## Ch 11.2 Significance testing for 2 way tables

A chi-square test for homogeneity

State: $H_0$: There is no difference in the distribution of a categorical variable for several populations or treatments. (think = 0)
$H_a$: There is a difference in the distribution of a categorical variable for several populations or treatments. (think≠0)
don't forget to mention $\alpha$

Plan: **Conditions for performing a Chi-Square Test for Homogeneity**

***Random:** The data come from independent random samples or from the groups in a randomized experiment.

***10%:** When sampling without replacement, check that $n \le 0.10N$.

***Large Counts:** All *expected* counts are at least 5.

Do: When $H_0$ is true, the expected count in any cell of a two-way table is

$$\text{expected count} = \frac{\text{row total} \bullet \text{column total}}{\text{table total}}$$

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

This time, the sum is over all cells (not including the totals!) in the two-way table.

Use Table C to find the *P*-value. Then use your calculator's $\chi^2$cdfcommand.

df = (# of rows - 1)(# of columns - 1)

Conclude:

Same as previous tests

Feb 21-11:20 AM

**Example 1:** Market researchers suspect that background music may affect the mood and buying behavior of customers. One study in a European restaurant compared three randomly assigned treatments: no music, French accordion music, and Italian string music. Under each condition, the researchers recorded the number of customers who ordered French, Italian, and other entrees. Here is a table that summarizes the data:

Observed Counts

| Entree ordered | None | French | Italian | Total |
|---|---|---|---|---|
| | Type of Music | | | |
| French | 30 | 39 | 30 | 99 |
| Italian | 11 | 1 | 19 | 31 |
| Other | 43 | 35 | 35 | 113 |
| Total | 84 | 75 | 84 | 243 |

State:
$H_o$ The null hypothesis in the restaurant experiment is that there's no difference in the distribution of entrees ordered when no music, French accordion music, or Italian string music is played.

$H_a$ The alternative is the restaurant experiment has a difference on the distribution of entrees ordered with music.

$$\text{expected count} = \frac{\text{row total} \bullet \text{column total}}{\text{table total}}$$

Expected Counts

| Entree ordered | None | French | Italian | Total |
|---|---|---|---|---|
| | Type of Music | | | |
| French | 34.22 | 30.56 | 34.22 | 99 |
| Italian | 10.72 | 9.57 | 10.72 | 31 |
| Other | 39.06 | 34.88 | 39.06 | 113 |
| Total | 84 | 75 | 84 | 243 |

Plan: A chi-square test for homogeneity if conditions are met.

- random
- sample sizes are greater than 5
- represents ≤ 10% of the population

Do: The $\chi^2$ statistic is the sum of nine such terms:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \frac{(30-34.22)^2}{34.22} + \frac{(39-30.56)^2}{30.56} + \ldots + \frac{(35-39.06)^2}{39.06}$$

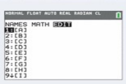
$$= 0.52 + 2.33 + \ldots + 0.42 = 18.28$$

In the calculator ----------------------------------------→

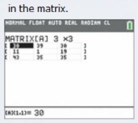
Feb 21-12:05 PM

TI-83/84

- Press [2nd] [$x^{-1}$] (MATRIX), arrow to EDIT, and choose A.
- Enter the dimensions of the matrix: 3 × 3.

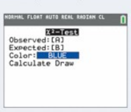



- Enter the observed counts from the two-way table in the same locations in the matrix.




**2.** Specify the chi-square test, the matrix where the observed counts are found, and the matrix where the expected counts will be stored.

- Press [STAT], arrow to TESTS, and choose $\chi^2$-Test.
- Adjust your settings as shown.




If you go to matrix B there you'll find the expected counts.

Feb 21-1:59 PM

**AP EXAM TIP** In the "Do" step, you aren't required to show every term in the chi-square statistic. Writing the first few terms of the sum followed by "…" is considered as "showing work." We suggest that you do this and then let your calculator tackle the computations.

As in the test for goodness of fit, you should think of the chi-square statistic $\chi^2$ as a measure of how much the observed counts deviate from the expected counts. Once again, large values of $\chi^2$ are evidence against $H_0$ and in favor of $H_a$. The *P*-value measures the strength of this evidence. When conditions are met, *P*-values for a chi-square test for homogeneity come from a chi-square distribution with df = (number of rows − 1) ×(number of columns − 1).

