Chapter 11

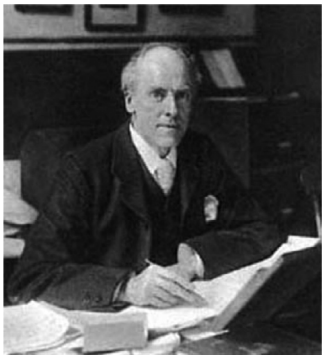
# Categorical data analysis and goodness-of-fit tests and applications

## Chapter outline



## Objective

In this chapter, we will study various methods of categorical data analysis, including goodness-of-fit tests, to determine if a given set of data follows a particular probability distribution.



Karl Pearson
*(Source: http://www-history.mcs.st-and.ac.uk/~history/PictDisplay/Pearson.html)*

Karl Pearson (1857–1936) is considered the founder of the 20th-century science of statistics. Pearson contributed in several different fields such as anthropology, biometry, genetics, scientific methods, and statistical theory. He applied


**Mathematical Statistics with Applications in R.** https://doi.org/10.1016/B978-0-12-817815-7.00011-7

statistics to biological problems of heredity and evolution. In 1911 he founded the world's first university statistics department at the University College London.

He is the author of *The Grammar of Science*, the three volumes of *The Life, Letters and Labors of Francis Galton*, and *The Ethic of Free Thought*. Pearson was the founder of the statistical journal *Biometrika*. In 1900, he published a paper on the chi-square goodness-of-fit test that we will study in this chapter. This is one of Pearson's most significant contributions to statistics. In 1893, Pearson coined the term "standard deviation."

## 11.1 Introduction

Techniques presented in the previous chapters are mostly designed for quantitative or numerical data that included both discrete and continuous data. In general, there are two types of data, namely quantitative and categorical. This chapter provides some introductory ideas on categorical data analysis. Categorical (or qualitative) data are the outcome of an experiment or a process that can be categorized into a finite number of mutually exclusive groups or categories. The categorical variables are measured on a scale that is nominal or ordinal. These data are represented through contingency tables. Examples of categorical variables are the political philosophy of a person such as liberal, conservative, or moderate; sex of an individual; make and model of a new auto; education level of an individual; or customer satisfaction surveys with categories such as poor, fair, good, great, excellent; etc. Categories may either be unordered (*nominal*) or ordered (*ordinal*). Telephone numbers, zip codes, blood types, occupation, gender, race/ethnicity, etc. have no particular order. Age group, degree of agreement with a statement on a questionnaire (strongly agree, agree, neutral, disagree, strongly disagree, etc.), grade in an exam (such as A, B, C, etc.), or patient condition (poor, fair, good, excellent) have a natural ordering of categories. Binary variables such as success and failure for nominal or ordinal distinction are unimportant. Categorical variables can be analyzed with a chi-square goodness-of-fit test. Counts and percentages are the basic statistics available for categorical variables. Hence, goodness-of-fit tests consist of determining whether the frequency counts in the categories of the variable agree with a specific distribution. For the regression analysis with contingency table, we will use the logistic regression.

Categorical data can be summarized using a frequency table. We can use a bar graph, Pareto chart, and pie chart for graphically representing the categorical data. There are many other effective graphical representations available in practice, however, they are beyond the level of this book. For a detailed account on categorical analysis, we refer to other books, such as Agresti's, on the topic.

## 11.2 Contingency tables and probability calculations

In categorical data, the observed frequencies are organized in rows and columns like a spreadsheet. The table of observed cell frequencies is called a *contingency table*. The basics of two-way contingency tables are introduced in this section. Categorical data are often summarized by reporting the proportion or percentage of each category. Contingency tables are used in recording counts or percentages for categorical data. We might be interested in if the new medicine's effectiveness depends on sex. Contingency tables are very useful for figuring out whether two events are dependent or independent. In this section, we will study a two-way contingency table with $N$ rows and $M$ columns. We will now give examples, where $N = 2 = M$, as well as, $M$ and $N$ both greater than two. $2 \times 2$ contingency tables are very common in many applications, where binary (yes−no, or success−failure) plays an important role. In the following example, the effectiveness is almost the same, hence the treatment can be considered gender neutral.

---

**EXAMPLE 11.2.1**

In a medical trial to study the effectiveness of a new medication for a specific illness, 180 patients were included in the study, among whom 80 were females and 100 were males. Out of these people, 55 females and 68 males responded positively to the medication.

**(a)** Create a contingency table.
**(b)** What is the probability that the medication gives a positive (success) result for males?
**(c)** What is the overall probability that the medication gives a positive result?
**(d)** Is a positive response of the new medication independent of gender?

**Solution**

**(a)** *The contingency table is given by* Table 11.1

**TABLE 11.1** Effect of Medication Based on Sex.

| | Male | Female | Totals |
|---|---|---|---|
| Positive | 68 | 55 | 123 |
| Negative | 32 | 25 | 57 |
| Totals | 100 | 80 | 180 |

**(b)** *The probability that the medication gives a positive result for males is*

$$P(\textit{positive if male}) = \frac{68}{100} = 0.68.$$

**(c)** *The overall probability that the medication gives a positive result is*

$$P(\text{overall } \textit{positive}) = \frac{123}{180} = 0.6833.$$

**(d)** *Recall that two events* A *and* B *are independent if and only if* $\mathrm{P(A} \cap \mathrm{B)} = \mathrm{P(A)P(B)}$. *Let the events* A *represent female, and* B *represent positive response of medication. Then,* $\mathrm{P(A)} = \frac{80}{180} = 0.44$, *and* $\mathrm{P(B)} = \frac{123}{180} = 0.68$. *Also,*

$$P(A \cap B) = \frac{55}{180} = 0.305$$

$$\neq P(A)P(B) = 0.300.$$

*Hence, positive response of the new medication and gender may not be independent events. However, to make a statistical conclusion, we need to perform a chi-square test, described later in the chapter.*

In general, a $2 \times 2$ contingency table can be written as in Table 11.2.

**TABLE 11.2** Notation for Joint Outcomes.

| | | Y | | |
|---|---|---|---|---|
| | | **1** | **2** | **Total** |
| X | 1 | $n_{11}$ | $n_{12}$ | $n_1$ |
| | 2 | $n_{21}$ | $n_{22}$ | $n_2$ |
| | | $n^1$ | $n^2$ | $n$ |

where $(n_{11}, n_{12}, n_{21}, n_{22})$ are random variables that have a multinomial distribution with sample size $n = (n_{11} + n_{12} + n_{21} + n_{22})$ and we can create a corresponding probabilities table of the joint distribution as in Table 11.3.

**TABLE 11.3** Notation for Joint Probabilities.

| | Y | | |
|---|---|---|---|
| | | **1** | **2** |
| X | 1 | $\pi_{11}$ | $\pi_{12}$ |
| | 2 | $\pi_{21}$ | $\pi_{22}$ |

Thus, $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ define the probability structure of the contingency table, where $\pi_{ij}'s$ can be estimated using observed data, $p_{ij} = n_{ij}/n$. From Table 11.2 we can estimate the following probabilities:

$$\widehat{P}(Y = 1) = \frac{n_{11} + n_{21}}{n} = \frac{n^1}{n}$$

$$\widehat{P}(X = 1) = \frac{n_{11} + n_{12}}{n} = \frac{n_1}{n}$$

$$\widehat{P}(Y = 1|X = 1) = \frac{n_{11}}{n_{11} + n_{12}} = \frac{n_{11}}{n_1}$$

$$\widehat{P}(X = 1|Y = 1) = \frac{n_{11}}{n_{11} + n_{21}} = \frac{n_{11}}{n^1}, \text{ etc.}$$

The marginal probability distributions of $X$ and $Y$ are the sums of cell probabilities across the columns and rows, respectively. In disease diagnostic tests, usually $Y$ is taken as the outcome of the test (positive (1), negative (2)), and $X$ is the actual condition (has disease (1), no disease (2)). With this interpretation in Table 11.2, we can define sensitivity and specificity.

**Definition 11.2.1**. *Sensitivity* (also called *true positive rate*) is defined as the probability that a patient gets a positive test result, when he has the disease. That is, the proportion of actual positives that are correctly identified:

$$Sensitivity = P(Y = 1|X = 1)$$

and *specificity* (or *true negative rate*) is the probability that the patient gets a negative test result, when he doesn't have the disease. That is, specificity measures the proportion of actual negatives that are correctly identified. Thus,

$$Specificity = P(Y = 2|X = 2).$$

In the diagnostic tests, it is better to rewrite the contingency table as in Table 11.4.

**TABLE 11.4** Notation for Probabilities.

| | | Test result (Y) | |
|---|---|---|---|
| | | **Positive (1)** | **Negative (2)** |
| True state ($X$) | Positive (1) | $\pi_1$ (sensitivity) | $1 - \pi_1$ (false negative) |
| | Negative (2) | $\pi_2$ (false positive) | $1 - \pi_2$ (specificity) |

In many of the categorical data analyses, *odds ratio* plays an important role, appearing as a parameter in the models as a measure of association or as a relative measure of effect. Thus, an odds ratio is a relative measure of effect of a treatment. In a $2 \times 2$ within row $i$, the odds of success instead of failure is $L_i = \pi_i/(1 - \pi_i)$. The ratio of odds $L_1$ and $L_2$ is the odds ratio given by

$$\theta = \frac{L_1}{L_2} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}.$$

For joint distributions with cell probabilities $\{\pi_{ij}\}$ in Table 11.3, the odds in row $i$ is $O_i = \frac{\pi_{i1}}{\pi_{i2}}, i = 1, 2$. Then the *odds ratio* is defined as

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Thus, we have the following interpretations based on the values of the odds ratio. $\theta = 1$ will correspond to independence of $X$ and $Y$. When $1 < \theta < \infty$, subjects in row 1 are more likely to have success than are the subjects in row 2, and for values of $0 < \theta < 1$, subjects in row 2 are more likely to have success than are subjects in row 2 (control [placebo, or no treatment] is better than intervention). Odds ratio is always nonnegative. When values of $\theta$ are further away from 1, this will represent stronger association in that direction.

**EXAMPLE 11.2.2** For the data of Example 11.2.1, we rewrite the table in terms of probabilities, and obtain the odds ratio.

**Solution**

*The contingency table in terms of probabilities is given by (after approximating to the second digit)*

| | Male | Female | Totals |
|---|---|---|---|
| Positive | 0.38 | 0.31 | 0.69 |
| Negative | 0.17 | 0.14 | 0.31 |
| Totals | 0.55 | 0.45 | 1.00 |

*Now, the odds ratio is given by*

$$\theta = \frac{(0.38)(0.14)}{(0.31)(0.17)} = 1.009.$$

*That is, for males the test is slightly more likely to give a correct result than for females.*

Contingency tables can have more than two categories as can be seen in Example 11.2.2.

