Tests of Significance—Chi-Square and Slope of Least Squares Line

TOPIC 15

- Chi-Square Test for Goodness of Fit
- Chi-Square Test for Independence
- Chi-Square Test for Homogeneity of Proportions
- Hypothesis Test for Slope of Least Squares Line

In this topic we continue our development of tools to analyze data. We learn about inference on distributions of counts using chi-square models. This can be used to solve such problems as "Do test results support Mendel's genetic principles?" (goodness-of-fit test); "Was surviving the Titanic sinking independent of a passenger's status?" (independence test); and "Do students, teachers, and staff show the same distributions in types of cars driven?" (homogeneity test). We then learn about inference with regard to linear association of two variables. This can be used to solve such problems as "Is there a linear relationship between the grade received on a term paper and the number of pages turned in?"

TIP

Unless you have counts, you cannot use χ^2 methods.

CHI-SQUARE TEST FOR GOODNESS OF FIT

A critical question is often whether or not an observed pattern of data fits some given distribution. A perfect fit cannot be expected, and so we must look at discrepancies and make judgments as to the *goodness of fit.*

One approach is similar to that developed earlier. There is the null hypothesis of a good fit, that is, the hypothesis that a given theoretical distribution correctly describes the situation, problem, or activity under consideration. Our observed data consist of one possible sample from a whole universe of possible samples. We ask about the chance of obtaining a sample with the observed discrepancies if the null hypothesis is really true. Finally, if the chance is too small, we reject the null hypothesis and say that the fit is not a good one.

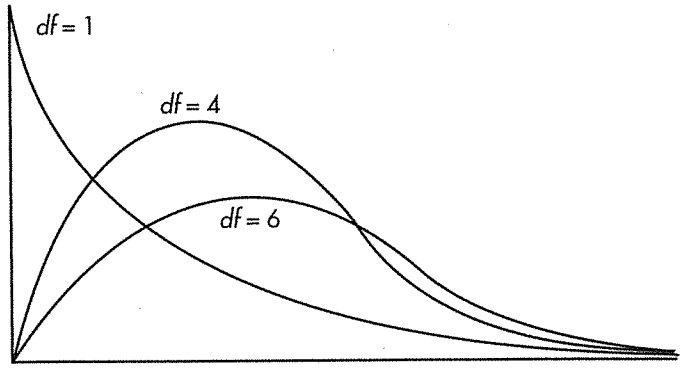
How do we decide about the significance of observed discrepancies? It should come as no surprise that the best information is obtained from squaring the discrepancy values, as this has been our technique for studying variances from the

beginning. Furthermore, since, for example, an observed difference of 23 is more significant if the original values are 105 and 128 than if they are 10,602 and 10,625, we must appropriately *weight* each difference. Such weighting is accomplished by dividing each difference by the expected values. The sum of these weighted differences or discrepancies is called *chi-square* and is denoted as χ^2 (χ is the lowercase Greek letter chi):

$$\chi^2 = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$$

The smaller the resulting χ^2-value, the better the fit. The *P*-value is the probability of obtaining a χ^2 value as extreme as the one obtained if the null hypothesis is assumed true. If the χ^2 value is large enough, that is, if the *P*-value is small enough, we say there is sufficient evidence to reject the null hypothesis and to claim that the fit is poor.

To decide how large a calculated χ^2-value must be to be significant, that is, to choose a critical value, we must understand how χ^2-values are distributed. A χ^2-distribution is not symmetric and is always skewed to the right. There are distinct χ^2-distributions, each with an associated number of degrees of freedom (*df*). The larger the *df* value, the closer the χ^2-distribution to a normal distribution. Note, for example, that squaring the often-used *z*-scores 1.645, 1.96, and 2.576 results in 2.71, 3.84, and 6.63, respectively, which are entries found in the first row of the χ^2-distribution table.



A large χ^2-value may or may not be significant—the answer depends on which χ^2-distribution we are using. A table is given of critical χ^2-values for the more commonly used percentages or probabilities. To use the χ^2-distribution for approximations in goodness-of-fit problems, the individual expected values cannot be too small. An often-used rule of thumb is that no expected value should be less than 5. Finally, as in all hypothesis tests we've looked at, the sample should be randomly chosen from the given population.