Feb 21-2:04 PM

DO:
Because the two-way table that summarizes the data from the study has three rows and three columns, we use a chi-square distribution with df = (3 − 1)(3 − 1) = 4 to find the *P*-value.

***Table:*** Look at the df = 4 row in **Table C**. The calculated value $\chi^2 = 18.28$ lies between the critical values 16.42 and 18.47. The corresponding *P*-value is between 0.001 and 0.0025.

| | *P* | |
|---|---|---|
| df | .0025 | .001 |
| 4 | 16.42 | 18.47 |

***Calculator:*** Chi squared/test statistic 18.28 P-value 0.0011.

Conclude:
Because the *P*-value, 0.0011, is less than our default $\alpha = 0.05$ significance level, we reject $H_0$. We have convincing evidence of a difference in the true distributions of entrees ordered at this restaurant when no music, French accordion music, or Italian string music is played. Furthermore, the random assignment allows us to say that the difference is caused by the music that's played.

Feb 21-2:05 PM

**Example 9** **Are Cell-Only Telephone Users Different?**

*The chi-square test for homogeneity*

Random digit dialing telephone surveys used to exclude cell phone numbers. If the opinions of people who have only cell phones differ from those of people who have landline service, the poll results may not represent the entire adult population. The Pew Research Center interviewed separate random samples of cell-only and landline telephone users who were less than 30 years old. Here's what the Pew survey found about how these people describe their political party affiliation:[7]

| | Cell-only sample | Landline sample |
|---|---|---|
| Democrat or lean Democratic | 49 | 47 |
| Refuse to lean either way | 15 | 27 |
| Republican or lean Republican | 32 | 30 |
| **Total** | **96** | **104** |

**PROBLEM:**

**(a)** Compare the distributions of political party affiliation for cell-only and landline phone users.

**(b)** Do these data provide convincing evidence at the $\alpha = 0.05$ level that the distribution of party affiliation differs in the under-30 cell-only and landline user populations?

State: $H_0$: There is no difference in the distribution of party affiliation in the under-30 cell-only and landline populations.

$H_a$: There is a difference in the distribution of party affiliation in the under-30 cell-only and landline populations. at the $\alpha = 0.05$ level.

PLAN: If conditions are met, we should use a chi-square test for homogeneity.

- ***Random:*** The data came from independent random samples of 96 cell-only and 104 landline users.
  - **10%:** Sampling without replacement was used, so there need to be at least 10(96) = 960 cell-only users under age 30 and at least 10(104) = 1040 landline users under age 30. This is safe to assume.
- ***Large Counts:*** We followed the steps in the Technology Corner on page 706 to get the expected counts. The calculator screen shot confirms that all expected counts are at least 5.

Feb 21-2:08 PM

**DO:** A chi-square test on the calculator gave

- ***Test statistic:***

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$= \frac{(49 - 46.08)^2}{46.08} + \frac{(47 - 49.92)^2}{49.92} + \cdots = 3.22$$

- ***P*-value:** Using df = (number of rows − 1)(number of columns − 1) = (3 − 1)(2 − 1) = 2, the *P*-value is 0.1999.

**CONCLUDE:** Because our *P*-value, 0.1999, is greater than $\alpha = 0.05$, we fail to reject $H_0$. There is not convincing evidence that the distribution of party affiliation differs in the under-30 cell-only and landline user populations.

Feb 21-3:18 PM

**Follow-up Analysis**

The chi-square test for homogeneity allows us to compare the distribution of a categorical variable for any number of populations or treatments. If the test allows us to reject the null hypothesis of no difference, we then want to do a follow-up analysis that examines the differences in detail.

Start by examining which cells in the two-way table show large deviations between the observed and expected counts. Then look at the individual components to see which terms contribute most to the chi-square statistic.