**EXAMPLE 11.2.2**

Fruit trees are subject to a bacteria-caused disease commonly called fire blight (because the resulting dead branches look like they have been burned). One can imagine several different treatments for this disease: treatment *A*: no action (a control group); treatment *B*: careful removal of clearly affected branches; and treatment *C*: frequent spraying of the foliage with an antibiotic in addition to careful removal of clearly affected branches. One can also imagine several different outcomes from the disease: outcome 1: tree dies in the same year as the disease was noticed; outcome 2: tree dies 2−4 years after the disease was noticed; and outcome 3: tree survives beyond 4 years. A group of *N* trees are assorted into one of the treatments (i.e., every tree falls into exactly one of the following treatment categories [ *A* | *B* | *C* ] ) and over the next few years the outcome is recorded (i.e., every tree falls into exactly one of the following outcome categories [ 1 | 2 | 3 ] ). If we count the number of trees in a particular treatment/outcome pair (e.g., the number of trees that received treatment *B* and lived beyond 4 years: #*B*3), we can display the results in a contingency table:

| | Treatment | | | |
|---|---|---|---|---|
| **Outcome** | **A** | **B** | **C** | **Totals** |
| 1 | 8 | 5 | 3 | 16 |
| 2 | 4 | 3 | 3 | 10 |
| 3 | 3 | 6 | 7 | 16 |
| Totals | 15 | 14 | 13 | 42 |

**(a)** What is the probability that a randomly selected tree was given treatment B?
**(b)** What is the probability that a randomly selected tree received treatment B given it had outcome 2?
**(c)** What is the probability that a randomly selected tree received treatment B or will have outcome 2?

**Solution**

**(a)** *From the table,* $\mathrm{P(B)} = \frac{14}{42} = 0.33.$

**(b)** *From the table,* $\mathrm{P(B|2)} = \frac{3}{10} = 0.3.$

**(c)** *Thus, we have*

$$P(B \cup 2) = P(B) + P(2) - P(B \cap 2)$$

$$= \frac{14}{42} + \frac{10}{42} - \frac{3}{42} = 0.5.$$

The general representation of a two-way $r \times c$ contingency table, with cells representing counts of outcomes with $n_{ij}$ representing observed cell frequency at cell $(i, j)$, can be represented by Table 11.5.

**TABLE 11.5** Two-Way $r \times c$ Contingency Table.

| $X \setminus Y \rightarrow$ $\downarrow$ | 1 | 2 | … | c | Total |
|---|---|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | … | $n_{1c}$ | $n_1$ |
| 2 | $n_{21}$ | $n_{22}$ | … | $n_{2c}$ | $n_2$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |
| r | $n_{r1}$ | $n_{r2}$ | … | $n_{rc}$ | $n_r$ |
| Total | $n^1$ | $n^2$ | … | $n^c$ | $n$ |

Here, $n = \sum_{i=1}^{r} n_i = \sum_{j=1}^{c} n^j$ is the total number of observations, $n_i = \sum_{j=1}^{r} n_{ij}$ is the marginal frequency of row $i$, and $n^j = \sum_{i=1}^{c} n_{ij}$ is the marginal frequency of column $j$, $i = 1, \ldots, r$, $j = 1, \ldots, c$. From this table, we can calculate various probabilities. For example, the joint distribution of $X$ and Y can be expressed using the multinomial distribution given by

$$P(N_{11} = n_{11}, \ldots, N_{rc} = n_{rc}) = \frac{n!}{\prod_{i=1}^{r}\prod_{j=1}^{c} n_{ij}!} \prod_{i=1}^{r}\prod_{j=1}^{c} p_{ij}^{n_{ij}}$$

where the probability of having an outcome with $X = i$ and $Y = j$ is denoted as

$$p_{ij} = P(X = i, Y = j) = \frac{n_{ij}}{n}, i = 1, 2, \ldots, r, j = 1, 2, \ldots, c.$$

By restricting to a particular column or row, we can also obtain the conditional probabilities.

## Exercises 11.2

**11.2.1.** In a random sample of 120 females and 110 males, 80 females and 70 males own iPhones, the rest own other brands. (a) Create a contingency table. (b) What is the probability that a chosen female doesn't own an iPhone? (c) What is the probability that a randomly chosen person from this group owns an iPhone? (d) Are gender and iPhone ownership independent?

**11.2.2.** In order to study the association between mortality and treatment, a sample of 150 mice was divided into two groups: 110 were given a standard dose of pathogenic bacteria followed by an antiserum, and a control group of 40, after receiving pathogenic bacteria, was not given the antiserum. After a month, the numbers of alive and dead mice in each group are given in Table 11.6.
Compare the probabilities of survival in the two groups.

**TABLE 11.6** Contingency: Treatment and Mortality Rate.

| | Outcome | | |
|---|---|---|---|
| | Alive | Dead | Total |
| Antiserum | 80 | 30 | 110 |
| Control | 15 | 25 | 40 |
| Total | 95 | 55 | 150 |

**11.2.3.** Among teen drivers, two major reasons for causing accidents are texting and driving, and drunk driving. In a sample of 1000 accidents, Table 11.7 below lists the fatal and nonfatal accidents based on the two reasons that caused the accident.

**TABLE 11.7** Contingency Table: Accidents due to Texting and Drunken Driving.

| | Reason | | |
|---|---|---|---|
| | Texting while driving | Drunken driving | Total |
| Fatal accident | 210 | 42 | 252 |
| Nonfatal accident | 140 | 608 | 748 |
| Total | 350 | 650 | 1000 |

**(a)** What is the probability that a randomly chosen teen involved in an accident was texting while driving, given that he is involved in a nonfatal accident?
**(b)** What is the probability that a randomly chosen teen was driving while drunk, given that she or he is involved in a nonfatal accident?
**(c)** What is the probability that a randomly chosen teen is involved in an accident?

**11.2.4.** In two simple random samples of 100 men and 100 women, the color of their eyes was recorded. Here, you are now sampling from two different populations that may have different response probabilities. The actual data of the experiment are summarized in Table 11.8.

**TABLE 11.8** Contingency Table: Sex by Eye Color.

| | Eye color | | | |
|---|---|---|---|---|
| Sex | Blue | Green | Brown | Total |
| Female | 40 | 25 | 35 | 100 |
| Male | 45 | 20 | 35 | 100 |
| Total | 85 | 45 | 70 | 200 |

**(a)** What is the probability that a chosen female doesn't have brown eye color?
**(b)** What is the probability that a randomly chosen person from this group has brown-colored eyes?
**(c)** What is the probability that a randomly chosen person from this group has brown-colored eyes given he is a male?
**(d)** Create a relative frequency table and interpret its contents.

## 11.3 Estimation in categorical data

Estimation in categorical data generally involves the proportion of "successes" in a given population. This may consist of estimating a single population proportion, comparing two population proportions, or investigating the potential relationship between two or more categorical variables. Thus, if $X$ is a binary response from a trial with two possible outcomes (success/failure), then the methods of Section 5.5.2, and Section 5.5.7 for single population and two populations cases, respectively, can be used for the estimation. We will summarize the results here.

### 11.3.1 Large sample confidence intervals for *p*

For a random sample of size $n$ from a given population, the point estimate of the population parameter $p$ is given by

$$\widehat{p} = \frac{\text{the number of "successes"}}{n} = \frac{X}{n}.$$

The statistic $\widehat{p}$ is the key entity in the binomial probability estimation, with true mean $p$ and variance $(p(1-p))/n$, respectively. For large sample size $n$ *(if both* $np \geq 5$ *and* $n(1-p) \geq 5$*)*, we use the normal pdf to obtain approximate $100(1-\alpha)\%$ *confidence interval for* $p$ which is given by

$$\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right).$$

That is,

$$P\left[\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq p \leq \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right] \approx (1-\alpha)$$

and we read it as "based on the random sample of size $n$, we are about $100(1-\alpha)\%$ certain that the true value of $p$ is in the interval $\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \widehat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$.

We now restate a procedure from Section 5.7 for a large sample confidence interval for the difference of the true proportions, $p_1 - p_2$, in two binomial distributed populations.

**Large sample confidence interval for $p_1 - p_2$**

The $(1-\alpha)100\%$ large sample confidence interval for $p_1 - p_2$ is given by

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\left(\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}\right)},$$

where $\widehat{p}_1$ and $\widehat{p}_2$ are the points estimators of $p_1$ and $p_2$. This approximation is applicable if $\widehat{p}_i n_i \geq 5, i = 1,2$ and $(1-\widehat{p}_i)n_i \geq 5, i = 1,2$. The two samples are independent.

The Wald confidence interval can be obtained using the following R-commands.
*1-sample proportions test* (*Wald*)
library(epitools)
binom.approx(9,20)
We will get the following output.

| *X* | *n* | Proportion | Lower | Upper conf. level | |
|---|---|---|---|---|---|
| 9 | 20 | 0.45 | 0.2319678 | 0.6680322 | 0.95 |

From this, we can see that about 95% confidence interval for the true proportion $p$ is (0.2319678, 0.6680322).

## 11.4 Hypothesis testing in categorical data analysis

Hypothesis testing on the population proportion is the same as in Sections 6.4 and 6.5 for one and two proportions, respectively, and we will refer to those sections. Now we will explain the chi-square tests. There are two kinds of chi-square tests: one-way and two-way analysis. For example, if we are interested in comparing the effectiveness of two or more types of drugs in treating a particular disease, we will have a one-way analysis. Note that in order to use ANOVA or a *t*-test, we need at least one of the variables to be continuous. Thus, we resort to Pearson's chi-square test. In addition, if we are interested to find out whether these drugs differently affect men and women, then we need two-way analysis, and in two-way analysis we will use data from contingency tables. The purpose of both is to determine if the observed frequencies

are significantly different from the frequencies that we would expect by chance or from a hypothesized distribution. In both one-way and two-way data, chi-square tests are most often used.

### 11.4.1 The chi-square tests for count data: one-way analysis

A chi-square test is useful to analyze categorical data and it is intended to test how likely it is that an observed probability distribution is due to chance, that is, to test whether a frequency distribution observed for categories fits an expected probability distribution. In this section, we will study several commonly used tests for count data, where observations are given by counting that assumes nonnegative integer values, {0,1,2,…} (this test can be considered as a one-way test). These are basically large sample tests based on a $\chi^2$-approximation. Suppose that we have outcomes of a multinomial experiment that consists of $k$ mutually exclusive and exhaustive events, $A_1, \ldots, A_k$. Let $P(A_i) = p_i, i = 1, 2, \ldots, k$. Then $\sum_{i=1}^{k} p_i = 1$. Let the experiment be repeated $n$ times, and let $X_i$ ($i = 1, 2, \ldots, k$), represent the number of times the event $A_i$ occurs. Then $(X_1, \ldots, X_k)$ has a multinomial distribution with parameters $n, p_1, \ldots, p_k$. Recall that if $n$ is the total number of trials, that is, $x_i \in \{0, 1, \ldots, n\}$ with $\sum_{i=1}^{k} x_i = n$, then the pmf of the multinomial distribution is given by

$$\frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k},$$

with $E(X_i) = np_i$.

Now, let

$$Q^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}.$$

It can be shown that for large $n$, the random variable $Q^2$ is approximately $\chi^2$-distributed with $(k-1)$ degrees of freedom. It is required that $np_i \geq 5$ ($i = 1, 2, \ldots, k$) for the approximation to be valid, although the approximation generally works well if we only have a few values of $i$ (no more than 20% of the total cells), $np_i \geq 1$ and the rest (about 80%) satisfy the condition that $np_i \geq 5$. This statistic was proposed by Karl Pearson in his 1900 paper.

