other since $Z_{R_{12\text{obs}}} = 4.7238 > Z_{1-\alpha/(k(k-1))} = 1.8339$, $Z_{R_{13\text{obs}}} = 2.0016 > Z_{1-\alpha/(k(k-1))} = 1.8339$, and $Z_{R_{23\text{obs}}} = 2.7222 > Z_{1-\alpha/(k(k-1))} = 1.8339$. In this case, the probability that all the statements are correct is $1-\alpha = 0.8$. Since `hwfat` is the accepted standard for measuring body fat, neither of the other two methods is an acceptable substitute for measuring body fat for high school wrestlers.

R Code 10.15 computes the multiple comparisons according to (10.20).

R Code 10.15

```
> alpha <- 0.2
> ZR12 <- abs(Rj[1] - Rj[2])/sqrt(b * k * (k + 1)/6)
> ZR13 <- abs(Rj[1] - Rj[3])/sqrt(b * k * (k + 1)/6)
> ZR23 <- abs(Rj[2] - Rj[3])/sqrt(b * k * (k + 1)/6)
> CV <- qnorm(1 - alpha/(k * (k - 1)))
> ZRij <- c(ZR12, ZR13, ZR23)
> names(ZRij) <- c("ZR12", "ZR13", "ZR23")
> ZRij

    ZR12     ZR13     ZR23
4.723781 2.001602 2.722179

> CV

[1] 1.833915

> which(ZRij > CV)

ZR12 ZR13 ZR23
   1    2    3
```

10.7 Goodness-of-Fit Tests

Many statistical procedures require knowledge of the population from which the sample is taken. For example, using Student's t-distribution for testing a hypothesis or constructing a confidence interval for μ assumes that the parent population is normal. In this section, **goodness-of-fit** (GOF) procedures are presented that will help to identify the distribution of the population from which the sample is drawn. The null hypothesis in a goodness-of-fit test is a statement about the form of the cumulative distribution. When all the parameters in the null hypothesis are specified, the hypothesis is called simple. Recall that in the event the null hypothesis does not completely specify all of the parameters of the distribution, the hypothesis is said to be composite. Goodness-of-fit tests are typically used when the form of the population is in question. In contrast to most of the statistical procedures discussed so far, where the goal has been to reject the null hypothesis, in a GOF test one hopes to retain the null hypothesis. Two general approaches, one designed primarily for discrete distributions (chi-square goodness-of-fit) and one designed primarily for continuous distributions (Kolmogorov-Smirnov), are presented.

10.7.1 The Chi-Square Goodness-of-Fit Test

Given a single random sample of size n from an unknown population F_X, one may wish to test the hypothesis that F_X has some known distribution $F_0(x)$ for all x. For example, using the data frame SOCCER from Example 4.4 on page 256, is it reasonable to assume the number of goals scored during regulation time for the 232 soccer matches has a Poisson distribution with $\lambda = 2.5$? Although the problem was previously analyzed, it will be considered again shortly in the context of the chi-square goodness-of-fit test. The chi-square goodness-of-fit test is based on a normalized statistic that examines the vertical deviations between what is observed and what is expected when H_0 is true in k mutually exclusive categories. At times, such as in surveys of brand preferences, where the categories/groups would be the brand names, the sample will lend itself to being divided into k mutually exclusive categories. Other times, the categories/groupings will be more arbitrary. Before applying the chi-square goodness-of-fit test, the data must be grouped according to some scheme to form k mutually exclusive categories. When the null hypothesis completely specifies the population, the probability that a random observation will fall into each of the chosen or fixed categories can be computed. Once the probabilities for a data point to fall into each of the chosen or fixed categories is computed, multiplying the probabilities by n produces the expected counts for each category under the null distribution. If the null hypothesis is true, the differences between the counts observed in the k categories and the counts expected in the k categories should be small. The test criterion for testing $H_0 : F_X(x) = F_0(x)$ for all x against the alternative $H_1 : F_X(x) \neq F_0(x)$ for some x when the null hypothesis is completely specified is

$$\chi^2_{\text{obs}} = \sum_{i=1}^{k} \frac{(O_k - E_k)^2}{E_k}, \tag{10.21}$$

where χ^2_{obs} is the sum of the squared deviations between what is observed (O_k) and what is expected (E_k) in each of the k categories divided by what is expected in each of the k categories. Large values of χ^2_{obs} occur when the observed data are inconsistent with the null hypothesis and thus lead to rejection of the null hypothesis. The exact distribution of χ^2_{obs} is very complicated; however, for large n, provided all expected categories are at least 5, χ^2_{obs} is distributed approximately χ^2 with $k - 1$ degrees of freedom. When the null hypothesis is composite, that is, not all of the parameters are specified, the degrees of freedom for the random variable χ^2_{obs} are reduced by one for each parameter that must be estimated.

Example 10.12 ▷ ***Soccer Goodness-of-Fit*** ◁ Test the hypothesis that the number of goals scored during regulation time for the 232 soccer matches stored in the data frame SOCCER has a Poisson **cdf** with $\lambda = 2.5$ with the chi-square goodness-of-fit test and an α level of 0.05. Produce a histogram showing the number of observed goals scored during regulation time and superimpose on the histogram the number of goals that are expected to be made when the distribution of goals follows a Poisson distribution with $\lambda = 2.5$.

Solution: Since the number of categories for a Poisson distribution is theoretically infinite, a table is first constructed of the observed number of goals to get an idea of reasonable categories.

```
> xtabs(~goals, data = SOCCER)

goals
 0  1  2  3  4  5  6  7  8
19 49 60 47 32 18  3  3  1
```

Based on the table, a decision is made to create categories for 0, 1, 2, 3, 4, 5, and 6 or more goals. Under the null hypothesis that $F_0(x)$ is a Poisson distribution with $\lambda = 2.5$, the probabilities of scoring 0, 1, 2, 3, 4, 5, and 6 or more goals are computed with R as follows:

```
> PX <- c(dpois(0:5, 2.5), ppois(5, 2.5, lower = FALSE))
> PX

[1] 0.08208500 0.20521250 0.25651562 0.21376302 0.13360189 0.06680094
[7] 0.04202104
```

Since there were a total of $n = 232$ soccer games, the expected number of goals for the six categories is simply $232 \times$ `PX`.

```
> EX <- 232*PX
> OB <- c(as.vector(xtabs(~goals, data = SOCCER)[1:6]),
+         sum(xtabs(~goals, data = SOCCER)[7:9]))
> OB

[1] 19 49 60 47 32 18  7

> ans <- cbind(PX, EX, OB)
> row.names(ans) <- c(" X=0"," X=1"," X=2"," X=3"," X=4"," X=5","X>=6")
> ans

             PX        EX OB
 X=0 0.08208500 19.043720 19
 X=1 0.20521250 47.609299 49
 X=2 0.25651562 59.511624 60
 X=3 0.21376302 49.593020 47
 X=4 0.13360189 30.995638 32
 X=5 0.06680094 15.497819 18
X>=6 0.04202104  9.748881  7
```

Step 1: **Hypotheses** — The null and alternative hypotheses for using the chi-square goodness-of-fit test to test the hypothesis that the number of goals scored during regulation time for the 232 soccer matches stored in the data frame `SOCCER` has a Poisson **cdf** with $\lambda = 2.5$ are

$$H_0 : F_X(x) = F_0(x) \sim Pois(\lambda = 2.5) \text{ for all } x \text{ versus}$$
$$H_1 : F_X(x) \neq F_0(x) \text{ for some } x.$$

Step 2: **Test Statistic** — The test statistic chosen is χ^2_{obs}.