EXAMPLE 15.1

In a recent year, at the 6 p.m. time slot, television channels 2, 3, 4, and 5 captured the entire audience with 30%, 25%, 20%, and 25%, respectively. During the first week of the next season, 500 viewers are interviewed.

a. If viewer preferences have not changed, what number of persons is expected to watch each channel?
Answer: .30(500) = 150, .25(500) = 125, .20(500) = 100, and .25(500) = 125, so we have

	Channel 2	3	4	5
Expected number	150	125	100	125

b. Suppose that the actual observed numbers are as follows:

	Channel 2	3	4	5
Observed number	139	138	112	111

Do these numbers indicate a change? Are the differences significant?
Answer: Check the conditions:

1. *Randomization*: We must assume that the 500 viewers are a representative sample.
2. We note that the expected values (150, 125, 100, 125) are all ≥ 5.

H_0: The television audience is distributed over channels 2, 3, 4, and 5 with percentages 30%, 25%, 20%, and 25%, respectively.
H_a: The audience distribution is not 30%, 25%, 20%, and 25%, respectively.

We calculate

$$\chi^2 = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$$

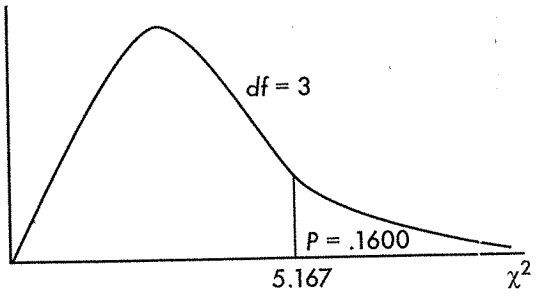
$$= \frac{(139-150)^2}{150} + \frac{(138-125)^2}{125} + \frac{(112-100)^2}{100} + \frac{(111-125)^2}{125}$$

$$= 5.167$$

Then the P-value is $P = P(\chi^2 > 5.167) = .1600$. [With $df = n - 1 = 3$, the TI-84 gives χ^2cdf(5.167,1000,3) = .1600, or a direct test for goodness of fit can be downloaded onto older TI-84+ calculators and comes preloaded on most!]

Conclusion:

With this large a *P*-value (.1600) there is not sufficient evidence to reject H_0. That is, there is not sufficient evidence that viewer preferences have changed.



Note: While the TI-83 does not have the goodness of fit download that is available on the TI-84, one can still use a TI-83 to calculate the above χ^2 by putting the observed values in list L1 and the expected values in L2, then calculating $(L1 - L2)^2/L2 \rightarrow L3$ and $\chi^2 = \text{sum}(L3)$, where "sum" is found under LIST → MATH.

EXAMPLE 15.2

A grocery store manager wishes to determine whether a certain product will sell equally well in any of five locations in the store. Five displays are set up, one in each location, and the resulting numbers of the product sold are noted.

	Location 1	2	3	4	5
Actual number sold	43	29	52	34	48

Is there enough evidence that location makes a difference? Test at both the 5% and 10% significance levels.

Answer:

H_0: Sales of the product are uniformly distributed over the five locations.
H_a: Sales are not uniformly distributed over the five locations.

A total of 43 + 29 + 52 + 34 + 48 = 206 units were sold. If location doesn't matter, we would expect $\frac{206}{5} = 41.2$ units sold per location (uniform distribution).

	Location 1	2	3	4	5
Expected number sold	41.2	41.2	41.2	41.2	41.2

Check the conditions:

1. *Randomization:* We must assume that the 206 units sold are a representative sample.
2. We note that the expected values (all 41.2) are all ≥ 5.

Thus

$$\chi^2 = \frac{(43-41.2)^2}{41.2} + \frac{(29-41.2)^2}{41.2} + \frac{(52-41.2)^2}{41.2} + \frac{(34-41.2)^2}{41.2} + \frac{(48-41.2)^2}{41.2}$$
$$= 8.903$$

The number of degrees of freedom is the number of classes minus 1; that is, $df = 5 - 1 = 4$.

The P-value is $P = P(\chi^2 > 8.903) = .0636$. [On the TI-84: χ^2cdf(8.903,1000,4).]

With $P = .0636$ there is sufficient evidence to reject H_0 at the 10% level but not at the 5% level. If the grocery store manager is willing to accept a 10% chance of committing a Type I error, there is enough evidence to claim location makes a difference.

CHI-SQUARE TEST FOR INDEPENDENCE

In the goodness-of-fit problems above, a set of expectations was based on an assumption about how the distribution should turn out. We then tested whether an observed sample distribution could reasonably have come from a larger set based on the assumed distribution.

