PH 702 Introduction to Biostatistics

Part 9: Analysis of Two-Way Tables

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

2026/11/18 11:21:28

Learning objectives

- Summarize two categorical variables using a two-way table and conditional percentages.
- Calculate expected counts and interpret the chi-square test of independence.
- Use a chi-square test to assess a claimed categorical distribution.
- Check sampling and expected-count conditions and interpret conclusions in context.

Scenario I: Two-way tables (introduced in Chapter 2)

- A two-way table records the observed counts for combinations of two categorical variables.
- A conditional distribution describes one variable within a specified category of the other.
- Different conditional distributions suggest an association between two categorical variables.
- Identical conditional distributions (homogeneity) indicate that the two categorical variables are independent.

Demo: Music and wine purchases

A supermarket experiment investigated whether background music influences customers' wine purchases. Customers were randomly assigned to three music conditions: no music, French accordion music, or Italian string music. The researchers examined whether the type of music is associated customers' choices of wine. The observed numbers of bottles purchased were:

Wine	No music	French music	Italian music	**Total**
French	30 (35.7%; 30.3%)	39 (52.0%; 39.4%)	30 (35.7%; 30.3%)	**99**
Italian	11 (13.1%; 35.5%)	1 (1.3%; 3.2%)	19 (22.6%; 61.3%)	**31**
Other	43 (51.2%; 38.1%)	35 (46.7%; 31.0%)	35 (41.7%; 31.0%)	**113**
Total	**84**	**75**	**84**	**243**

Note: Each cell shows **count (column %; row %)**.

- Column %: Conditional distribution of wine given music type.
- Row %: Conditional distribution of music type given wine purchased.

Expected counts under independence

If the two categorical variables are independent, the conditional distributions should be the same across groups.

The **expected count** is the number of observations we would expect in each cell of a two-way table **if the two categorical variables are independent**.

Expected counts are calculated using the row and column totals. Specifically, for the cell in row i and column j, the expected count is

$$E_{ij} = \frac{(\text{row } i \text{ total})(\text{column } j \text{ total})}{\text{grand total}}.$$

—

Demo: Expected counts for the music and wine example

Suppose **music type and wine choice are independent**. We can calculate how many purchases of each wine type we would expect under each music treatment.

For example, the expected number of French wine purchases under no music is

$$E_{\text{French, no music}} = \frac{99 \times 84}{243} = 34.22.$$

See below for the expected count for each cell:

Wine	No music	French music	Italian music
French	34.22	30.56	34.22
Italian	10.72	9.57	10.72
Other	39.06	34.88	39.06

What is the chi-square (χ^2) distribution?

- The chi-square (χ^2) distribution is a **family of right-skewed distributions** that take only nonnegative values.
- Its shape depends on the **degrees of freedom** (df).
- As df increases, the distribution becomes less right-skewed and more flat.

```
/* Chi-square density curves for different degrees of freedom */
data chisquare_density;
    do x = 0.1 to 30 by 0.1;
        df1  = pdf("CHISQUARE", x, 1);
        df4  = pdf("CHISQUARE", x, 4);
        df10 = pdf("CHISQUARE", x, 10);
        output;
    end;
run;

proc sgplot data=chisquare_density;
    series x=x y=df1 /
        legendlabel="df = 1"
        lineattrs=(color=blue thickness=2);
    series x=x y=df4 /
        legendlabel="df = 4"
```

```
        lineattrs=(color=red thickness=2);
    series x=x y=df10 /
        legendlabel="df = 10"
        lineattrs=(color=green thickness=2);
    xaxis label="Chi-square value";
    yaxis label="Probability density";
    keylegend / title="Degrees of freedom";
    title "Chi-square Distributions with Different Degrees of Freedom";
run;
```

Chi-square test for independence

Hypotheses:

- H_0: The two categorical variables are **independent**.
- H_a: The two categorical variables are **associated** (not independent).

Test statistic:

$$\chi^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- O_{ij} is the observed count in cell (i, j).
- E_{ij} is the expected count in cell (i, j) under independence (H_0).
- Under H_0, the above chi-square statistic approximately follows the chi-square distribution with

$$\text{df} = (r-1)(c-1),$$

  where r and c are the numbers of rows and columns.

P-value:

$$\text{P-value} = \Pr(\chi^2_{\text{df}} \geq \chi^2_{\text{obs}})$$

- χ^2_{df}: χ^2 distribution with $\text{df} = (r-1)(c-1)$.

Decision: Reject H_0 if P-value $\leq \alpha$; otherwise, do not reject H_0.

Conclusion: Testing hypotheses H_0 : ___ vs. H_a : ___ , we carried out the ___ test and obtained P-value ___ . So, at level ___ , there was/wasn't sufficient statistical evidence against H_0, i.e., we believed that ___ .

Conditions (for the chi-square approximation to be appropriate):

1. **Independent observations:** Data come from an appropriate random sample or randomized study, with independent observations.
2. **Sufficiently large expected counts:** For the chi-square approximation to be appropriate:
   - All expected counts should be at least 1, with an average expected count of at least 5.
   - For a 2×2 table, all four expected counts should be at least 5.

Demo: Is the music associated with wine choice?

H_0 : The music is NOT associated with wine choice versus H_a : The music is associated with wine choice.

