14.1.4 The article **"Application of Methods for Central Statistical Monitoring in Clinical Trials" (*Clinical Trials*, 2013: 783–806)** made a strong case for central statistical monitoring as an alternative to more expensive onsite data verification. It suggested various methods for identifying data characteristics such as outliers, incorrect dates, anomalous data patterns, unusual correlation structures, and digit preferences. Exercise 3.21 of this book introduced Benford's Law, which gives a probability model for the first significant digit in many large data sets: $p(x) = \log_{10}((x+1)/x)$ for x = 1, 2, …, 9. The cited article gave the following frequencies for the first significant digit in a variety of variables whose values were determined in one particular clinical trial:

Digit	1	2	3	4	5	6	7	8	9
Freq.	342	180	164	155	86	65	54	47	56

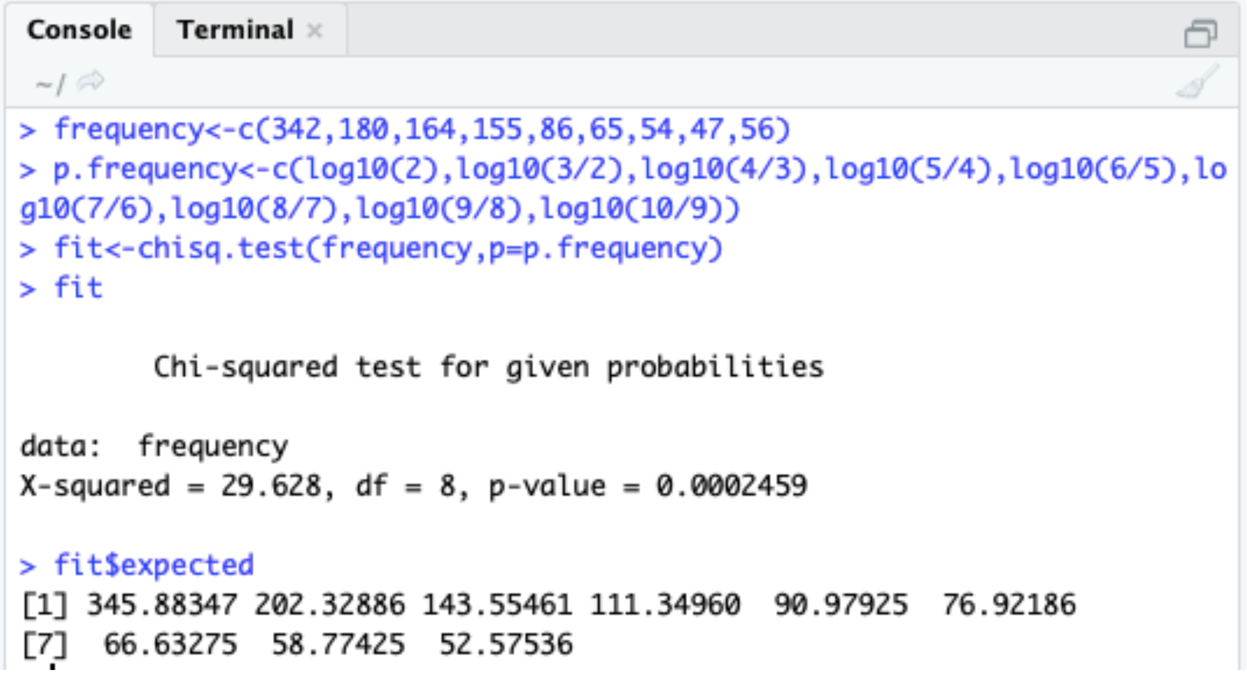
Carry out a test of hypotheses to see whether or not these frequencies are consistent with Benford's Law (the cited article gave P-value information).

$$n = 342 + 180 + 164 + 155 + 86 + 66 + 54 + 47 + 56 = 1149$$

$H_0: p(1) = \log_{10}\left(\frac{2}{1}\right), p(2) = \log_{10}\left(\frac{3}{2}\right), p(3) = \log_{10}\left(\frac{4}{3}\right), \ldots p(9) = \log_{10}\left(\frac{10}{9}\right)$

H_a: At least one of the $p(i)'$sis different

$\chi^2_{0.05,9-1} = \chi^2_{0.05,8} = 15.507 < 29.6282$. p-value = 0.0002459 (Reject null hypothesis)



```
> frequency<-c(342,180,164,155,86,65,54,47,56)
> p.frequency<-c(log10(2),log10(3/2),log10(4/3),log10(5/4),log10(6/5),lo
g10(7/6),log10(8/7),log10(9/8),log10(10/9))
> fit<-chisq.test(frequency,p=p.frequency)
> fit

        Chi-squared test for given probabilities

data:  frequency
X-squared = 29.628, df = 8, p-value = 0.0002459

> fit$expected
[1] 345.88347 202.32886 143.55461 111.34960  90.97925  76.92186
[7]  66.63275  58.77425  52.57536
```

Conclusion: There is sufficient evidence to reject the claim that the proportions are consistent with Benford's Law.

