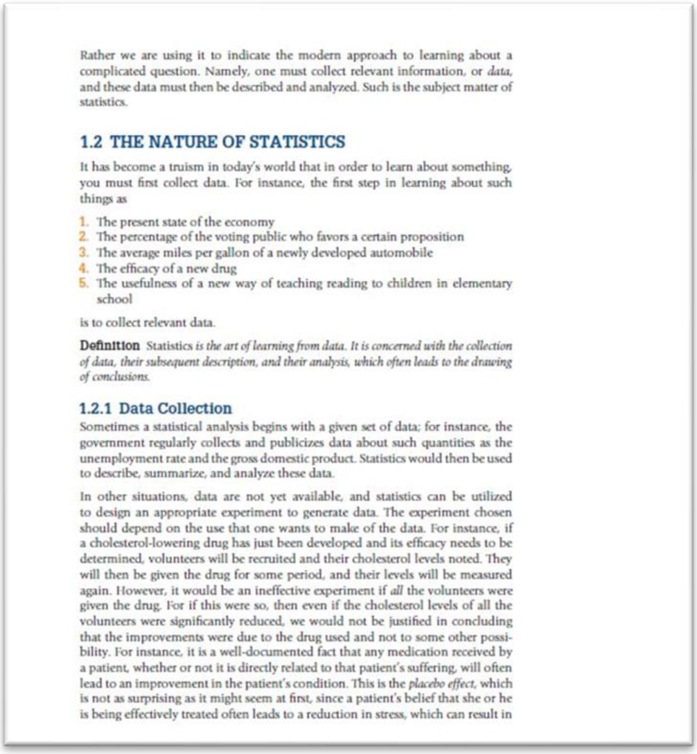
This chapter introduces the subject matter of statistics, the art of learning from data. It describes the two branches of statistics, descriptive and inferential. The idea of learning about a population by sampling and studying certain of its members is discussed. Some history is presented.

1.1 INTRODUCTION

Is it better for children to start school at a younger or older age? This is certainly a question of interest to many parents as well as to people who set public policy. How can we answer it?

It is reasonable to start by thinking about this question, relating it to your own experiences, and talking it over with friends. However, if you want to convince

Introductory Statistics. DOI: 10.1016/B978-0-12-374388-6.00001-6



Rather we are using it to indicate the modern approach to learning about a complicated question. Namely, one must collect relevant information, or *data*, and these data must then be described and analyzed. Such is the subject matter of statistics.

1.2 THE NATURE OF STATISTICS

It has become a truism in today's world that in order to learn about something, you must first collect data. For instance, the first step in learning about such things as

1. The present state of the economy
2. The percentage of the voting public who favors a certain proposition
3. The average miles per gallon of a newly developed automobile
4. The efficacy of a new drug
5. The usefulness of a new way of teaching reading to children in elementary school

is to collect relevant data.

Definition Statistics *is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions.*

1.2.1 Data Collection

Sometimes a statistical analysis begins with a given set of data; for instance, the government regularly collects and publicizes data about such quantities as the unemployment rate and the gross domestic product. Statistics would then be used to describe, summarize, and analyze these data.

In other situations, data are not yet available, and statistics can be utilized to design an appropriate experiment to generate data. The experiment chosen should depend on the use that one wants to make of the data. For instance, if a cholesterol-lowering drug has just been developed and its efficacy needs to be determined, volunteers will be recruited and their cholesterol levels noted. They will then be given the drug for some period, and their levels will be measured again. However, it would be an ineffective experiment if *all* the volunteers were given the drug. For if this were so, then even if the cholesterol levels of all the volunteers were significantly reduced, we would not be justified in concluding that the improvements were due to the drug used and not to some other possibility. For instance, it is a well-documented fact that any medication received by a patient, whether or not it is directly related to that patient's suffering, will often lead to an improvement in the patient's condition. This is the *placebo effect*, which is not as surprising as it might seem at first, since a patient's belief that she or he is being effectively treated often leads to a reduction in stress, which can result in