It should be noted that the $\chi^2$-test that we are studying in this section is an approximate test valid for large samples. Often $X_i$ is called the observed frequency and is denoted by $O_i$ (this is the observed value in class $i$), and $np_i$ is called the expected frequency and is denoted by $E_i$ (this is the theoretical distribution frequency under the null hypothesis). Thus, with these notations, we can calculate

$$Q^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum \frac{(Observed - Expected)^2}{Expected}.$$

The example given below illustrates how we apply this goodness-of-fit test.

---

**EXAMPLE 11.4.1**

A plant geneticist grows 200 progeny from a cross that is hypothesized to result in a 3:1 phenotypic ratio of red-flowered to white-flowered plants. Suppose the cross process produces 170 red- to 30 white-flowered plants. (a) Calculate $Q^2$ for this experiment. (b) Do the given data support the 3:1 ratio at $\alpha = 0.05$?

**Solution**

*There are two categories of data totaling* $n = 200$. *Hence,* $k = 2$. *Let* $i = 1$ *represent red-flowered and* $i = 2$ *represent white-flowered plants. Then* $O_1 = 170$, *and* $O_2 = 30$.

*Here, we want to test the hypothesis to answer the posed question.*

$H_0$: *The flower color population ratio is* 3 : 1,

*Vs.*

$H_a$: *The flower color population sampled has a flower color ratio that is not* 3 *red* : 1 *white.*

**(a)** *We are given that the probability of red flowers is* $p_1 = 3/4$, *and the probability of white flowers is* $p_2 = 1/4$ *and the condition that* $np_1 \geq 5$ *and* $np_2 \geq 5$, *are satisfied. Thus, we can proceed to calculate* $Q^2$ *for the information that is given. Thus,*

$$E_1 = np_1 = (200)(3/4) = 150, \text{ and } E_2 = np_2 = (200)(1/4) = 50$$

*and*

$$Q^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(170 - 150)^2}{150} + \frac{(30 - 50)^2}{50} = 10.667.$$

*Since* $k = 2$, *from the* $\chi^2$*- table with 1 degree of freedom and* $\alpha = 0.05$, *the rejection region is* $\left\{Q^2 > \chi^2_{1,0.05} = 3.841\right\}$. *Since 10.667 is greater than 3.841, we reject the null hypothesis and conclude that the color ratio is not 3:1. The data support the alternative hypothesis that the ratio is not 3 red: 1 white.*

---

The type of calculation in Example 11.4.1 gives a measure of how close our observed frequencies are compared to the expected frequencies. Smaller values of $Q^2$ indicate better fit of the data. The test is also called a "**goodness-of-fit**" test statistic, because this measures how well the observed distribution of the data fits with the distribution that is expected if data are consistent with the assumed distribution. Note that this is equivalent to testing the parameters of a multinomial distribution. Let an experiment have $k$ mutually exclusive and exhaustive outcomes $A_1$, $A_2$, …, $A_k$. We would like to test the null hypothesis that all the $p_i = p(A_i)$, $i = 1, 2, \ldots, k$ are equal to known numbers $p_{i0}, i = 1, \ldots, k$.

The test procedure that we use to test the subject hypothesis is summarized below.

**Testing the parameters of a multinomial distribution (summary)**

To test

$H_0: p_1 = p_{10}, \ldots, p_k = p_{k0}$

*Vs.*

$H_a$: At least one of the probabilities is different from the hypothesized values

The test is always a one-sided upper tail test.

Let $O_i$ be the observed frequency, $E_i = np_{i0}$ be the expected frequency (frequency under the null hypothesis), and $k$ be the number of classes. The test statistic is

$$Q^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

The test statistic $Q^2$ has an approximate chi-square probability distribution with $k - 1$ degrees of freedom.

The rejection region is given by

$$Q^2 \geq \chi^2_{k-1,\,\alpha}.$$

**Assumption:** $E_i \geq 5$ for all $k$ and no more than 20% cells have $5 > E_i \geq 1$.

Note that the chi-square test will tell us if there is a significant difference between the observed data and the hypothesis distribution. However, it cannot test the strength of dependence or direction of the difference. This test is known as the $\chi^2-$ *goodness-of-fit test*. It implies that if the observed data are very close to the expected data, we have a very good fit and we do not reject the null hypothesis. That is, for small $Q^2$ values, we don't have enough evidence to reject $H_0$ and hence, we will not reject $H_0$.

The following examples illustrate how we apply the chi-square goodness-of-fit test.

---

**EXAMPLE 11.4.2**

A TV station broadcasts a series of programs on the ill effects of smoking marijuana. After the series, the station wants to know whether people have changed their opinion about legalizing marijuana. Historical data from before the series showing the proportions of different categories of opinions is shown in the first table below, and the second table shows the sample proportion from the 500 randomly selected people.

**Before the Series Was Shown**

| For legalization | Decriminalization | Existing law (fine or imprisonment) | No opinion |
|---|---|---|---|
| 7% | 18% | 65% | 10% |

**After the Series Was Shown**

| For legalization | Decriminalization | Existing law (fine or imprisonment) | No opinion |
|---|---|---|---|
| 39% | 9% | 36% | 16% |

Here, $k = 4$, and we wish to test the following hypothesis

$$H_0: p_1 = 0.07; \quad p_2 = 0.18; \quad p_3 = 0.65; \quad p_4 = 0.1$$

*Vs.*

$H_a$: At least one of the probabilities is different from the hypothesized value.

The test is always an upper tail test. We will test this hypothesis using $\alpha = 0.01$.

**Solution**

*We have the expected frequencies,*

$$E_1 = (500)(0.07) = 35; \quad E_2 = 90; \quad E_3 = 325; \quad E_4 = 50.$$

*The observed frequencies are*

$$O_1 = (500)(0.39) = 195; \quad O_2 = 45; \quad O_3 = 180; \quad O_4 = 80.$$

*The value of the test statistic is given by*

$$Q^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

$$= \left[ \frac{(195-35)^2}{35} + \frac{(45-90)^2}{90} + \frac{(180-325)^2}{325} + \frac{(80-50)^2}{50} \right]$$

$$= 836.62.$$

*From the $\chi^2$-table, $\chi^2_{0.01,\,3} = 11.3449$. Because the test statistic $Q^2 = 836.62 > 11.3449$, we reject $H_0$ at $\alpha = 0.01$. Hence, the data suggest that people have changed their opinion after watching the series on the ill effects of smoking marijuana was shown. That is, the TV station broadcast did not change the opinion of the audience.*

---

**EXAMPLE 11.4.3**

A die is rolled 60 times and the face values are recorded. The results of this experiment are:

| Up face | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Frequency | 8 | 11 | 5 | 12 | 15 | 9 |

Is the die balanced fair? Test this question using $\alpha = 0.05$.

**Solution**

*If the die is fair, we must have*

$$p_1 = p_2 = \ldots = p_6 = \frac{1}{6}$$

*where* $p_i = P$ *(face value on the die is* i), i = 1, 2, …, 6. *This experimental outcome follows the discrete uniform probability distribution.*

*Hence,*

$$H_0: p_1 = p_2 = \ldots = p_6 = \frac{1}{6}$$

*Vs.*

$H_a$: At least one of the probabilities is different from the hypothesized value of 1/6

*Note that* $E^1 = n_1 p_1 = (60)\ (1/6) = 10, \ldots, E_6 = 10$, *and the condition of using this test is satisfied.*
*We summarize the calculations in the following table:*

| Face value | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Frequency, $O_i$ | 8 | 11 | 5 | 12 | 15 | 9 |
| Expected value, $E_i$ | 10 | 10 | 10 | 10 | 10 | 10 |

*The test statistic value is given by*

$$Q^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = 6.$$

*From the chi-square table with 5 d.f.,* $\chi^2_{0.05,\ 5} = 11.070$.
*Thus,* $\chi^2_{0.05,5} = 11.070 = 11.07 > Q^2 = 6$ *and since the value of the test statistic does not fall in the rejection region, we do not reject* $H_0$. *Therefore, we do not have enough evidence to conclude that the die is not fair.*

The tests that we will study here are approximate tests, but very useful in performing statistical analysis. Let the random variables $(X_1, \ldots, X_k)$ have a multinomial distribution with parameters $n, p_1, \ldots, p_k$. Let $n$ be known. We will now present some important tests based on the chi-square $\chi^2$-statistic.

### 11.4.2 Two-way contingency table: test for independence

Another important use of the $\chi^2$-statistic is testing for dependencies or associations between the rows and columns in a contingency table. That is, if we have two categorical variables, is there convincing evidence of association between the variables in the population? Here, we have seen that $n$ randomly selected items are classified according to two different criteria, or two factors (row factor and column factor), where the row factor has $r$ levels and the column factor has $c$ levels. The obtained data are displayed in a contingency table as shown in Table 11.9, where $n_{ij}$ represents the number of data values in row $i$ and column $j$. Our interest here is to test for independence of the two-way classifications of observed events. For example, we might classify a sample of students by male or female and by their grade on a statistics course in order to test the hypothesis that the grades are independent of gender. More generally the problem is to investigate a *dependency* (or *contingency*) between two classification criteria.

In the present study the given data of a problem are presented in a tabular form as illustrated by Table 11.9.

**TABLE 11.9 Two-Way Contingency Table.**

| | Levels of column factor | | | | |
|---|---|---|---|---|---|
| | **1** | **2** | **…** | **C** | **Row total** |
| Row 1 | $n_{11}$ | $n_{12}$ | | $n_{1c}$ | $n_{1.}$ |
| levels 2 | $n_{21}$ | $n_{21}$ | | $n_{2c}$ | $n_{2.}$ |
| . | | | | | |
| . | | | | | |
| $r$ | $n_{r1}$ | $n_{r2}$ | | $n_{rc}$ | $n_{r.}$ |
| Column totals | $n_{.1}$ | $n_{.2}$ | | $n_{.c}$ | $N$ |

where $N = \sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r} n_{i.} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$ is the grand total.

Here, we wish to test the hypothesis that the two factors (rows and columns) are independent. We summarize the procedure in the following table for testing that the factors represented by the rows are independent of those represented by the columns.

**Testing for the independence of two factors**

To test

$H_0$: The factors are independent

*Vs.*

$H_a$: The factors are dependent

the test statistic is,

$$Q^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where

$$O_{ij} = n_{ij}$$

and

$$E_{ij} = \frac{n_{i.} n_{.j}}{N}.$$

Then under the null hypothesis the test statistic $Q^2$ has an approximate chi-square probability distribution with $(r-1)(c-1)$ degrees of freedom.

Hence, the rejection region is $Q^2 > \chi^2_{\alpha,\,(r-1)(c-1)}$.

**Assumption:** $E_{ij} \geq 5$.

**EXAMPLE 11.4.4**

Table 11.10 gives a classification according to religious affiliation and marital status for 500 randomly selected individuals.

**TABLE 11.10 Marital Status and Religious Affiliation.**

| | | Religious affiliation | | | | | |
|---|---|---|---|---|---|---|---|
| | | **A** | **B** | **C** | **D** | **None** | **Total** |
| Marital status | Single | 39 | 19 | 12 | 28 | 18 | 116 |
| | With spouse | 172 | 61 | 44 | 70 | 37 | 384 |
| | **Total** | 211 | 80 | 56 | 98 | 55 | **500** |

Using a level of significance, $\alpha = 0.01$, test the null hypothesis that marital status and religious affiliation are independent.