14.2.12 Consider a large population of families in which each family has exactly three children. If the genders of the three children in any family are independent of one another, the number of male children in a randomly selected family will have a binomial distribution based on three trials.

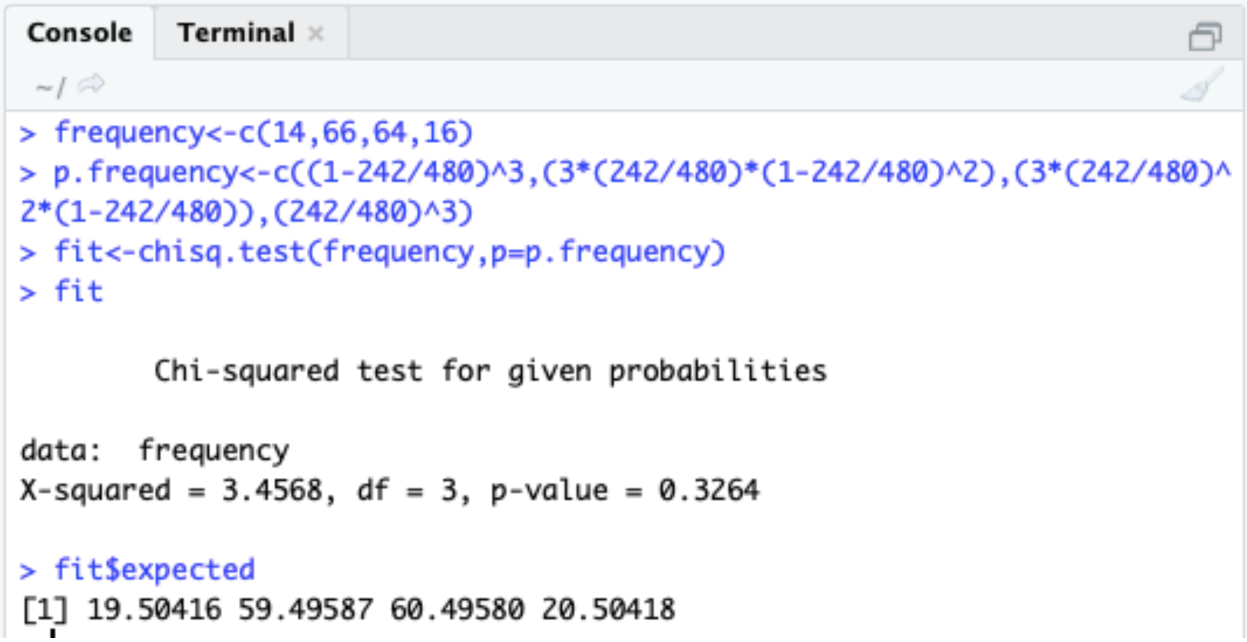
(a) Suppose a random sample of 160 families yields the following results. Test the relevant hypotheses by proceeding as in Example 14.5.

Number of Male Children	0	1	2	3
Frequency	14	66	64	16

$H_0: p_0 = \pi_0(\boldsymbol{\theta}), p_1 = \pi_1(\boldsymbol{\theta}), p_2 = \pi_2(\boldsymbol{\theta}), p_3 = \pi_3(\boldsymbol{\theta})$
H_a: at least one p_i differs
Maximum likelihood:

$$f(n_0, n_1, n_2, n_3; \theta) = \left[(1-\hat{\theta})^3\right]^{n_0} \cdot \left[3\hat{\theta}\left(1-\hat{\theta}\right)^2\right]^{n_1} \cdot \left[3\hat{\theta}^2(1-\hat{\theta})\right]^{n_2} \cdot \left[\hat{\theta}^3\right]^{n_3}$$
$$= 3^{n_1+n_2} \cdot \left(1-\hat{\theta}\right)^{3n_0+n_1+n_2} \cdot \hat{\theta}^{n_1+2n_2+3n_3}$$
$$\hat{\theta} = \frac{n_1 + 2n_2 + 3n_3}{3(n_0 + n_1 + n_2 + n_3)} = \frac{n_1 + 2n_2 + 3n_3}{3n} = \frac{66 + 2 \cdot 64 + 3 \cdot 16}{3 \cdot 160} = \frac{242}{480} = 0.5042$$

