[10-09-23-Tr 12]

Table 11-2
Last Digits of Weights

Last Digit	Frequency
0	35
1	0
2	2
3	1
4	4
5	24
6	1
7	4
8	7
9	2

SOLUTION Here is the verification that the four conditions of a multinomial experiment are all satisfied:

1. The number of trials (last digits) is the fixed number 80.
2. The trials are independent, because the last digit of any individual weight does not affect the last digit of any other weight.
3. Each outcome (last digit) is classified into exactly 1 of 10 different categories. The categories are identified as 0, 1, 2, . . . , 9.
4. In testing the claim that the 10 digits are equally likely, each possible digit has a probability of 1/10, and by assumption, that probability remains constant for each subject.

In this section we are presenting a method for testing a claim that in a multinomial experiment, the frequencies observed in the different categories fit some claimed distribution. Because we test for how well an observed frequency distribution fits some specified theoretical distribution, this method is often called a *goodness-of-fit test.*

Definition

A **goodness-of-fit test** is used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution.

For example, using the data in Table 11-2, we can test the hypothesis that the data fit a uniform distribution, with all of the digits being equally likely. Our goodness-of-fit tests will incorporate the following notation.

Notation

O represents the *observed frequency* of an outcome.
E represents the *expected frequency* of an outcome.
k represents the *number of different categories* or outcomes.
n represents the total *number of trials.*

Finding Expected Frequencies

In Table 11-2 the observed frequencies O are 35, 0, 2, 1, 4, 24, 1, 4, 7, and 2. The sum of the observed frequencies is 80, so $n = 80$. If we assume that the 80 digits were obtained from a population in which all digits are equally likely, then we *expect* that each digit should occur in 1/10 of the 80 trials, so each of the 10 expected frequencies is given by $E = 8$. If we generalize this result, we get an easy procedure for finding expected frequencies whenever we are assuming that all of the expected frequencies are equal: Simply divide the total number of observations by the number of different categories ($E = n/k$). In other cases where the expected frequencies are not all equal, we can often find the expected frequency for each category by multiplying the sum of all observed frequencies and the probability p for the category, so $E = np$. We summarize these two procedures here.

Note to Instructor

It would be helpful to briefly review the concept of *expected value,* introduced in Section 5-2. Present a few simple and obvious examples such as this one: Find the expected number of girls born in groups of 100 babies. When students respond with the correct answer of 50, ask them to describe the exact thought process that led to the answer. They will respond that they found 1/2 of 100, which can be generalized as $p \times n$, which leads to $E = np$. Also point out that the expected number of girls among 3 babies is 1.5, so the expected value need not be an integer.

Also, emphasize that the methods of this section require that each *expected frequency* must be at least 5, but there is no requirement that *observed frequencies* must be at least 5.

- **If all expected frequencies are equal, then each expected frequency is the sum of all observed frequencies divided by the number of categories, so that $E = n/k$.**
- **If the expected frequencies are not all equal, then each expected frequency is found by multiplying the sum of all observed frequencies by the probability for the category, so $E = np$ for each category.**

As good as these two formulas for E might be, it would be better to use an informal approach based on an understanding of the circumstances. Just ask, "How can the observed frequencies be split up among the different categories so that there is perfect agreement with the claimed distribution?" Also, recognize that the *observed* frequencies must all be whole numbers because they represent actual counts, but *expected* frequencies need not be whole numbers. For example, when rolling a single die 33 times, the expected frequency for each possible outcome is $3/6 = 5.5$. The expected frequency for the number of 3s occurring is 5.5, even though it is impossible to have the outcome of 3 occur exactly 5.5 times.

We know that sample frequencies typically deviate somewhat from the values we theoretically expect, so we now present the key question: Are the differences between the actual *observed* values O and the theoretically *expected* values E statistically significant? We need a measure of the discrepancy between the O and E values, so we use the test statistic that is given with the requirements and critical values. (Later, we will explain how this test statistic was developed, but you can see that it has differences of $O - E$ as a key component.)

