10

Goodness of fit

In this chapter we return to problems involving categorical data. We previously summarized such data using tables. Here we discuss a significance test for the distribution of the values in a table. The test statistic will be based on how well the actual counts for each category fit the expected counts.

Such tests are called goodness-of-fit tests, as they measure how well the distribution of the data fits a probability model. In this chapter we will also discuss goodness-of-fit tests for continuous data. For example, we will learn a significance test for investigating whether a data set is normally distributed.

10.1 The chi-squared goodness-of-fit test

In a public-opinion poll, there are often more than two possible opinions. For example, suppose a survey of registered voters is taken to see which candidate is likely to be elected in an upcoming election. For simplicity, we assume there are two candidates, a Republican and a Democrat. A prospective voter may choose one of these or may be undecided. If 100 people are surveyed, and the results are 35 for the Republican, 40 for the Democrat, and 25 undecided, is the difference between the Republican and Democratic candidate significant?

The multinomial distribution

Before answering a question about significance, we need a probability model, so that p-value calculations can be made. The above example is a bit different from the familiar polling model. When there are just two categories to choose from we use the binomial model as our probability model; in this case, with more categories, we generalize and use the *multinomial model*.

Assume we have k categories to choose from, labeled 1 through k. We pick one of the categories at random, with probabilities specified by $p_1, p_2, \ldots, p_k$; p_i gives the probability of selecting category i. We must have $p_1 + p_2 + \cdots + p_k = 1$. If all the p_i equal $1/k$, then each category is equally likely (like rolling a die). Picking a category with these probabilities produces a single random value; repeat this selection n times, with each pick being independent, to get n values. A table of values will report the frequencies. Call

these table entries $y_1, y_2, \ldots, y_k$. These k numbers sum to n. The joint distribution of these random variables is called the *multinomial distribution*.

• Example 10.1: Using `sample` to simulate multinomial data
We can create multinomial data in R with the `sample` function. For example, an M&M's bag is filled using colors drawn from a fixed ratio. A bag of 30 can be filled as follows:[1]

```
cols <- c("blue", "brown", "green", "orange", "red", "yellow",
          "purple")
prob <- c(1, 1, 1, 1, 2, 2, 2)       # ratio of colors
prob <- prob / sum(prob)
n <- 30
bagful <- sample(cols, n, replace=TRUE, prob=prob)
table(bagful)

## bagful
##   blue  brown  green orange purple    red yellow
##      3      3      4      4      3      8      5
```

••

A formula for the multinomial distribution is similar to that for the binomial distribution except that more factors are involved, as there are more categories to choose from. The distribution can be specified as follows:

$$\mathsf{P}(y_1 = y_1, \ldots, y_k = y_k) = \binom{n}{y_1}\binom{n-y_1}{y_2}\cdots\binom{n-y_1-y_2-\cdots-y_{k-1}}{y_k} p_1^{y_i}\cdots p_k^{y_k}.$$

As an example, consider the voter survey. Suppose we expected the percentages to be 35% Republican, 35% Democrat, and 30% undecided. What is the probability in a survey of 100 likely voters that we see 35, 40, and 25, respectively? It is

$$\mathsf{P}(y_1 = 35, y_2 = 40, y_3 = 25) = \binom{100}{35}\binom{65}{40}\binom{25}{25}(0.35)^{35}(0.35)^{40}(0.3)^{25}.$$

This value can be found directly with

```
choose(100,35)*choose(65,40)*choose(25,25) * .35^35 * .35^40 * .30^25

## [1] 0.00386
```

[1] The `sample` function needs to be called with `replace=TRUE` to sample with replacement. The `mosaic` package provides a convenience wrapper `resample` which uses this default.

(We could skip the last choose factor, as $\binom{j}{j} = 1$ for any j.) The `dmultinom` function can also be used for the above computation. This small value is the probability of the observed value, but it is not a p-value. A p-value also includes the probability of seeing more extreme values than the observed one. We would still need to specify what that means to compute a p-value.

Pearson's χ^2-statistic

Trying to use the multinomial distribution directly to answer a problem about the p-value is difficult, as the variables y_i are correlated—they add to n. If one value is large the others are more likely to be small, so the meaning of "extreme" in calculating a p-value is not immediately clear. As an alternative, the problem of testing whether a given set of probabilities could have produced the data is done as before: by comparing the observed value with the expected value and then normalizing to get something with a known distribution.

Each y_i is a random variable telling us how many of the n choices were in category i. If we focus on a single i, then y_i is seen to be $\text{Binomial}(n, p_i)$ with an expected value of np_i. Based on this, a good statistic might be

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

This gives the total discrepancy between the observed and the expected. We use the square as $\sum(y_i - np_i) = 0$. This sum gets larger when a category is larger or smaller than expected. So a larger-than-expected value contributes, and any correlated smaller-than-expected values do, too. As usual, we scale this by the right amount to yield a test statistic with a known distribution. In this case, each term is divided by the expected amount, producing *Pearson's chi-squared statistic* (written using the Greek letter "*chi*"):

$$\chi^2 = \sum_{i=1}^{k} \frac{(y_i - np_i)^2}{np_i} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}. \tag{10.1}$$

If the multinomial model is correct, then the *asymptotic* distribution of y_i is known to be the chi-squared distribution with $k - 1$ degrees of freedom. The number of degrees of freedom coincides with the number of free ways we can specify the values for p_i in the null hypothesis. Here we are free to choose $k - 1$ of the values but not k, as the values must sum to 1.

The chi-squared distribution is a good fit to the sampling distribution of the statistic if the expected cell counts are all five or more. Figure 10.1 shows a simulation and a histogram of the corresponding χ^2-statistic, along with a theoretical density, when $n = 20$ and np is 5 or more for all the p's.