$$\pi_0(\hat{\boldsymbol{\theta}}) = \binom{3}{0}\theta^0(1-\theta)^{3-0} = (1-\theta)^3 = 0.1219, n\pi_0(\hat{\boldsymbol{\theta}}) = 160 \cdot 0.1219 = 19.5042$$
$$\pi_1(\hat{\boldsymbol{\theta}}) = \binom{3}{1}\theta^1(1-\theta)^{3-1} = 3\theta(1-\theta)^2 = 0.3718, n\pi_1(\hat{\boldsymbol{\theta}}) = 160 \cdot 0.3718 = 59.4959$$
$$\pi_2(\hat{\boldsymbol{\theta}}) = \binom{3}{2}\theta^2(1-\theta)^{3-2} = 3\theta^2(1-\theta) = 0.3781, n\pi_2(\hat{\boldsymbol{\theta}}) = 160 \cdot 0.3781 = 60.4958$$
$$\pi_3(\hat{\boldsymbol{\theta}}) = \binom{3}{3}\theta^3(1-\theta)^{3-3} = \theta^3 = 0.1282, n\pi_3(\hat{\boldsymbol{\theta}}) = 160 \cdot 0.1282 = 20.5042$$



```
> frequency<-c(14,66,64,16)
> p.frequency<-c((1-242/480)^3,(3*(242/480)*(1-242/480)^2),(3*(242/480)^
2*(1-242/480)),(242/480)^3)
> fit<-chisq.test(frequency,p=p.frequency)
> fit

	Chi-squared test for given probabilities

data:  frequency
X-squared = 3.4568, df = 3, p-value = 0.3264

> fit$expected
[1] 19.50416 59.49587 60.49580 20.50418
```

$\chi^2_{0.05,4-1-1} = \chi^2_{0.05,2} = 5.992 > 3.4568$ (do not reject null hypothesis)
Conclusion: There is no significant evidence that at least one of the proportions differ.

(b) Suppose a random sample of families in a nonhuman population resulted in observed frequencies of 15, 20, 12, and 3, respectively. Would the chi-squared test be based on the same number of degrees of freedom as the test in part (a)? Explain.
$H_0: p_0 = \pi_0(\boldsymbol{\theta}), p_1 = \pi_1(\boldsymbol{\theta}), p_2 = \pi_2(\boldsymbol{\theta}), p_3 = \pi_3(\boldsymbol{\theta})$
H_a: at least one p_i differs

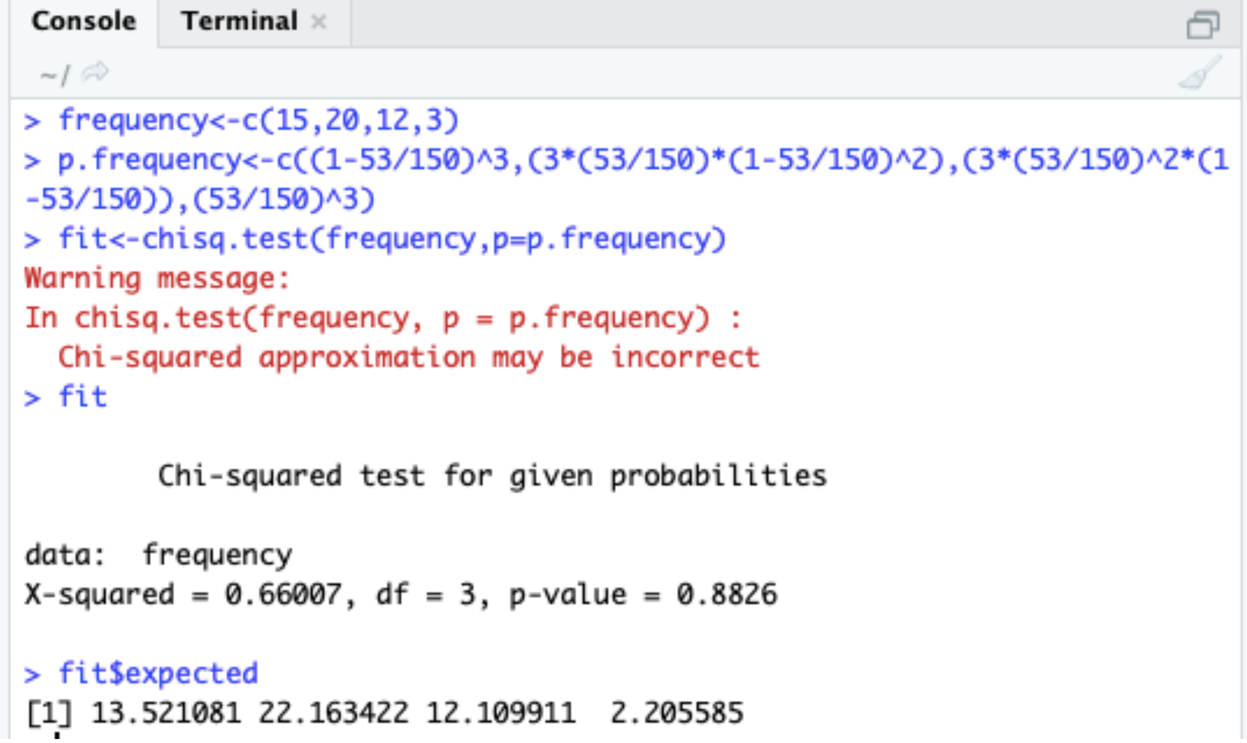
$$n = 15 + 20 + 12 + 3 = 50, \hat{\theta} = \frac{n_1 + 2n_2 + 3n_3}{3n} = \frac{20 + 2 \cdot 12 + 3 \cdot 3}{3 \cdot 50} = \frac{53}{150} = 0.3533$$