Our earlier restaurant study found significant differences among the true distributions of entrees ordered under each of the three music conditions. We entered the two-way table for the study into Minitab software and requested a chi-square test. The output appears in the figure below. Minitab repeats the two-way table of observed counts and puts the expected count for each cell below the observed count. Finally, the software prints the 9 individual components that contribute to the $\chi^2$ statistic.

```
Chi-Square Test: None, French, Italian

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

          None   French  Italian   Total
1           30       39       30      99
         34.22    30.56    34.22
         0.521    2.334    0.521

2           11        1       19      31
         10.72     9.57    10.72
         0.008    7.672    6.404

3           43       35       35     113
         39.06    34.88    39.06
         0.397    0.000    0.422

Total       84       75       84     243

Chi-Sq = 18.279, DF = 4, P-Value = 0.001
```

Looking at the output, we see that just two of the nine components that make up the chi-square statistic contribute about 14 (almost 77%) of the total $\chi^2 = 18.28$. Comparing the observed and expected counts in these two cells, we see that orders of Italian entrees are much below expectation when French music is playing and well above expectation when Italian music is playing. We are led to a specific conclusion: orders of Italian entrees are strongly affected by Italian and French music. More advanced methods provide tests and confidence intervals that make this follow-up analysis more complete.

Feb 21-3:20 PM

**Comparing Several Proportions**

**Homogeneity** - tests that all categories are the same amount.

*The chi-square test for homogeneity allows us to test $H_0$: $p_1 = p_2 = \ldots = p_k$. This null hypothesis says that there is no difference in the proportions of successes for the $k$ populations or treatments. The alternative hypothesis is $H_a$: at least two of the $p_i$'s are different.

Many students *incorrectly state* $H_a$ as "all the proportions are different." Think about it this way: the opposite of "all the proportions are equal" is "some of the proportions are not equal."

Feb 21-3:25 PM

**Example 6:** A study followed a random sample of 8474 people with normal blood pressure for about four years. All the individuals were free of heart disease at the beginning of the study. Each person took the Spielberger Trait Anger Scale test, which measures how prone a person is to sudden anger. Researchers also recorded whether each individual developed coronary heart disease (CHD). This includes people who had heart attacks and those who needed medical treatment for heart disease. Here is a two-way table that summarizes the data:

| | Low anger | Moderate anger | High anger | Total |
|---|---|---|---|---|
| CHD | 53 | 110 | 27 | **190** |
| No CHD | 3057 | 4621 | 606 | **8284** |
| **Total** | **3110** | **4731** | **633** | **8474** |

a) Is this an observational study or an experiment? Justify your answer.

This is an observational study. Researchers did not deliberately impose any treatments. They just recorded data about two variables—anger level and whether or not the person developed CHD—for each randomly chosen individual.

Feb 22-12:10 PM

b) Make a well-labeled bar graph that compares CHD rates for the different anger levels. Describe what you see.

In this setting, anger level is the explanatory variable and whether or not a person gets heart disease is the response variable. So we compare the percents of people who did and did not get heart disease in each of the three anger categories:

(divided by the total from the column)

| | CHD | no CHD |
|---|---|---|
| Low anger: | $\frac{53}{3110} = 0.0170 = 1.70\%$ | $\frac{3057}{3110} = 0.9830 = 98.30\%$ |
| Moderate anger: | $\frac{110}{4731} = 0.0233 = 2.33\%$ | $\frac{4621}{4731} = 0.9767 = 97.67\%$ |
| High anger: | $\frac{27}{633} = 0.0427 = 4.27\%$ | $\frac{606}{633} = 0.9573 = 95.73\%$ |


The bar graph shows the percent of people in each of the three anger categories who developed CHD. There is a clear trend: as the anger score increases, so does the percent who suffer heart disease. A much higher percent of people in the high anger category developed CHD (4.27%) than in the moderate (2.33%) and low (1.70%) anger categories.

Feb 22-12:12 PM

**The Chi-Square Test for Independence** Day 2

We often gather data from a random sample and arrange them in a two-way table to see if two categorical variables are associated. The sample data are easy to investigate: turn them into percents and look for a relationship between the variables.

$H_0$: There is no association between anger level and heart-disease status in the population of people with normal blood pressure. ( think = 0)

$H_a$: There is an association between anger level and heart-disease status in the population of people with normal blood pressure.

No association between two variables means that knowing the value of one variable does not help us predict the value of the other. That is, the variables are *independent*. An equivalent way to state the hypotheses is therefore

$H_0$: Anger and heart-disease status are independent in the population of people with normal blood pressure.