**Solution**

*We need to test the hypothesis*

$H_0$: *Marital status and religious affiliation are independent*

*Vs.*

$H_a$: *Marital status and religious affiliation are dependent.*

*Here,* $c = 5$ *and* $r = 2$. *For* $\alpha = 0.01$, *and for* $(c-1)(r-1) = 4$ *degrees of freedom, we have*

$$\chi^2_{0.01,4} = 13.2767.$$

*Hence, the rejection region is* $Q^2 > 13.2767$.

*We have* $E_{ij} = \frac{n_i n_j}{N}$. *Thus,*

$$E_{11} = \frac{(116)(211)}{500} = 48.952;\ E_{12} = \frac{(116)(80)}{500} = 18.5;$$
$$E_{13} = \frac{(116)(56)}{500} = 12.992,\ E_{14} = \frac{(116)(98)}{500} = 22.736;$$
$$E_{15} = \frac{(116)(55)}{500} = 12.76,\ E_{21} = \frac{(384)(211)}{500} = 162.05;$$
$$E_{22} = \frac{(384)(80)}{500} = 61.44;\ E_{23} = \frac{(384)(56)}{500} = 43.008;$$

*and*

$$E_{24} = \frac{(384)(98)}{500} = 75.264; E_{25} = \frac{(384)(55)}{500} = 42.24.$$

*The value of the test statistic is*

$$Q^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$
$$= \left[\frac{(39-48.952)^2}{48.952} + \frac{(19-18.5)^2}{18.5} + \frac{(12-12.992)^2}{12.992} + \frac{(28-22.736)^2}{22.736} + \frac{(18-12.76)^2}{12.76} + \frac{(172-162.05)^2}{162.05}\right.$$
$$\left. + \frac{(61-61.44)^2}{61.44} + \frac{(44-43.08)^2}{43.08} + \frac{(70-75.264)^2}{75.264} + \frac{(37-42.24)^2}{42.24}\right] = 7.1351.$$

*Because the observed value of $Q^2$ does not fall in the rejection region, we do not reject the null hypothesis at $\alpha = 0.01$. Therefore, based on the given data, the marital status and religious affiliation are independent. Note that the assumption of $E_{ij} \geq 5$ is satisfied.*

It should be noted that the chi-square test becomes inaccurate when used to analyze $2 \times 2$ contingency tables, and when the large sample conditions, $E_i \geq 5$ for all cells and no more than 20% cells with $E_i \geq 1$, are not met. Fisher's exact test is used in these cases, and we refer the reader to the book by Agresti, among other places.

## Exercises 11.4

**11.4.1.** If we toss a coin a few times, we expect half heads and half tails. Suppose we tossed a coin 200 times and obtained 104 heads. Can we assume the coin is fair? Use $\alpha = 0.05$.

**11.4.2.** The following table gives the opinion on collective bargaining by a random sample of 200 employees of a school system, belonging to a teachers' union.

**Opinion on Collective Bargaining by Teachers' Union**.

| | For | Against | Undecided | Total |
|---|---|---|---|---|
| Staff | 30 | 15 | 15 | 60 |
| Faculty | 50 | 10 | 40 | 100 |
| Administration | 10 | 25 | 5 | 40 |
| Column totals | 90 | 50 | 60 | **200** |

Test the hypotheses

$H_0$: Opinion on collective bargaining is independent of employee classification

*Vs*.

$H_a$: Opinion on collective bargaining is dependent on employee classification using $\alpha = 0.05$.

**11.4.3.** A random sample was taken of 300 undergraduate students from a university. The students in the sample were classified according to their gender and according to the choice of their major. The results are given in Table 11.11.

**TABLE 11.11** Gender and Major Contingency Table.

| Gender | Arts and sciences | Engineering | Business | Other | Total |
|---|---|---|---|---|---|
| Male | 75 | 40 | 24 | 66 | 205 |
| Female | 45 | 12 | 15 | 23 | 95 |
| Total | 120 | 52 | 39 | 89 | 300 |

Test the hypothesis that the choice of the major by undergraduate students in this university is independent of their gender. Use $\alpha = 0.01$.

**11.4.4.** A presidential candidate advertises on TV by comparing his positions on some important issues with those of his opponent. After a series of advertisements, a pollster wants to know whether people have changed their opinion about the candidate. Historical data from before the advertisement show the proportions of different categories of opinions in the first table below, and then the second table below shows the data based on a survey of 950 randomly chosen people:

**Before the Advertisement Was Shown.**

| Support the candidate | Oppose the candidate | Need to know more about the candidate | Undecided |
|---|---|---|---|
| 40% | 20% | 5% | 35% |

**After the Advertisement Was Shown.**

| Support the candidate | Oppose the candidate | Need to know more about the candidate | Undecided |
|---|---|---|---|
| 45% | 25% | 2% | 28% |

Let $p_i$, $i = 1, 2, 3, 4$, represent the respective true proportions.
Test

$$H_0: p_1 = 0.35;\ p_2 = 0.20;\ p_3 = 0.15;\ p_4 = 0.3$$

*Vs.*

$H_a$: At least one of the probabilities is different from the hypothesized value.

Test this hypothesis using $\alpha = 0.05$.

**11.4.5.** A survey of footwear preferences of a random sample of 100 undergraduate students (50 females and 50 males) from a large university resulted in Table 11.12.

**TABLE 11.12** Gender and Footwear Table.

| | Boots | Leather shoes | Sneakers | Sandals | Other |
|---|---|---|---|---|---|
| Female | 12 | 9 | 12 | 10 | 7 |
| Male | 10 | 12 | 17 | 7 | 4 |

**(a)** Let $p_i$, $i = 1, 2, 3, 4, 5$ represent the respective true proportions of students with a particular footwear preference, and let

$$H_0: p_1 = 0.20;\ p_2 = 0.20;\ p_3 = 0.30;\ p_4 = 0.20;\ p_5 = 0.10$$

*Vs.*

$H_a$: At least one of the probabilities is different from the hypothesized value.

Test this hypothesis using $\alpha = 0.05$.

**(b)** Test the hypothesis that the choice of footwear by undergraduate students in this university is independent of their gender, using $\alpha = 0.05$.

**11.4.6.** A casino game involves rolling three dice. The winning is directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 150 times, with the following observed counts:

| Number of sixes | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Number of rolls | 72 | 51 | 21 | 6 |

Assuming that roll of one die does not affect the roll of others, test to determine if the dice are fair, at $\alpha = 0.05$.

**11.4.7.** Criminologists are interested to know if there is any relationship between homicides and seasons of the year. In the paper "Is Crime Seasonal" (https://bjs.gov/content/pub/pdf/ics.pdf), the following data for 1361 homicides are given in terms of seasons.

| Winter | Spring | Summer | Fall |
|---|---|---|---|
| 328 | 334 | 372 | 327 |

Do these data support the theory that the homicide rate is not the same over the seasons?

**11.4.8.** In order to find out the relationship between packaging preferences (in terms of size) to economic status, a manufacturing company of pain medication conducted a survey. Table 11.13 gives the result of this survey.

**TABLE 11.13 Economic Status and Size of Purchase.**

| | Lower | Middle | Upper |
|---|---|---|---|
| Small | 23 | 24 | 19 |
| Medium | 22 | 26 | 20 |
| Large | 16 | 28 | 18 |
| Jumbo | 15 | 21 | 30 |

Is there a significant relationship between packaging preferences and economic status? Use $\alpha = 0.05$.

## 11.5 Goodness-of-fit tests to identify the probability distribution

In studying various real-world phenomena, we begin with a random sample of data $X_1, \ldots, X_n$ that represents values of some sort of a subject of interest. These measurements could represent the amount of carbon dioxide, $CO_2$, in the atmosphere on a daily basis, the sizes of cancerous breast tumors, the monthly average rainfall in the state of Florida, the average monthly unemployment rate in the United States, the hourly wind forces of a hurricane, etc. In order for us to probabilistically understand the behavior of these phenomena, we will need to identify the probability distribution that characterizes the probabilistic behavior of the given data, that is, the pdf of the random sample they were drawn from. For example, at a certain time point we say that these data follow or come from the normal or exponential probability distribution. One of the important questions then is whether the observed data are representative or follow a particular probability distribution. The goodness-of-fit tests are used to test if a sample fits a particular distribution. In fact, there is nothing we can do parametrically or statistically unless through goodness-of-fit testing we identify the probability density functions, which probabilistically characterize the behavior of the given data, the phenomenon of interest.

To accomplish this objective of identifying the underlying probability distribution, we will discuss four statistical tests (methods) that we can use to determine how good the data fit a particular well-defined probability distribution. These four tests are: *Pearson's chi-square test*, *Kolmogorov–Smirnov test*, *Anderson–Darling test*, and *Shapiro–Wilk test*. Even though we will give theoretical steps of how to calculate the quantities for most of the tests, it should be noted that, in practice, most of the goodness-of-fit tests will be done using statistical software. For large data sets, it is tedious to do goodness-of-fit analysis by hand. There are other methods we can follow if we are not able to identify the appropriate pdf, such as nonparametric or probability distribution free analysis, which will be discussed in Chapter 12.

### 11.5.1 Pearson's chi-square test

When we are interested in studying the behavior of a given unknown phenomenon, we begin by obtaining thorough experimentation or other means a set of data, the random sample. The initial step of studying this phenomenon is to try to identify the probability distribution that characterizes the behavior of the given data. The methods that we use are called goodness-of-fit tests. That is, if we assume that a given set of data follows the normal or Gaussian probability distribution, the data must be a good-fit to this distribution with a high degree of assurance. Historically, the first statistical method to test the fit of a particular distribution to a given set of data was Person's chi-square goodness-of-fit test.

In hypothesis-testing problems we often assume that the form of the population distribution is known. For example, in a $\chi^2$-test for variance, we assume that the population is normal. The goodness-of-fit test examines the validity of such an assumption if we have a large enough sample. We now describe the goodness-of-fit test procedure for such an application. This test uses a measure of goodness of fit, which is the mean of the differences between the observed and expected outcome frequencies (counts of observations), each squared and divided by the expected frequencies. That is, the test statistic is given by:

$$Q^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Here, $O_i$ is the $i$th observed outcome frequency (in class $i$), $E_i$ is the $i$th expected (theoretical) frequency, and $i = 1, 2, \ldots, k$ is the number of classes. The expected frequency, $E_i$, is calculated by

$$E_i = [F_0(y_u) - F_0(y_l)]n,$$

where $F_0$ is the cumulative probability distribution that is being tested (assumed) to determine if the given data follow (fit) this probability distribution; $Y_u$ and $Y_l$ are the upper and lower limits of class $i$, respectively; and $n$ is the sample size. Thus, we proceed to set up the hypothesis,

$H_0$: *The given data follow a specific probability distribution* $(F)$

*Vs*.

$H_a$: *The data do not follow the specified probability distribution*.