$$\pi_0(\widehat{\boldsymbol{\theta}}) = \binom{3}{0}\theta^0(1-\theta)^{3-0} = (1-\theta)^3 = 0.2704, n\pi_0(\widehat{\boldsymbol{\theta}}) = 50 \cdot 0.2704 = 13.5211$$

$$\pi_1(\widehat{\boldsymbol{\theta}}) = \binom{3}{1}\theta^1(1-\theta)^{3-1} = 3\theta(1-\theta)^2 = 0.4433, n\pi_1(\widehat{\boldsymbol{\theta}}) = 50 \cdot 0.4433 = 22.1634$$

$$\pi_2(\widehat{\boldsymbol{\theta}}) = \binom{3}{2}\theta^2(1-\theta)^{3-2} = 3\theta^2(1-\theta) = 0.2422, n\pi_2(\widehat{\boldsymbol{\theta}}) = 50 \cdot 0.2422 = 12.1099$$

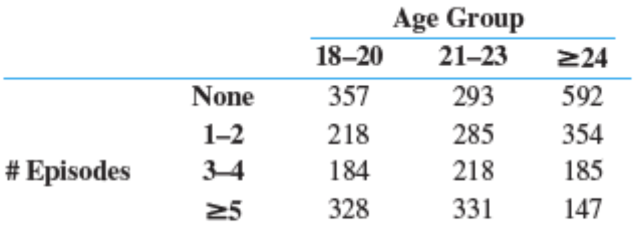
$$\pi_3(\widehat{\boldsymbol{\theta}}) = \binom{3}{3}\theta^3(1-\theta)^{3-3} = \theta^3 = 0.0441, n\pi_3(\widehat{\boldsymbol{\theta}}) = 50 \cdot 0.0441 = 2.2056$$



```
> frequency<-c(15,20,12,3)
> p.frequency<-c((1-53/150)^3,(3*(53/150)*(1-53/150)^2),(3*(53/150)^2*(1
-53/150)),(53/150)^3)
> fit<-chisq.test(frequency,p=p.frequency)
Warning message:
In chisq.test(frequency, p = p.frequency) :
  Chi-squared approximation may be incorrect
> fit

	Chi-squared test for given probabilities

data:  frequency
X-squared = 0.66007, df = 3, p-value = 0.8826

> fit$expected
[1] 13.521081 22.163422 12.109911  2.205585
```

Since $n\pi_3(\widehat{\boldsymbol{\theta}}) = 2.2056 < 5$, the mentioned test is not appropriate and *df* = 3 s not a good choice (as warned by R Studio).

