Chapter 9

9.1 Commentary

This section ends with a discussion of some issues related to the meaning of the χ^2 goodness-of-fit test for readers who want a deeper understanding of the procedure.

9.1 Solutions to Exercises

1. Let $Y = N_1$, the number of defective items, and let $\theta = p_1$, the probability that each item is defective. The level α_0 test requires us to choose c_1 and c_2 such that $\Pr(Y \le c_1|\theta = 0.1) + \Pr(Y \ge c_2|\theta = 0.1)$ is close to α_0. We can compute the probability that $Y = y$ for each $y = 0, \ldots, 100$ and arrange the numbers from smallest to largest. The smallest values correspond to large values of y down to $y = 25$, then some values corresponding to small values of y start to appear in the list. The sum of the values reaches 0.0636 when $c_1 = 4$ and $c_2 = 16$. So $\alpha_0 = 0.0636$ is the smallest α_0 for which we would reject $H_0 : \theta = 0.1$ using such a test.

2.
$$\begin{aligned} Q &= \sum_{i=1}^{k} \frac{(N_i - n/k)^2}{n/k} = \frac{k}{n}\sum_{i=1}^{k}\left(N_i^2 - 2\frac{n}{k}N_i + \frac{n^2}{k^2}\right) = \frac{k}{n}\left(\sum_{i=1}^{k} N_i^2 - 2\frac{n}{k}\sum_{i=1}^{k} N_i + \frac{n^2}{k}\right) \\ &= \frac{k}{n}\left(\sum_{i=1}^{k} N_i^2 - 2\frac{n^2}{k} + \frac{n^2}{k}\right) = \left(\frac{k}{n}\sum_{i=1}^{k} N_i^2\right) - n. \end{aligned}$$

3. We obtain the following frequencies:

i	0	1	2	3	4	5	6	7	8	9
N_i	25	16	19	20	20	22	24	15	14	25

Since $P_i^0 = 1/10$ for every value of i, and $n = 200$, we find from Exercise 2 or Eq. (9.1.2) that $Q = 7.4$. If Q has a χ^2 distribution with 9 degrees of freedom, $\Pr(Q \ge 7.4) = 0.6$.

4. We obtain the following table:

	AA	Aa	aa
N_i	10	10	4
np_i^0	6	12	6

It is found from Eq. (9.1.2) that $Q = 11/3$. If Q has a χ^2 distribution with 2 degrees of freedom, then the value of $\Pr(Q \ge 11/3)$ is between 0.1 and 0.2.

5. (a) The number of successes is $n\overline{X}_n$ and the number of failures is $n(1-\overline{X}_n)$. Therefore,

$$\begin{aligned} Q &= \frac{(n\overline{X}_n - np_0)^2}{np_0} + \frac{[n(1-\overline{X}_n) - n(1-p_0)]^2}{n(1-p_0)} \\ &= n(\overline{X}_n - p_0)^2 \left(\frac{1}{p_0} + \frac{1}{1-p_0}\right) \\ &= \frac{n(\overline{X}_n - p_0)^2}{p_0(1-p_0)} \end{aligned}$$

(b) If $p = p_0$, then $E(\overline{X}_n) = p_0$ and $\text{Var}(\overline{X}_n) = p_0(1-p_0)/n$. Therefore, by the central limit theorem, the d.f. of

$$Z = \frac{\overline{X}_n - p_0}{[p_0(1-p_0)/n]^{1/2}}$$

converges to the d.f. of the standard normal distribution. Since $Q = Z^2$, the d.f. of Q will converge to the d.f. of the χ^2 distribution with 1 degree of freedom.

6. Here, $p_0 = 0.3$, $n = 50$, and $\overline{X}_n = 21/50$. By Exercise 5, $Q = 3.44$. If Q has a χ^2 distribution with 1 degree of freedom, then $\Pr(Q \geq 3.4)$ is slightly greater than 0.05.

7. We obtain the following table:

	$0 < x < 0.2$	$0.2 < x < 0.5$	$0.5 < x < 0.8$	$0.8 < x < 1.$
N_i	391	490	580	339
np_i^0	360	540	540	360

If Q has a χ^2 distribution with 3 degrees of freedom, then $\Pr(Q \geq 11.34) = 0.01$. Therefore, we should reject H_0 if $Q \geq 11.34$. It is found from Eq. (9.1.2) that Q = 11.5.