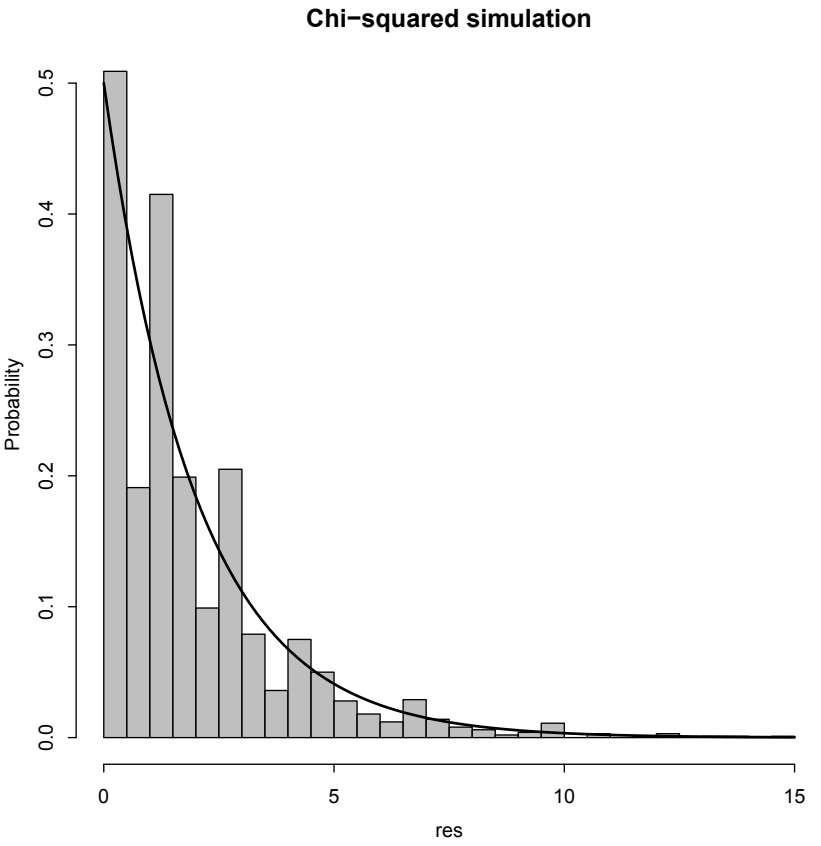

Figure 10.1: Simulation of χ^2-statistic with $n = 20$ and probabilities 3/12, 4/12, and 5/12. The chi-squared density with $2 = k - 1$ degrees of freedom is added.

Using this statistic as a test statistic allows us to construct a significance test. Larger values are now considered more extreme, as they imply more discrepancy from the predicted amount.

The chi-squared significance test for goodness of fit

Let $y_1, y_2, \ldots, y_k$ be the observed cell counts in a table that arise from random sampling. Suppose their joint distribution is described by the multinomial model with probabilities p_1, p_2, ..., p_k. A significance test of

$$H_0 : p_1 = \pi_1, \cdots, p_k = \pi_k, \quad H_A : p_i \neq \pi_i \text{ for at least one } i$$

can be performed with the χ^2-statistic. The π_i are specified probabilities. Under H_0 the sampling distribution is asymptotically the chi-squared distribution with $k-1$ degrees of freedom. This is a good approximation, provided that the expected cell counts are all five or more. Large values of the statistic support the alternative.

This test is implemented by the `chisq.test` function. The function is called with

```
chisq.test(x, p=...)
```

The data is given in tabulated form in `x`; the null hypothesis is specified with the argument `p` as a vector of probabilities. The default is a uniform probability assumption. This should be given as a named argument, as it is not the second position in the list of arguments. The alternative hypothesis is not specified, as it does not change. A warning message will be returned if any category has fewer than five expected counts.

For example, suppose we wanted to know whether the voter data was generated according to the probabilities $p_1 = .35$, $p_2 = .35$, and $p_3 = .30$. To investigate, we can perform a significance test. This can be done directly with the `chisq.test` function or "by hand." We illustrate both approaches, as we'll see soon that knowing how to do it the long way allows us to do more problems.

To do this by hand, we specify the counts in `y` and the probabilities in `p`, then form the test statistic:

```
y <- c(35, 40, 25)
p <- c(35, 35, 30)              # ratios
p <- p/sum(p)                   # proportions
n <- sum(y)
chi2 <- sum( (y - n*p)^2 / (n*p) )
chi2

## [1] 1.548

pchisq(chi2, df=3 - 1, lower.tail=FALSE)

## [1] 0.4613
```

In contrast, the above could have been done with

```
chisq.test(y, p=p)

##
##  Chi-squared test for given probabilities
```

```
##
## data:  y
## X-squared = 1.548, df = 2, p-value = 0.4613
```

The function returns the value of the test statistic (after `X-squared`), the degrees of freedom, and the p-value.

• Example 10.2: Teen smoking
The `samhda` (`UsingR`) data set contains information about health behavior for school-age children. For example, the variable `amt.smoke` measures how often a child smoked in the previous month. There are seven levels: a `1` means he smoked every day and a `7` means not at all. Values `98` and `99` indicate missing data. See `?samhda` for a description. We investigate whether the sample proportions are statistically different from the probabilities:

$$p_1 = 0.15, p_2 = 0.05, p_3 = 0.05, p_4 = 0.05, p_5 = 0.10, p_6 = 0.20, p_7 = 0.40.$$

A test of significance can be constructed as follows:

```
library(UsingR)
amt <- with(samhda, amt.smoke[amt.smoke < 98])
y <- table(amt)
y

## amt
##   1   2   3   4   5   6   7
##  32   7  13  10  14  43 105

ps <- c(0.15, 0.05, 0.05, 0.05, 0.10, 0.20, 0.40)
chisq.test(y, p=ps)

##
##  Chi-squared test for given probabilities
##
## data:  y
## X-squared = 7.938, df = 6, p-value = 0.2427
```

The p-value of 0.2427 is not significant. There is no evidence that the population proportions differ from those specified by the null hypothesis. ••

Partially specified null hypotheses

In the example with voting data, we might be interested in assessing whether the Republican candidate differences from the Democrat can be attributed to

sampling variation. That is, we would want to test the hypotheses

$$H_0 : p_1 = p_2, \quad H_A : p_1 \neq p_2.$$

These, too, can be tested with the χ^2-statistic, but we need to specify what we mean by "expected," as under H_0 this is not fully specified. (The value of p_1 and p_2 depends on p_3, which isn't specified.)

To do so, we will use any values completely specified by the null hypothesis; for those values that aren't (e.g., p_3 above), we estimate using the null hypothesis to pool our data as appropriate. For this problem, none of the p_i values are fully specified. To estimate $\widehat{p}_1 = \widehat{p}_2$, we use both of the cell counts through $(y_1 + y_2)/(2n)$. This leaves $\widehat{p}_3 = y_3/n = 1 - \widehat{p}_1 - \widehat{p}_2$. Then the χ^2-statistic in this case becomes

$$\chi^2 = \sum_{i=1}^{k} \frac{(y_i - n\widehat{p}_i)^2}{n\widehat{p}_i}.$$

Again, if all the expected counts are large enough, this will have an approximately chi-squared distribution. There is only one degree of freedom in this problem, as only one thing is left to estimate, namely the value $p = p_1 = p_2$. Once we specify a value of p, then, by the assumptions in the null hypothesis, all the p_i are decided.

We get the p-value in our example as follows:

```
y <- c(35, 40, 25)
n <- sum(y)
phat1 <- phat2 <- sum(y[1:2])/(2*n)
phat3 <- 1 - phat1 - phat2
phat <- c(phat1, phat2, phat3)
#
obs <- sum((y - n*phat)^2/(n*phat))
obs

## [1] 0.3333

pchisq(obs, df =1 , lower.tail=FALSE)

## [1] 0.5637
```

The difference is not statistically significant.