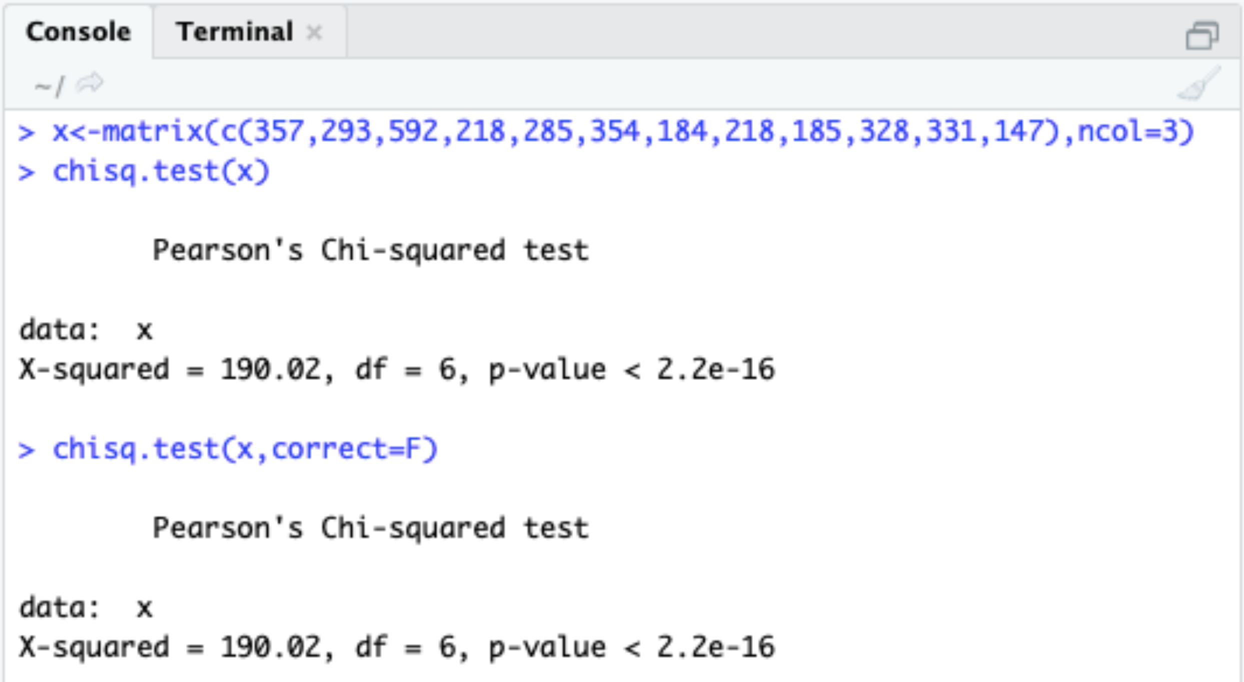
14.3.25 In an investigation of alcohol use among college students, each male student in a sample was categorized both according to age group and according to the number of heavy drinking episodes during the previous 30 days **("Alcohol Use in Students Seeking Primary Care Treatment at University Health Services,"** ***J. of Amer. College Health*****, 2012: 217–225)**.



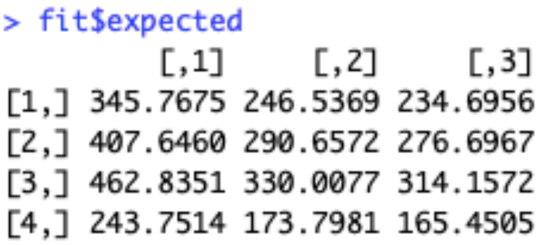
		Age Group		
		18–20	21–23	≥24
# Episodes	None	357	293	592
	1–2	218	285	354
	3–4	184	218	185
	≥5	328	331	147

Does there appear to be an associated between extend of binge drinking and age group in the population from which the sample was selected? Carry out a test of hypotheses at significance level .01.

H_0: the variables are independent; H_a: the variables are dependent; $\alpha = 0.01$



```
> x<-matrix(c(357,293,592,218,285,354,184,218,185,328,331,147),ncol=3)
> chisq.test(x)

        Pearson's Chi-squared test

data:  x
X-squared = 190.02, df = 6, p-value < 2.2e-16

> chisq.test(x,correct=F)

        Pearson's Chi-squared test

data:  x
X-squared = 190.02, df = 6, p-value < 2.2e-16
```



```
> fit$expected
         [,1]     [,2]     [,3]
[1,] 345.7675 246.5369 234.6956
[2,] 407.6460 290.6572 276.6967
[3,] 462.8351 330.0077 314.1572
[4,] 243.7514 173.7981 165.4505
```

Conclusion: Reject null hypothesis. There is sufficient evidence to support the claim of an association between the variables.