Requirements

1. The data have been randomly selected.
2. The sample data consist of frequency counts for each of the different categories.
3. For each category, the *expected* frequency is at least 5. (The expected frequency for a category is the frequency that would occur if the data actually have the distribution that is being claimed. There is no requirement that the *observed* frequency for each category must be at least 5.)

Test Statistic for Goodness-of-Fit Tests in Multinomial Experiments

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Critical values

1. Critical values are found in Table A-4 by using $k - 1$ degrees of freedom, where k = number of categories.
2. Goodness-of-fit hypothesis tests are always *right-tailed.*

Note to Instructor

About the χ^2 notation: We usually use Greek letters for population parameters, but here we use χ^2 for a test statistic. Although not consistent, this is very common notation in this context. (Ralph Waldo Emerson said that "a foolish consistency is the hobgoblin of little minds.")

Try to lead the class to developing their own reasoning process for why the tests of this section are all right-tailed. Ask the class these questions: If there are large discrepancies between the observed frequencies and those that are expected, what do we know about the $O - E$ values? The $(O - E)^2$ values? The value of the χ^2 test statistic? Where on the χ^2 distribution do large discrepancies fall?

TATISTICS N THE NEWS

afest Airplane eats

Many of us believe that the ear seats are safest in an airplane crash. Safety experts do not agree that any particular part of an airplane is safer than others. Some planes crash nose first when they come down, but others crash tail first on takeoff. Matt McCormick, a survival expert for the National Transportation Safety Board, told *Travel* magazine that "there is no one safe place to sit." Goodness-of-fit tests can be used with a null hypothesis that all sections of an airplane are equally safe. Crashed airplanes could be divided into the front, middle, and rear sections. The observed frequencies of fatalities could then be compared to the frequencies that would be expected with a uniform distribution of fatalities. The χ^2 test statistic reflects the size of the discrepancies between observed and expected frequencies, and it would reveal whether some sections are safer than others.

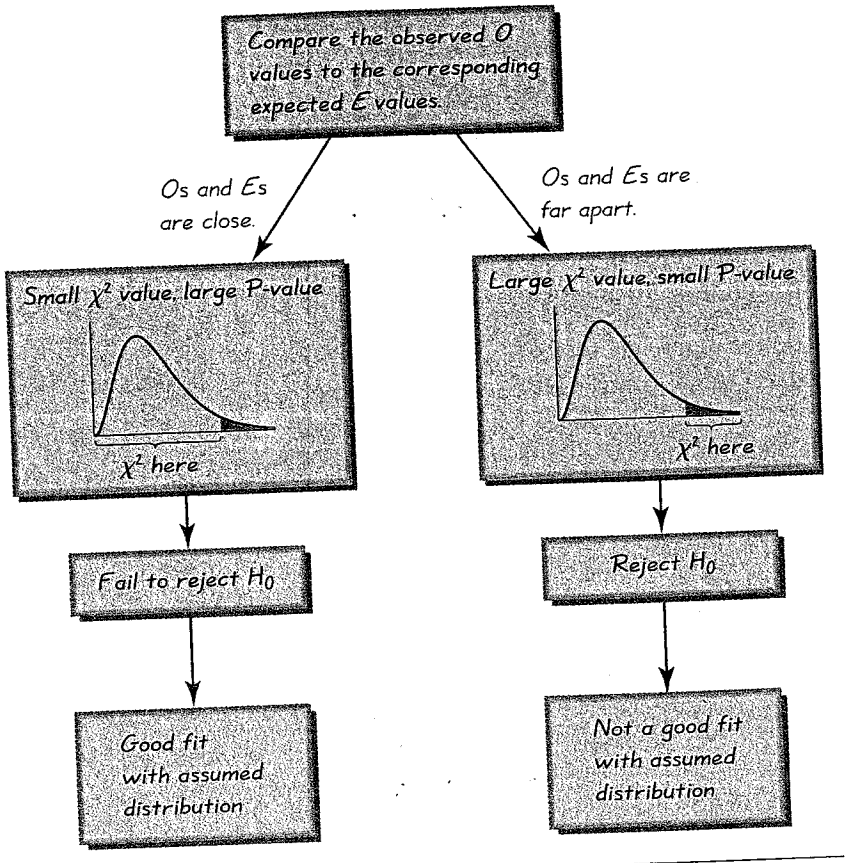

Figure 11-3 Relationships Among the χ^2 Test Statistic, *P*-Value, and Goodness-of-Fit