In general, the χ^2-statistic can be used in significance tests where the null specifies some relationship among the multinomial probabilities. The asymptotic distribution of the test statistic under the null hypothesis will be chi-squared. The degrees of freedom will depend on the number of values that we are free to specify.

Candidate	party	poll amount	actual
Schwarzenegger	Republican	315	48.6
Bustamante	Democrat	197	31.5
McClintock	Republican	141	12.5
Camejo	Green	39	2.8
Huffington	Independent	16	0.6
Other	–	79	4.0

Table 10.1: California gubernatorial recall election.

Problems

10.1 A die is rolled 100 times and yields these frequencies

	1	2	3	4	5	6
count	13	17	9	17	18	26

Is this a fair die? Answer using a significance test with $H_0 : p_i = 1/6$ for each i, and $H_A : p_i \neq 1/6$ for at least one i.

10.2 Table 10.1 contains the results of a poll of 787 registered voters and the actual race results (in percentages of total votes) in the 2003 gubernatorial recall election in California.

Is the sample data consistent with the actual results? Answer this using a test of significance.

10.3 A package of M&M's candies is filled from batches that contain a specified percentage of each of six colors. These percentages are given in the `mandms` (`UsingR`) data frame. Assume a package of candies contains the following color distribution: 15 blue, 34 brown, 7 green, 19 orange, 29 red, and 24 yellow. Perform a chi-squared test with the null hypothesis that the candies are from a `milk chocolate` package. Repeat assuming the candies are from a `Peanut` package. Based on the p-values, which would you suspect is the true source of the candies?

10.4 The `pi2000` (`UsingR`) data set contains the first 2,000 digits of π. Perform a chi-squared significance test to see if the digits appear with equal probability.

10.5 A simple trick for determining what language a document is written in is to compare the letter distributions (e.g., the number of z's) to the known proportions for a language. For these proportions, we use the familiar letter frequencies given in the `frequencies` variable of the `scrabble` (`UsingR`) data set. These are an okay approximation to those in the English language.

	a	e	i	o	u
Count	28	39	23	22	11
Scrabble frequency	9	12	9	8	4

Table 10.2: Vowel distribution and Scrabble frequency.

For simplicity (see `?scrabble` for more details), we focus on the vowel distribution of a paragraph from R's webpage appearing below. The counts and Scrabble frequencies are given in Table 10.2.

> R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

Perform a chi-squared goodness-of-fit test to see whether the distribution of vowels appears to be from English.

10.6 The names of common stars are typically Greek or Arab in derivation. The `bright.stars` (`UsingR`) data set contains 96 names of common stars. Perform a significance test on the letter distribution to see whether they could be mistaken for English words.

The letter distribution can be found with:

```
all.names <- paste(bright.stars$name, sep="", collapse="")
x <- unlist(strsplit(tolower(all.names), ""))
letter.dist <- sapply(letters, function(i) sum(x == i))
```

The English-letter frequency is found using the `scrabble` (`UsingR`) data set with:

```
ps <- scrabble$frequency[1:26]
ps <- ps/sum(ps)
```

10.7 The number of murders by day of week in New Jersey during 2011 and 2003 is shown in Table 10.3.

1. For the 2011 data, perform a significance test to test the null hypothesis that a murder is equally likely to occur on any given day.

Year	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
2003	53	42	51	45	36	37	65
2011	63	53	50	51	55	52	56

Table 10.3: Number of murders by day of week in New Jersey in 2003 and 2011. (Source: `http://www.njsp.org`.)

2. Again, for the 2011 data perform a significance test of the null hypothesis that murders happen on each weekday with equal probability; similarly on the weekends, but not necessarily with the same probability.

For each test, write down explicitly the null and alternative hypotheses.

10.8 A large bag of M&M's is opened and some of the colors are counted: 41 brown, 48 orange, 105 yellow, and 58 green. Test the partially specified null hypothesis that the probability of brown is equal to the probability of orange. What do you conclude?

10.9 The data for Figure 10.1 was simulated using the following commands:

```
M <- 2000; n <- 20
p <- c(3,4,5)/12

res <- replicate(M, {
  x <- sample(1:3, n, replace=TRUE, prob=p)
  y <- sapply(1:3, function(i) sum(x==i))
  expected <- n * p
  chi <- sum( (y - expected)^2/expected )
  chi
})

col <- rgb(.7,.7,.7,.75)
hist(res, prob=TRUE, breaks=50,
     col=col, ylab="Probability", xlab="res",
     main="Chi-squared simulation")
curve(dchisq(x, df=length(p)-1), add=TRUE, lwd=2)
```

The sampling distribution of χ^2 is well approximated by the chi-squared distribution, with $k-1$ degrees if the expected cell counts are all five or more. Do a simulation like the above, only with $n = 5$. Does the fit seem right? Repeat with $n = 20$ using the different probabilities `p=c(1,19,20)/40`. Does the fit seem right?

10.10 When $k = 2$ you can algebraically simplify the χ^2-statistic. Show that it simplifies to

$$\chi^2 = \left(\frac{\widehat{p}_1 - p_1}{\sqrt{p_1(1-p_1)/n}} \right)^2.$$

This is the square of the statistic used in the one-sample test of proportion and is asymptotically a single-squared normal or a chi-squared random variable with 1 degree of freedom. Thus, in this case, the chi-squared test is equivalent to the test of proportions.

10.2 The chi-squared test of independence

In a two-way contingency table we are interested in the relationship between the variables. In particular, we ask whether the levels of one variable effect the distribution of the other variable. That is, are they independent random variables in the same sense that we defined an independent sequence of random variables?

For example, in the seat-belt-usage data from Table 3.3, does the fact that a parent has her seat belt buckled effect the chance that the child's seat belt will be buckled?

The differences appear so dramatic that the answer seems to be obvious. We can set up a significance test to help make the decision formal, using a method that can be used when the data does not tell such a clear story.

To approach this question with a significance test, we need to state the null and alternative hypotheses, a test statistic, and a probability model.

First, our model for the sampling is that each observed car follows some specified probability that is recorded in any given cell. These probabilities don't change from observation to observation, and the outcome of one does not effect the distribution of another. That is, we have an *i.i.d.* sequence. Then a multinomial model applies. Fix some notation. Let n_r be the number of rows in the table (the number of levels of the row variable), n_c be the number of columns, and y_{ij} record the frequency of the (i, j) cell. Let p_{ij} be the cell probability for the ith row and jth column from a model for the data. The marginal probabilities are denoted p_i^r and p_j^c where, for example, $p_i^r = p_{i1} + p_{i2} + \cdots + p_{in_j}$.

Our null hypothesis is that the column variable should be *independent* of the row variable. When stated in terms of the cell probabilities, this says that $p_{ij} = p_i^r p_j^c$. This is consistent with the notion that independence means multiply.

Our hypotheses can be stated informally as:

H_0: the variables are independent, $\quad H_A$: the variables are not independent.