Step 3: **Rejection Region Calculations** — Reject if $\chi^2_{\text{obs}} > \chi^2_{1-\alpha;k-1}$. The χ^2_{obs} is computed with (10.21) in R Code 10.16.

R Code 10.16

```
> chi.obs <- sum((OB - EX)^2/EX)
> chi.obs

[1] 1.39194
```

$$1.3919 = \chi^2_{\text{obs}} \overset{?}{>} \chi^2_{0.95;6} = 12.5916.$$

Step 4: **Statistical Conclusion** — The $\wp$-value is 0.9663.

```
> p.val <- pchisq(chi.obs, 7 - 1, lower = FALSE)
> p.val

[1] 0.9663469
```

I. Since $\chi^2_{\text{obs}} = 1.3919$ is not greater than $\chi^2_{0.95;6} = 12.5916$, fail to reject H_0.

II. Since the $\wp$-value $= 0.9663$ is greater than 0.05, fail to reject H_0.

Fail to reject H_0.

Step 5: **English Conclusion** — There is no evidence to suggest that the true **cdf** does not equal the Poisson distribution with $\lambda = 2.5$ for at least one x.

To perform a goodness-of-fit test with the function `chisq.test()`, one may specify a vector of observed values for the argument `x=`, and a vector of probabilities of the same length as the vector passed to `x=` to the argument `p=`.

```
> chisq.test(x = OB, p = PX)


Chi-squared test for given probabilities

data:  OB
X-squared = 1.3919, df = 6, p-value = 0.9663
```

R Code 10.17 uses base graphics to create a histogram with superimposed expected goals and the result is shown in Figure 10.11 on the next page.

R Code 10.17

```
> hist(SOCCER$goals, breaks = c((-0.5 + 0):(8 + 0.5)), col = "lightblue",
+      xlab = "Goals scored", ylab = "", freq = TRUE, main = "")
> x <- 0:8
> fx <- (dpois(0:8, lambda = 2.5))*232
> lines(x, fx, type = "h")
> lines(x, fx, type = "p", pch = 16)
```

Note that the histogram does not reflect the category ≥ 6, but rather depicts the observed categories of 6, 7, and 8.

Although the chi-square goodness-of-fit test is primarily designed for discrete distributions, it can also be used with a continuous distribution if appropriate categories are defined.

Example 10.13 ▷ ***Goodness-of-Fit for SAT Scores*** ◁ Use the chi-square goodness-of-fit test with $\alpha = 0.05$ to test the hypothesis that the SAT scores stored in the data frame `GRADES` have a normal **cdf**. Use categories $(-\infty, \mu - 2\sigma]$, $(\mu - 2\sigma, \mu - \sigma]$, $(\mu - \sigma, \mu]$, $(\mu, \mu + \sigma]$,

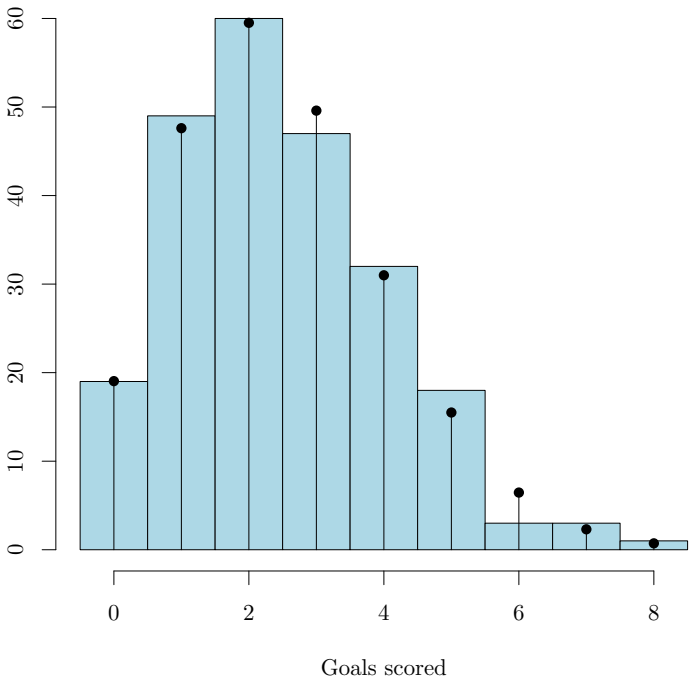

FIGURE 10.11: Histogram of observed goals for SOCCER with a superimposed Poisson distribution with $\lambda = 2.5$ (vertical lines)

$(\mu + \sigma, \mu + 2\sigma]$, and $(\mu + 2\sigma, \infty]$. Produce a histogram using the categories specified and superimpose on the histogram the expected number of SAT scores in each category when $F_0(x) \sim N(\mu = \bar{x}, \sigma = s)$.

Solution: The test follows:

Step 1: **Hypotheses** — The null and alternative hypotheses for using the chi-square goodness-of-fit test to test the hypothesis that the SAT scores stored in the data frame GRADES have a Normal **cdf** are

$$H_0 : F_X(x) = F_0(x) \sim N(\mu = \bar{x},\ \sigma = s) \text{ for all } x \text{ versus}$$
$$H_1 : F_X(x) \neq F_0(x) \text{ for some } x.$$

Step 2: **Test Statistic** — Since the mean and standard deviation are unknown, the first step is to estimate the unknown parameters μ and σ using $\bar{x} = 1134.65$ and $s = 145.6087$.

```
> mu <- mean(GRADES$sat)
> sig <- sd(GRADES$sat)
> c(mu, sig)

[1] 1134.6500  145.6087
```

Because a normal distribution is continuous, it is necessary to create categories that include all the data. The categories $\mu - 3\sigma$ to $\mu - 2\sigma, \ldots, \mu + 2\sigma$ to $\mu + 3\sigma$

are 697.824 to 843.4326, 843.4326 to 989.0413, 989.0413 to 1134.65, 1134.65 to 1280.2587, 1280.2587 to 1425.8674, and 1425.8674 to 1571.476. These particular categories include all of the observed SAT scores; however, the probabilities actually computed for the largest and smallest categories will be all of the area to the right and left, respectively, of $\bar{x} \pm 2s$. This is done so that the total area under the distribution in the null hypothesis is one.

```
> bin <- seq(from = mu - 3*sig, to = mu + 3*sig, by = sig)
> round(bin, 0)                       # vector of bin cut points

[1]  698  843  989 1135 1280 1426 1571

> T1 <- table(cut(GRADES$sat, breaks = bin))
> T1                                  # count of observations in bins


          (698,843]           (843,989]      (989,1.13e+03]
                  4                   27                  65
(1.13e+03,1.28e+03] (1.28e+03,1.43e+03] (1.43e+03,1.57e+03]
                 80                   21                   3

> OB <- as.vector(T1)
> OB                                  # vector of observations

[1]  4 27 65 80 21  3

> PR <- c(pnorm(-2), pnorm(-1:2) - pnorm(-2:1),
+         pnorm(2, lower = FALSE))  # area under curve
> EX <- 200*PR                        # Expected count in bins
> ans <- cbind(PR, EX, OB)            # column bind values in ans
> ans

             PR        EX OB
[1,] 0.02275013  4.550026  4
[2,] 0.13590512 27.181024 27
[3,] 0.34134475 68.268949 65
[4,] 0.34134475 68.268949 80
[5,] 0.13590512 27.181024 21
[6,] 0.02275013  4.550026  3
```

Step 3: **Rejection Region Calculations** — Reject if $\chi^2_{\text{obs}} > \chi^2_{1-\alpha;k-p-1}$.