The χ^2 test statistic is based on differences between observed and expected values, so *close agreement* between observed and expected values will lead to a *small* value of χ^2 and a *large* *P*-value. A large discrepancy between observed and expected values will lead to a *large* value of χ^2 and a *small* *P*-value. The hypothesis tests of this section are therefore always right-tailed, because the critical value and critical region are located at the extreme right of the distribution. These relationships are summarized and illustrated in Figure 11-3.

Once we know how to find the value of the test statistic and the critical value, we can test hypotheses by using the same general procedures introduced in Chapter 8.

EXAMPLE Last Digit Analysis of Weights: Equal Expected Frequencies See Table 11-2 for the last digits of 80 weights. Test the claim that the digits do *not* occur with the same frequency. Based on the results, what can we conclude about the procedure used to obtain the weights?

SOLUTION

REQUIREMENT ✔ We require that the sample data are randomly selected, they consist of frequency counts, the data come from a multinomial experiment, and each expected frequency must be at least 5. We have noted earlier

that the data come from randomly selected students. The data do consist of frequency counts. The preceding example established that the conditions for a multinomial experiment are satisfied. The preceding discussion of expected values included the result that each expected frequency is 8, so each expected frequency does satisfy the requirement of being a value of at least 5. All of the requirements are satisfied and we can proceed with the hypothesis test. ✔

The claim that the digits do not occur with the same frequency is equivalent to the claim that the relative frequencies or probabilities of the 10 cells (p_0, $p_1, \ldots, p_9$) are not all equal. We will use the traditional method for testing hypotheses (see Figure 8-9).

Step 1: The original claim is that the digits do not occur with the same frequency. That is, at least one of the probabilities $p_0, p_1, \ldots, p_9$ is different from the others.

Step 2: If the original claim is false, then all of the probabilities are the same. That is, $p_0 = p_1 = \cdots = p_9$.

Step 3: The null hypothesis must contain the condition of equality, so we have

H_0: $p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9$

H_1: At least one of the probabilities is different from the others.

Step 4: No significance level was specified, so we select $\alpha = 0.05$, a very common choice.

Step 5: Because we are testing a claim about the distribution of the last digits being a uniform distribution, we use the goodness-of-fit test described in this section. The χ^2 distribution is used with the test statistic given earlier.

Step 6: The observed frequencies O are listed in Table 11-2. Each corresponding expected frequency E is equal to 8 (because the 80 digits would be uniformly distributed through the 10 categories). Table 11-3 shows the computation of the χ^2 test statistic. The test statistic is $\chi^2 = 156.500$. The critical value is $\chi^2 = 16.919$ (found in Table A-4 with $\alpha = 0.05$ in the right tail and degrees of freedom equal to $k - 1 = 9$). The test statistic and critical value are shown in Figure 11-4.

Step 7: Because the test statistic falls within the critical region, there is sufficient evidence to reject the null hypothesis.

Step 8: There is sufficient evidence to support the claim that the last digits do not occur with the same relative frequency. We now have very strong evidence suggesting that the weights were not actually measured. It is reasonable to speculate that they were reported values instead of actual measurements.

The preceding example dealt with the null hypothesis that the probabilities for the different categories are all equal. The methods of this section can also be used when the hypothesized probabilities (or frequencies) are different, as shown in the next example.