Chapter 2: Describing Data Sets

Example 2.1 – Codable

Example 2.2 – Codable

Example 2.3 – Codable

Example 2.4 – Codable

Example 2.5 – Codable

Example 2.6 – Codable

Example 2.7 – Codable

Chapter 3: Using Statistics to Summarize Data Sets

Example 3.1 – Codable

Example 3.2– Codable

Example 3.3 – Codable

Example 3.4 – Codable

Example 3.5 – Codable

Example 3.6 – Codable

Example 3.7 – Codable

Example 3.8 – Codable

Example 3.9 – Codable

Example 3.10 – Codable

Example 3.11 – Codable

Example 3.12 – Codable

Example 3.13 – Codable

Example 3.14– Codable

Example 3.15 – Codable

Example 3.16 – Codable

Example 3.17 – Codable

Example 3.18 – Codable

Example 3.19 – Codable

Example 3.20 – Codable

Example 3.21 – Codable

Example 3.22 – Codable

Example 3.23 – Codable

Example 3.24 – Codable

Chapter 4: Probability

Example 4.1 – Codable

Example 4.2– Codable

Example 4.3 – Codable

Example 4.4 – Codable

Example 4.5 – Codable

Example 4.6 – Codable

Example 4.7 – Codable

Example 4.8 – Codable

Example 4.9 – Codable

Example 4.10 – Codable

Example 4.11 – Codable

Example 4.12 – Codable

Example 4.13 – Codable

Example 4.14– Codable

Example 4.15 – Codable

Example 4.16 – Codable

Example 4.17 – Codable

Example 4.18 – Codable

Example 4.19 – Codable

Example 4.20 – Codable

Example 4.21 – Codable

Example 4.22 – Codable

Example 4.23 – Codable

Example 4.24 – Codable

Example 4.25 – Codable

Chapter 5: Discrete Random Variables

Example 5.1 – Codable

Example 5.2– Codable

Example 5.3 – Codable

Example 5.4 – Codable

Example 5.5 – Codable

Example 5.6 – Codable

Example 5.7 – Codable

Example 5.8 – Codable

Example 5.9 – Codable

Example 5.10 – Codable

Example 5.11 – Codable

Example 5.12 – Codable

Example 5.13 – Codable

Example 5.14– Codable

Example 5.15 – Codable

Example 5.16 – Codable

Example 5.17 – Codable

Example 5.18 – Codable

Example 5.19 – Codable

Example 5.20 – Codable

Example 5.21 – Codable

Example 5.22 – Codable

Example 5.23 – Codable

Chapter 6: Normal Random Variables

Example 6.1 – Codable

Example 6.2– Codable

Example 6.3 – Codable

Example 6.4 – Codable

Example 6.5 – Codable

Example 6.6 – Codable

Example 6.7 – Codable

Example 6.8 – Codable

Example 6.9 – Codable

Example 6.10 – Codable

Example 6.11 – Codable

Chapter 7: Distributions of Sampling Statistics

Example 7.1– Codable

Example 7.2– Codable

Example 7.3 – Codable

Example 7.4 – Codable

Example 7.5 – Codable

Example 7.6 – Codable

Example 7.7 – Codable

Chapter 8: Estimation

Example 8.1 – Codable

Example 8.2– Codable

Example 8.3 – Codable

Example 8.4 – Codable

Example 8.5 – Codable

Example 8.6 – Codable

Example 8.7 – Codable

Example 8.8 – Codable

Example 8.9 – Codable

Example 8.10 – Codable

Example 8.11 – Codable

Example 8.12 – – Non-Codable (Reason: The example problem is variable based and hence not codeable.)

Example 8.12- Find $t_{8,0.05}$.

■ **Example 8.12**

Find $t_{8,0.05}$.

Solution

The value of $t_{8,0.05}$ can be obtained from Table D.2. The following is taken from that table.

Values of $t_{n,\alpha}$

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$
6	1.440	1.943	2.447
7	1.415	1.895	2.365
→ 8	1.397	1.860	2.306
9	1.383	1.833	2.262

Reading down the $\alpha = 0.05$ column for the row $n = 8$ shows that $t_{8,0.05} = 1.860$. ■

By the symmetry of the t distribution about zero, it follows (see Fig. 8.10) that

$$P\{|T_n| \le t_{n,\alpha/2}\} = P\{-t_{n,\alpha/2} \le T_n \le t_{n,\alpha/2}\} = 1 - \alpha$$

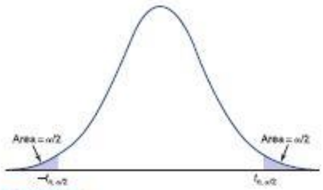


FIGURE 8.10
$P\{|T_n| \le t_{n,\alpha/2}\} = P\{-t_{n,\alpha/2} \le T_n \le t_{n,\alpha/2}\} = 1 - \alpha$.

Hence, upon using the result that $\sqrt{n}(\overline{X} - \mu)/S$ has a t distribution with $n - 1$ degrees of freedom, we see that

$$P\left\{\sqrt{n}\frac{|\overline{X} - \mu|}{S} \le t_{n-1,\alpha/2}\right\} = 1 - \alpha$$

In exactly the same manner as we did when σ was known, we can show that the preceding equation is equivalent to

$$P\left\{\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right\} = 1 - \alpha$$

Therefore, we showed the following.

A $100(1 - \alpha)$ percent confidence interval estimator for the population mean μ is given by the interval

$$\overline{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

Program 8-3 will compute the desired confidence interval estimate for a given data set.

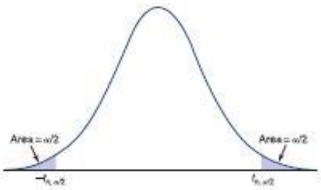


FIGURE 8.10
$P\{|T_n| \le t_{n,\alpha/2}\} = P\{-t_{n,\alpha/2} \le T_n \le t_{n,\alpha/2}\} = 1 - \alpha$.

Hence, upon using the result that $\sqrt{n}(\overline{X} - \mu)/S$ has a t distribution with $n - 1$ degrees of freedom, we see that

$$P\left\{\sqrt{n}\frac{|\overline{X} - \mu|}{S} \le t_{n-1,\alpha/2}\right\} = 1 - \alpha$$

In exactly the same manner as we did when σ was known, we can show that the preceding equation is equivalent to

Example 8.13 – Codable

Example 8.14– Codable

Example 8.15 – Codable

Example 8.16 – Codable

Example 8.17 – Codable

Example 8.18 – Codable

Chapter 9: Testing Statistical Hypotheses

Example 9.1 – Codable

Example 9.2– Codable

Example 9.3 – Codable

Example 9.4 – Codable

Example 9.5 – Codable

Example 9.6 – Codable

Example 9.7 – Codable

Example 9.8 – Codable

Example 9.9 – Codable

Example 9.10 – Codable

Example 9.11 – Codable

Chapter 10: Hypothesis Tests Concerning Two Populations

Example 10.1 – Codable

Example 10.2– Codable

Example 10.3 – Codable

Example 10.4 – Codable

Example 10.5 – Codable

Example 10.6 – Codable

Example 10.7 – Non-Codable (Reason: The example problem is variable based and the final answer is also in the form of variables.)