More equivalent formulations of the hypotheses:

Perspective	H_0	H_a
Homogeneity	Wine-purchase distributions are identical across music treatments.	At least one music treatment has a different wine-purchase distribution.
Independence	Music treatment and wine choice are independent.	Music treatment and wine choice are associated.
Conditional distribution of wine	The distribution of wine choice given music treatment is the same across all music treatments.	The distribution of wine choice differs across at least two music treatments.
Conditional distribution of music	The distribution of music treatment given wine choice is the same across all wine categories.	The distribution of music treatment differs across at least two wine categories.

All above formulations express the same underlying null hypothesis: music treatment and wine choice are independent, or equivalently, wine-purchase distributions are homogeneous across music treatments.

```
data wine;
  input wine $ music $ count;
  datalines;
French None 30
Italian None 11
Other None 43
French FrenchMusic 39
Italian FrenchMusic 1
Other FrenchMusic 35
French ItalianMusic 30
Italian ItalianMusic 19
Other ItalianMusic 35
;
run;

proc freq data=wine;
  tables music*wine / chisq expected nocol norow nopercent;
  weight count;
run;
```

Conclusion: Testing hypotheses H_0 : ___ vs. H_a : ___ , we carried out the ___ test and obtained P-value ___ . So, at level ___ , there was/wasn't sufficient statistical evidence against H_0, i.e., we believed that ___ .

Exercise: Party affiliation and phone type

Traditional telephone surveys often contacted people through landlines, excluding those who used only cell phones. If cell-only users have different political opinions from landline users, surveys that exclude cell-only users may not accurately represent the population.

To investigate this issue, the Pew Research Center interviewed separate random samples of people under age 30:

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

Party affiliation	Cell-only	Landline

1. State H_0 and H_a for comparing party-affiliation distributions.
2. Calculate the expected count for cell-only respondents who refuse to lean either way.
3. Identify the degrees of freedom.
4. Presumably, the P-value is about 0.20. Interpret the result at level $\alpha = 0.05$.

Exercise: Anger and coronary heart disease

One SRS of 8,474 people was classified by anger level and coronary heart disease (CHD):

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

1. Calculate the CHD percentage in each anger group and describe the pattern.
2. State the independence hypotheses and compute df.
3. Presumably, $\chi^2 = 16.077$ and P-value ≈ 0.00032. State a conclusion in context.

```
data anger;
  input outcome $ anger $ count;
  datalines;
CHD Low 53
CHD Moderate 110
CHD High 27
NoCHD Low 3057
NoCHD Moderate 4621
NoCHD High 606
;
run;

proc freq data=anger;
  tables anger*outcome / chisq expected nopercent norow;
  weight count;
run;
```

Scenario II: Comparing an observed distribution with a specific distribution

Suppose we have **one categorical variable** and want to determine whether its distribution agrees with a **specific distribution**.

For example, a candy manufacturer claims that its candies' colors come in certain proportions. Do the observed color counts support this claim?

Chi-square goodness-of-fit test

Hypotheses:

- $H_0 : p_1 = p_{10}, \ldots, p_k = p_{k0}$. (The population distribution agrees with the claimed distribution.)
  - p_i: the true population proportion in category i
  - p_{i0}: the claimed proportion for category i
  - k: the number of categories
- H_a: The population distribution differs from the claimed distribution in at least one category.

Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- O_i is the observed count in category i.
- E_i is the expected count in category i.

P-value:

$$\text{P-value} = \Pr(\chi^2_{\text{df}} \geq \chi^2_{\text{obs}})$$

- χ^2_{df}: χ^2 distribution with $\text{df} = k - 1$.

Conclusion: Testing hypotheses H_0 : ___ vs. H_a : ___ , we carried out the ___ test and obtained P-value ___ . So, at level ___ , there was/wasn't sufficient statistical evidence against H_0, i.e., we believed that ___ .

Conditions (for the chi-square approximation to be appropriate):

1. **Categorical data:** Each observation belongs to exactly one of the k categories.
2. **Independent observations:** Data come from an appropriate random sample or randomized design, with independent observations.
3. **Expected-count condition:**
   - Every expected count should be at least 1.
   - The average expected count should be at least 5.

Demo: M&M's color distribution

The manufacturer's claimed proportions and one SRS of 60 candies are:

Color	Blue	Orange	Green	Yellow	Red	Brown
Claimed proportions	0.24	0.20	0.16	0.14	0.13	0.13
Observed counts	9	8	12	15	10	6

- $\alpha = 0.05$
- H_0: all six color proportions match the claimed distribution
- H_a: at least one differs.

```
data candies;
    input color $ count;
    datalines;
Blue    9
Orange  8
Green  12
Yellow 15
Red    10
Brown   6
;
run;

proc freq data=candies order=data;
```

```
    tables color / testp=(24 20 16 14 13 13);
    weight count;
run;
```

Conclusion: Testing hypotheses H_0 : ___ vs. H_a : ___ , we carried out the ___ test and obtained P-value ___ . So, at level ___ , there was/wasn't sufficient statistical evidence against H_0, i.e., we believed that ___ .

Exercise: M&M's

1. What is the expected blue count?
2. Compute the contribution of **yellow** candies to the chi-square statistic.
3. What is the df?
4. What would you conclude at $\alpha = 0.10$? Compare with the conclusion at $\alpha = 0.05$.