In many real-world problems we want to compare two or more observed samples without any prior assumptions about an expected distribution. In what is called a *test of independence*, we ask whether the two or more samples might reasonably have come from some larger set. For example, do nonsmokers, light smokers, and heavy smokers all have the same likelihood of being eventually diagnosed with cancer, heart disease, or emphysema? Is there a relationship (association) between smoking status and being diagnosed with one of these diseases?

We classify our observations in two ways and then ask whether the two ways are independent of each other. For example, we might consider several age groups and within each group ask how many employees show various levels of job satisfaction. The null hypothesis is that age and job satisfaction are independent, that is, that the proportion of employees expressing a given level of job satisfaction is the same no matter which age group is considered.

Analysis involves calculating a table of *expected* values, assuming the null hypothesis about independence is true. We then compare these expected values with the observed values and ask whether the differences are reasonable if H_0 is true. The significance of the differences is gauged by the same χ^2-value of weighted squared differences. The smaller the resulting χ^2-value, the more reasonable the null hypothesis of independence. If the χ^2-value is large enough, that is, if the P-value is small enough, we can say that the evidence is sufficient to reject the null hypothesis and to claim that there *is* some relationship between the two variables or methods of classification.

In this type of problem,

$$df = (r-1)(c-1)$$

where df is the number of degrees of freedom, r is the number of rows, and c is the number of columns.

A point worth noting is that even if there is sufficient evidence to reject the null hypothesis of independence, we cannot necessarily claim any direct *causal* relationship. In other words, although we can make a statement about some link or relationship between

two variables, we are *not* justified in claiming that one causes the other. For example, we may demonstrate a relationship between salary level and job satisfaction, but our methods would not show that higher salaries cause higher job satisfaction. Perhaps an employee's higher job satisfaction impresses his superiors and thus leads to larger increases in pay. Or perhaps there is a third variable, such as training, education, or personality, that has a direct causal relationship to both salary level and job satisfaction.

EXAMPLE 15.3

In a nationwide telephone poll of 1000 adults representing Democrats, Republicans, and Independents, respondents were asked two questions: their party affiliation and if their confidence in the U.S. banking system had been shaken by the savings and loan crisis. The answers, cross-classified by party affiliation, are given in the following *contingency table.*

	Confidence Shaken		
Observed	Yes	No	No opinion
Democrats	175	220	55
Republicans	150	165	35
Independents	75	105	20

Test the null hypothesis that shaken confidence in the banking system is independent of party affiliation. Use a 10% significance level.

Answer:

H_0: Party affiliation and shaken confidence in the banking system are independent.
H_a: Party affiliation and shaken confidence in the banking system are not independent.

The above table gives the observed results. To determine the expected values, we must first determine the row and column totals:

Row totals: 175 + 220 + 55 = 450,
150 + 165 + 35 = 350,
75 + 105 + 20 = 200.

Column totals: 175 + 150 + 75 = 400,
220 + 165 + 105 = 490,
55 + 35 + 20 = 110.

	Yes	No	No opinion	
Democrats				450
Republicans				350
Independents				200
	400	490	110	

To calculate, for example, the expected value in the upper left box, we can proceed in any of several equivalent ways. First, we could note that the proportion of Democrats is $\frac{450}{1000} = .45$; and so, if independent, the expected number of Democrat *yes* responses is .45(400) = 180. Instead, we could note that the proportion of *yes* responses is $\frac{400}{1000} = .4$; and

so, if independent, the expected number of Democrat *yes* responses is .4(450) = 180. Finally, we could note that both these calculations simply involve $\frac{(450)(400)}{1000} = 180$.

In other words, *the expected value of any box can be calculated by multiplying the corresponding row total by the appropriate column total and then dividing by the grand total.* Thus, for example, the expected value for the middle box, which corresponds to Republican *no* responses, is $\frac{(350)(490)}{1000} = 171.5$.

Continuing in this manner, we fill in the table as follows:

Expected

	Yes	No	No opinion	
Democrats	180	220.5	49.5	450
Republicans	140	171.5	38.5	350
Independents	80	98	22	200
	400	490	110	

[An appropriate check at this point is that each expected cell count is at least 5.]

Next we calculate the value of chi-square:

$$\chi^2 = \frac{(175-180)^2}{180} + \frac{(220-220.5)^2}{220.5} + \frac{(55-49.5)^2}{49.5}$$
$$+ \frac{(150-140)^2}{140} + \frac{(165-171.5)^2}{171.5} + \frac{(35-38.5)^2}{38.5}$$
$$+ \frac{(75-80)^2}{80} + \frac{(100-98)^2}{98} + \frac{(20-22)^2}{22}$$
$$= 3.024$$

[On the TI-84, go to MATRIX and EDIT. Put the data into a matrix. Then STAT, TESTS, χ^2-Test, will give $\chi^2 = 3.0243$. Note also that the expected values are automatically stored in a second matrix.]