We proceed to calculate the value of the $Q^2$ statistic and if it is greater than the value we obtain from the $\chi^2_{\alpha,\, k-1}$ tables for a given level of significance $\alpha$ and $k - 1$ degrees of freedom, we reject the hypothesis. Note that for $(k - 1)$ degrees of freedom, we need to know that the $F$ distribution is completely defined. If there are any unknown parameters that need to be estimated, we need to reduce that many degrees of freedom. That is, the data do not follow or fit the specified probability distribution. Thus, if the calculated value of the chi-square test statistic is less than the $\chi^2_{\alpha,\, k-1}$ value that we obtain from the tables, indeed the specified data fit the specified probability distribution at a level of significance $\alpha$. That is, the rejection region is given by

$$P\left(Q^2 \geq \chi^2_{\alpha,k-1}\right) = \alpha.$$

The basic assumptions for applying this test are

**i.** The observed frequencies in the $k$ classes should be independent.

**ii.** $\sum_{i=1}^{k} E_i = \sum_{i=1}^{k} O_i = n$.

**iii.** The total frequency, $n$, should be more than 50.

**iv.** Each expected frequency, $E_i$, in each class should be at least 5.

In testing the above hypothesis, we usually assume a value of the level of significance $\alpha$, like $\alpha = 0.01, 0.05, 0.1$, etc. and proceed to make the decision of accepting or rejecting the null hypothesis based on the assumed $\alpha$. However, by using statistical packages such as R, it gives you a $p$ value, in contrast to a fixed $\alpha$ value, that is calculated based on the test statistic, and denotes the threshold value of the significance level in the sense that the null hypothesis will be accepted at all significance $\alpha$ levels less than the calculated $p$ value. For example, if $p$ value $= 0.05$, the null hypothesis will not be rejected for all values of assumed $\alpha < p$ value of 0.05, and will be rejected for higher levels. Recall that the $p$ value is the probability of observing a sample statistic as extreme as the test statistic. Since here the test statistic has chi-square distribution, use the chi-square table to calculate the $p$ value. Note that recently using the $p$ value has created some useful criticism of its applicability but we will not discuss these issues here. Following is a summary of a step-by-step procedure for applying the subject test.

**Goodness-of-fit test procedures for identifying the probability distributions**

Let $X_1, \ldots X_n$ be a sample from a population with cdf $F(x)$. We wish to test $H_o$: $F(x) = F_o(x)$, where $F_o(x)$ is completely specified (assumed) pdf.

1. Divide the range of values of the random variables $X_1$ into $k$ nonoverlapping intervals $I_1$, $I_2$, … $I_k$. Let $O_j$ be the number of sample values that fall in the interval $I_j$ ($j = 1, 2, \ldots, k$).
2. Assuming the probability distribution of $X$ to be $F_o(x)$, find $P(X \in I_j)$. Let $P(X \in I_j) = \pi_i$. Let $E_j = n\pi_j$ be the expected frequency.
3. Compute the test statistic $Q^2$ given by

$$Q^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

   The test statistic $Q^2$ has an approximate $\chi^2$-distribution with $(k - 1)$ degrees of freedom.
4. Reject the $H_0$ if $Q^2 \geq \chi^2_{\alpha,\,(k-1)}$.
5. **Assumptions:** $E_j \geq 5$, $j = 1, 2, \ldots, k$.

It should be noted that when the hypothesis distribution does not involve any extra parameters, the degrees of freedom is $(k - 1)$. If the hypothesis distribution involves extra parameters (e.g., in the exponential distribution example given below, because the exponential distribution has one rate parameter involved which needs to be estimated from the data), the degrees of freedom for the chi-square test need to be adjusted to subtract one degree of freedom used for estimating each of the unknown parameters. Also note that if the observed data, $O_i$, is very close to the expected value, $E_i$, the difference $O_i - E_i$ is going to be very small, which implies the $Q^2$ statistic will be small and, thus, a good fit of the given data to the assumed pdf. It should be noted that when data are numerical, we don't have natural categories. We need to create categories (similar to the way we create intervals for histogram) such that for each category the condition $E_j = n\pi_j \geq 5$ is satisfied. Example 11.5.1 is given only for demonstration purposes, for more accuracy, our sample size should be at least 50.

**EXAMPLE 11.5.1**

We are given a random sample of $n = 30$ observations of a given experiment of a certain phenomenon of interest, that is.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.79 | 2.62 | 11.92 | 9.77 | 12.13 | 15.04 | 16.14 | 20.74 | 22.73 | 23.29 | 24.97 | 26.12 |
| 211.06 | 29.60 | 32.47 | 36.32 | 42.18 | 45.06 | 45.64 | 48.34 | 48.87 | 64.99 | 66.28 | 68.00 |
| 68.60 | 75.34 | 99.32 | 162.48 | 164.38 | 235.95 | | | | | | |

We believe that these data may follow the exponential pdf. Test our belief at $\alpha = 0.05$.

**Solution**

*We need to test*

$H_0$: *The given data follow an expontential probability distribution*

*Vs.*

$H_a$: *The data do not follow the specified probability distribution*

*We will now give steps of how we can solve this problem analytically and then illustrate how this can be implemented in R.*

*Recall that the pdf of the exponential distribution with the rate parameter* $\lambda$ *is* $f(x) = \lambda \exp(-\lambda x)$, *for* $x \geq 0$, *and* $\lambda > 0$. *The MLE of* $\lambda$ *is given by* $\widehat{\lambda} = \frac{1}{\bar{x}}$. *Since* $\lambda$ *is unknown, we can calculate bin probabilities using* $\widehat{\lambda}$ *in place of* $\lambda$.

*For the exponential random variable* X *with the rate parameter* $\lambda$, *we know that*

$$P(a \leq X \leq b) = \int_a^b \lambda e^{-\lambda x} dx = e^{-\lambda a} - e^{-\lambda b}.$$

*Based on this, we can now calculate the probability of the exponential random variable falling in each individual interval (bin).*

*Note that the minimum value for these data is 1.79, and the maximum is 235.16. The sample mean is* $\bar{x} = 57.738$. *Thus,* $\widehat{\lambda} = \frac{1}{57.738} \approx 0.017$. *Considering the observed range of data and the size of each bin to ensure the large sample approximation is valid for the grouped data, we divide the data into four unequal-width bins (intervals) as* [0, 25], [25, 50], [50, 80], and [80, ∞] *(since the exponential is continuous, open or closed intervals will not change the probabilities). Then we can calculate each cell/bin probabilities as follows (you could also use R to calculate the cumulative density function at a certain value, a, by using pexp(a, rate = λ)), and then calculate the difference between the CDFs at the lower and upper bounds of each interval to calculate these cell probabilities:*

$$P(0 \leq X \leq 25) = 0.351, P(25 \leq X \leq 50) = 0.228, P(50 \leq X \leq 80) = 0.170 \text{ and}$$
$$P(X \geq 80) = 0.250.$$

*Thus, the expected cell frequencies under the assumed exponential distribution are calculated as* $E_1 = 0.351 \times 30 = 10.54, E_2 = 0.228 \times 30 = 6.84, E_3 = 0.170 \times 30 = 5.11$, *and* $E_4 = 0.250 \times 30 = 7.51$.

*Note that the condition,* $E_i \geq 5$, *for each* i, *is satisfied, and hence, the test is appropriate as the approximate chi-square distribution is satisfied. Now the observed and expected frequencies as well as* $\frac{(O_i - E_i)^2}{E_i}$, $i = 1, ..., 4$, *needed for calculating the chi-square test statistic are given in the following table.*

| Data interval (bin) | Observed frequency ($O_i$) | Expected frequency ($E_i$) | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|
| 0–25 | 11 | 10.54 | 0.0198 |
| 25–50 | 9 | 6.84 | 0.6837 |
| 50–80 | 5 | 5.11 | 0.0025 |
| $\geq 80$ | 5 | 7.51 | 0.8364 |

Thus, the test statistic $Q^2$ is given by

$$Q^2 = \sum_{i=1}^{k=4} \frac{(O_i - E_i)^2}{E_i} = 1.5424.$$

*From the* $\chi^2$*-table with* $k - 2 = 2$ *degrees of freedom (one additional degree of freedom is lost because we had to estimate* $\lambda$*), and with* $\alpha = 0.05$, *rejection region is* $\{Q^2 \geq 5.991\}$. *Since 1.5424 is less than 5.991, we fail to reject* $H_0$. *Thus, we can conclude that the observed data fit well with the exponential distribution.*

*Below we provide the R-code and output for implementing the above-described goodness-of-fit test.*

**R-code and output**

```
> x = c(1.79,2.62,11.92,9.77,12.13,15.04,16.14,20.74,22.73,23.29,24.97.
+ 26.12,211.06,29.60,32.47,36.32,42.18,45.06,45.64,48.34,48.87,64.99.
+ 66.28,68.00,68.60,75.34,99.32,162.48,164.38,235.95)
> # estimate the rate parameter
> lambda <- 1/mean(x)
> lambda
[1] 0.01731962
> # define the bin boundaries
> bounds <- c(25,50,80,Inf)
> # the cummulative bin frequencies
> Ocum <- c(sum(x<=bounds[1]),sum(x<=bounds[2]),sum(x<=bounds[3]),sum(x<=bounds[4]))
> # the observed bin frequencies
> O <- - Ocum-c(0,Ocum[-4])
> # CDF
> cp <-pexp(bounds, rate = lambda)
> # bin probabilities
> bps <-cp-c(0,cp[-4])
> # chi-square test.
> res <-chisq.test(x = 0,p = bps)
> res
# p-value
> pv <-1-pchisq(as.numeric(res[1]),2)
> pv
[1] 0.4624616.
```

*Thus, for a* p *value of* 0.4624616, *we do not reject the null hypothesis and we conclude that the given data are consistent with the exponential distribution.*

**EXAMPLE 11.5.2**

The grades of students in a class of 200 are given in the following table. Test the hypothesis that the grades are normally distributed with a mean of 75 and a standard deviation of 8. Use $\alpha = 0.05$.

| Range | 0–59 | 60–69 | 70–79 | 80–89 | 90–100 |
|---|---|---|---|---|---|
| Number of students | 12 | 36 | 90 | 44 | 18 |

**Solution**

*To test the hypothesis,*

$$H_0: \textit{Student grades follow a } N(\mu = 75,\ \sigma^2 = 64) \textit{ distribution.}$$

*Vs.*

$$H_a: \textit{Student grades do not follow the } N(\mu = 75,\ \sigma^2 = 64) \textit{ distribution.}$$

*We have* $O_1 = 12$, $O_2 = 36$, $O_3 = 90$, $O_4 = 44$, $O_5 = 18$.

*We now compute* $\pi_i (i = 1, 2, \ldots, 5)$, *using the continuity correction factor,*

$$\pi_1 = P\{X \le 59.5 | H_0\} = P\left\{z \le \frac{59.5 - 75}{8}\right\} = 0.0262,$$

$$\pi_2 = 0.2189, \pi_3 = 0.4722, \pi_4 = 0.2476, \pi_5 = 0.0351,$$

*and*

$$E_1 = 5.24, E_2 = 43.78, E_3 = 94.44, E_4 = 49.52, E_5 = 7.02.$$

*The test statistic results in*

$$Q^2 = \sum_{i=1}^{n} \frac{(O_i - e_i)^2}{e_i}$$

$$= \frac{(12 - 5.74)^2}{5.74} + \frac{(36 - 43.78)^2}{43.78} + \frac{(90 - 94.44)^2}{94.44} + \frac{(44 - 49.52)^2}{49.52} + \frac{(18 - 7.02)^2}{7.02}$$

$$= 26.22.$$

*$Q^2$ has a chi-square distribution with $(5 - 1) = 4$ degrees of freedom. The critical value is $\chi^2_{0.05,\,4} = 7.11$. Hence, the rejection region is $Q^2 > 11.11$. Because the observed value of $Q^2 = 26.22 > 11.11$, we reject $H_0$ at $\alpha = 0.05$. Thus, we conclude that the given data do not follow (or are drawn) from the normal pdf.*

## 11.5.2 The Kolmogorov–Smirnov test: (one population)

Let $X_i, i = 1, 2, \ldots, n$ be a random sample of $n$ observations and we shall assume is drawn (it follows) from a probability distribution whose cumulative distribution is specified to be $F_0(x)$. Our objective now is to determine if the actual (correct) cumulative probability is $F(x)$ based on the assumed $F_0(x)$. That is, we wish to test the following hypothesis:

$H_0$: *The true probability distribution that follows the given data,* $F(x)$, *is actually the assumed distribution* $F_0(x)$,

for all x.