Now that the expected and observed counts for each of the categories are computed, the χ^2_{obs} value can be computed according to (10.21) and is 4.1737.

```
> chi.obs <- sum((OB - EX)^2/EX)
> chi.obs

[1] 4.173654
```

Step 4: **Statistical Conclusion** — In this problem, two parameters were estimated, and as a consequence, the degrees of freedom are computed as $6-2-1=3$. The $\wp$-value is 0.2433.

```
> p.val <- pchisq(chi.obs, 6 - 2 - 1, lower = FALSE)
> p.val

[1] 0.2433129
```

I. Since $\chi^2_{\text{obs}} = 4.1737$ is not greater than $\chi^2_{0.95;3} = 7.8147$, fail to reject H_0.

II. Since the $\wp$-value $= 0.2433$ is greater than 0.05, fail to reject H_0.

Fail to reject H_0.

Step 5: **English Conclusion** — There is no evidence to suggest that the true **cdf** of SAT scores is not a normal distribution.

If one uses the R function `chisq.test()`, the degrees of freedom and the subsequent $\wp$-value will be incorrect, as illustrated next in R Code 10.18.

R Code 10.18

```
> chisq.test(x = OB, p = PR)  # returns incorrect dof and p-value


Chi-squared test for given probabilities

data:  OB
X-squared = 4.1737, df = 5, p-value = 0.5247
```

Since it is not feasible to produce a histogram that extends from $-\infty$ to ∞, a histogram is created where the categories will simply cover the range of observed values. In this problem, the range of the SAT scores is 720 to 1550. The histogram with categories $(\mu-3\sigma, \mu-2\sigma]$, $(\mu-2\sigma, \mu-\sigma]$, $(\mu-\sigma, \mu]$, $(\mu+, \mu+\sigma]$, $(\mu+\sigma, \mu+2\sigma]$, and $(\mu+2\sigma, \mu+3\sigma]$, superimposed with the expected number of SAT scores for the categories $(-\infty, \mu-2\sigma]$, $(\mu-2\sigma, \mu-\sigma]$, $(\mu-\sigma, \mu]$, $(\mu, \mu+\sigma]$, $(\mu+\sigma, \mu+2\sigma]$, and $(\mu+2\sigma, \infty]$ is computed in R Code 10.19 and depicted in Figure 10.12 on the facing page.

R Code 10.19

```
> hist(GRADES$sat, breaks = bin, col = "lightblue", xlab = "SAT scores",
+     ylab = "", freq = TRUE, main = "")
> x <- bin[2:7] - sig/2
> fx <- PR * 200
> lines(x, fx, type = "h")
> lines(x, fx, type = "p", pch = 16)
```

10.7.2 Kolmogorov-Smirnov Goodness-of-Fit Test

In Section 10.7.1, the chi-square goodness-of-fit test worked by measuring the vertical distance between what was observed in a particular category and what was expected in

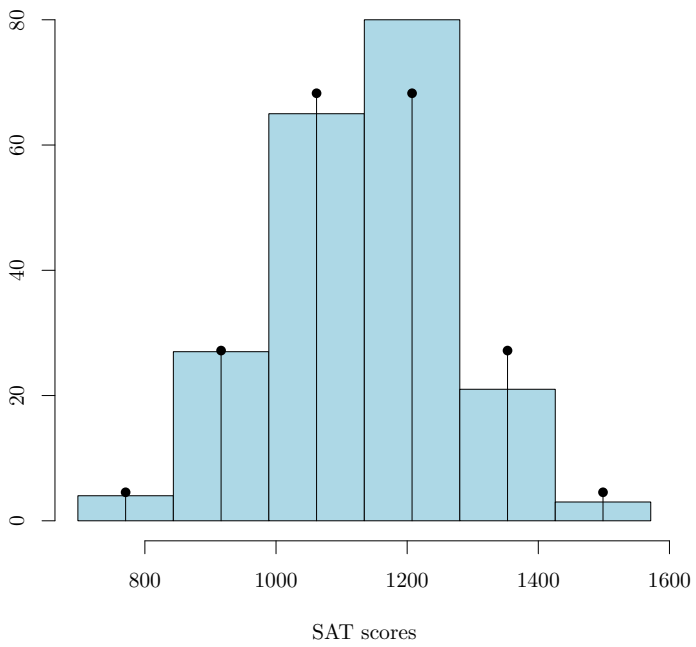

FIGURE 10.12: Histogram of SAT scores in GRADES superimposed with the expected number of SAT scores for the categories $(-\infty, \mu - 2\sigma]$, $(\mu - 2\sigma, \mu - \sigma]$, $(\mu - \sigma, \mu]$, $(\mu, \mu + \sigma]$, $(\mu + \sigma, \mu + 2\sigma]$, and $(\mu + 2\sigma, \infty]$ (vertical lines)

that same category under the null hypothesis for each of the k categories. In contrast to the chi-square goodness-of-fit test, the Kolmogorov-Smirnov goodness-of-fit test uses all n observations and measures vertical deviations between the cumulative distribution function (**cdf**), $F_0(x)$ (where all parameters are specified), and the empirical cumulative distribution function (**ecdf**), $\hat{F}_n(x)$, for all x. For large n, the deviations between $F_0(x)$ and $\hat{F}_n(x)$ should be small for all values of x. The statistic D_n, called the Kolmogorov-Smirnov one-sample statistic, is defined as

$$D_n = \sup_x \left| \hat{F}_n(x) - F_0(x) \right|. \tag{10.22}$$

The statistic D_n does not depend on $F_0(x)$ as long as $F(x)$ is continuous. The derivation of the sampling distribution of D_n is beyond the scope of this text. The curious reader can refer to Gibbons and Chakraborti (2003), page 114, for the derivation of the sampling distribution of D_n. The statistic and sampling distribution of D_n should only be used with simple hypotheses. When the null hypothesis is composite, the critical values for the Kolmogorov-Smirnov test (based on the sampling distribution of D_n) are extremely conservative. The Kolmogorov-Smirnov test can be used to assess normality provided the distribution is completely specified. In a test of normality where the null hypothesis is not completely specified, the statistic D_n can still be used by estimating the unknown parameters of $F_0(x)$ using maximum likelihood $\left(\hat{F}_0(x)\right)$ and substituting $\hat{F}_0(x)$ for $F_0(x)$ in (10.22); however, this further complicates the sampling distribution of D_n. When testing a composite normal hypothesis with unknown μ and σ, the test that uses $D_n = \sup_x \left| \hat{F}_n(x) - \hat{F}_0(x) \right|$ is called Lilliefors's normality test (explained more fully starting on page 643). Lilliefors used simulation to study the sampling distribution of D_n for composite hypotheses and subsequently to publish critical values for using D_n with composite hypotheses. Simulation will be used to show the differences in the distribution of D_n for a simple null hypothesis

versus the distribution of D_n with a composite null hypothesis.