In terms of our notation, we can rewrite the null hypothesis as $H_0 : p_{ij} = p_i^r p_j^c$.

	Child		
Parent	buckled	unbuckled	marginal
buckled	56	8	64
unbuckled	2	16	18
marginal	58	24	82

Table 10.4: Seat-belt usage in California with marginal distributions

The χ^2-statistic,

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}},$$

can still be used as a test statistic after we have estimated each p_{ij} in order to compute the "expected" counts. Again we use the data and the assumptions to estimate the p_{ij}. Basically, the data is used to estimate the marginal probabilities, and the assumption of independence allows us to estimate the p_{ij} from there.

The marginal probabilities are estimated by the marginal distributions of the data. For our example these are added to Table 3.3 to give Table 10.4.

The estimate for $p_1^r = \mathsf{P}(\text{parent is buckled})$ is $\widehat{p}_1^r = 64/82$, and for $p_2^r = \mathsf{P}(\text{parent is unbuckled})$ it is $\widehat{p}_2^r = 18/82$. Similarly, for p_j^c we have $\widehat{p}_1^c = 58/82$ and $\widehat{p}_2^c = 24/82$. As usual, we've used a "hat" for estimated values.

With these estimates, we can use the relationship $p_{ij} = p_i^r p_j^c$ to find the estimate $\widehat{p}_{ij} = \widehat{p}_i^r \widehat{p}_j^c$. For our seat-belt data we have the estimates in Table 10.5. In order to show where the values come from, the values have not been simplified.

With this table we can compute the expected amounts in the ijth cell with $n\widehat{p}_{ij}$. This is often written $R_i C_j / n$, where R_i is the row sum and C_j the column sum, as this simplifies computations by hand.

With the expected amounts now known, we form the χ^2-statistic as:

$$\chi^2 = \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} \frac{(y_{ij} - n\widehat{p}_{ij})^2}{n\widehat{p}_{ij}}. \tag{10.2}$$

Under the hypothesis of multinomial data and the independence of the variables, the sampling distribution of χ^2 will be the chi-squared distribution with $(n_r - 1) \cdot (n_c - 1)$ degrees of freedom. Why this many? In general, we subtract one degree of freedom from $n_r \cdot n_c - 1$ for each estimated parameter. As there are $n_r - 1 + n_c + 1$ estimated parameters, the value for the degrees of freedom is $n_r \cdot n_c - 1 - (n_r - 1 + n_c + 1) = n_r \cdot n_c - n_r - n_c + 1 = (n_r - 1) \cdot (n_c - 1)$.

	Child		
Parent	buckled	unbuckled	marginal
buckled	$\frac{64}{82} \cdot \frac{58}{82}$	$\frac{64}{82} \cdot \frac{24}{82}$	$\frac{64}{82}$
unbuckled	$\frac{18}{82} \cdot \frac{58}{82}$	$\frac{18}{82} \cdot \frac{24}{82}$	$\frac{18}{82}$
marginal	$\frac{58}{82}$	$\frac{24}{82}$	1

Table 10.5: Seat-belt usage in California with estimates $\widehat{p}_{ij}$ for the corresponding p_{ij}.

We now have all the pieces to formulate the problem in the language of a significance test.

The chi-squared test of independence

Let $y_{ij}, i = 1, \ldots, n_r, j = 1, \ldots, n_c$ be the cell frequencies in a two-way contingency table for which the multinomial model applies. A significance test of

$$H_0 : \text{the two variables are independent}$$
$$H_A : \text{the two variables are not independent}$$

can be performed using the chi-squared test statistic (10.2). Under the null hypothesis, this statistic has a sampling distribution that is approximated by the chi-squared distribution with $(n_r - 1)(n_c - 1)$ degrees of freedom. The p-value is computed using $\mathsf{P}(\chi^2 \geq \text{ observed value } | H_0)$.

In R this test is performed by the `chisq.test` function. If the data is summarized in a table or a matrix in the variable `x` the usage is

```
chisq.test(x).
```

If the data is unsummarized and is stored in two variables `x` and `y` where the ith entries match up, then the function can be used as

```
chisq.test(x, y).
```

Alternatively, the data could be summarized first using `table`. (For `table` and `xtabs` objects, the `summary` method for these objects will also report this test.)

For each usage, the null and alternative hypotheses are not specified, as they are the same each time the test is used.

The argument `simulate.p.value=TRUE` will return a p-value estimated using a Monte Carlo simulation. This is used if the expected counts in some cells are too small to use the chi-squared distribution to approximate the sampling distribution of χ^2.

To illustrate, the following will do the chi-squared test on the seat-belt data. This data is summarized, so we first need to make a table. We use `rbind` to combine rows.

```
seatbelt <- rbind(c(56,8), c(2,16))
seatbelt

##      [,1] [,2]
## [1,]   56    8
## [2,]    2   16

chisq.test(seatbelt)

##
##  Pearson's Chi-squared test with Yates' continuity
##  correction
##
## data:  seatbelt
## X-squared = 36, df = 1, p-value = 1.978e-09
```

The minuscule p-value is consistent with our observation that the two variables are not independent.

• Example 10.3: Teen smoking and gender
The `samhda` (`UsingR`) data set contains survey data on 590 children. The variables `gender` and `amt.smoke` contain information about the gender of the participant and how often the participant smoked in the last month. Are the two variables independent?

We compute a p-value for the hypotheses

$$H_0 : \text{the two variables are independent,}$$
$$H_A : \text{the two variables are not independent}$$

using the χ^2-statistic.

In this example we use `xtabs` to make a table, then apply `chisq.test`. The `xtabs` function allows us to use the convenient `subset` argument to eliminate the data for which the values are not applicable.

```
tbl <- xtabs( ~ gender + amt.smoke,    # no left side in formula
              subset = amt.smoke < 98 & gender !=7,
              data=samhda)
tbl
```

```
##       amt.smoke
## gender  1  2  3  4  5  6  7
##      1 16  3  5  6  7 24 64
##      2 16  4  8  4  7 19 40

chisq.test(tbl)

## Warning: Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 4.147, df = 6, p-value = 0.6568
```

The significance test shows no reason to doubt the hypothesis that the two variables are independent.

The warning message is due to some expected counts being small. Could this significantly change the p-value reported? A p-value based on a simulation may be computed.

```
chisq.test(tbl,simulate.p.value=TRUE)

##
##  Pearson's Chi-squared test with simulated p-value (based
##  on 2000 replicates)
##
## data:  tbl
## X-squared = 4.147, df = NA, p-value = 0.6707
```

The p-value is not changed significantly. ••

The chi-squared test of homogeneity

How can we assess the effectiveness of a drug treatment? Typically, there is a clinical trial, with each participant randomly allocated to either a treatment group or a placebo group. If the results are measured numerically, a t-test may be appropriate to investigate whether any differences in means are significant. When the results are categorical, we see next how to use the χ^2-statistic to test whether the distributions of the results are the same.