8. If Z denotes a random variable having a standard normal distribution and X denotes the height of a man selected at random from the city, then

$$\begin{aligned} \Pr(X < 66) &= \Pr(Z < -2) = 0.0227, \\ \Pr(66 < X < 67.5) &= \Pr(-2 < Z < -0.5) = 0.2858, \\ \Pr(67.5 < X < 68.5) &= \Pr(-0.5 < Z < 0.5) = 0.3830, \\ \Pr(68.5 < X < 70) &= \Pr(0.5 < Z < 2) = 0.2858, \\ \Pr(X > 70) &= \Pr(Z > 2) = 0.0227. \end{aligned}$$

Therefore, we obtain the following table:

	N_i	np_i^0
$x < 66$	18	11.35
$66 < x < 67.5$	177	142.9
$67.5 < x < 68.5$	198	191.5
$68.5 < x < 70$	102	142.9
$x > 70$	5	11.35

It is found from Eq. (9.1.2) that $Q = 27.5$. If Q has a χ^2 distribution with 4 degrees of freedom, then $\Pr(Q \geq 27.5)$ is much less than 0.005.

The statistic Q is then 12.96. The two p-values for 10 and 11 degrees of freedom are 0.2258 and 0.2959.

7. There is no single correct answer to this problem. The M.L.E.'s $\hat{\mu} = \overline{X}_n$ and $\hat{\sigma}^2 = S_n^2/n$ should be calculated from the given observations. These observations should then be grouped into intervals and the observed number in each interval compared with the expected number in that interval if each of the 50 observations had a normal distribution with mean $\overline{X}_n$ and variance S_n^2/n. If the number of intervals is k, then when H_0 is true, the approximate distribution of the statistic Q will lie between the χ^2 distribution with $k-3$ degrees of freedom and the χ^2 distribution with $k-1$ degrees of freedom.

8. There is no single correct answer to this problem. The M.L.E. $\hat{\beta} = 1/\overline{X}_n$ of the parameter of the exponential distribution should be calculated from the given observations. These observations should then be grouped into intervals and the observed number in each interval compared with the expected number in that interval if each of the 50 observations had an exponential distribution with parameter $1/\overline{X}_n$. If the number of intervals is k, then when H_0 is true, the approximate distribution of the statistic Q will lie between a χ^2 distribution with $k-2$ degrees of freedom and a χ^2 distribution with $k-1$ degrees of freedom.

9.3 Solutions to Exercises.

1. Table S.9.1 contains the expected counts for this example. The value of the χ^2 statistic Q calculated

Table S.9.1: Expected cell counts for Exercise 1 of Section 9.3.

	Good grades	Athletic ability	Popularity
Boys	117.3	42.7	67.0
Girls	129.7	47.3	74.0

from these data is $Q = 21.5$. This should be compared to a χ^2 distribution with two degrees of freedom. The tail area can be calculated using statistical software as 2.2×10^{-5}.

2. $$Q = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(N_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = \sum_{i=1}^{R}\sum_{j=1}^{C}\left(\frac{N_{ij}^2}{\hat{E}_{ij}} - 2N_{ij} + \hat{E}_{ij}\right) = \left(\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{N_{ij}^2}{\hat{E}_{ij}}\right) - 2n + n$$
$$= \left(\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{N_{ij}^2}{\hat{E}_{ij}}\right) - n.$$

3. By Exercise 2,

$$Q = \sum_{i=1}^{R}\frac{N_{i1}^2}{\hat{E}_{i1}} + \sum_{i=1}^{R}\frac{N_{i2}^2}{\hat{E}_{i2}} - n.$$

But

$$\sum_{i=1}^{R}\frac{N_{i2}^2}{\hat{E}_{i2}} = \sum_{i=1}^{R}\frac{(N_{i+}-N_{i1})^2}{\hat{E}_{i2}} = \sum_{i=1}^{R}\frac{N_{i+}^2}{\hat{E}_{i2}} - 2\sum_{i=1}^{R}\frac{N_{i+}N_{i1}}{\hat{E}_{i2}} + \sum_{i=1}^{R}\frac{N_{i1}^2}{\hat{E}_{i2}}.$$

In the first two sums on the right, we let $\hat{E}_{i2} = N_{i+}N_{+2}/n$, and in the third sum we let $\hat{E}_{i2} = N_{+2}\hat{E}_{i1}/N_{+1}$. We then obtain

$$\sum_{i=1}^{R}\frac{N_{i2}^2}{\hat{E}_{i2}} = \frac{n}{N_{+2}}\sum_{i=1}^{R}N_{i+} - \frac{2n}{N_{+2}}\sum_{i=1}^{R}N_{i1} + \frac{N_{+1}}{N_{+2}}\sum_{i=1}^{R}\frac{N_{i1}^2}{\hat{E}_{i1}} = \frac{n^2}{N_{+2}} - 2n\frac{N_{+1}}{N_{+2}} + \frac{N_{+1}}{N_{+2}}\sum_{i=1}^{R}\frac{N_{i1}^2}{\hat{E}_{i1}}.$$