Recall that the **ecdf** was defined in (3.5) to be:

$$\hat{F}_n(t) = \sum_{i=1}^{n} \boldsymbol{I}\{x_i \le t\}/n.$$

An equivalent expression for the **ecdf** is

$$\hat{F}_n(x) = \begin{cases} 0 & x < X_{(1)} \\ \frac{i}{n} & X_{(i)} \le x \le X_{(i+1)} \\ 1 & x > X_{(n)}, \end{cases} \tag{10.23}$$

which will prove useful in computing D_n. When all n observations are distinct, D_n can be computed as

$$D_n = \max_{i=1,\ldots,n} M_i \tag{10.24}$$

where

$$M_i = \max\left\{\left|\hat{F}_n\left(X_{(i)}\right) - F_0\left(X_{(i)}\right)\right|, \left|F_0\left(X_{(i)}\right) - \hat{F}_n\left(X_{(i-1)}\right)\right|\right\}. \tag{10.25}$$

Since $\hat{F}_n\left(X_{(i)}\right) = \frac{i}{n}$ and $\hat{F}_n\left(X_{(i-1)}\right) = \frac{i-1}{n}$, (10.25) can be expressed as

$$M_i = \max\left\{\left|\frac{i}{n} - F_0\left(X_{(i)}\right)\right| = D_i^+, \left|F_0\left(X_{(i)}\right) - \frac{i-1}{n}\right| = D_i^-\right\}. \tag{10.26}$$

Stated formally, the null and alternative hypotheses for the Kolmogorov-Smirnov test for goodness-of-fit are

$$H_0 : F(x) = F_0(x) \text{ for all } x \text{ versus } H_1 : F(x) \ne F_0(x) \text{ for some } x. \tag{10.27}$$

The null hypothesis is rejected when $D_n > D_{n;1-\alpha}$ or when the test's $\wp$-value is less than the largest acceptable α value. Since R will compute the $\wp$-value for the Kolmogorov-Smirnov test, critical values for various n and α are not presented. R uses the function `ks.test(x, y, ...)`, where `x` is a numeric vector of observations and `y` is either a numeric vector of data values or a character string naming a cumulative distribution function.

Example 10.14 ▷ ***Kolmogorov-Smirnov GOF Test*** ◁ Test whether the observations 5, 6, 7, 8, and 9 are from a normal distribution with $\mu = 6.5$ and $\sigma = \sqrt{2}$. That is, the hypothesized distribution is $F_0(x) \sim N\left(6.5, \sqrt{2}\right)$.

Solution: Since $F_0(x) \sim N\left(6.5, \sqrt{2}\right)$, it follows that

$$F_0\left(X_{(i)}\right) = P\left(Y \le X_{(i)}\right) = P\left(\frac{Y - 6.5}{\sqrt{2}} \le \frac{X_{(i)} - 6.5}{\sqrt{2}}\right) = P\left(Z \le \frac{X_{(i)} - 6.5}{\sqrt{2}}\right).$$

To compute $F_0\left(X_{(i)}\right)$ with R, key in

```
> x <- 5:9
> mu <- 6.5
> sig <- sqrt(2)
> x <- sort(x)
> n <- length(x)
> FoX <- pnorm(x, mean = mu, sd = sig)
> FoX

[1] 0.1444222 0.3618368 0.6381632 0.8555778 0.9614501
```

The quantities $\hat{F}_n(X_{(i)}) = \frac{i}{n}$, $\hat{F}_n(X_{(i-1)}) = \frac{i-1}{n}$, D_i^+, D_i^-, and M_i are computed and stored in the R objects `FnX`, `Fn1X`, `Dp`, `Dm`, and `Mi`, respectively. The Kolmogorov-Smirnov statistic $D_n = \max_{i=1,\ldots,n} M_i$ is 0.25558. The values from the R code are shown in Table 10.20.

```
> FnX <- seq(1:n)/n
> Fn1X <- (seq(1:n) - 1)/n
> DP <- (FnX - FoX)
> DM <- FoX - Fn1X
> Dp <- abs(DP)
> Dm <- abs(DM)
> EXP <- cbind(x, FnX, Fn1X, FoX, Dp, Dm)
> Mi <- apply(EXP[, c(5, 6)], 1, max)
> TOT <- cbind(EXP, Mi)
> TOT

     x FnX Fn1X       FoX         Dp        Dm        Mi
[1,] 5 0.2  0.0 0.1444222 0.05557782 0.1444222 0.1444222
[2,] 6 0.4  0.2 0.3618368 0.03816320 0.1618368 0.1618368
[3,] 7 0.6  0.4 0.6381632 0.03816320 0.2381632 0.2381632
[4,] 8 0.8  0.6 0.8555778 0.05557782 0.2555778 0.2555778
[5,] 9 1.0  0.8 0.9614501 0.03854994 0.1614501 0.1614501

> Dn <- max(Mi)
> Dn

[1] 0.2555778
```

Table 10.20: Calculating D_n

i	$X_{(i)}$	$\frac{i}{n} - F_0\left(X_{(i)}\right)$	$F_0\left(X_{(i)}\right) - \frac{i-1}{n}$	D^+	D^-	M_i
1	5	$\frac{1}{5} - 0.14442$	$0.14442 - 0$	0.055578	0.14442	0.14442
2	6	$\frac{2}{5} - 0.36184$	$0.36184 - \frac{1}{5}$	0.038163	0.16184	0.16184
3	7	$\frac{3}{5} - 0.63816$	$0.63816 - \frac{2}{5}$	0.038163	0.23816	0.23816
4	8	$\frac{4}{5} - 0.85558$	$0.85558 - \frac{3}{5}$	0.055578	0.25558	0.25558
5	9	$\frac{5}{5} - 0.96145$	$0.96145 - \frac{4}{5}$	0.038550	0.16145	0.16145
					$D_n =$	0.25558

The computation of the Komolgorov-Smirnov statistic D_n and its $\wp$-value are shown in R Code 10.20.

R Code 10.20

```
> ks.test(x, y = "pnorm", mean = mu, sd = sig)


One-sample Kolmogorov-Smirnov test
```

```
data:  x
D = 0.25558, p-value = 0.8269
alternative hypothesis: two-sided
```

The Komolgorov-Smirnov statistic is labeled `D` in the output produced by `ks.test()`. The value $D_n = 0.2556$ with a corresponding $\wp$-value of 0.8269 provides no evidence to reject the null hypothesis that $F_0(x) \sim N\left(6.5, \sqrt{2}\right)$. Figure 10.13 provides a graphical illustration of the vertical deviations used to compute the statistic D_n for this problem.

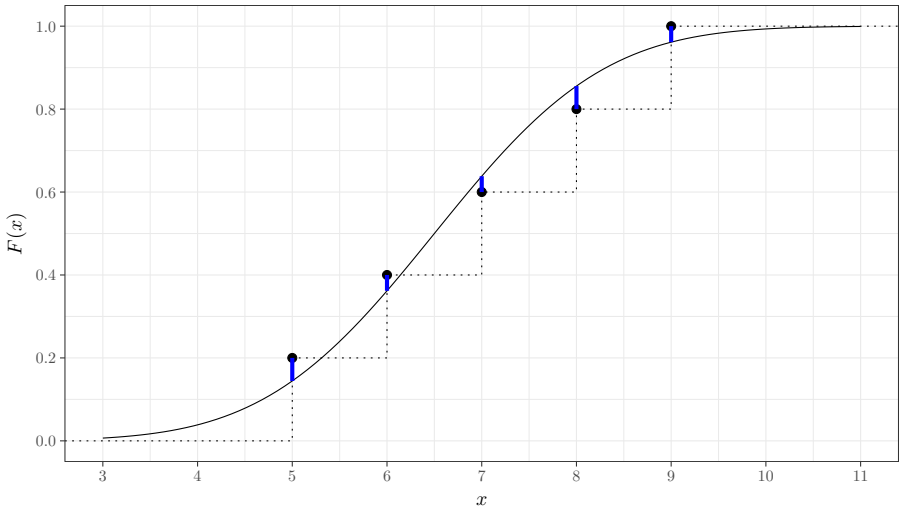

FIGURE 10.13: Graphical illustration of the vertical deviations used to compute the statistic D_n for Example 10.14 on page 640. The solid "S"-shaped line is the hypothesized distribution $F_0(x) \sim N\left(6.5, \sqrt{2}\right)$. The vertical solid segments between the solid circles and $F_0(x)$ represents the D^+ values. The vertical dotted distance represents the D^- values and the dotted stair shaped values represent the **ecdf**.