Stanford University Medical Center conducted a study to determine whether the antidepressant Celexa can help stop compulsive shopping. Twenty-four compulsive shoppers participated in the study: twelve were given a placebo and twelve a dosage of Celexa daily for seven weeks. After

	Much worse	Worse	Same	Much improved	Very much so
Celexa	0	2	3	5	2
placebo	0	2	8	2	0

Table 10.6: Does Celexa treatment cut down on compulsive shopping?

this time the individuals were surveyed to determine whether their desires to shop had been curtailed. Data simulated from a preliminary report is given in Table 10.6.

Does this indicate that the two samples have different distributions?

We formulate this as a significance test using hypotheses:

$$H_0 : \text{the two distributions are the same}$$
$$H_A : \text{the two distributions are different.}$$

We use the χ^2-statistic. Again we need to determine the expected amounts, as they are not fully specified by H_0.

Let the random variable be the column variable, and the category that breaks up the data be the row variable in our table of data. For row i of the table, let p_{ij} be the probability that the random variable (the survey result) will be in the jth level of the random variable. We can rephrase the hypotheses as

$$H_0 : p_{ij} = p_j \text{ for all rows } i, \quad H_A : p_{ij} \neq p_j \text{ for some } i, j.$$

If we let n_i be the number of counts in each row (R_i before), then the expected amount in the (i,j) cell under H_0 should be $n_i p_j$. We don't specify the value of p_j in the null hypothesis, so it is estimated. Under H_0 all the data in the jth column of our table is binomial with n and p_j, so an estimator for p_j would be the column sum divided by n: C_j/n. Based on this, the expected number in the (i,j)-cell would be

$$e_{ij} = n_i \widehat{p}_j = \frac{R_i C_j}{n}.$$

This is the same formula as the chi-squared test of independence.

As the test statistic and its sampling distribution under H_0 are the same as with the test of independence, the chi-squared significance tests of homogeneity and independence are identical in implementation despite the differences in the hypotheses.

Before proceeding, let's combine the data so that there are three outcomes: "worse," "same," and "better."

```
celexa <- c(2, 3, 7)
placebo <- c(2, 8, 2)
x <- rbind(celexa, placebo)
colnames(x) <- c("worse", "same", "better")
x

##         worse same better
## celexa      2    3      7
## placebo     2    8      2

chisq.test(x)

## Warning: Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 5.051, df = 2, p-value = 0.08004
```

The warning notes that one or more of the expected cell counts is less than five, indicating a possible discrepancy with the asymptotic distribution used to find the p-value. We can use a simulation to find the p-value, instead of using the chi-squared distribution approximation, as follows:

```
chisq.test(x, simulate.p.value=TRUE)

##
##  Pearson's Chi-squared test with simulated p-value (based
##  on 2000 replicates)
##
## data:  x
## X-squared = 5.051, df = NA, p-value = 0.1014
```

In both cases, the p-value is small but not tiny.

Problems

10.11 A number of drivers were surveyed to see whether they had been in an accident during the previous year, and, if so, whether it was a minor or major accident. The results are tabulated by age group in Table 10.7. Do a chi-squared hypothesis test of independence for the two variables.

10.12 The `airquality` data set contains measurements of air quality in New York City. We wish to see if ozone levels are independent of temperature. First we gather the data, using `complete.cases` to remove missing data from our data set.

Age	Accident type		
	none	minor	major
under 18	67	10	5
18–25	42	6	5
26–40	75	8	4
41–65	56	4	6
over 65	57	15	1

Table 10.7: Type of accident by age.

```
aq <- airquality[complete.cases(airquality),]
aq <- transform(aq,
  te = cut(Temp, quantile(Temp)),
  oz = cut(Ozone,quantile(Ozone))
)
xtabs(~ te + oz, data=aq)

##            oz
## te          (1,18] (18,31] (31,62] (62,168]
##   (57,71]       15       9       3        0
##   (71,79]       10      10       7        1
##   (79,84.5]      4       6      11        5
##   (84.5,97]      0       0       6       22
```

Perform a chi-squared test of independence on the two variables `te` and `oz`. Does the data support an assumption of independence?

10.13 The following table contains data on the severity of injuries sustained during car crashes.

		Injury level			
		none	minimal	minor	major
Seat belt	yes	12,813	647	359	42
	no	65,963	4,000	2,642	303

The data is tabulated by whether or not the passenger wore a seat belt. Are the two variables independent?

10.14 The data set `oral.lesion` (`UsingR`) contains data on location of an oral lesion for three geographic locations. This data set appears in an article by Mehta and Patel about differences in p-values in tests for independence when the exact or asymptotic distributions are used. Compare the p-values found

by `chisq.test` when the asymptotic distribution of the sampling distribution is used to find the p-value and when a simulated value is used. Are the p-values similar? If not, which do you think is more accurate? Why?

10.15 In an effort to increase student retention, many colleges have tried "block programs." Assume that 100 students are broken into two groups of "50" at random. Fifty are in a block program; the others are not. The number of years each student attends the college is then measured. The following records the data:

Program	1 year	2 year	3 year	4 year	5+ years
nonblock	18	15	5	8	4
block	10	5	7	18	10

We wish to test whether a block program makes a difference in retention, assuming this sample is representative of a future population. Perform a chi-squared test of significance to investigate whether the distributions are homogeneous.

10.16 Table 10.3 lists the number of murders in New Jersey by day of week for the years 2003 and 2011. Is there a statistically significant difference in the distributions?

10.3 Goodness-of-fit tests for continuous distributions

When finding confidence intervals for a sample we were concerned about whether or not the data was sampled from a normal distribution. To investigate, we made a quantile plot or histogram and eyeballed the result. In this section, we see how to compare a continuous distribution with a theoretical one using a significance test.

The chi-squared test is used for categorical data. We can try to make it work for continuous data by "binning." That is, as in a construction of a histogram, we can choose some bins and count the number of data points in each. Now the data can be thought of as categorical, and the test can be used for goodness of fit.

This is fine in theory but works poorly in practice. The Kolmogorov-Smirnov test will be a better alternative in the continuous distribution case.