Example 10.7- Suppose we are interested in learning about the effect of a newly developed gasoline detergent additive on automobile mileage. To gather information, seven cars have been assembled, and their gasoline mileages (in units of miles per gallon) have been determined. For each car this determination is made both when gasoline without the additive is used and when gasoline with the additive is used. The data can be represented as follows:

■ Example 10.7

Suppose we are interested in learning about the effect of a newly developed gasoline detergent additive on automobile mileage. To gather information, seven cars have been assembled, and their gasoline mileages (in units of miles per gallon) have been determined. For each car this determination is made both when gasoline without the additive is used and when gasoline with the additive is used. The data can be represented as follows:

Car	Mileage without additive	Mileage with additive
1	24.2	23.5
2	30.4	29.6
3	32.7	32.3
4	19.8	17.6
5	25.0	25.3
6	24.9	25.4
7	22.2	20.6

For instance, car 1 got 24.2 miles per gallon by using gasoline without the additive and only 23.5 miles per gallon by using gasoline with the additive, whereas car 4 obtained 19.8 miles per gallon by using gasoline without the additive and 17.6 miles per gallon by using gasoline with the additive.

Now, it is easy to see that two factors will determine a car's mileage per gallon. One factor is whether the gasoline includes the additive, and the second factor is the car itself. For this reason we should not treat the two samples as being independent; rather, we should consider paired data. ■

Suppose we want to test

$$H_0: \mu_x = \mu_y \quad \text{against} \quad H_1: \mu_x \neq \mu_y$$

where the two samples consist of the paired data $X_i, Y_i, = 1, \ldots, n$. We can test this null hypothesis that the population means are equal by looking at the differences between the data values in a pairing. That is, let

$$D_i = X_i - Y_i \quad i = 1, \ldots, n$$

Now,

$$E[D_i] = E[X_i] - E[Y_i]$$

or, with $\mu_d = E[D_i]$,

$$\mu_d = \mu_x - \mu_y$$

The hypothesis that $\mu_x = \mu_y$ is therefore equivalent to the hypothesis that $\mu_d = 0$. Thus we can test the hypothesis that the population means are equal by testing

$$H_0: \mu_d = 0 \quad \text{against} \quad H_1: \mu_d \neq 0$$

Assuming that the random variables $D_1, \ldots, D_n$ constitute a sample from a normal population, we can test this null hypothesis by using the t test described in Sec. 9.4. That is, if we let $\overline{D}$ and S_d denote, respectively, the sample mean and sample standard deviation of the data $D_1, \ldots, D_n$, then the test statistic TS is given by

$$\text{TS} = \sqrt{n}\frac{\overline{D}}{S_d}$$

The significance-level-α test will be to

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } |\text{TS}| \geq t_{n-1,\alpha/2} \\ \text{Not reject } H_0 & \text{otherwise} \end{array}$$

where the value of $t_{n-1,\alpha/2}$ can be obtained from Table D.2.

Equivalently, the test can be performed by computing the value of the test statistic TS, say it is equal to v, and then computing the resulting p value, given by

$$p \text{ value} = P\{|T_{n-1}| \geq |v|\} = 2P\{T_{n-1} \geq |v|\}$$

where T_{n-1} is a t random variable with $n-1$ degrees of freedom. If a personal computer is available, then Program 9-1 can be used to determine the value of the test statistic and the resulting p value. The successive data values entered in this program should be $D_1, D_2, \ldots, D_n$ and the value of μ_0 (the null hypothesis value for the mean of D) entered should be 0.

Example 10.8 – Codable

Example 10.9 – Codable

Example 10.10 – Codable

Example 10.11 – Codable

Example 10.12 – Non-Codable (Reason: The example problem is variable based and the final answer is also in the form of variables.)

Example 10.12- In 1970, the researchers Herbst,Ulfelder, and Poskanzer (H-U-P) suspected that vaginal cancer in young women, a rather rare disease, might be caused by one's mother having taken the drug diethylstilbestrol (usually referred to as DES) while pregnant…….

■ **Example 10.12**

In 1970, the researchers Herbst, Ulfelder, and Poskanzer (H-U-P) suspected that vaginal cancer in young women, a rather rare disease, might be caused by one's mother having taken the drug diethylstilbestrol (usually referred to as DES) while pregnant. To study this possibility, the researchers could have performed an observational study by searching for a (treatment) group of women whose mothers took DES when pregnant and a (control) group of women whose mothers did not. They could then observe these groups for a period of time and use the resulting data to test the hypothesis that the probabilities of contracting vaginal cancer are the same for both groups. However, because vaginal cancer is so rare (in both groups), such a study would require a large number of individuals in both groups and would probably have to continue for many years to obtain significant results. Consequently, H-U-P decided on a different type of observational study. They uncovered 8 women between the ages of 15 and 22 who had vaginal cancer. Each of these women (called cases) was then matched with 4 others, called *referents* or *controls*. Each of the referents of a case was free of the cancer and was born within 5 days in the same hospital and in the same type of room (either private or public) as the case. Arguing that if DES had no effect on vaginal cancer then the probability, call it p_c, that the mother of a case took DES would be the same as the probability, call it p_r, that the mother of a referent took DES, the researchers H-U-P decided to test

$$H_0: p_c = p_r \quad \text{against} \quad H_1: p_c \neq p_r$$

Discovering that 7 of the 8 cases had mothers who took DES while pregnant whereas none of the 32 referents had mothers who took the drug, the researchers concluded that there was a strong association between DES and vaginal cancer (see Herbst, A., Ulfelder, H., and Poskanzer, D., "Adenocarcinoma of the Vagina: Association of Maternal Stilbestrol Therapy with Tumor Appearance in Young Women," *New England Journal of Medicine*, **284**, 878–881, 1971). (The p value for these data is approximately 0.) ■