In Example 10.14 on page 640, the statistic $D_n = 0.2556$ returned a $\wp$-value of 0.8269. To visualize the sampling distribution of D_n and to find simulated critical values, one can use R Code 10.21.

R Code 10.21

```
> ksdist <- function (n = 10, sims = 10000, alpha = 0.05){
+     Dn <- replicate(sims, ks.test(rnorm(n), pnorm)$statistic)
+     cv <- quantile(Dn, 1 - alpha)
+     plot(density(Dn), col = "blue", lwd = 2, main = "",
+     xlab = paste("Simulated critical value =", round(cv, 3),
+     "for n =", n, "when the alpha value =", alpha))
+     title(
+     main = list(expression(paste("Simulated Sampling Distribution of ",
+     D[n]))))
+ }
```

The graph from running `ksdist(n = 5, sims = 10000, alpha = 0.05)` when using a seed of 13 is shown in Figure 10.14. This simulation indicates a value of 0.567 or greater would be required to reject the null hypothesis in Example 10.14 on page 640 at the $\alpha = 0.05$ level. The simulated $\wp$-value for the value $D_n = 0.2556$ in Figure 10.14 is 0.8292, very close to the 0.8269 reported from using `ks.test()`.

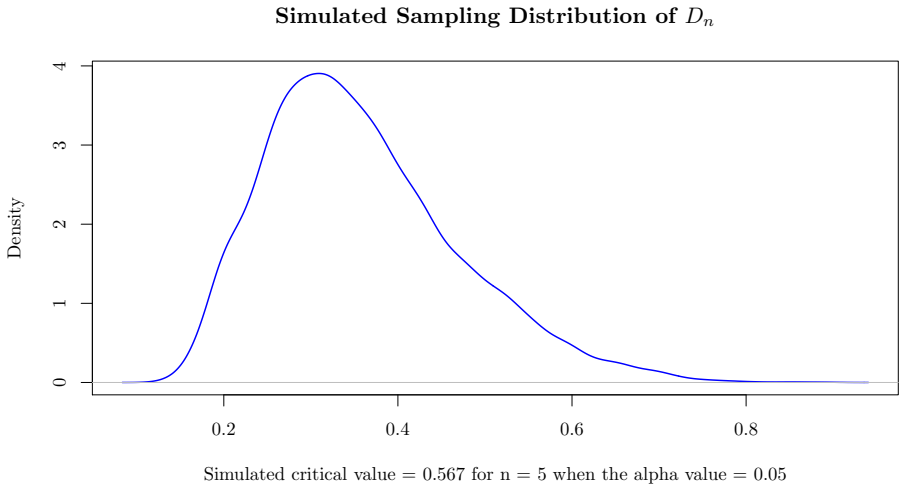

FIGURE 10.14: Graphical illustration of `ksdist(n = 5, sims = 10000, alpha = 0.05)`

Lilliefors's Test of Normality

Expanding on the simulation for the sampling distribution for D_n, consider what happens when the null hypothesis changes from simple to composite using the code for the function `ksLdist()`. Note that the D_n values stored in `D_n[i]` are for a simple null hypothesis of normality while the D_n values stored in `DnL[i]` are for a composite hypothesis of normality. The critical values reported by Lilliefors (1967) were based on simulations using 1000 or more samples using logic similar to the R Code 10.22 used to create the function `ksLdist()`.

R Code 10.22

```
> ksLdist <- function (n = 10, sims = 1000, alpha = 0.05){
+     Dn <- c()
+     DnL <- c()
+     for (i in 1:sims) {
+         x <- rnorm(n)
+         mu <- mean(x)
+         sig <- sd(x)
+         Dn[i] <- ks.test(x, pnorm)$statistic
+         DnL[i] <- ks.test(x, pnorm, mean = mu, sd = sig)$statistic
+     }
+     ys <- range(density(DnL)$y)
+     xs <- range(density(Dn)$x)
+     cv <- quantile(Dn, 1 - alpha)
```

```
+     cvp <- quantile(DnL, 1 - alpha)
+     plot(density(Dn, bw = 0.02), col="blue", lwd=2, ylim=ys, xlim=xs,
+         main = "", , xlab="", sub = paste("Simulated critical value =",
+         round(cv, 3), "(simple hypothesis) and ", round(cvp, 3),
+           "(composite hypothesis)\n for n =", n,"when the alpha value =",
+           alpha))
+     title(
+     main = list(expression(paste("Simulated Sampling Distribution of ",
+     D[n]))))
+     lines(density(DnL, bw = 0.02), col = "red", lwd = 2, lty = 2)
+     legend(mean(xs), max(ys), legend = c("Simple Hypothesis",
+         "Composite Hypothesis"), col = c("blue", "red"), xjust = 0,
+         text.col = c("black", "black"), lty = c(1, 2), bg = "gray95",
+         cex = 1, lwd = 2)
+     box()
+     abline(h = 0)
+ }
```

The function `ksLdist()` allows the user to choose the number of samples with the argument `sims=` and easily to verify the results given by Lilliefors (1967). Dallal and Wilkinson (1986) duplicated the work by Lilliefors (1967) using much larger samples as well as deriving an analytic approximation for the upper tail $\wp$-values for $D_n = \sup_x \left|\hat{F}_n(x) - \hat{F}_0(x)\right|$. For $\wp$-values less than 0.100 and sample sizes ranging from 5 to 100, the Dallal-Wilkinson approximation is

$$\widehat{\wp\text{-value}} = \exp(-7.01256 \cdot D_n^2 \cdot (n + 2.78019) + 2.99587 \cdot D_n \cdot \sqrt{n + 2.78019} - 0.122119 + 0.974598/\sqrt{n} + 1.67997/n) \quad (10.28)$$

The estimated densities from running `ksLdist(sims = 10000, n =10)` with a seed of 13 are shown in Figure 10.15 on the facing page, which highlights how much less variability is present in the sampling distribution of D_n when the null hypothesis is composite. To test a composite hypothesis of normality correctly, one should use the R function `lillie.test()` available in the R package `nortest`. That is, one should not use the R function `ks.test()`.

Example 10.15 ▷ ***Long-Distance Phone Calls*** ◁ Calculate the $\wp$-value and state the English conclusion for testing whether the times spent on long-distance phone calls (`call.time`) in the data frame `PHONE` have a normal distribution using the R function `lillie.test` from the `nortest` package Verify the reported $\wp$-value using (10.28).

Solution: Note that the function `nortest()` labels the statistic D_n with a `D`. The value `nortest()` computes for D_n is 0.191 with a $\wp$-value of 0.0291.