*Vs.*

$H_a$: *The actual cumulative distribution,* $F(x)$ *is not* $F_0(x)$, for at least one x.

The Kolmogorov–Smirnov goodness-of-fit test to test the above hypothesis is based on the following test statistic:

$$D = \underset{-\infty < x < \infty}{Max} \{|F_n(x) - F_0(x)|\},$$

where $F_n(x)$ is the sample (empirical) distribution function given by

$$F_n(x) = \frac{\textit{number of } X'\textit{s in the sample} \le x}{n}.$$

Note that for an ordered data $X_{(1)}, \ldots, X_{(n)}$, $F_n(x)$ can be given as

$$F_n(x) = \begin{cases} 0, & x < X_{(1)}, \\ \dfrac{i}{n}, & X_{(i)} \le x < X_{(i+1)}, \\ 1, & x > X_{(n)}. \end{cases}$$

If $F_0(x)$ and $F_n(x)$ are plotted against the $x$-axis, $D$ is the value of the largest vertical distance between $F_0(x)$ and $F_n(x)$. In order to compute $D$, we can use the following. If the $n$ observations are distinct, then define

$$K_i = \max\left\{\left|\frac{i}{n} - F_0(X_{(i)})\right|, \left|\frac{(i-1)}{n} - F_0(X_{(i)})\right|\right\},$$

and

$$D = \max_{i=1,\ldots,n} K_i.$$

If there are tied observations, let $l$ be the number of distinct observations and let $Y_{(1)} < \ldots < Y_{(l)}$ be ordered distinct observations. Then, let

$$K_i^{'} = \max\left\{\left|F_n\left(Y_{(i)}\right) - F_0\left(Y_{(i)}\right)\right|, \left|F_n\left(Y_{(i-1)}\right) - F_0\left(Y_{(i)}\right)\right|\right\},$$

and

$$D = \max_{i=1,\ldots,l} K_i^{'}.$$

**Procedure to calculate D**

To calculate the value of the test statistic D, we follow the following three steps:

1. We calculate the assumed cumulative distribution, $F_0(x)$, based on the given data of observations and the specified population distribution.
2. We proceed to obtain the cumulative distribution of the sample, $F_n(x)$, is the empirical distribution function defined as a step function,

$$F_n(x) = \frac{\#X_i \leq x}{n},$$

the number of observations $X_i \leq x$ divided by $n$.

3. We find the absolute difference

$$|F_0(x) - F_n(x)|.$$

Thus, we have a value of the test statistic D, and if

$$D \leq D_\alpha,$$

we will not reject the hypothesis, $H_0$ at level of significance $\alpha$.

That is, we accept the hypothesis, where $D_\alpha$ is the critical value from the Kolmogorov–Smirnov tables that is based on a given $\alpha$ and $n$. The following example illustrates how we apply this test.

**EXAMPLE 11.5.3** From a large statistics class, we have taken a random sample of 55 students, $n = 55$, and recorded their ages. The resulting data are:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 27 | 25 | 24 | 24 | 22 | 20 | 21 | 22 | 21 | 25 | 24 |
| 26 | 25 | 24 | 23 | 22 | 20 | 21 | 19 | 21 | 25 | 24 |
| 26 | 25 | 22 | 23 | 22 | 22 | 21 | 19 | 21 | 23 | 21 |
| 26 | 24 | 22 | 23 | 22 | 22 | 20 | 19 | 21 | 23 | 21 |
| 26 | 24 | 22 | 23 | 21 | 19 | 20 | 18 | 20 | 20 | 18 |

We believe that these data follow the normal pdf and wish to use the Kolmogorov–Smirnov goodness-of-fit test, given above, to test our belief. That is, test

$H_0$: *The ages of the students follows the normal probability distribution*

*Vs.*

$H_a$: *The ages of students does not follow the normal probability distribution*

**Solution**

*It usually helps to obtain a possible visual indication of the pdf by structuring a histogram of the given data (see Figure 11.1). That is,*

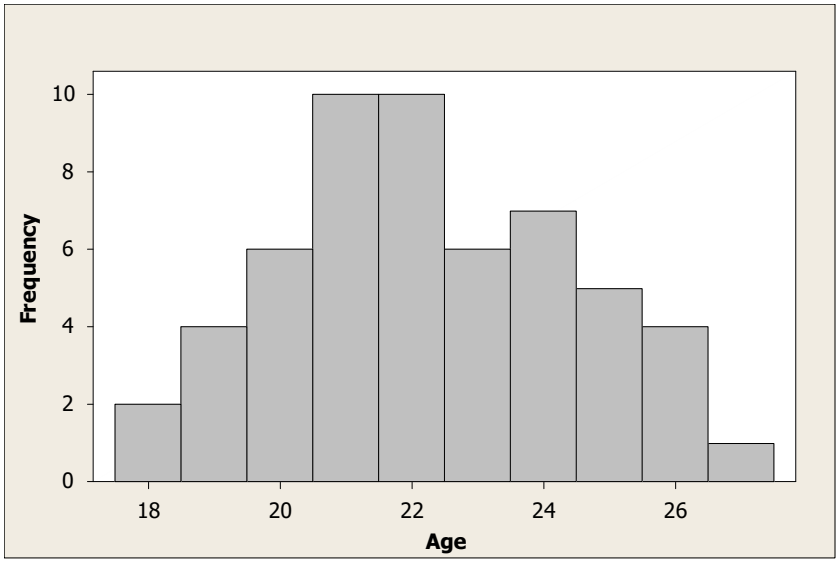



**FIGURE 11.1** Histogram of ages.

*Visually it seems that the normal pdf is a good possibility. We shall now test it statistically.*

*The sample mean is* $\bar{x} = 22$ *and the sample standard deviation is* $s = 2.08$. *The three-step procedure of the subject test to obtain the value of the test statistic D can be easily calculated using the following table and letting D as max of the column,* $|F_0(x) - F_n(x)|$

| Row | Age | $F_0(x)$ | $F_n(x)$ | $\lvert F_0(x) - F_n(x)\rvert$ | D | Critical value |
|---|---|---|---|---|---|---|
| 1 | 18 | 0.028 | 0.018 | 0.010 | 0.127 | 0.183 |
| 2 | 18 | 0.028 | 0.036 | 0.009 | | |
| 3 | 19 | 0.071 | 0.055 | 0.017 | | |
| 4 | 19 | 0.071 | 0.073 | 0.001 | | |
| 5 | 19 | 0.071 | 0.091 | 0.019 | | |
| 6 | 19 | 0.071 | 0.109 | 0.038 | | |
| 7 | 20 | 0.155 | 0.127 | 0.028 | | |
| 8 | 20 | 0.155 | 0.145 | 0.010 | | |
| 9 | 20 | 0.155 | 0.164 | 0.009 | | |
| 10 | 20 | 0.155 | 0.182 | 0.027 | | |
| 11 | 20 | 0.155 | 0.200 | 0.045 | | |
| 12 | 20 | 0.155 | 0.218 | 0.063 | | |
| 13 | 21 | 0.286 | 0.236 | 0.050 | | |
| 14 | 21 | 0.286 | 0.255 | 0.032 | | |
| 15 | 21 | 0.286 | 0.273 | 0.013 | | |
| 16 | 21 | 0.286 | 0.291 | 0.005 | | |
| 17 | 21 | 0.286 | 0.309 | 0.023 | | |
| 18 | 21 | 0.286 | 0.327 | 0.041 | | |
| 19 | 21 | 0.286 | 0.345 | 0.059 | | |
| 20 | 21 | 0.286 | 0.364 | 0.078 | | |
| 21 | 21 | 0.286 | 0.382 | 0.096 | | |
| 22 | 21 | 0.286 | 0.400 | 0.114 | | |
| 23 | 22 | 0.454 | 0.418 | 0.036 | | |
| 24 | 22 | 0.454 | 0.436 | 0.018 | | |
| 25 | 22 | 0.454 | 0.455 | 0.000 | | |
| 26 | 22 | 0.454 | 0.473 | 0.018 | | |

| | | | | |
|---|---|---|---|---|
| 27 | 22 | 0.454 | 0.491 | 0.037 |
| 28 | 22 | 0.454 | 0.509 | 0.055 |
| 29 | 22 | 0.454 | 0.527 | 0.073 |
| 30 | 22 | 0.454 | 0.545 | 0.091 |
| 31 | 22 | 0.454 | 0.564 | 0.109 |
| 32 | 22 | 0.454 | 0.582 | 0.127 |
| 33 | 23 | 0.631 | 0.600 | 0.031 |
| 34 | 23 | 0.631 | 0.618 | 0.013 |
| 35 | 23 | 0.631 | 0.636 | 0.005 |
| 36 | 23 | 0.631 | 0.655 | 0.023 |
| 37 | 23 | 0.631 | 0.673 | 0.041 |
| 38 | 23 | 0.631 | 0.691 | 0.059 |
| 39 | 24 | 0.784 | 0.709 | 0.075 |
| 40 | 24 | 0.784 | 0.727 | 0.057 |
| 41 | 24 | 0.784 | 0.745 | 0.039 |
| 42 | 24 | 0.784 | 0.764 | 0.020 |
| 43 | 24 | 0.784 | 0.782 | 0.002 |
| 44 | 24 | 0.784 | 0.800 | 0.016 |
| 45 | 24 | 0.784 | 0.818 | 0.034 |
| 46 | 25 | 0.892 | 0.836 | 0.055 |
| 47 | 25 | 0.892 | 0.855 | 0.037 |
| 48 | 25 | 0.892 | 0.873 | 0.019 |
| 49 | 25 | 0.892 | 0.891 | 0.001 |
| 50 | 25 | 0.892 | 0.909 | 0.017 |
| 51 | 26 | 0.954 | 0.927 | 0.027 |
| 52 | 26 | 0.954 | 0.945 | 0.009 |
| 53 | 26 | 0.954 | 0.964 | 0.010 |
| 54 | 26 | 0.954 | 0.982 | 0.028 |
| 55 | 27 | 0.984 | 1.000 | 0.016 |

*Since the D-statistic = 0.127 <* $D_{\alpha=0.05} = 0.183$ *(from the K–S table), we fail to reject the null hypothesis at the level of significance* $\alpha = 0.05$. *Thus, the ages of the students in the class indeed follow the normal pdf.*

*Also, we can easily calculate the Kolmogorov–Smirnov test statistics and the* p *value using R code, and the output is given below:*

```
x = c(27,25,24,24,22,20, 21,22,21,25,24.
+ 26,25,24,23,22,20,21,19,21,25,24.
+ 26,25,22,23,22,22,21,19,21,23,21.
+ 26,24,22,23,22,22,20,19,21,23,21.
+ 26,24,22,23,21,19,20,18,20,20,18)

ks.test(x,pnorm, mean(x),sd(x))
```

**Output**

*One-sample Kolmogorov–Smirnov test*

```
data: x
D = 0.1274, P-value = 0.3336
alternative hypothesis: two-sided
```

*Since the* p *value is large, we cannot reject the null hypothesis.*

### 11.5.3 The Anderson–Darling test

The Anderson–Darling goodness-of-fit test is also used to determine if a given set of data is drawn from a population that follows a specific probability distribution. This is a modification of the Kolmogorov–Smirnov (K–S) test and gives more weight to the tails than the K–S test. However, critical values for the Anderson–Darling test make use of particular

distribution resulting in the need for calculating critical values for each distribution. As a result, we will not give critical values for this test, instead we will use the software. There are Anderson−Darling tables available for many popular distributions, such as, normal, lognormal, exponential, Weibull, etc. Let $X_i, i = 1, 2, \ldots, n$ be a random sample of observations and $Y_i, i = 1, 2, \ldots, n$ is the corresponding ordered value according to size. The hypothesis that we wish to test is:

$H_0$: *The given data follow a specific probability distribution*

*Vs.*

$H_a$: *The given data do not follow the specified probability distributio*n.