Kolmogorov-Smirnov test

Suppose we have a random sample $x_1, x_2, \ldots, x_n$ from some continuous distribution. (There should be no ties in the data.) Let $f(x)$ be the density and X some other random variable with this density. The *cumulative distribution*

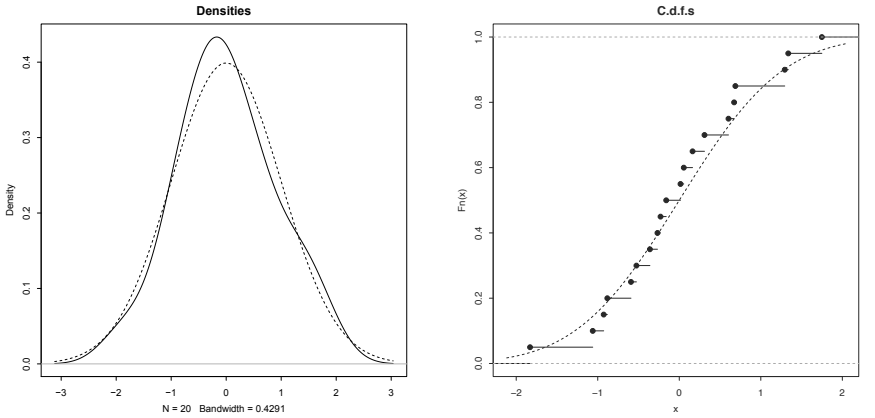

Figure 10.2: For a sample of size 20 from a normally distributed population, both sample and theoretical densities and cumulative distribution functions are drawn.

function for X is $F(x) = \mathsf{P}(X \leq x)$, or the area to the left of x under the density of X.

The c.d.f. can be defined the same way when X is discrete. In that case it is computed from the p.d.f. by summing: $\mathsf{P}(X \leq x) = \sum_{y \leq x} f(y)$.

For a sample, $x_1, x_2, \ldots, x_n$, the empirical distribution is the distribution generated by sampling from the data points. This becomes:

$$F_n(x) = \frac{\#\{i : x_i \leq x\}}{n}.$$

The function $F_n(x)$ can easily be plotted in R (e.g, Figure 10.2) using the `ecdf` function in the `stats` package. This function is used in a manner similar to the `density` function: the return value is plotted in a new figure using `plot` or may be added to the existing plot using `lines`.

If the data is from the population with c.d.f. F, then we would expect that F_n is close to F is some way. But what does "close" mean? In this context, we have two different functions of x. Define the distance between them as the largest difference they have:

$$D = \text{maximum in } x \text{ of } |F_n(x) - F(x)|.$$

The surprising thing is that with only the assumption that F is continuous, D has a known sampling distribution called the Kolmogorov-Smirnov distribution. This is illustrated in Figure 10.3, where the sampling distribution of the statistic for $n = 25$ is simulated for several families of random data. In each case, we see the same distribution. This fact allows us to construct a significance test using the test statistic D. In addition, a similar test can be done to compare two independent samples.

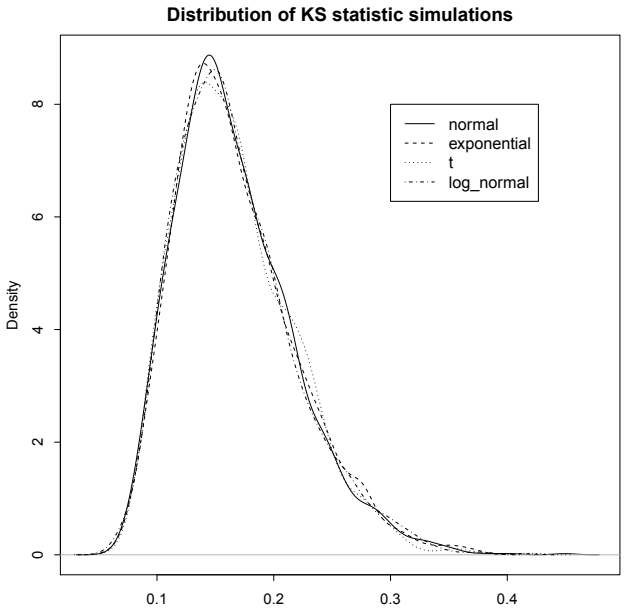

Figure 10.3: Density estimates for sampling distribution of the Kolmogorov-Smirnov statistic with $n = 25$ for normal, exponential, t, and log-normal data.

The Kolmogorov-Smirnov goodness-of-fit test

Assume $x_1, x_2, \ldots, x_n$ is an *i.i.d.* sample from a continuous distribution with c.d.f. $F(x)$. Let $F_n(x)$ be the empirical c.d.f. A significance test of

$$H_0 : F(x) = F_0(x), \qquad H_A : F(x) \neq F_0(x)$$

can be constructed with test statistic D. Large values of D support the alternative hypothesis.

In R, this test is implemented in the function `ks.test`. Its usage follows this pattern:

```
ks.test(x, y="name", ...)
```

The variable `x` stores the data. The argument `y` is used to set the family name of the distribution in H_0. It has a character value of `"name"` containing the "p" function that returns the c.d.f. for the family (e.g., `"pnorm"` or `"pt"`). The ...

argument allows the specification of the assumed parameter values. These depend on the family name and are specified as named arguments, as in `mean=1, sd=1`. The parameter values should not be estimated from the data, as this effects the sampling distribution of D.

If we have two *i.i.d.* independent samples, $x_1,\ldots,x_n$ and $y_1,\ldots,y_m$, from two continuous distributions F^X and F^Y, then a significance test of

$$H_0 : F^X = F^Y, \qquad H_A : F^X \neq F^Y$$

can be constructed with a similar test statistic:

$$D = \text{maximum in } x \text{ of } |F_n^X(x) - F_m^Y(x)|.$$

In this case, the `ks.test` can be used as

```
ks.test(x, y)
```

where `x` and `y` store the data.

We illustrate with some simulated data.

```
x <- rnorm(100, mean=5, sd=2)
ks.test(x, "pnorm", mean=0, sd=2)          # "wrong" parameters

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.7991, p-value < 2.2e-16
## alternative hypothesis: two-sided

ks.test(x, "pnorm", mean=5, sd=2)$p.value  # correct parameters

## [1] 0.02613

x = runif(100, min=0, max=5)
ks.test(x, "punif", min=0, max=6)$p.value  # "wrong" parameters

## [1] 0.0014

ks.test(x, "punif", min=0, max=5)$p.value  # correct parameters

## [1] 0.4299
```

The p-values are significant only when the parameters do not match the known population ones.

• Example 10.4: Difference in SAT scores
The data set `stud.recs` (`UsingR`) contains math and verbal SAT scores for

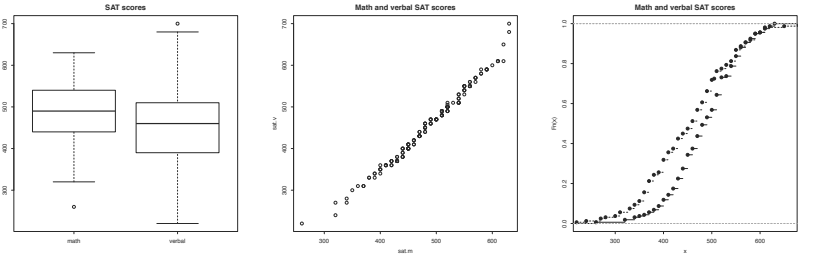

Figure 10.4: Three plots comparing the distribution of math and verbal SAT scores in the `stud.recs` (`UsingR`) data set.

some students (`sat.m` and `sat.v`). Assuming naively that the two samples are independent, are the samples from the same population of scores?