Table 11-3 Calculating the χ^2 Test Statistic for the Last Digits of Weights

Last Digit	Observed Frequency O	Expected Frequency E	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
0	35	8	27	729	91.1250
1	0	8	−8	64	8.0000
2	2	8	−6	36	4.500
3	1	8	−7	49	6.125
4	4	8	−4	16	2.000
5	24	8	16	256	32.000
6	1	8	−7	49	6.125
7	4	8	−4	16	2.000
8	7	8	−1	1	0.125
9	2	8	−6	36	4.500
	80	80			$\chi^2 = \sum \frac{(O - E)^2}{E} = 156.500$

(Except for rounding errors, these two totals must agree.)

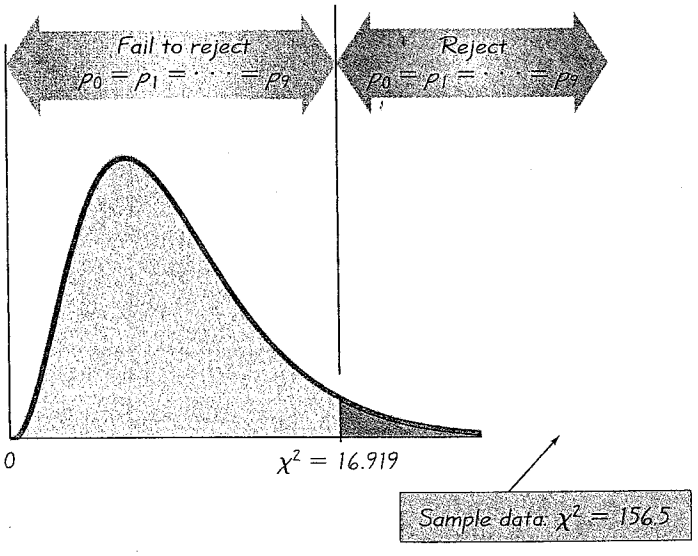

Figure 11-4 Test of $p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9$

EXAMPLE Detecting Fraud: Unequal Expected Frequencies In the Chapter Problem, it was noted that statistics is sometimes used to detect fraud. The second row of Table 11-1 lists percentages for leading digits as expected from Benford's law, and the third row lists the frequency counts expected when the Benford's law percentages are applied to 784 leading digits. The bottom row of Table 11-1 lists the observed frequencies of the leading digits from amounts on 784 checks issued by seven different companies. Test the claim that there is a significant discrepancy between the leading digits expected from Benford's law and the leading digits observed on the 784 checks. Use a significance level of 0.01.

Note to Instructor
Students often have difficulty finding the expected frequencies when the proportions are not all equal, as in this example. It would be helpful to carefully explain how this is done.

SOLUTION

REQUIREMENTS ✔ In checking the three requirements listed earlier, we begin by noting that the leading digits from the checks are not actually random. However, we treat them as random for the purpose of determining whether they are typical results that might be obtained from a random sample following Benford's law. The data are listed as frequency counts. They satisfy the requirements of a multinomial experiment. Each expected frequency (shown in Table 11-1) is at least 5. All of the requirements are satisfied and we can proceed with the hypothesis test. ✔

Step 1: The original claim is that the leading digits do not have the same distribution as claimed by Benford's law. That is, at least one of the following equations is wrong: $p_1 = 0.301$ and $p_2 = 0.176$ and $p_3 = 0.125$ and $p_4 = 0.097$ and $p_5 = 0.079$ and $p_6 = 0.067$ and $p_7 = 0.058$ and $p_8 = 0.051$ and $p_9 = 0.046$. (The proportions are the decimal equivalent values of the percentages listed for Benford's law in Table 11-1.)

Step 2: If the original claim is false, then the following are all true: $p_1 = 0.301$ and $p_2 = 0.176$ and $p_3 = 0.125$ and $p_4 = 0.097$ and $p_5 = 0.079$ and $p_6 = 0.067$ and $p_7 = 0.058$ and $p_8 = 0.051$ and $p_9 = 0.046$.