The Anderson−Darling test statistic for testing the above hypothesis is given by

$A^2 = -n - S$ where $S = \sum_{i=1}^{n} \frac{(2i-1)}{n}[\ln F(Y_i) + \ln(1 - F(Y_{n+1-i})]$, $n$ is the random sample size, $Y_i$ the ordered data, and $F$ the specified probability distribution that we are testing. For a given level of significance $\alpha$, the hypothesis is rejected if the value of the test statistic $A$ is greater than the critical value $A_\alpha$, that is, if

$$A > A_\alpha.$$

Thus, we reject the null hypothesis in favor of the alternative hypothesis; the specified probability distribution does not fit the distribution of the drawn data from the population. The $A_\alpha$ is obtained from the Anderson−Darling tables for a given $\alpha$. The following example illustrates how we apply the subject test.

---

**EXAMPLE 11.5.4**

Use ages of the 55 students given in Example 11.5.3 to illustrate the applicability of the Anderson−Darling goodness-of-fit test.

**Solution**

*The data are given in Example 11.3.3 and we proceed to test our belief that the students' ages follow the normal pdf.*

```
install.packages('nortest')
library(nortest)
 ad.test(x,"pnorm")
```

```
Output
Anderson−Darling normality test
```

```
data: x
A = 0.6456, p-value = 0.08743
```

*Thus, the Anderson−Darling statistic is* A = *0.6456 with a* p *value of 0.08743. Thus, at a 5% level of significance we fail to reject the null hypothesis. The data fit the normal probability distribution with mean 22 and standard deviation 2.*

---

## 11.5.4 Shapiro−Wilk normality test

The Shapiro−Wilk goodness-of-fit test is used to determine if a random sample, $X_i, i = 1, 2, \ldots, n$, is drawn from a normal Gaussian probability distribution with true mean and variance, $\mu$ and $\sigma^2$, respectively. That is, $X \sim N(\mu, \sigma^2)$. Thus, we wish to test the following hypothesis:

$H_0$: *The random sample was drawn from a normal population,* $\mathrm{N}(\mu, \sigma^2)$

*Vs.*

$H_a$: *The random sample does not follow* $\mathrm{N}(\mu, \sigma^2)$.

To test this hypothesis, we use the Shapiro−Wilk test statistic, which is given by

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

where $x_{(i)}$ are the ordered sample values and $a_i$ are constants that are generated by the expression,

$$(a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} m)^{1/2}}$$

with $m = (m_1, m_2, \ldots, m_n)^T$ being the expected values of the ordered statistics that are independent and identically distributed random variables that follow the standard normal, $N(0, 1)$, and $V$ is the covariance matrix of the order statistics.

**EXAMPLE 11.5.5 Proceed to use the Shapiro–Wilk normality test for the data of Example 11.5.3 that we used the Anderson–Darling goodness-of-fit test to see if the ages of the students follow the normal pdf.**

Use $\alpha = 0.05$.

**Solution**
*The R code for the subject test is*
*Shapiro.test(x)*

**Output**
*Shapiro–Wilk normality test*

*Data:* x
W = *0.9683,* p *value =0 .1551*
*Thus, since the* p *value is larger than 0.05, we fail to reject the null hypothesis and the ages of the students indeed follow the normal pdf. This result is the same as that obtained using the Anderson–Darling test.*

## 11.5.5 The P–P plots and Q–Q plots

We commonly use a visual interpretation of graphs (plots) to determine if a given random sample of data follows or is drawn from a well-known probability distribution. These graphs are the probability, P–P plots and the quantile, Q–Q plots.

The *P–P plot* is a graphical tool used to determine how well a given data set fits a specific probability distribution that we are testing. This plot compares the empirical cumulative distribution functions of the given data with that of the assumed true cumulative probability distribution functions. If the plot of these two distributions is approximately linear, it indicates that the assumed true pdf gives a reasonably good fit to the given data that we seek to find its true pdf.

### *11.5.5.1 Steps to construct the P–P plot*

Let $F(x)$ be the cumulative pdf of the random variable, $X$, with a random sample $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ of ordered data values with associated probabilities $\widehat{C}_{(i)} = \frac{i}{n+1}$, the scattered P–P plot is the plot of $\widehat{C}_{(i)}$ versus $C_{(i)} = F\left[X = x_{(i)}\right]$, of the possibly true cumulative pdf that we are testing. The step-by-step procedure that we follow to structure the P–P plot is given below.

**Steps for P–P plot**

**Step 1.** Given a random sample $x_1, x_2, \ldots, x_n$, sort the data in ascending order,

$$\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}, \ldots, \boldsymbol{x}_{(n)}.$$

**Step 2.** Associate with each of the ordered data value $x_{(1)}$ a cumulative probability,

$$\widehat{\boldsymbol{C}}_{(i)} = \frac{\boldsymbol{i}}{\boldsymbol{n}+1}.$$

**Step 3.** Determine the hypothetical probabilities associated with the probability distribution we are testing:

$$\boldsymbol{C}_{(i)} = \boldsymbol{F}\left[\boldsymbol{X} = \boldsymbol{x}_{(i)}\right],$$

$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{P}[\boldsymbol{X} \leq \boldsymbol{x}],$$

where $F(x)$ is the cumulative pdf.

**Step 4.** Construct the scatter plot of $\widehat{\boldsymbol{C}}_{(i)}$ versus $\boldsymbol{C}_{(i)} = \boldsymbol{F}\left[\boldsymbol{X} = \boldsymbol{x}_{(i)}\right]$.

**Step 5.** Interpret the plot; if the overall pattern follows approximately a straight line, then the data follow the assumed probability distribution, and if the overall pattern has curvature or shelves, then the data have skewed behavior and therefore they do not follow the assumed pdf.

The following example illustrates how we obtain and interpret the subject plot.

**EXAMPLE 11.5.6** **Using the data of Example 11.5.3, obtain the P–P plot as in Figure 11.2**

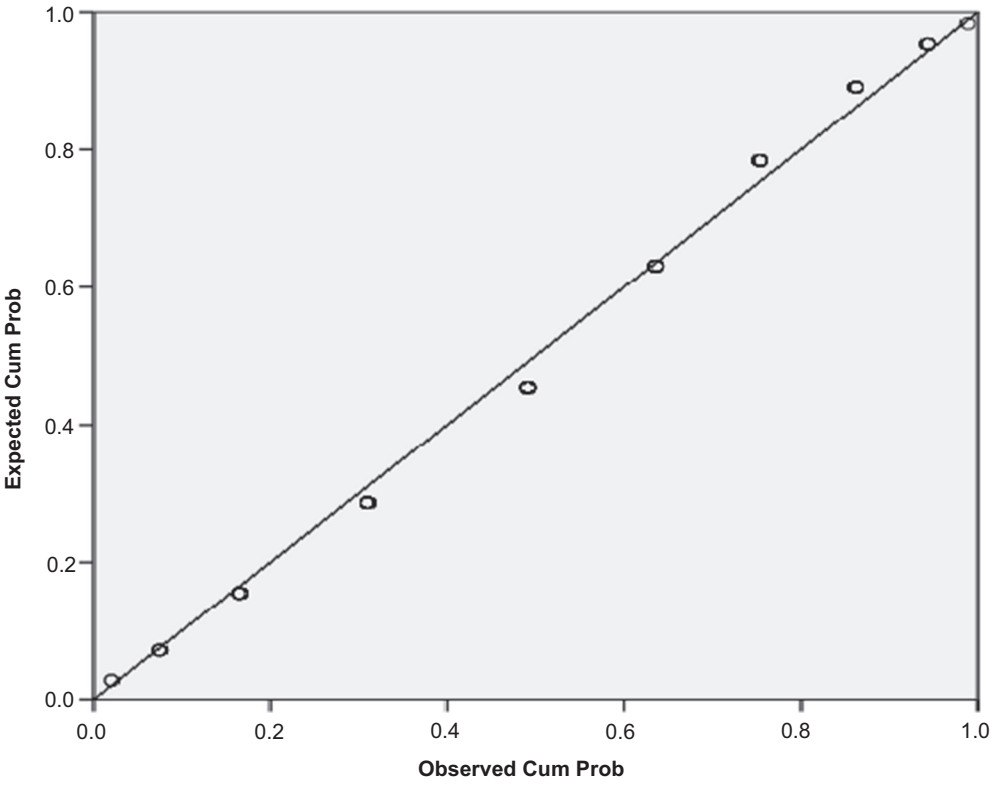



**FIGURE 11.2** P–P plot of the ages.

Thus, the data fall on a straight line and we can conclude that the information of the ages of the students follows the normal pdf, which is consistent with our previous test. Again, the P–P plot is a visual decision and we cannot associate with it a degree of confidence.

The Q–Q plot is another graphical method that is commonly used to obtain a graphical (visual) indication of the true pdf that the given data come from. This method is a graph of the quantiles of the empirical distribution of the given data versus the quantiles of the assumed true pdf that we are testing. If the resulting graph of these two distributions follows a linear pattern, it indicates that the assumed pdf fits the given data reasonably well. A step-by-step procedure of obtaining the Q–Q plot is given below.

**Steps to obtain Q–Q plots**

Let $F(x)$ be the assumed cumulative pdf of the random variable $X$, with a random sample $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ of ordered data values with associated probabilities $\widehat{C}_{(i)} = \frac{i}{n+1}$, the Q–Q plot is the $x_{(i)} = F^{-1}(\widehat{C}_{(i)})$, the inverse function of $F(x)$.

**Step 1.** Given a random sample $x_1, x_2, \ldots, x_n$, sort the data in ascending order:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}.$$

**Step 2.** Associate with each of the order data value $x_{(1)}$ a cumulative probability,

$$\widehat{C}_{(i)} = \frac{i}{n+1}.$$

**Step 3.** Determine the estimated value of the random variable associated with the assumed probability distribution

$$x_{(i)} = F^{-1}(\widehat{C}_{(i)})$$

where $F(x)$ is the cumulative density function.