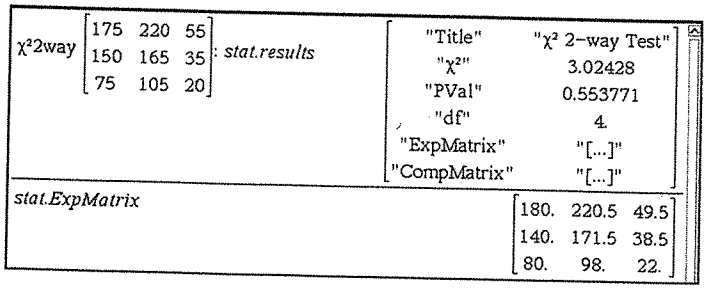
Note that, once the 180, 220.5, 140, and 171.5 boxes are calculated, the other expected values can be found by using the row and column totals. Thus the number of degrees of freedom here is 4. Or we calculate

$$df = (r-1)(c-1) = (3-1)(3-1) = 4$$

The *P*-value is calculated to be $P = P(\chi^2 > 3.024) = .5538$.

With such a large *P*-value, there is *no* evidence of any relationship between party affiliation and shaken confidence in the banking system.

On the TI-Nspire the result shows as:



As for conditions to check for chi-square tests for independence, we should check that the sample is randomly chosen and that the expected values for all cells are at least 5. If a category has one of its expected cell count less than 5, we can combine categories that are logically similar (for example "disagree" and "strongly disagree"), or combine numerically small categories collectively as "other."

EXAMPLE 15.4

To determine whether men with a combination of childhood abuse and a certain abnormal gene are more likely to commit violent crimes, a study is run on a simple random sample of 575 males in the 25 to 35 age group. The data are summarized in the following table:

	Not abused, normal gene	Abused, normal gene	Not abused, abnormal gene	Abused, abnormal gene
Criminal behavior	48	21	32	26
Normal behavior	201	79	118	50

a. Is there evidence of a relationship between the four categories (based on childhood abuse and abnormal genetics) and behavior (criminal versus normal)? Explain.

b. Is there evidence that among men with the normal gene, the proportion of abused men who commit violent crimes is greater than the proportion of nonabused men who commit violent crimes? Explain.

c. Is there evidence that among men who were not abused as children, the proportion of men with the abnormal gene who commit violent crimes is greater than the proportion of men with the normal gene who commit violent crimes? Explain.

d. Is there a contradiction in the above results? Explain.

Answers:

a. A chi-square test for independence is indicated. The expected cell counts are as follows:

55.0	22.1	33.1	16.8
194.0	77.9	116.9	59.2

The condition that all cell counts are greater than 5 is met.

H_0: The four categories (based on childhood abuse and abnormal genetics) and behavior (criminal versus normal) are independent.

H_a: The four categories (based on childhood abuse and abnormal genetics) and behavior (criminal versus normal) are not independent (there is a relationship).

Running a chi-square test gives $\chi^2 = 7.752$. With $df = (r - 1)(c - 1) = 3$, we get $P = .0514$. Since $.0514 < .10$, the data do provide some evidence (at least at the 10% significance level) to reject H_0 and conclude that there is evidence of a relationship between the four categories (based on childhood abuse and abnormal genetics) and behavior (criminal versus normal).

b. A two-proportion z-test is indicated. We must check that n is large enough: $n_1\hat{p}_1 = 21 > 10$, $n_1(1 - \hat{p}_1) = 79 > 10$, $n_2\hat{p}_2 = 48 > 10$, $n_2(1 - \hat{p}_2) = 201 > 10$. We must assume simple random samples from the target population, since this is not given.