R Code 10.23

```
> library(nortest)
> lillie.test(PHONE$call.time)


Lilliefors (Kolmogorov-Smirnov) normality test

data:  PHONE$call.time
D = 0.19102, p-value = 0.0291
```

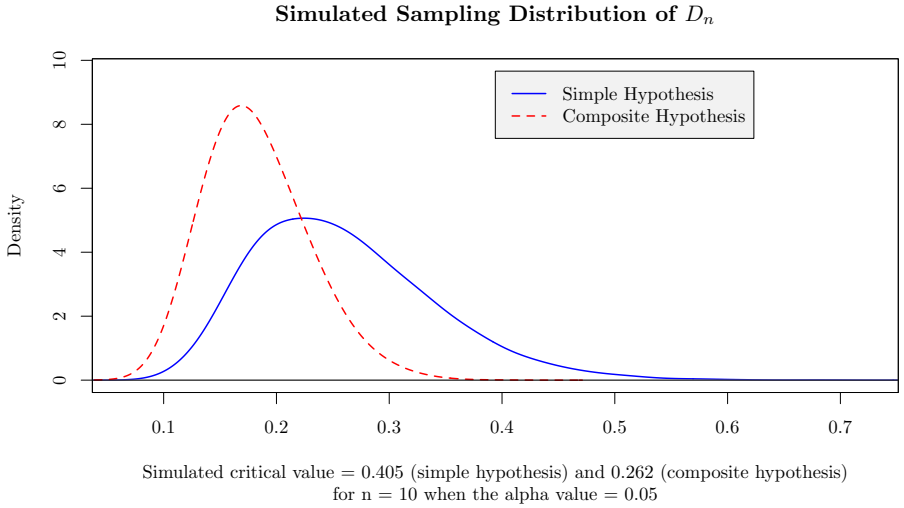

FIGURE 10.15: Estimated densities for simple and composite hypotheses from running `ksLdist(sims = 10000, n = 10)`

To compute the $\wp$-value using (10.28), a small function `DWA()` is written in R Code 10.24. Running the function `DWA()` with the arguments `Dn = 0.191` and `n = 23` returns an estimated $\wp$-value of 0.0291.

R Code 10.24

```
> DWA <- function(Dn = 0.3, n = 10) {
+     p.value <- exp(-7.01256 * Dn^2 * (n + 2.78019) + 2.99587 *
+         Dn * (n + 2.78019)^0.5 - 0.122119 + 0.974598/n^0.5 +
+         1.67997/n)
+     names(p.value) <- NULL
+     round(p.value, 4)
+ }
> DWA(Dn = 0.191, n = 23)

[1] 0.0291
```

With a $\wp$-value of 0.0291, the null hypothesis is rejected. There is evidence that phone call length is not normally distributed. ■

10.7.3 Shapiro-Wilk Normality Test

The Shapiro-Wilk test is appropriate for testing normality. More specifically, the test allows for a composite hypothesis of normality. That is, the parameters of the normal distribution do not need to be specified in the null hypothesis of the test (as they must be for the Lilliefors test). Although the test is known to be conservative, it is useful for testing normality with small samples. The test statistic measures how closely the empirical quantiles of the sample follow the corresponding theoretical quantiles of a normal distribution. This means that small values of the test statistic lead to the rejection of the null hypothesis (that the distribution is normal).

To calculate the test statistic for a random sample of size n, $x_1, x_2, \ldots, x_n$, the sample

must be sorted: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The Shapiro-Wilk test statistic takes the form

$$W = \frac{b^2}{nS_u^2}, \tag{10.29}$$

where S_u^2 is the uncorrected sample variance, $b = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} a_{n-i+1}(x_{(n-i+1)} - x_{(i)})$, and $\lfloor \frac{n}{2} \rfloor$ is the integer part of $\frac{n}{2}$. The coefficients a_{n-i+1} that are calculated automatically by the function `shapiro.test()` are tabulated in Table 6 of Shapiro and Wilk (1965).

The critical region of the test is given by

$$\mathbb{P}(W \leq K | H_0) = \alpha,$$

where α is the significance level. The critical values K can be found in Shapiro and Wilk (1965, Table 5), but they are not displayed in the output for `shapiro.test()`. The vector of weights $\mathbf{a}' = (a_1, \ldots, a_n)$, where $a_i = -a_{n-i+1}$, is calculated as

$$\mathbf{a} = \frac{\mathbf{w}'\mathbf{V}^{-1}}{\mathbf{w}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{w}}, \tag{10.30}$$

where the elements of the vector $\mathbf{w}$ are $w_i = E\left[x_{(i)}\right]$ and $\mathbf{V}$ is the covariance matrix of the order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

Example 10.16 ▷ ***Shapiro-Wilk Normality Test*** ◁ Use the Shapiro-Wilk test with the random sample $\{47, 50, 57, 54, 52, 54, 53, 65, 62, 67, 69, 74, 51, 57, 57, 59\}$ to test for normality using $\alpha = 0.05$.

Solution: First, order the data:

$$47 \leq 50 \leq 51 \leq 52 \leq 53 \leq 54 = 54 \leq 57 = 57 = 57 \leq 59 \leq 62 \leq 65 \leq 67 \leq 69 \leq 74.$$

Next, calculate the differences $x_{(n-i+1)} - x_{(i)}$ for $i = 1, 2, \ldots, \lfloor \frac{n}{2} \rfloor = 8$:

$x_{(16)} - x_{(1)} = 74 - 47 = 27$	$x_{(13)} - x_{(4)} = 65 - 52 = 13$	$x_{(10)} - x_{(7)} = 57 - 54 = 3$
$x_{(15)} - x_{(2)} = 69 - 50 = 19$	$x_{(12)} - x_{(5)} = 62 - 53 = 9$	$x_{(9)} - x_{(8)} = 57 - 57 = 0$
$x_{(14)} - x_{(3)} = 67 - 51 = 16$	$x_{(11)} - x_{(6)} = 59 - 54 = 5$	

Looking at Table 6 from Shapiro and Wilk (1965) ($n = 16$ and $i = 1, \ldots, 8$), one obtains

$a_{16} = 0.5056$	$a_{14} = 0.2521$	$a_{12} = 0.1447$	$a_{10} = 0.0593$
$a_{15} = 0.3290$	$a_{13} = 0.1939$	$a_{11} = 0.1005$	$a_9 = 0.0196$

which means $b = \sum_{i=1}^{8} a_{n-i+1}(x_{(n-i+1)} - x_{(i)}) = 28.4392$ and $nS_u^2 = 854$.
The Shapiro-Wilk test statistic value is then

$$W = \frac{b^2}{nS_u^2} = \frac{808.7881}{854} = 0.9471.$$

The critical value K with $\alpha = 0.05$ and $n = 16$ is 0.887. As $W_{obs} = 0.9471 > 0.887$, one fails to reject the null hypothesis of normality.

```
> x <- c(47, 50, 57, 54, 52, 54, 53, 65, 62, 67, 69, 74, 51, 57, 57, 59)
> shapiro.test(x)


Shapiro-Wilk normality test

data:  x
W = 0.94705, p-value = 0.4445
```

10.8 Categorical Data Analysis

This section provides an overview of two common scenarios where categorical data are generated and explains how each scenario is analyzed. The basic 2×2 contingency table with fixed row totals was introduced in Section 9.9.3, Testing Equality of Proportions with Fisher's exact test. The 2×2 contingency table can be generalized for I rows and J columns and is referred to as an $I \times J$ contingency table. The sampling scheme employed to acquire the information in the table will determine the type of hypothesis that can be tested. Consider the following two scenarios:

SCENARIO ONE: Is there an association between gender and a person's happiness? To investigate whether happiness depends on gender, one might use information collected from the General Social Survey (GSS) (http://sda.berkeley.edu/GSS). In each survey, the GSS asks, "Taken all together, how would you say things are these days — would you say that you are very happy, pretty happy, or not too happy?" Respondents to each survey are coded as either male or female. The information in Table 10.21 shows how a subset of respondents (26-year-olds) were classified with respect to the variables HAPPY and SEX.

Table 10.21: Twenty-six-year-olds' happiness

	HAPPY		
SEX	Very happy	Pretty happy	Not too happy
Male	110	277	50
Female	163	302	63

SCENARIO TWO: In a double blind randomized drug trial (neither the patient nor the physician evaluating the patient knows the treatment, drug or placebo, the patient receives), 400 male patients with mild dementia were randomly divided into two groups of 200. One group was given a placebo over three months while the second group received an experimental drug for three months. At the end of the three months, the physicians (all psychiatrists) classified the 400 patients into one of three categories: improved, no change, or worse. The information in Table 10.22 shows how the pschiatrists classified the patients. Are the proportions in the three status categories the same for the two treatments?