First, we make a q-q plot, a side-by-side boxplot, and a plot of the e.c.d.f.'s for the data, to see whether there is any merit to the question.

```
library(UsingR)
sat.m <- stud.recs$sat.m; sat.v <- stud.recs$sat.v
```

```
boxplot(list(math=sat.m, verbal=sat.v), main="SAT scores")
qqplot(sat.m, sat.v, main="Math and verbal SAT scores")
plot(ecdf(sat.m), main="Math and verbal SAT scores")
lines(ecdf(sat.v), lty=2)
```

The graphics are in Figure 10.4. The q-q plot shows similarly shaped distributions, but boxplots show that the centers appear to be different. Consequently, the cumulative distribution functions do not look that similar. The Kolmogorov-Smirnov test detects this and returns a small p-value.

```
ks.test(sat.m, sat.v)

## Warning: p-value will be approximate in the presence of ties

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  sat.m and sat.v
## D = 0.2125, p-value = 0.001456
## alternative hypothesis: two-sided
```

●●

The Shapiro-Wilk test for normality

The Kolmogorov-Smirnov test for a univariate data set works when the distribution in the null hypothesis is fully specified prior to our looking at the data. In particular, any assumptions on the values for the parameters should not depend on the data, as this can change the sampling distribution. Figure 10.5 shows the sampling distribution of the Kolmogorov-Smirnov statistic for Normal(0,1) data and the sampling distribution of the Kolmogorov-Smirnov statistic for the same data when the sample values of $\bar{x}$ and s are used for the parameters of the normal distribution (instead of 0 and 1). The figure was generated with this simulation:

```
res <- replicate(2000, {
  x <- rnorm(25, mean=0, sd=1)
  c(ks.test(x, pnorm, mean=mean(x), sd=sd(x))$statistic,
    ks.test(x, pnorm, mean=0,       sd=1)$statistic)
})
plot(density(res[1,]), main="K-S sampling distribution", ylab="")
lines(density(res[2,]), lty=2)
legend(0.2, 12, legend=c("estimated", "exact"), lty=1:2)
```

(To retrieve just the value of the test statistic from the output of `ks.test` we take advantage of the fact that its return value is a list with one component named `statistic` containing the desired value. This is why the syntax `ks.test(...)$statistic` is used.)

A consequence is that we can't use the Kolmogorov-Smirnov test to test for normality of a data set unless we know the parameters of the underlying distribution.[2] The Shapiro-Wilk test allows us to perform that analysis. This test statistic is based on the ideas behind the quantile-quantile plot, which we've used to gauge normality. Its definition is a bit involved, but its usage in R is not.

The Shapiro-Wilk test for normality

If $x_1, x_2, \ldots, x_n$ is an *i.i.d.* sample from a continuous distribution, a significance test of

$$H_0 : \text{parent distribution is normal,}$$
$$H_A : \text{the parent distribution is not normal}$$

can be carried out with the Shapiro-Wilk test statistic.

[2] The Lilliefors test, implemented by `lillie.test` in the contributed package `nortest`, will make the necessary adjustments to use this test statistic. As well, the `nortest` package implements other tests of normality.

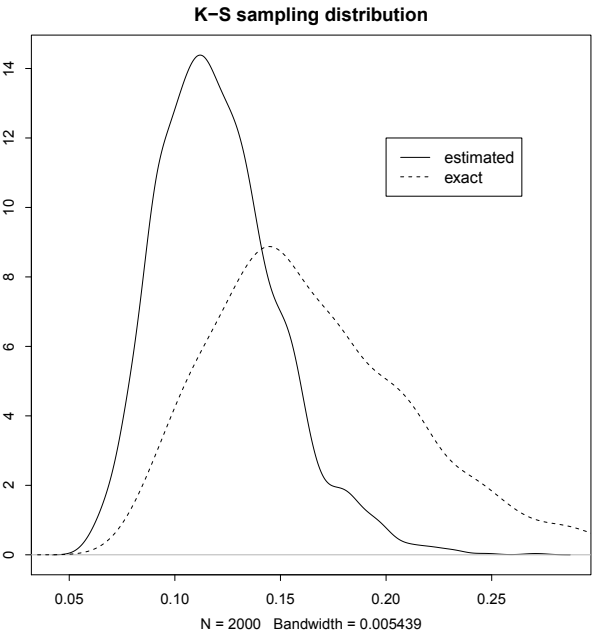

Figure 10.5: The sampling distribution for the Kolmogorov-Smirnov statistic when the parameters are estimated (solid line) and when not.

In R, the function `shapiro.test` will perform the test. The usage is simply

```
shapiro.test(x)
```

where the data vector `x` contains the sample data.

• **Example 10.5: Normality of SAT scores**
For the SAT data in the `stud.recs` (`UsingR`) data set, we saw in Example 10.3 that the two distributions are different. Are they normally distributed? We can answer with the Shapiro-Wilk test:

```
shapiro.test(stud.recs$sat.m)

##
##  Shapiro-Wilk normality test
##
## data:  stud.recs$sat.m
## W = 0.9898, p-value = 0.3055

shapiro.test(stud.recs$sat.v)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  stud.recs$sat.v
## W = 0.994, p-value = 0.752
```

In each case, the p-value is not statistically significant. There is no evidence in the data that the assumption of it being a random sample from a normal population should be doubted. ••

• Example 10.6: Is on-base percentage normally distributed?
The data set `OBP` (`UsingR`) records the on-base percentage for all players in the 2002 Major League Baseball season. It appears bell shaped except for one outlier. Does the data come from a normally distributed population?

Using the Shapiro-Wilk test gives us

```
shapiro.test(OBP)$p.value

## [1] 1.206e-07
```

The difference from normality is statistically significant. Perhaps this is due to the one outlier. We investigate:

```
shapiro.test(OBP[OBP<.5])$p.value

## [1] 0.006404
```

The conclusion is the same. However, note the dramatic difference in the p-value that just one outlier makes. The statistic is not very resistant. ••

In defining the t-test, it was assumed that the data is sampled from a normal population. This is because the sampling distribution of the t-statistic is known under this assumption. However, this would not preclude us from using the t-test to perform statistical inference on data that has failed a formal test for normality. For small samples the t-test may apply, as the distribution of the t-statistic is robust to small changes in the assumptions on the parent distribution. If the parent distribution is not normal but also not too skewed, then a t-test can be appropriate. For large samples, the central limit theorem may apply, making a t-test valid.