Step 3: The null hypothesis must contain the condition of equality, so we have

H_0: $p_1 = 0.301$ and $p_2 = 0.176$ and $p_3 = 0.125$ and $p_4 = 0.097$ and $p_5 = 0.079$ and $p_6 = 0.067$ and $p_7 = 0.058$ and $p_8 = 0.051$ and $p_9 = 0.046$

H_1: At least one of the proportions is not equal to the given claimed value.

Step 4: The significance level of $\alpha = 0.01$ was specified.

Step 5: Because we are testing a claim about the distribution of digits conforming to the distribution from Benford's law, we use the goodness-of-fit test described in this section. The χ^2 distribution is used with the test statistic given earlier.

Step 6: The observed frequencies O and the expected frequencies E are shown in Table 11-1. Adding the nine $(O - E)^2/E$ values results in the test statistic of $\chi^2 = 3650.251$. The critical value is $\chi^2 = 20.090$ (found in Table A-4 with $\alpha = 0.01$ in the right tail and degrees of freedom equal to $k - 1 = 8$). The test statistic and critical value are shown in Figure 11-5.

continued

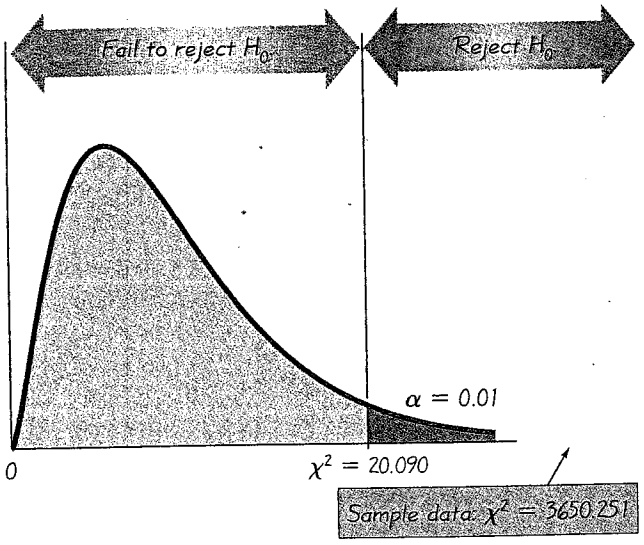

Figure 11-5
Testing for Agreement Between Observed Frequencies and Frequencies Expected with Benford's Law

Step 7: Because the test statistic falls within the critical region, there is sufficient evidence to reject the null hypothesis.

Step 8: There is sufficient evidence to support the claim that there is a discrepancy between the distribution expected from Benford's law and the observed distribution of leading digits from the checks.

In Figure 11-6(a) we graph the claimed proportions of 0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051, and 0.046 along with the observed proportions of 0.000, 0.019, 0.000, 0.097, 0.611, 0.233, 0.010, 0.029, and 0.000, so that we can visualize the discrepancy between the Benford's law distribution that was claimed and the frequencies that were observed. The points along the red line represent the claimed proportions, and the points along the green line represent the observed proportions. The corresponding pairs of points are far apart, showing that the expected frequencies are very different from the corresponding observed frequencies. The great disparity between the green line for observed frequencies and the red line for expected frequencies suggests that the check amounts are not the result of typical transactions. It appears that fraud may be involved. In fact, the Brooklyn District Attorney charged fraud by using this line of reasoning. For comparison, see Figure 11-6(b), which is based on the leading digits from the amounts on the last 200 checks written by the author. Note how the observed proportions from the author's checks agree quite well with the proportions expected with Benford's law. The author's checks appear to be typical instead of showing a pattern that might suggest fraud. In general, graphs such as Figure 11-6 are helpful in visually comparing expected frequencies and observed frequencies, as well as suggesting which categories result in the major discrepancies.

P-Values

The examples in this section used the traditional approach to hypothesis testing, but the *P*-value approach can also be used. *P*-values are automatically provided by STATDISK or the TI-83/84 Plus calculator, or they can be obtained by using