If we are interested in verifying the one-sided hypothesis that p_1 is larger than p_2, then we should take that to be the alternative hypothesis and so test

$$H_0: p_1 \le p_2 \quad \text{against} \quad H_1: p_1 > p_2$$

The same test statistic TS as used before is still employed, but now we reject H_0 only when TS is large (since this occurs when $\hat{p}_1 - \hat{p}_2$ is large). Thus, the one-sided significance-level-α test is to

Reject H_0	if $TS \ge z_\alpha$
Not reject H_0	otherwise

Alternatively, if the value of the test statistic TS is v, then the resulting p value is

$$p \text{ value} = P\{Z \ge v\}$$

where Z is a standard normal.

Remark *The test of*

$$H_0: p_1 \le p_2 \quad \textit{against} \quad H_1: p_1 > p_2$$

is the same as

$$H_0: p_1 = p_2 \quad \textit{against} \quad H_1: p_1 > p_2$$

This is so because in both cases we want to reject H_0 when $\hat{p}_1 - \hat{p}_2$ is so large that such a large value would have been highly unlikely if p_1 were not greater than p_2.

DES and vaginal cancer (see Herbst, A., Ulfelder, H., and Poskanzer, D., "Adenocarcinoma of the Vagina: Association of Maternal Stilbestrol Therapy with Tumor Appearance in Young Women," *New England Journal of Medicine*, **284**, 878–881, 1971). (The p value for these data is approximately 0.) ■

Example 10.13 – Codable

Chapter 11: Analysis of Variance

Example 11.1 – Codable

Example 11.2– Non-Codable (Reason: The example is the derivation of the value of Test Statistic and the result is in the form of variables/formula.)

Example 11.2- Let us do the computations of Example 11.1 by using Program 11-1. After the

data have been entered, we get the following output.

■ Example 11.2

Let us do the computations of Example 11.1 by using Program 11-1. After the data have been entered, we get the following output.

The denominator estimate is 165.967

The numerator estimate is 431.667

The value of the f-statistic is 2.6009

The *p*-value is 0.11525 ■

Table 11.2 summarizes the results of this section.

Remark *When $m = 2$, the preceding is a test of the null hypothesis that two independent samples, having a common population variance, have the same mean. The reader might*

Table 11.2 One-Factor ANOVA Table

Variables $\overline{X}_i$ and S_i^2, $i = 1, \ldots, m$, are the sample means and sample variances, respectively, of independent samples of size n from normal populations having means μ_i and a common variance σ^2.

Source of estimator	Estimator of σ^2	Value of test statistic
Between samples	$n\overline{S}^2 = \frac{n\sum_{i=1}^{m}(\overline{X}_i - \overline{X})^2}{m-1}$	$\text{TS} = \frac{n\overline{S}^2}{\sum_{i=1}^{m}\frac{S_i^2}{m}}$
Within samples	$\sum_{i=1}^{m}\frac{S_i^2}{m}$	

Significance-level-α test of H_0: all μ_i values are equal:

Reject H_0 if $\text{TS} \geq F_{m-1,m(n-1),\alpha}$

Table 11.2 One-Factor ANOVA Table

Variables $\overline{X}_i$ and S_i^2, $i = 1, \ldots, m$, are the sample means and sample variances, respectively, of independent samples of size n from normal populations having means μ_i and a common variance σ^2.

Source of estimator	Estimator of σ^2	Value of test statistic
Between samples	$n\overline{S}^2 = \frac{n\sum_{i=1}^{m}(\overline{X}_i - \overline{X})^2}{m-1}$	$\text{TS} = \frac{n\overline{S}^2}{\sum_{i=1}^{m}\frac{S_i^2}{m}}$
Within samples	$\sum_{i=1}^{m}\frac{S_i^2}{m}$	

Significance-level-α test of H_0: all μ_i values are equal:

Reject H_0	if $\text{TS} \geq F_{m-1,m(n-1),\alpha}$
Do not reject H_0	otherwise

If $\text{TS} = v$, then p value $= P\{F_{m-1,m(n-1)} \geq v\}$

where $F_{m-1,m(n-1)}$ is an F random variable with $m-1$ numerator and $m(n-1)$ denominator degrees of freedom.

wonder how this compares with the one presented in Chap. 10. It turns out that the tests are exactly the same. That is, assuming the same data are used, they always give rise to exactly the same p value.

Example 11.3 – Non-Codable (Reason: The example problem explains the concept of Parameter Estimation using variables and the final result is also in the form of variables.)