It follows that

$$Q = \left(1 + \frac{N_{+1}}{N_{+2}}\right)\sum_{i=1}^{R}\frac{N_{i1}^2}{\hat{E}_{i1}} + \frac{n}{N_{+2}}(n - 2N_{+1} - N_{+2}).$$

Since $n = N_{+1} + N_{+2}$,

$$Q = \frac{n}{N_{+2}}\sum_{i=1}^{R}\frac{N_{i1}^2}{\hat{E}_{i1}} - \frac{n}{N_{+2}}N_{+1}.$$

4. The values of $\hat{E}_{ij}$ are as given in the following table:

8	32
12	48

The value of Q is found from Eq. (9.3.4) or Exercise 3 to be 25/6. If Q has a χ^2 distribution with 1 degree of freedom, then $\Pr(Q \geq 25/6)$ lies between 0.025 and 0.05.

5. The values of $\hat{E}_{ij}$ are as given in the following table.

77.27	94.35	49.61	22.77
17.73	21.65	11.39	5.23

The value of Q is found from Eq. (9.3.4) or an adaptation of Exercise 3 to be 8.6. If Q has a χ^2 distribution with $(2-1)(4-1) = 3$ degrees of freedom, then $\Pr(Q \geq 8.6)$ lies between 0.025 and 0.05.

6. The values of $\hat{E}_{ij}$ are as given in the following table:

7.5	7.5
14.5	14.5

The value of Q is found from Eq. (9.3.4) or Exercise 3 to be 0.91. If Q has a χ^2 distribution with 1 degree of freedom, then $\Pr(Q \geq 0.91)$ lies between 0.3 and 0.4.

7. (a) The values of p_{i+} and p_{+j} are the marginal totals given in the following table:

			0.3
			0.3
			0.4
0.5	0.3	0.2	1.0

It can be verified that $p_{ij} = p_{i+}p_{+j}$ for each of the 9 entries in the table. It can be seen in advance that this relation will be satisfied for every entry in the table because it can be seen that the three rows of the table are proportional to each other or, equivalently, that the three columns are proportional to each other.

(b) Here is one example of a simulated data set

44	32	16	92
45	25	15	85
63	33	27	123
152	90	58	300

(c) The statistic Q calculated by any student from Eq. (9.3.4) will have a χ^2 distribution with $(3-1)(3-1)=4$ degrees of freedom. For the data in part (b), the table of $\hat{E}_{ij}$ values is

46.6	27.6	17.8
43.1	25.5	16.4
62.3	36.9	23.8

The value of Q is then 2.105. The p-value 0.7165.

8. To test whether the values obtained by n different students form a random sample of size n from a χ^2 distribution with 4 degrees of freedom, follow these steps: (1) Partition the positive part of the real line into k intervals; (2) Determine the probabilities $p_1^0, \ldots, p_k^0$ of these intervals for the χ^2 distribution with 4 degrees of freedom; (3) Calculate the value of the statistic Q given by Eq. (9.1.2). If the hypothesis H_0 is true, this statistic Q will have approximately a χ^2 distribution with $k-1$ degrees of freedom.

9. Let N_{ijk} denote the number of observations in the random sample that fall into the (i,j,k) cell, and let

$$N_{i++} = \sum_{j=1}^{C}\sum_{k=1}^{T} N_{ijk}, N_{+j+} = \sum_{i=1}^{R}\sum_{k=1}^{T} N_{ijk},$$
$$N_{++k} = \sum_{i=1}^{R}\sum_{j=1}^{C} N_{ijk}.$$

Then the M.L.E.'s are

$$\hat{p}_{i++} = \frac{N_{i++}}{n}, \hat{p}_{+j+} = \frac{N_{+j+}}{n}, \hat{p}_{++k} = \frac{N_{++k}}{n}.$$

Therefore, when H_0 is true,

$$\hat{E}_{ijk} = n\hat{p}_{i++}\hat{p}_{+j+}\hat{p}_{++k} = \frac{N_{i++}N_{+j+}N_{++k}}{n^2}.$$

Since $\sum_{i=1}^{R}\hat{p}_{i++} = \sum_{j=1}^{C}\hat{p}_{+j+} = \sum_{k=1}^{T}\hat{p}_{++k} = 1$, the number of parameters that have been estimated is $(R-1)+(C-1)+(T-1) = R+C+T-3$. Therefore, when H_0 is true, the approximate distribution of

$$Q = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{T}\frac{(N_{ijk}-\hat{E}_{ijk})^2}{\hat{E}_{ijk}}$$

will be a χ^2 distribution for which the number of degrees of freedom is $RCT-1-(R+C+T-3) = RCT-R-C-T+2$.