**Step 4.** Construct the scatter plot of $x_{(i)}$ versus

**Steps to obtain Q–Q plots—cont'd**

$\widehat{x}_{(i)} = F^{-1}\left[\widehat{C}_{(i)}\right]$.

**Step 5.** If the overall pattern follows approximately a straight line, then the data follow the assumed probability distribution. If the overall pattern has curvature or shelves, then the data have skewed behavior and they do not follow the assumed probability distribution.

The following example illustrates how we structure a Q–Q plot.

**EXAMPLE 11.5.7 We shall use the data given in Example 11.5.3, the ages of 55 students to construct the Q–Q plot to verify normality.**

**Solution**

*The results are given in Figure 11.3 (created using Minitab).*

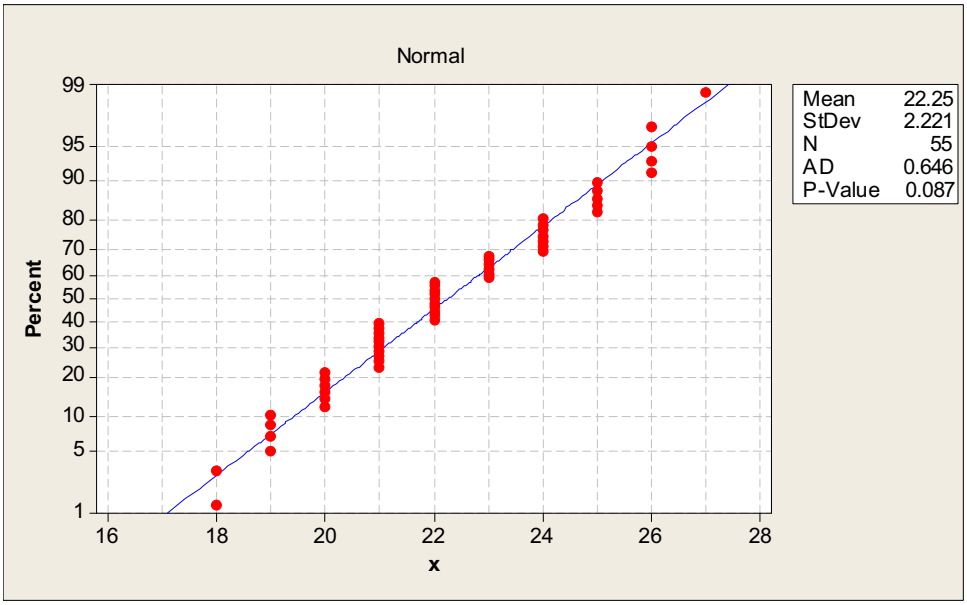



**FIGURE 11.3** Q–Q plot for the ages.

*Note that the plot follows approximately a straight line, which suggests that the data follow the normal pdf, which we have also proven using two other goodness-of-fit tests.*

## Exercises 11.5

**11.5.1** The speeds of vehicles (in mph) passing through a section of Highway 75 are recorded for a random sample of 150 vehicles and are given below. Test the hypothesis using the Anderson–Darling test that the speeds are normally distributed with a mean of 70 mph and a standard deviation of 4. Use $\alpha = 0.01$.

| Range | 40–55 | 56–65 | 66–75 | 76–85 | >85 |
|---|---|---|---|---|---|
| Number | 12 | 14 | 78 | 40 | 6 |

**11.5.2** The temperature in degrees Fahrenheit is recorded for a randomly selected 50 days in the city of Tampa, Florida, in 2018. The data collected are given below.

**City of Tampa**

| Temperature | 46–55 | 56–65 | 66–75 | 76–85 | 86–95 |
|---|---|---|---|---|---|
| Number of days | 4 | 6 | 13 | 23 | 4 |

Using one of the tests introduced in this section, test the hypothesis that the data follow normal pdf with mean 77°F and variance 6. Use $\alpha = 0.05$.

**11.5.3** A sample of 30 electronic circuit components is randomly selected from a production process. The lifetime, in hours, of each component is precisely measured by testing it until it fails. The time in hours that it took the component to fail is given below:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 268.276 | 420.559 | 6.590 | 78.389 | 14.123 | 85.507 | 216.594 | 39.892 | 9.468 | 83.088 |
| 519.682 | 315.754 | 139.046 | 4.522 | 81.480 | 209.099 | 170.128 | 711.794 | 115.778 | 108.640 |
| 226.053 | 443.029 | 35.662 | 115.668 | 5.032 | 111.357 | 331.462 | 184.734 | 79.502 | 611.019 |

Using the Pearson's chi-square goodness-of-fit test, test the hypothesis that the lifetimes of the components follow an exponential probability distribution with a mean of 200 h. Use $\alpha = 0.05$.

**11.5.4** For the data given in Example 11.5.3, test the goodness of fit that the data follow:
**(a)** the gamma pdf.
**(b)** the Weibull pdf.

**11.5.5** Using the data given in Example 11.5.1, construct the P–P plot and interpret the meaning of the graph.

**11.5.6** For the data given in Example 11.5.2, construct the P–P plot and interpret its meaning.

**11.5.7** Using the data given in Example 11.5.1, construct the graph of the Q–Q plot and interpret its meaning.

## 11.6 Chapter summary

In this chapter, we learned different aspects of categorical data analysis, including estimation and hypothesis testing problems. We also looked at goodness-of-fit methods and how we use them to attempt to identify the pdf that characterizes probabilistic behavior of a given set of data. These are the methods: chi-square, Kolmogorov–Smirnov, Anderson–Darling, and Shapiro–Wilk tests.

A list of some of the key definitions introduced in this chapter is given below:

- Categorical data analysis
- Estimation in categorical data
- Hypothesis testing in categorical data
- Test of independence
- Chi-square tests for count data
- Goodness-of-fit test
- Test for independence
- Contingency table
- P–P plot
- Q–Q plot
- Shapiro–Wilk normality test

In this chapter, we have also learned the following important concepts and procedures:

- Pearson's chi-square test procedure
- Kolmogorov–Smirnov test procedure
- Anderson–Darling test procedure
- Shapiro–Wilk test procedure
- P–P plot construction procedure
- Q–Q plot construction procedure

## 11.7 Computer examples

### 11.7.1 R-commands

Since most of the R-codes are already given in the chapter, we will only give the R-code for selecting a random sample from a large data set.

In R, sample() function can be used to take a random sample of size $n$. Suppose we want to take a random sample of size 40 from a data set named *mydata* without replacement.

**R-code**

Mysample < - mydata[sample(1:nrow(mydata), 40, replace = FALSE),]

When multiple distributions fit well with a data set based on the goodness-of-fit tests, then we may select the best-fitted distribution based on maximizing the log-likelihood value. The fitdistr() function in the MASS package in R can be used to calculate maximum likelihood fitting of a univariate distribution. Then the distribution with largest log likelihood can be chosen as best fit. Download package "MASS." Then do the following:

```
library(MASS)
fitdistr(mydata, 't')$loglik
  > fitdistr(mydata, 'normal')$loglik
> fitdistr(mydata, 'logistic')$loglik
  > fitdistr(mydata, 'weibull')$loglik
> fitdistr(mydata, 'gamma')$loglik
  > fitdistr(mydata, 'lognormal')$loglik
> fitdistr(mydata, 'exponential')$loglik
```

Some other distributions such as beta may need specification of additional parameters. We suggest you look at R-help. It should be noted that there are other packages, such as "fitdistrplus" that will provide functions for fitting univariate distributions to different types of data. We will not go into details.

### 11.7.2 Minitab examples

**EXAMPLE 11.8.1 (Contingency Table): Consider the following data with five levels and two factors. Test for dependence of the factors.**

| Factors | Levels | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 39 | 19 | 12 | 28 | 18 |
| 2 | 172 | 61 | 44 | 70 | 37 |

**Solution**

*In **C1** enter the data in column 1 (39 and 172), and continue to **C5.** Then,*

***Stat** > **Tables** > **Chi-Square-Test** … >* in ***Columns containing the table:*** Type ***C1 C2 C3 C4 C5*** > click ***OK***.

We will obtain the following output.

#### *11.7.2.1 Chi-square test*

Expected counts are printed below the observed counts.

| | | C1 | C2 | C3 | C4 | C5 | Total |
|---|---|---|---|---|---|---|---|
| 1 | | 39 | 19 | 12 | 28 | 18 | 116 |
| | | 48.95 | 18.56 | 12.99 | 22.74 | 12.76 | |
| 2 | | 172 | 61 | 44 | 70 | 37 | 384 |
| | | 162.05 | 61.44 | 43.01 | 75.26 | 42.24 | |
| Total | | 211 | 80 | 56 | 98 | 55 | 500 |

Chi-Sq = 2.023 + 0.010 + 0.076 + 1.219 + 2.152 + 0.611 + 0.003 + 0.023 + 0.368 + 0.650 = 11.135

DF = 4, $p$ value = 0.129

# Projects for Chapter 11

## 11A Fitting a distribution to data

A common problem in statistical modeling is fitting a probability distribution to a set of observations (data set) for a given variable. By doing this graphically (like a histogram), we may have some rough idea. If we do goodness-of-fit tests, with say two different distributions, it can happen that both hypotheses may not be rejected. So which one should we choose? This is mainly important in forecasting. Do a short paper on fitting a distribution to data and apply your results to each of the data in Section 11.4 to check if the chosen distributions are best possible. Some references are:

**(1)** *Fitting distributions With R*, http://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf
**(2)** *Fitting distributions to data and why you are probably doing it wrong*, by David Vose, http://www.vosesoftware.com/whitepapers/Fitting%20distributions%20to%20data.pdf.

## 11B Simpson's paradox

Simpson's paradox refers to a phenomenon whereby the association between a pair of variables $(X, Y)$ reverses sign upon conditioning of a third variable, $Z$, regardless of the values taken by $Z$. Confounding factors play a very important role in categorical data, resulting in Simpson's paradox, if we are not careful. As an example, consider data from two hospitals on an emergency surgical procedure:

| | Lived | Died | Survival rate |
|---|---|---|---|
| Hospital 1 | 120 | 180 | 40% |
| Hospital 2 | 60 | 140 | 30% |

From this contingency table, it looks like hospital 1 is significantly better than hospital 2.

**(a)** Use a hypothesis-testing procedure to test if the survival rates are different at the two hospitals. Use $\alpha = 0.05$.
Now, we are given the information that hospital 1 is situated in a wealthy area and, as a result, patients arrive there in relatively good condition. Whereas hospital 2 is in a poor neighborhood, thus, resulting in patients arriving in much worse condition. Now let us see what happens when we break down the above data by patient condition when they reached the hospital.

| | **Good condition** | | | **Bad condition** | | |
|---|---|---|---|---|---|---|
| | Lived | Died | Survival rate | Lived | Died | Survival rate |
| Hospital 1 | 120 | 150 | 44.44% | 0 | 30 | 0% |
| Hospital 2 | 30 | 30 | 50% | 30 | 110 | 21.42% |

Now, hospital 2 is better in both good and bad conditions! This is an example of Simpson's paradox. Here the confounding factor is the patient condition.

**(b)** Find two more real examples for Simpson's paradox.