Example 11.3- Four different standardized reading achievement tests were administered to each of five students. Their scores were as follows:

■ Example 11.3

Four different standardized reading achievement tests were administered to each of five students. Their scores were as follows:

	Student				
Examination	**1**	**2**	**3**	**4**	**5**
1	75	73	60	70	86
2	78	71	64	72	90
3	80	69	62	70	85
4	73	67	63	80	92

Each value in this set of 20 data points is affected by two factors: the examination and the student whose score on that examination is being recorded. The examination factor has four possible values, or *levels*, and the student factor has five possible levels. ■

In general, let us suppose that there are m possible levels of the first factor and n possible levels of the second. Let X_{ij} denote the value of the data obtained when the first factor is at level i and the second factor is at level j. We often portray the data set in the following array of rows and columns:

$$\begin{array}{ccccccc} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{in} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{m1} & X_{m2} & \cdots & X_{mj} & \cdots & X_{mn} \end{array}$$

Because of this we refer to the first factor as the *row* factor and the second factor as the *column* factor. Also, the data value X_{ij} is the value in row i and column j.

As in Sec. 11.2, we suppose that all the data values X_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$, are independent normal random variables with common variance σ^2. However, whereas in Sec. 11.2 we supposed that only a single factor affected the mean value of a data point—namely, the sample to which it belonged—in this section we will suppose that the mean value of the data point depends on both its row and its column. However, before specifying this model, we first recall the model of Sec. 11.2. If we let X_{ij} represent the value of the jth member of sample i, then this model supposes that

$$E[X_{ij}] = \mu_i$$

If we now let μ denote the average value of the μ_i, that is,

$$\mu = \frac{\sum_{i=1}^{m}\mu_i}{m}$$

then we can write the preceding as

$$E[X_{ij}] = \mu + \alpha_i$$

where $\alpha_i = \mu_i - \mu$. With this definition of α_i equal to the deviation of μ_i from the average of the means μ, it is easy to see that

$$\sum_{i=1}^{m}\alpha_i = 0$$

In the case of two factors, we write our model in terms of row and column deviations. Specifically, we suppose that the expected value of variable X_{ij} can be expressed as follows:

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

The value μ is referred to as the *grand mean*, α_i is the *deviation from the grand mean due to row* i, and β_j is the *deviation from the grand mean due to column* j. In addition, these quantities satisfy the following equalities:

$$\sum_{i=1}^{m} \alpha_i = \sum_{j=1}^{n} \beta_j = 0$$

Let us start by determining estimators for parameters μ, α_i, and β_j, $i = 1, \ldots, m$, $j = 1, \ldots, n$. To do so, we will find it convenient to introduce the following "dot" notation. Let

$$X_{i.} = \frac{\sum_{j=1}^{n} X_{ij}}{n} = \text{average of all values in row } i$$

$$X_{.j} = \frac{\sum_{i=1}^{m} X_{ij}}{m} = \text{average of all values in column } j$$

$$X_{..} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}}{nm} = \text{average of all } nm \text{ data values}$$

It is not difficult to show that

$$E[X_{i.}] = \mu + \alpha_i$$
$$E[X_{.j}] = \mu + \beta_j$$
$$E[X_{..}] = \mu$$

Since the preceding is equivalent to

$$E[X_{..}] = \mu$$
$$E[X_{i.} - X_{..}] = \alpha_i$$
$$E[X_{.j} - X_{..}] = \beta_j$$

we see that unbiased estimators of μ, α_i and β_j—call them $\hat{\mu}, \hat{\alpha}_i$, and $\hat{\beta}_j$—are given by

$$\hat{\mu} = X_{..}$$
$$\hat{\alpha}_i = X_{i.} - X_{..}$$
$$\hat{\beta}_j = X_{.j} - X_{..}$$

Example 11.4 – Codable

Example 11.5 – Codable

Chapter 12: Linear Regression

Example 12.1 – Codable

Example 12.2– Codable

Example 12.3 – Codable

Example 12.4 – Codable

Example 12.5 – Codable

Example 12.6 – Codable

Example 12.7 – Codable

Example 12.8 – Codable

Example 12.9 – Codable

Example 12.10 – Codable

Example 12.11 – – Non-Codable (Reason: The problem is definition based and uses the given values to illustrate the theoretical concept of multiple linear regression.)

Example 12.11-In laboratory experiments two factors that often affect the percentage yield of the experiment are the temperature and the pressure at which the experiment is conducted. The following data detail the results of four independent experiments. For each experiment, we have the temperature (in degrees Fahrenheit) at which the experiment is run, the pressure (in pounds per square inch), and the percentage yield.

■ **Example 12.11**

In laboratory experiments two factors that often affect the percentage yield of the experiment are the temperature and the pressure at which the experiment is conducted. The following data detail the results of four independent experiments. For each experiment, we have the temperature (in degrees Fahrenheit) at which the experiment is run, the pressure (in pounds per square inch), and the percentage yield.

Experiment	Temperature	Pressure	Percentage yield
1	140	210	68
2	150	220	82
3	160	210	74
4	130	230	80

■

Suppose that we are interested in predicting the response value Y on the basis of the values of the k input variables $x_1, x_2, \ldots, x_k$.

Definition *The multiple linear regression model supposes that the response* Y *is related to the input values* x_i, $i = 1, \ldots, k$, *through the relationship*

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

Definition *The multiple linear regression model supposes that the response* Y *is related to the input values* x_i, $i = 1, \ldots, k$, *through the relationship*

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

In this expression, $\beta_0, \beta_1, \ldots, \beta_k$ are *regression parameters* and e is an *error* random variable that has mean 0. The regression parameters will not be initially known and must be estimated from a set of data.

Suppose that we have at our disposal a set of n responses corresponding to n different sets of the k input values. Let y_i denote the ith response, and let the k input values corresponding to this response be $x_{i1}, x_{i2}, \ldots, x_{ik}$, $i = 1, \ldots, n$. Thus, for instance, y_1 was the response when the k input values were $x_{11}, x_{12}, \ldots, x_{1k}$. The data set is presented in Fig. 12.10.