Table 10.22: Mild dementia treatment results

	Status		
Treatment	Improve	No Change	Worse
Drug	67	76	57
Placebo	48	73	79

The two scenarios illustrate two different sampling schemes that both result in $I \times J$ contingency tables. In the first scenario, there is a single population (Americans) and individuals are sampled from this single population and classified into one of the IJ cells of the $I \times J$ contingency table based on the $I = 2$ SEX categories and the $J = 3$ HAPPY categories. The format of an $I \times J$ contingency table when sampling from a single population is shown in Table 10.23. The number of observations from the i^{th} row classified into the j^{th} column is denoted by n_{ij}. It follows that the number of observations in the j^{th} column $(1 \leq j \leq J)$ is $n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{Ij}$, while the number of observations in the i^{th} row $(1 \leq i \leq I)$ is $n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{iJ}$.

The true population proportion of individuals in cell (i, j) will be denoted π_{ij}. Under the assumption of independence between row and column variables (SEX and HAPPY in this example), $\pi_{ij} = \pi_{i\bullet} \times \pi_{\bullet j}$, where $\pi_{i\bullet} = \sum_{j=1}^{J} \pi_{ij}$ and $\pi_{\bullet j} = \sum_{i=1}^{I} \pi_{ij}$. That is, $\pi_{i\bullet}$ is the proportion of observations in the population classified in category i of the row variable and $\pi_{\bullet j}$ is the proportion of observations in the population classified in category j of the column variable. Since $\pi_{i\bullet}$ and $\pi_{\bullet j}$ are marginal population proportions, it follows that $\hat{\pi}_{i\bullet} = p_{i\bullet} = \frac{n_{i\bullet}}{n}$ and $\hat{\pi}_{\bullet j} = p_{\bullet j} = \frac{n_{\bullet j}}{n}$, where n is the sample size. Under the assumption of independence the expected count for cell (i, j) is $\mu_{ij} = n\pi_{ij} = n\pi_{i\bullet}\pi_{\bullet j}$ and $\hat{\mu}_{ij} = n\hat{\pi}_{ij} = n\hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j} = n\frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet}n_{\bullet j}}{n}$.

Table 10.23: Contingency table when sampling from a single population

	Col 1	Col 2	$\cdots$	Col J	*Totals*
Row 1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\bullet}$
Row 2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Row I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I\bullet}$
Totals	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet J}$	n

In the second scenario, there are two distinct populations from which samples are taken. The first population is the group of all patients receiving the experimental drug while the second population is the group of all patients receiving a placebo. In this scenario, there are $I = 2$ separate populations and $J = 3$ categories for the $I = 2$ populations. Individuals sampled from the $I = 2$ distinct populations are classified into one of the $J = 3$ status categories. This scenario has fixed row totals whereas the first scenario does not. In the first scenario, only the total sample size, n, is fixed. That is, neither the row nor the column totals are fixed. This is in contrast to scenario two, where the number of patients in each treatment group (row) was fixed. The notation used for an $I \times J$ contingency table when I samples from I distinct populations differs slightly from the notation used in Table 10.23

on the facing page with a contingency table from a single sample.

Since the sample sizes of the I distinct populations are denoted $n_{i\bullet}$, the total for all I samples is denoted by $n_{\bullet\bullet}$ rather than the notation n used for a single sample in Table 10.23 on the preceding page. Table 10.24 shows the general form and notation used for an $I \times J$ contingency table when sampling from I distinct populations. Each observation in each sample is classified into one of J categories. If $n_{i\bullet}$ denotes the number of observations in the i^{th} sample ($1 \le i \le I$) and n_{ij} denotes the number of observations from the i^{th} sample classified into the j^{th} category ($1 \le j \le J$), it follows that the number of observations in the j^{th} column is $n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{Ij}$, while the number of observations in the i^{th} row is $n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{iJ}$.

Table 10.24: General form and notation used for an $I \times J$ contingency table when sampling from I distinct populations

	Category 1	Category 2	$\cdots$	Category J	Totals
Population 1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\bullet}$
Population 2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Population I	n_{I1}	n_{I2}	$\ldots$	n_{IJ}	$n_{I\bullet}$
Totals	$n_{\bullet 1}$	$n_{\bullet 2}$	$\ldots$	$n_{\bullet J}$	$n_{\bullet\bullet}$

10.8.1 Test of Independence

Scenario one asks if there is an association between gender and a person's happiness. In Section 3.3.6 on page 211, two events, A and B, were defined as independent when $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$ or, equivalently, when $\mathbb{P}(A|B) = \mathbb{P}(A)$. If, instead of having a random sample from a single population, an $I \times J$ contingency table consisted of entries from the population, association could be mathematically verified by showing that $\mathbb{P}(n_{ij}) \neq \mathbb{P}(n_{i\bullet}) \times \mathbb{P}(n_{\bullet j})$ for some i and j. If by chance $\mathbb{P}(n_{ij}) = \mathbb{P}(n_{i\bullet}) \times \mathbb{P}(n_{\bullet j})$ for all i and j, then one would conclude there is no association between gender and a person's happiness. That is, the variables gender and happiness would be considered mathematically independent. Since the entire population is not given but rather a sample from a population, the values in the $I \times J$ contingency table can be expected to change from sample to sample. The question is, "By how much can the variables deviate from the mathematical definition of independence and still be considered statistically independent?"

The null and alternative hypotheses to test for independence between row and column variables is written $H_0 : \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$ versus $H_1 : \pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j}$. The test statistic is

$$\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{10.31}$$

It compares the observed frequencies in the table with the expected frequencies when H_0 is true. Under the assumption of independence, and when the observations in the cells are sufficiently large (usually greater than 5), $\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \dot{\sim} \chi^2_{(I-1)(J-1)}$, where $\hat{\mu}_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n} = E_{ij}$ and $n_{ij} = O_{ij}$. The null hypothesis of independence is rejected

when $\chi^2_{\text{obs}} > \chi^2_{1-\alpha;(I-1)(J-1)}$.

The chi-squared approximation is generally satisfactory if the E_{ij}s ($\hat{\mu}_{ij}$s) in the test statistic are not too small. Various rules of thumb exist for what might be considered too small. A very conservative rule is to require all E_{ij}s to be 5 or more. This can be accomplished by combining cells with small E_{ij}s and reducing the overall degrees of freedom. At times, it may be permissible to let the E_{ij} of a cell be as low as 0.5.

Test for SCENARIO ONE:

Step 1: **Hypotheses** — $H_0 : \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$ (Row and column variables are independent.) versus $H_1 : \pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j}$ for at least one i, j (Row and column variables are dependent.)

Step 2: **Test Statistic** — The test statistic is

$$\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \overset{\cdot}{\sim} \chi^2_{(I-1)(J-1)} = \chi^2_{(2-1)(3-1)} = \chi^2_2$$

under the assumption of independence. The χ^2_{obs} value is 4.3215.