$H_a$: Anger and heart-disease status are not independent in the population of people with normal blood pressure.

Feb 22-12:14 PM

**Example 7:** The null hypothesis is that there is no association between anger level and heart-disease status in the population of interest. If we assume that $H_0$ is true, then anger level and CHD status are independent. We can find the expected cell counts in the two-way table using the definition of independent events from Chapter 5: $P(A \mid B) = P(A)$.The chance process here is randomly selecting a person and recording his or her anger level and CHD status.

| | Low anger | Moderate anger | High anger | Total |
|---|---|---|---|---|
| CHD | 53 | 110 | 27 | **190** |
| No CHD | 3057 | 4621 | 606 | **8284** |
| **Total** | **3110** | **4731** | **633** | **8474** |

Let's start by considering the events "CHD" and "low anger." We see from the two-way table that 190 of the 8474 people in the study had CHD. If we imagine choosing one of these people at random, $P(\text{CHD}) = 190/8474$. Because anger level and CHD status are independent, knowing that the selected individual is low anger does not change the probability that this person develops CHD. That is to say, $P(\text{CHD} \mid \text{low anger}) = P(\text{CHD}) = 190/8474 = 0.02242$.

Of the 3110 low-anger people in the study, we'd expect $3110 \bullet \frac{190}{8474} = 3110(0.02242) = 69.73$ To get CHD.

You can see that the general formula we developed earlier for a test for homogeneity applies in this situation also:

$$\text{expected count} = \frac{\text{row total} \bullet \text{column total}}{\text{table total}} = \frac{190 \bullet 3110}{8474} = 69.73$$

| CHD, Low | CHD, Moderate | CHD, High |
|---|---|---|
| 3110(0.02242) = 69.73 | 4731(0.02242) = 106.08 | 633(0.02242) = 14.19 |
| **no CHD, Low** | **no CHD, Moderate** | **no CHD, High** |
| 3110(0.97758) = 3040.27 | 4731(0.97758) = 4624.92 | 633(0.97758) = 618.81 |

Feb 22-12:21 PM

**Example 8:** Here is the complete table of observed and expected counts for the CHD and anger study side by side:

| | Observed | | | Expected | | |
|---|---|---|---|---|---|---|
| | Low | Moderate | High | Low | Moderate | High |
| CHD | 53 | 110 | 27 | 69.73 | 106.08 | 14.19 |
| No CHD | 3057 | 4621 | 606 | 3040.27 | 4624.92 | 618.81 |

Do the data provide convincing evidence of an association between anger level and heart disease in the population of interest?

**State:** We want to perform a test of

$H_0$: There is no association between anger level and heart disease status in the population of people with normal blood pressure.

$H_a$: There is an association between anger level and heart disease status in the population of people with normal blood pressure.

Because no significance level was stated, we will use $\alpha = 0.05$.

**Plan:** If conditions are met, we should carry out a chi-square test for independence.

**Random:** The data came from a random sample of 8474 people with normal blood pressure.

**10%:** Because the researchers sampled without replacement, we need to check that the total number of people in the population with normal blood pressure is at least 10(8474) = 84,740. This seems reasonable to assume.

**Large Counts:** All the expected counts are at least 5, (the smallest is 14.19), so this condition is met.

Feb 22-12:27 PM

**Do:** We perform calculations assuming $H_0$ is true.

Test statistic:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$= \frac{(53-69.73)^2}{69.73} + \frac{(110-106.08)^2}{106.08} + \ldots + \frac{(606-618.81)^2}{618.81}$$

$$= 4.014 + 0.145 + \ldots + 0.265 = 16.077$$

***P*-Value:** The two-way table of anger level versus heart disease has 2 rows and 3 columns. We will use the chi-square distribution with df = (2 − 1)(3 − 1) = 2 to find the *P*-value. Look at the df = 2 line in **Table C**. The observed statistic $\chi^2 = 16.077$ is larger than the critical value 15.20 for $\alpha = 0.0005$. So the *P*-value is less than 0.0005.