Example 12.12 –Non-Codable (Reason: The problem uses the given values to illustrate the theoretical concept to estimate regression parameters.)

Example 12.12-In Example 12.11 there are two input variables, the temperature and the pressure, and so k = 2. There are four experimental results, and so n = 4. The value

■ **Example 12.12**

In Example 12.11 there are two input variables, the temperature and the pressure, and so $k = 2$. There are four experimental results, and so $n = 4$. The value

Set	Input 1	Input 2	...	Input k	Response
1	x_{11}	x_{12}	...	x_{1k}	y_1
2	x_{21}	x_{22}	...	x_{2k}	y_2
3	x_{31}	x_{32}	...	x_{3k}	y_3
⋮					
n	x_{n1}	x_{n2}	...	x_{nk}	y_n

FIGURE 12.10
Data on n experiments.

x_{i1} refers to the temperature and x_{i2} to the pressure of experiment i. The value y_i is the percentage yield (response) of experiment i. Thus, for instance,

$$x_{31} = 160 \quad x_{32} = 210 \quad y_3 = 74$$ ■

To estimate the regression parameters again, as in the case of simple linear regression, we use the method of least squares. That is, we start by noting that if $B_0, B_1, \ldots, B_k$ are estimators of the regression parameters $\beta_0, \beta_1, \ldots, \beta_k$, then the estimate of the response when the input values are $x_{i1}, x_{i2}, \ldots, x_{ik}$ is given by

$$\text{Estimated response} = B_0 + B_1 x_{i1} + B_2 x_{i2} + \cdots + B_k x_{ik}$$

Since the actual response was y_i, we see that the difference between the actual response and what would have been predicted if we had used the estimators $B_0, B_1, \ldots, B_k$ is

$$\epsilon_i = y_i - (B_0 + B_1 x_{i1} + B_2 x_{i2} + \cdots + B_k x_{ik})$$

Thus, ϵ_i can be regarded as the *error* that would have resulted if we had used the estimators B_i, $i = 0, \ldots, k$. The estimators that make the sum of the squares of the errors as small as possible are called the *least-squares estimators*.

The least-squares estimators of the regression parameters are the choices of B_i that make

$$\sum_{i=1}^{n} \epsilon_i^2$$

as small as possible.

The actual computations needed to obtain the least-squares estimators are algebraically messy and will not be presented here. Instead we refer to Program 12-2 to do the computations for us. The outputs of this program are the estimates of the regression parameters. In addition, the program provides predicted response values for specified sets of input values. That is, if the user enters the values $x_1, x_2, \ldots, x_k$, then the computer will print out the value of $B(0) + B(1)x_1 + \cdots + B(k)x_k$, where $B(0), B(1), \ldots, B(k)$ are the least-squares estimators of the regression parameters.

Example 12.13 – Codable

Chapter 13: Chi-Squared Goodness-of-Fit Tests

Example 13.1 – Non-Codable (Reason: The example problem verifies the null hypothesis theoretically and the final answer is in the form of a variable.)

■ Example 13.1

It is known that 41 percent of the U.S. population has type A blood, 9 percent has type B, 4 percent has type AB, and 46 percent has type O. Suppose that we suspect that the blood type distribution of people suffering from stomach cancer is different from that of the overall population.

To verify that the blood type distribution is different for those suffering from stomach cancer, we could test the null hypothesis

$$H_0: P_1 = 0.41,\ P_2 = 0.09,\ P_3 = 0.04,\ P_4 = 0.46$$

where P_1 is the proportion of all those with stomach cancer who have type A blood, P_2 is the proportion of those who have type B blood, P_3 is the proportion who have type AB blood, and P_4 is the proportion who have type O blood. A rejection of H_0 would then enable us to conclude that the blood type distribution is indeed different for those suffering from stomach cancer.

In the preceding scenario, each member of the population of individuals who are suffering from stomach cancer is given one of four possible values according to his or her blood type. We are interested in testing the hypothesis that $P_1 = 0.41, P_2 = 0.09, P_3 = 0.04, P_4 = 0.46$ represent the proportions of this population having each of the different values. ■

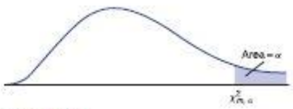


FIGURE 13.3
Chi-squared percentile $P(\chi^2_m \geq \chi^2_{m,\alpha}) = \alpha$.

Table 13.1 Some Values of $\chi^2_{m,\alpha}$

m	$\alpha = 0.99$	$\alpha = 0.95$	$\alpha = 0.05$	$\alpha = 0.01$
1	0.000157	0.00393	3.841	6.635
2	0.0201	0.103	5.991	9.210
3	0.115	0.352	7.815	11.345
4	0.297	0.711	9.488	13.277
5	0.554	1.145	11.070	15.086
6	0.872	1.635	12.592	16.812
7	1.239	2.167	14.067	18.475

Values of $\chi^2_{m,\alpha}$ for various values of m and α are given in App. Table D.3. A portion of this table is represented in Table 13.1.