Finding parameter values using `fitdistr`

If we know a data set comes from a known distribution and would like to estimate the parameter values, we can use the convenient `fitdistr` function

from the MASS library. This function estimates the parameters for a wide family of distributions. The function is called with these arguments:

```
fitdistr(x, densfun=family.name, start=list(...))
```

We specify the data as a data vector, x; the family is specified by its full name (unlike that used in ks.test); and, for many of the distributions, reasonable starting values are specified using a named list. The fitdistr function fits the parameters by a method called maximum-likelihood. Often this coincides with using the sample mean or standard deviation to estimate the parameters, but in general it allows for a uniform approach to this estimation problem and associated inferential problems.

• **Example 10.7: Exploring** fitdistr
The data set babyboom (UsingR) contains data on the births of 44 children in a one-day period at a hospital in Brisbane, Australia. The variable wt records the weights of each newborn. A histogram suggests that the data comes from a normally distributed population. We can use fitdistr to find estimates for the parameters μ and σ, which for the normal distribution are the population mean and standard deviation.

```
library(MASS)
fitdistr(babyboom$wt,"normal")

##     mean       sd
##   3275.95    522.00
##  (  78.69) (  55.65)
```

These estimates include standard errors in parentheses computed using a normal approximation. These can be employed to give confidence intervals for the estimates.

This estimate for the mean and standard deviation could also be done directly, as it coincides with the sample mean and sample standard deviation. However, the standard errors are new. To give a different usage, we look at the variable running.time, which records the time of day of each birth. The time differences between successive births are called the inter-arrival times. We first find the inter-arrival times using diff:

```
inter = diff(babyboom$running.time)
```

We fit the gamma distribution to the data. The gamma distribution generalizes the exponential distribution. It has two parameters, a shape and a rate. A value of 1 for the shape coincides with the exponential distribution. The fitdistr function does not need starting values for the gamma distribution.

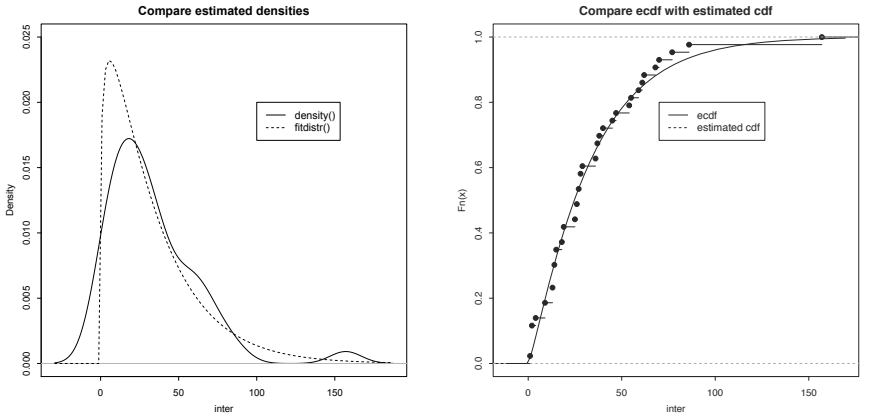

Figure 10.6: Empirical and theoretical densities and cumulative distribution functions for the inter-arrival times in the babyboom data set.

```
out <- fitdistr(inter, "gamma")
out

##     shape       rate
##   1.208846   0.036350
##  (0.233207) (0.008632)
```

Finally, we look at density estimates and cumulative distribution functions with the following commands (Figure 10.6):

```
plot(density(inter), ylim=c(0, 0.025),
     main="Compare estimated densities", xlab="inter")
curve(dgamma(x, shape=out$estimate["shape"],
             rate=out$estimate["rate"]), add=TRUE, lty=2)
legend(100, 0.020, legend=c("density()", "fitdistr()"),lty=1:2)
#
plot(ecdf(inter),
 main="Compare ecdf with estimated cdf", xlab="inter")
curve(pgamma(x, shape=1.208593, rate=0.036350), add=TRUE)
legend(70, 0.8, legend=c("ecdf", "estimated cdf"), lty=1:2)
```

●●

Problems

10.17 Carry out a Shapiro-Wilk test for the mother's height, `ht`, and weight, `wt`, in the `babies` (`UsingR`) data set. Remember to exclude the cases when `ht==99` and `wt==999`. Does the sample data appear to come from a normal population in each case?

10.18 The `brightness` (`UsingR`) data set contains brightness measurements for 966 stars from the Hipparcos catalog. Is the data normal? Compare the result of a significance test with the graphical investigation done by

```
hist(brightness, prob=TRUE)
lines(density(brightness))
curve(dnorm(x, mean(brightness), sd(brightness)), add=TRUE)
```

10.19 The variable `temperature` in the data set `normtemp` (`UsingR`) contains normal body temperature measurements for 130 healthy, randomly selected individuals. Test the assumption that the sample data of normal body temperature comes from a normal distribution?

10.20 The `rivers` data set contains the length of 141 major rivers in North America. Fit this distribution using the gamma distribution and `fitdistr`. How well does the gamma distribution fit the data? Answer by graphing the empirical and estimated densities.

10.21 Find parameter estimates for μ and σ for the variables `sat.m` and `sat.v` in the `stud.recs` (`UsingR`) data set. Assume the respective populations are normally distributed.

10.22 How good is the Kolmogorov-Smirnov test at rejecting the null when it is false? The following command will do 1000 simulations of the test when the data is not normal, but long-tailed and symmetric.

```
res <- replicate(1000,  ks.test(rt(25, df=3),"pnorm")$p.value)
```

(The syntax above is using the fact that `ks.test` returns a list of values with one component named `p.value`.) What percentage of the trials have a p-value less than 0.05?

Try this with the exponential distribution (that is, replace `rt(25,df=3)` with `rexp(25)-1`). Is it better when the data is skewed?

10.23 A key to understanding why the Kolmogorov-Smirnov statistic has a sampling distribution that does not depend on the underlying parent population (as long as it is continuous) is the fact that if $F(x)$ is the c.d.f. for a random variable X, then $F(X)$ is uniformly distributed on $[0,1]$.

This can be proved algebraically using inverse functions, but instead we see how to simulate the problem to gain insight. The following line will illustrate this for the normal distribution:

```
qqplot(pnorm(rnorm(100)), runif(100))
```

The `qqplot` should be nearly straight if the distribution is uniform. Change the distribution to some others and see that you get a nearly straight line in each case. For example, the *t*-distribution with 5 degrees of freedom would be investigated with

```
qqplot(pt(rt(100,df=5),df=5),runif(100))
```

Try the uniform distribution, the exponential distribution, and the lognormal distribution (`lnorm`).

10.24 Is the Shapiro-Wilk test resistant to outliers? Run the following commands and decide whether the presence of a single large outlier (the 5) changes the ability of the test to detect normality.

```
shapiro.test(c(rnorm( 100), 5))
shapiro.test(c(rnorm(1000), 5))
shapiro.test(c(rnorm(4000), 5))
```