Step 3: **Rejection Region Calculations** — The rejection region is

$$\chi^2_{\text{obs}} > \chi^2_{1-\alpha;(I-1)(J-1)} = \chi^2_{0.95;2} = 5.9915.$$

Before the statistic $\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ can be computed, the expected counts for each of the ij cells must be calculated. Note that $O_{ij} = n_{ij}$ and $E_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n}$.

```
> HA <- c(110, 277, 50, 163, 302, 63)
> HAT <- matrix(data = HA, nrow = 2, byrow = TRUE)
> dimnames(HAT) <- list(SEX = c("Male", "Female"),
+     Category = c("Very Happy", "Pretty Happy", "Not To Happy"))
> HAT

        Category
SEX      Very Happy Pretty Happy Not To Happy
  Male          110          277           50
  Female        163          302           63

> E <- outer(rowSums(HAT), colSums(HAT), "*")/sum(HAT)
> E

       Very Happy Pretty Happy Not To Happy
Male      123.628        262.2     51.17202
Female    149.372        316.8     61.82798

> # OR
> chisq.test(HAT)$expected

        Category
SEX      Very Happy Pretty Happy Not To Happy
  Male      123.628        262.2     51.17202
  Female    149.372        316.8     61.82798
```

$$\chi^2_{\text{obs}} = \frac{(110 - 123.6280)^2}{123.6280} + \frac{(277 - 262.2)^2}{262.2} + \cdots + \frac{(63 - 61.828)^2}{61.828} = 4.3215.$$

The value of the test statistic is $\chi^2_{obs} = 4.3215$. This can be done with code by entering

```
> chi.obs <- sum((HAT - E)^2/E)
> chi.obs

[1] 4.321482
```

$4.3215 = \chi^2_{\text{obs}} \overset{?}{>} \chi^2_{0.95,2} = 5.9915.$

Step 4: **Statistical Conclusion** — The $\wp$-value is 0.1152.

```
> p.val <- pchisq(chi.obs, 2, lower = FALSE)
> p.val

[1] 0.1152397
```

I. From the rejection region, since $\chi^2_{\text{obs}} = 4.3215 < \chi^2_{0.95;2} = 5.9915$, fail to reject the null hypothesis of independence.

II. Since the $\wp$-value $= 0.1152$ is greater than 0.05, fail to reject the null hypothesis of independence.

Fail to reject $\mathbf{H_0}$.

Step 5: **English Conclusion** — There is not sufficient evidence to suggest the variables gender and happiness are statistically dependent.

The function `chisq.test()` can also be used to test the null hypothesis of independence.

```
> chisq.test(HAT)


Pearson's Chi-squared test

data:  HAT
X-squared = 4.3215, df = 2, p-value = 0.1152
```

10.8.2 Test of Homogeneity

The question of interest in scenario two is whether the proportions in each of the $j = 3$ categories for the $i = 2$ populations are equivalent. Specifically, is $\pi_{1j} = \pi_{2j}$ for all j? This question is answered with a test of homogeneity. In general, the null hypothesis for a test of homogeneity with $i = I$ populations is written

$$H_0 : \pi_{1j} = \pi_{2j} = \cdots = \pi_{Ij} \text{ for all } j \text{ versus } H_1 : \pi_{ij} \neq \pi_{i+1,j} \text{ for some } (i, j). \qquad (10.32)$$

Expressed in words, the null hypothesis is that the I populations are homogeneous with respect to the J categories versus the I populations are not homogeneous with respect to the J categories. An equivalent interpretation is that for each population $i = 1, 2, \ldots, I$, the proportion of people in the j^{th} category is the same. When H_0 is true, $\pi_{1j} = \pi_{2j} = \cdots = \pi_{Ij}$ for all j. Under the null hypothesis, $\mu_{ij} = n_{i\bullet}\pi_{ij}$, $\hat{\pi}_{ij} = p_{ij} = \frac{n_{\bullet j}}{n_{\bullet\bullet}}$, and $\hat{\mu}_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}} = E_{ij}$. When H_0 is true, all the probabilities in the j^{th} column are equal, and a pooled estimate of π_{ij} is obtained by adding all the frequencies in the j^{th} column ($n_{\bullet j}$) and dividing the total by $n_{\bullet\bullet}$. The statistic used in this type of problem has the same form as the one used for the test of independence in (10.31). Substituting the homogeneity expressions for O_{ij} and E_{ij}, the statistic is expressed as

$$\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i\bullet}n_{\bullet j}/n_{\bullet\bullet})^2}{n_{i\bullet}n_{\bullet j}/n_{\bullet\bullet}} \doteq \chi^2_{(I-1)(J-1)}.$$

The null hypothesis of homogeneity is rejected when $\chi^2_{\text{obs}} > \chi^2_{1-\alpha;(I-1)(J-1)}$.

When row and column totals are not fixed, the numbers in the i, j cells can be used to estimate their corresponding population proportions without assuming the null hypothesis is true. With fixed row or column totals, this estimation cannot be accomplished. That is, $\hat{\pi}_{ij} = p_{ij} \neq \frac{n_{ij}}{n_{\bullet\bullet}}$ when H_0 is false.

Test for SCENARIO TWO:

Step 1: **Hypotheses** — $H_0 : \pi_{1j} = \pi_{2j}$ for all j versus $H_1 : \pi_{i,\,j} \neq \pi_{i+1,\,j}$ for some (i, j). That is, all the probabilities in the same column are equal to each other versus at least two of the probabilities in the same column are not equal to each other.

Step 2: **Test Statistic** — The test statistic is

$$\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(I-1)(J-1)} = \chi^2_{(2-1)(3-1)} = \chi^2_2$$

under the null hypothesis. The χ^2_{obs} value is 6.7584.

Step 3: **Rejection Region Calculations** — The rejection region is

$$\chi^2_{\text{obs}} > \chi^2_{1-\alpha;(I-1)\cdot(J-1)} = \chi^2_{0.95;2} = 5.9915.$$

Before the statistic $\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ can be computed, the expected counts for each of the ij cells must be determined. Recall that $O_{ij} = n_{ij}$ and $E_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}$.

```
> DT <- c(67, 76, 57, 48, 73, 79)
> DTT <- matrix(data = DT, nrow = 2, byrow = TRUE)
> dimnames(DTT) <- list(Treatment = c("Drug", "Placebo"),
+    Category = c("Improve", "No Change", "Worse"))
> DTT

         Category
Treatment Improve No Change Worse
  Drug         67        76    57
  Placebo      48        73    79
```

```
> E <- chisq.test(DTT)$expected
> E

          Category
Treatment Improve No Change Worse
  Drug       57.5      74.5    68
  Placebo    57.5      74.5    68
```

$$\chi^2_{\text{obs}} = \frac{(67-57.5)^2}{57.5} + \frac{(76-74.5)^2}{74.5} + \cdots + \frac{(79-68)^2}{68} = 6.7584.$$

The value of the test statistic is $\chi^2_{obs} = 6.7584$. This can be done with code by entering

```
> chi.obs <- sum((DTT - E)^2/E)
> chi.obs

[1] 6.758357
```

$6.7584 = \chi^2_{\text{obs}} \overset{?}{>} \chi^2_{.95,2} = 5.9915.$

Step 4: **Statistical Conclusion** — The $\wp$-value is 0.03408.

```
> p.val <- pchisq(chi.obs, 2, lower = FALSE)
> p.val

[1] 0.03407544
```

I. From the rejection region, since $\chi^2_{\text{obs}} = 6.7584 > \chi_{0.95;2} = 5.9915$, reject the null hypothesis of homogeneity.

II. Since the $\wp$-value $= 0.0341$ is less than 0.05, reject the null hypothesis of homogeneity.

Reject H_0.

Step 5: **English Conclusion** — There is sufficient evidence to suggest that not all of the probabilities for the $i = 2$ populations with respect to each of the J categories are equal.

Using `chisq.test()` directly produces the same results.

```
> chisq.test(DTT)


Pearson's Chi-squared test

data:  DTT
X-squared = 6.7584, df = 2, p-value = 0.03408
```