To test the null hypothesis that $P_i = p_i, i = 1, \ldots, k$, first we need to draw a random sample of elements from the population. Suppose this sample is of size n. Let N_i denote the number of elements of the sample that have value i, for $i = 1, \ldots, k$. Now, if the null hypothesis is true, then each element of the sample will have value i with probability p_i. Also, since the population is assumed to be very large, it follows that the successive values of the members of the sample will be independent. Thus, if the null hypothesis is true, then N_i will have the same distribution as the number of successes in n independent trials, when each trial is a success with probability p_i. That is, if H_0 is true, then N_i will be a binomial random variable with parameters n and p_i. Since the expected value of a binomial is the product of its parameters, we see that when H_0 is true,

$$E[N_i] = np_i \qquad i = 1, \ldots, k$$

For each i, let e_i denote this expected number of outcomes that equal i when H_0 is true. That is,

$$e_i = np_i$$

Thus, when H_0 is true, we expect that N_i would be relatively close to e_i. That is, when the null hypothesis is true, the quantity $(N_i - e_i)^2$ should not be too large, say, in relation to e_i. Since this is true for each value of i, a reasonable way of testing H_0 would be to compute the value of the test statistic

$$TS = \sum_{i=1}^{k} \frac{(N_i - e_i)^2}{e_i}$$

and then reject H_0 when TS is sufficiently large.

To determine how large TS need be to justify rejection of the null hypothesis, we use a result that was proved by Karl Pearson in 1900. This result states that for large values of n, TS will have an approximately chi-squared distribution with $k-1$ degrees of freedom. Let $\chi^2_{k-1,\alpha}$ denote the $100(1-\alpha)$th percentile of this distribution; that is, a chi-squared random variable having $k-1$ degrees of freedom will exceed this value with probability α (Fig. 13.3). Then the approximate significance-level-α test of the null hypothesis H_0 against the alternative H_1 is as follows:

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } TS \geq \chi^2_{k-1,\alpha} \\ \text{Do not reject } H_0 & \text{otherwise} \end{array}$$

The preceding is called the *chi-squared goodness-of-fit test*. For reasonably large values of n, it results in a hypothesis test of H_0 whose significance level is approximately equal to α. An accepted rule of thumb is that this approximation will be quite good provided n is large enough so that $e_i \geq 1$ for each i and at least 80 percent of the values e_i exceed 5.

Example 13.2– Codable

Example 13.3 – Codable

Example 13.4 – Codable

Example 13.5 – Non-Codable (Reason: The problem depicts testing for independents in population and the final answer is in the form of variables.)

Example 13.5- Consider a population of voting-age adults, and suppose that each adult is classified according to both gender—female or male—and political affiliation—
Democrat, Republican, or Independent.

■ **Example 13.5**

Consider a population of voting-age adults, and suppose that each adult is classified according to both gender—female or male—and political affiliation—Democrat, Republican, or Independent. Let the X characteristic represent gender and the Y characteristic represent political affiliation. Since there are two possible genders and three possible political affiliations, $r = 2$ and $s = 3$. Let us say that a person's X characteristic is 1 if the person is a woman and 2 if the person is a man. Also, say that a person's Y characteristic is 1 if the person is a

Democrat, 2 if the person is a Republican, and 3 if he or she is an Independent. Thus, for instance, a woman who is a Republican would have X characteristic 1 and Y characteristic 2. ■

Let P_{ij} denote the proportion of the population that has both X characterization i and Y characterization j, for i being any of the values $1, 2, \ldots, r$ and j being any of the values $1, 2, \ldots, s$. Also, let P_i denote the proportion of the population who have X characteristic i, and let Q_j be the proportion who have Y characteristic j. Thus if X and Y denote the values of the X characteristic and Y characteristic of a randomly chosen member of the population, then

$$P\{X = i, Y = j\} = P_{ij}$$
$$P\{X = i\} = P_i$$

Example 13.6 – Non-Codable (Reason: Non-Codable (Reason: The problem depicts testing for independents in population and the final answer is in the form of variables.)

Example 13.6- For the situation described in Example 13.5, P_{11} represents the proportion of the population consisting of women who classify themselves as Democrats, P_{12} is the proportion of the population consisting of women who classify themselves as Republicans, and P_{13} is the proportion of the population consisting of women who classify themselves as Independents

■ Example 13.6

For the situation described in Example 13.5, P_{11} represents the proportion of the population consisting of women who classify themselves as Democrats, P_{12} is the proportion of the population consisting of women who classify themselves as Republicans, and P_{13} is the proportion of the population consisting of women who classify themselves as Independents. The proportions P_{21}, P_{22}, and P_{23} are defined similarly, with *men* replacing *women* in the definitions. The quantities P_1 and P_2 are the proportions of the population that are, respectively, women and men; Q_1, Q_2, and Q_3 are the proportions of the population that are, respectively, Democrats, Republicans, and Independents. ■

We will be interested in developing a test of the hypothesis that the X characteristic and Y characteristic of a randomly chosen member of the population are independent. Recalling that X and Y are independent if

$$P\{X = i, Y = j\} = P\{X = i\}P\{Y = j\}$$

it follows that we want to test the null hypothesis

$$H_0: P_{ij} = P_i Q_j \quad \text{for all } i = 1, \ldots, r,\ j = 1, \ldots, s$$

against the alternative

$$H_1: P_{ij} \neq P_i Q_j \quad \text{for some values of } i \text{ and } j$$

To test this hypothesis of independence, we start by choosing a random sample of size n of members of the population. Let N_{ij} denote the number of elements of the sample that have both X characteristic i and Y characteristic j.

Example 13.7 – Non-Codable (Reason: Non-Codable (Reason: Non-Codable (Reason: The problem depicts testing for independents in population and the final answer is in the form of variables.)