***Using Technology:*** The calculator's $\chi^2$-Test gives $\chi^2 = 16.077$ and *P*-value = 0.00032 using df = 2.

**Conclude:** Reject $H_0$. Since the *P*-value, 0.00032 is clearly less than $\alpha = 0.05$, there is enough evidence to conclude that there is an association between anger level and heart-disease status in the population of people with normal blood pressure.

Feb 22-12:28 PM

**Using Chi-Square Tests Wisely**

Both the chi-square test for homogeneity and the chi-square test for association/independence start with a two-way table of observed counts. They even calculate the test statistic, degrees of freedom, and *P*-value in the same way. *The questions that these two tests answer are different, however.*

- A chi-square test for homogeneity tests whether the distribution of a categorical variable is the same for each of several populations or treatments.
- The chi-square test for independence tests whether two categorical variables are associated in some population of interest.

Unfortunately, it is quite common to see questions asking about association when a test for homogeneity applies and questions asking about differences between proportions or the distribution of a variable when a test of independence applies. Sometimes, people avoid the distinction altogether and pose questions about the "relationship" between two variables.

Instead of focusing on the question asked, it's much easier to look at how the data were produced.

✓If the data come from two or more independent random samples or treatment groups in a randomized experiment, then do a chi-square test for homogeneity.

✓If the data come from a single random sample, with the individuals classified according to two categorical variables, use a chi-square test for independence.

**AP EXAM TIP**: If you have trouble distinguishing the two types of chi-square tests for two-way tables, you're better off just saying "chi-square test" than choosing the wrong type. Better yet, learn to tell the difference!

Feb 22-12:28 PM

**Example 9:** Are men and women equally likely to suffer lingering fear from watching scary movies as children? Researchers asked a random sample of 117 college students to write narrative accounts of their exposure to scary movies before the age of 13. More than one-fourth of the students said that some of the fright symptoms are still present when they are awake. The following table breaks down these results by gender.

| | Gender | | |
|---|---|---|---|
| Fright symptoms? | Male | Female | Total |
| Yes | 7 | 29 | **36** |
| No | 31 | 50 | **81** |
| **Total** | **38** | **79** | **117** |

Minitab output for a chi-square test using these data is shown below.

```
Chi-Square Test: Male, Female
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts
        Male    Female    Total
1          7        29       36
       11.69     24.31
       1.883     0.906

2         31        50       81
       26.31     54.69
       0.837     0.403

Total     38        79      117
Chi-Sq = 4.028, DF = 1, P-Value = 0.045
```

a) Explain why a chi-square test for independence and not a chi-square test for homogeneity should be used in this setting.

The data were produced using a single random sample of college students, who were then classified by gender and whether or not they had lingering fright symptoms. The chi-square test for homogeneity requires independent random samples from each population.

Feb 22-7:49 PM

Minitab output for a chi-square test using these data is shown below.

```
Chi-Square Test: Male, Female
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts
        Male    Female    Total
1          7        29       36
       11.69     24.31
       1.883     0.906

2         31        50       81
       26.31     54.69
       0.837     0.403

Total     38        79      117
Chi-Sq = 4.028, DF = 1, P-Value = 0.045
```


b) State an appropriate pair of hypotheses for researchers to test in this setting.

The null hypothesis is $H_0$: There is no association between gender and ongoing fright symptoms in the population of college students.

The alternative hypothesis is $H_a$: There is an association between gender and ongoing fright symptoms in the population of college students.

c) Which cell contributes most to the chi-square statistic? In what way does this cell differ from what the null hypothesis suggests?

Men who admit to having lingering fright symptoms account for the largest component of the chi-square statistic: 1.883 of the total 4.028. Far fewer men in the sample admitted to fright symptoms (7) than we would expect if $H_0$ were true (11.69).

Feb 22-8:00 PM

d) Interpret the $P$-value in context. What conclusion would you draw at $\alpha = 0.01$?

If gender and ongoing fright symptoms really are independent in the population of interest, there is a 0.045 chance of obtaining a random sample of 117 students that gives a chi-square statistic of 4.028 or higher. Because the $P$-value, 0.045, is greater than 0.01, we would fail to reject $H_0$. We do not have convincing evidence that there is an association between gender and fright symptoms in the population of college students.

Feb 22-8:04 PM