H_0: $p_1 - p_2 = 0$ (where p_1 is the proportion of abused men with normal genetics who commit violent crimes and p_2 is the proportion of nonabused men with normal genetics who commit violent crimes)

H_a: $p_1 - p_2 > 0$ (the proportion of abused men with normal genetics who commit violent crimes is greater than the proportion of nonabused men with normal genetics who commit violent crimes)

$$\hat{p}_1 = \frac{21}{100} = .21,\ \hat{p}_2 = \frac{48}{249} = .193,\ \text{and}\ \hat{p} = \frac{21+48}{100+249} = .198$$

$$\sigma_d = \sqrt{(.198)(.802)\left(\tfrac{1}{100}+\tfrac{1}{249}\right)} = .0472$$

$$\text{So } z = \frac{.21-.193}{.0472} = 0.360 \text{ and } P = .359.$$

With such a large P-value there is no evidence to reject H_0, and thus we conclude that there is *no* evidence that the proportion of abused men with normal genetics who commit violent crimes is greater than the proportion of nonabused men with normal genetics who commit violent crimes.

c. A two-proportion z-test is indicated. We must check that n is large enough: $n_1\hat{p}_1 = 32 > 10$, $n_1(1-\hat{p}_1) = 118 > 10$, $n_2\hat{p}_2 = 48 > 10$, $n_2(1-\hat{p}_2) = 201 > 10$. We must assume simple random samples from the target population, since this is not given.

H_0: $p_1 - p_2 = 0$ (where p_1 is the proportion of nonabused men with abnormal genetics who commit violent crimes and p_2 is the proportion of nonabused men with normal genetics who commit violent crimes)

H_a: $p_1 - p_2 > 0$ (the proportion of nonabused men with abnormal genetics who commit violent crimes is greater than the proportion of nonabused men with normal genetics who commit violent crimes)

$$\hat{p}_1 = \frac{32}{150} = .213,\ \hat{p}_2 = \frac{48}{249} = .193,\ \text{and}\ \hat{p} = \frac{32+48}{150+249} = .2005$$

$$\sigma_d = \sqrt{(.2005)(.7995)\left(\tfrac{1}{150}+\tfrac{1}{249}\right)} = .0414$$

$$\text{So } z = \frac{.213-.193}{.0414} = 0.483 \text{ and } P = .315.$$

With such a large P-value there is no evidence to reject H_0, and thus we conclude that there is *no* evidence that the proportion of nonabused men with the abnormal gene who commit violent crimes is greater than the proportion of nonabused men with the normal gene who commit violent crimes.

d. There is no contradiction. It is possible to have evidence of an overall relationship without significant evidence showing in a subset of the categories.

CHI-SQUARE TEST FOR HOMOGENEITY OF PROPORTIONS

In chi-square goodness-of-fit tests we work with a single variable in comparing a single sample to a population model. In chi-square independence tests we work with a single sample classified on two variables. Chi-square procedures can also be used with a single variable to compare samples from two or more populations. It is important that the samples be *simple random samples*, that they be taken *independ-*

ently of each other, that the original populations be large compared to the sample sizes, and that the expected values for all cells be at least 5. The contingency table used has a row for each sample.

EXAMPLE 15.5

In a large city, a group of AP Statistics students work together on a project to determine which group of school employees has the greatest proportion who are satisfied with their jobs. In independent simple random samples of 100 teachers, 60 administrators, 45 custodians, and 55 secretaries, the numbers satisfied with their jobs were found to be 82, 38, 34, and 36, respectively. Is there evidence that the proportion of employees satisfied with their jobs is different in different school system job categories?

Answer:

H_0: The proportion of employees satisfied with their jobs is the same across the various school system job categories.

H_a: At least two of the job categories differ in the proportion of employees satisfied with their jobs.

The observed counts are as follows:

	Satisfied	Not satisfied
Teachers	82	18
Administrators	38	22
Custodians	34	11
Secretaries	36	19

Just as we did in the previous section, we can calculate the expected value of any cell by multiplying the corresponding row total by the appropriate column total and then dividing by the grand total. In this case, this results in the following expected counts:

	Satisfied	Not satisfied	
Teachers	73.1	26.9	100
Administrators	43.8	16.2	60
Custodians	32.9	12.1	45
Secretaries	40.2	14.8	55
	190	70	260

We note that all expected cell counts are >5, and then calculate chi-square:

$$\chi^2 = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}} = \frac{(82 - 73.1)^2}{73.1} + \cdots + \frac{(19 - 14.8)^2}{14.8} = 8.640$$

With 4 – 1 = 3 degrees of freedom, we calculate the *P*-value to be $P(\chi^2 > 8.640) = .0345$. [On the TI-84: χ^2cdf(8.640,1000,3).] With this small a *P*-value, there is sufficient evidence to reject H_0, and we can conclude that there is evidence that the proportion of employees satisfied with their jobs is *not* the same across all the school system job categories.

Note: On the TI-84 we could also have put the observed data into a matrix and used χ^2-Test, resulting in $\chi^2 = 8.707$ and $P = .0335$.