Example 13.7- Consider Example 13.5, and suppose that a random sample of 300 people were chosen from the population, with the following data resulting:

■ Example 13.7

Consider Example 13.5, and suppose that a random sample of 300 people were chosen from the population, with the following data resulting:

	j			
i	Democrat	Republican	Independent	Total
Women	68	56	32	156
Men	52	72	20	144
Total	120	128	52	300

Thus, for instance, the random sample of size 300 contained 68 women who classified themselves as Democrats, 56 women who classified themselves as Republicans, and 32 women who classified themselves as Independents; that is, $N_{11} = 68$, $N_{12} = 56$, and $N_{13} = 32$. Similarly, $N_{21} = 52$, $N_{22} = 72$, and $N_{23} = 20$.

This table, which specifies the number of members of the sample that fall in each of the rs cells, is called a *contingency table*. ■

If the hypothesis is true that the X and Y characteristics of a randomly chosen member of the population are independent, then each element of the sample will have X characteristic i and Y characteristic j with probability P_iQ_j. Hence, if these probabilities were known then, from the results of Sec. 13.2, we could test H_0 by using the test statistic

$$TS = \sum_i \sum_j \frac{(N_{ij} - e_{ij})^2}{e_{ij}}$$

where

$$e_{ij} = nP_iQ_j$$

The quantity e_{ij} represents the expected number, when H_0 is true, of elements in the sample that have both X characteristic i and Y characteristic j. In computing TS we must calculate the sum of the terms for all rs possible values of the pair i, j. When H_0 is true, TS will have an approximately chi-squared distribution with $rs - 1$ degrees of freedom.

The trouble with using this approach directly is that the $r + s$ quantities P_i and Q_j, $i = 1, \ldots, r$, $j = 1, \ldots, s$, are not specified by the null hypothesis. Thus, we need first to estimate them. To do so, let N_i and M_j denote the number of elements of the sample that have, respectively, X characteristic i and Y characteristic j. Because N_i/n and M_j/n are the proportions of the sample having, respectively, X characteristic i and Y characteristic j, it is natural to use them as estimators of P_i and Q_j.

That is, we estimate P_i and Q_j by

$$\hat{P}_i = \frac{N_i}{n} \quad \hat{Q}_j = \frac{M_j}{n}$$

This leads to the following estimate of e_{ij}:

$$\hat{e}_{ij} = n\hat{P}_i\hat{Q}_j = \frac{N_iM_j}{n}$$

In words, $\hat{e}_{ij}$ is equal to the product of the number of members of the sample that have X characteristic i (that is, the sum of row i of the contingency table) and the number of members of the sample that have Y characteristic j (that is, the sum of column j of the contingency table) divided by the sample size n.

Thus, it seems that a reasonable test statistic to use in testing the independence of the X characteristic and the Y characteristic is the following:

$$TS = \sum_i \sum_j \frac{(N_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

where $\hat{e}_{ij}$, $i = 1, \ldots, r$, $j = 1, \ldots, s$, are as just given.

To specify the set of values of TS that will result in rejection of the null hypothesis, we need to know the distribution of TS when the null hypothesis is true. It can be shown that when H_0 is true, the distribution of the test statistic TS is approximately a chi-squared distribution with $(r-1)(s-1)$ degrees of freedom. From this, it follows that the significance-level-α test of H_0 is as follows:

$$\begin{array}{ll} \text{Reject } H_0 & \text{if TS} \geq \chi^2_{(r-1)(s-1),\alpha} \\ \text{Do not reject } H_0 & \text{otherwise} \end{array}$$

A technical remark: It is not difficult to see why the test statistic TS should have $(r-1)(s-1)$ degrees of freedom. Recall from Sec. 13.2 that if all the values P_i and Q_j are specified in advance, then the test statistic has $rs - 1$ degrees of freedom. (This is so since k, the number of different types of elements in the population, is equal to rs.) Now, at first glance it may seem that we have to use the data to estimate $r + s$ parameters. However, since the P_i's and the Q_j's both sum to 1—that is, $\Sigma_i P_i = \Sigma_j Q_j = 1$—we really only need to estimate $r - 1$ of the P_i's and $s - 1$ of the Q_j's. (For instance, if r is equal to 2, then an estimate of P_1 will automatically provide an estimate of P_2 since $P_2 = 1 - P_1$.) Hence, we actually need to estimate $r - 1 + s - 1 = r + s - 2$ parameters. Since a degree of freedom is lost for each parameter estimated, it follows that the resulting test statistic has $rs - 1 - (r + s - 2) = rs - r - s + 1 = (r-1)(s-1)$ degrees of freedom.

Example 13.8 – Codable

Example 13.9 – Codable

Example 13.10 – Codable

Example 13.11 – Codable

Chapter 14: Nonparametric Hypotheses Tests

Example 14.1 – Codable

Example 14.2– Codable

Example 14.3 – Codable

Example 14.4 – Codable

Example 14.5 – Codable

Example 14.6 – Codable

Example 14.7 – Codable

Example 14.8 – Codable

Example 14.9 – Codable

Example 14.10 – Codable

Example 14.11 – Codable

Example 14.12 –Codable

Example 14.13 – Codable

Example 14.14– Codable

Example 14.15 – Codable

Example 14.16 –Codable

Example 14.17 – Codable

Example 14.18 – Codable

Example 14.19 – Codable

Chapter 15: Quality Control

Example 15.1 – Codable

Example 15.2– Codable

Example 15.3 – Codable

Example 15.4 – Codable

Example 15.5 – Codable

Example 15.6 – Codable

Example 15.7 – Codable