# Chapter 9: Analysis of Two-Way Tables

MATH 560-01
Statistical Data Analysis

March 15, 2021

These slides are based on material from *Introduction to the Practice of Statistics* by David S. Moore, George P. McCabe, and Bruce A. Craig, 9th edition.

# Sections

## 9.1 Inference for Two-Way Tables

After completing this section, students should be able to:

- compute *joint, marginal, and conditional distributions* given a *two-way table*

## 9.1 Inference for Two-Way Tables

After completing this section, students should be able to:

- compute *joint, marginal, and conditional distributions* given a *two-way table*
- use a chi-square table to obtain bounds on probabilities for random variables following $\chi^2$-distributions

## 9.1 Inference for Two-Way Tables

After completing this section, students should be able to:

- compute *joint, marginal, and conditional distributions* given a *two-way table*
- use a chi-square table to obtain bounds on probabilities for random variables following $\chi^2$-distributions
- perform *chi-square tests* to determine if there is an association between the row and column variables of a two-way table when all expected cell counts are large enough

## 9.1 Inference for Two-Way Tables

After completing this section, students should be able to:

- compute *joint, marginal, and conditional distributions* given a *two-way table*
- use a chi-square table to obtain bounds on probabilities for random variables following $\chi^2$-distributions
- perform *chi-square tests* to determine if there is an association between the row and column variables of a two-way table when all expected cell counts are large enough
- perform *chi-square tests* to determine if the distribution of the response variable is the same for all populations when all expected cell counts are large enough

# 9.1 Inference for Two-Way Tables

- Two variables are **associated** if knowing the value of one of the variables for a case tells you something about the value of the other variable for that case.

# 9.1 Inference for Two-Way Tables

- Two variables are **associated** if knowing the value of one of the variables for a case tells you something about the value of the other variable for that case.
- **Response variable**: variable that measures the outcome of the study

# 9.1 Inference for Two-Way Tables

- Two variables are **associated** if knowing the value of one of the variables for a case tells you something about the value of the other variable for that case.
- **Response variable**: variable that measures the outcome of the study
- **Explanatory variable**: variable that explains or causes a change in the response

# 9.1 Inference for Two-Way Tables

- Two variables are **associated** if knowing the value of one of the variables for a case tells you something about the value of the other variable for that case.
- **Response variable**: variable that measures the outcome of the study
- **Explanatory variable**: variable that explains or causes a change in the response
- In this section, we consider the situation where both variables are *categorical*.

# 9.1 Inference for Two-Way Tables

- When both variables are categorical, it is useful to summarize the data in a **two-way table** which gives the counts for every combination of the values for the two variables.

# 9.1 Inference for Two-Way Tables

- When both variables are categorical, it is useful to summarize the data in a **two-way table** which gives the counts for every combination of the values for the two variables.
- In this chapter, we consider two-way tables of counts with $r$ rows and $c$ columns. We use the term $\boldsymbol{r \times c}$ **table** to describe such a table.

# 9.1 Inference for Two-Way Tables

- When both variables are categorical, it is useful to summarize the data in a **two-way table** which gives the counts for every combination of the values for the two variables.
- In this chapter, we consider two-way tables of counts with $r$ rows and $c$ columns. We use the term $\boldsymbol{r \times c}$ **table** to describe such a table.
- Two-way tables can be used to estimate probabilities for **joint**, **marginal**, and **conditional** distributions. (These distributions are also discussed in Chapter 2 in the textbook.)

# 9.1 Inference for Two-Way Tables

## Joint, Marginal, and Conditional Distributions

| | Second Variable | | | | |
|---|---|---|---|---|---|
| First Variable | Category 1 | Category 2 | $\cdots$ | Category $c$ | Total |
| Category 1 | $p_{11}$ | $p_{12}$ | $\cdots$ | $p_{1c}$ | $\sum_{j=1}^{c} p_{1j}$ |
| Category 2 | $p_{21}$ | $p_{22}$ | $\cdots$ | $p_{2c}$ | $\sum_{j=1}^{c} p_{2j}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| Category r | $p_{r1}$ | $p_{r2}$ | $\cdots$ | $p_{rc}$ | $\sum_{j=1}^{c} p_{rj}$ |
| Total | $\sum_{i=1}^{r} p_{i1}$ | $\sum_{i=1}^{r} p_{i2}$ | $\cdots$ | $\sum_{i=1}^{r} p_{ic}$ | 1 |

# 9.1 Inference for Two-Way Tables

## Joint, Marginal, and Conditional Distributions

- **Joint distribution**: the probabilities of each possible pair of categories for the First Variable and the Second Variable; these are the values $p_{11}, p_{12}, \ldots, p_{1c}$, $p_{21}, p_{22}, \ldots, p_{2c}, \cdots, p_{r1}, p_{r2}, \ldots, p_{rc}$.

# 9.1 Inference for Two-Way Tables

## Joint, Marginal, and Conditional Distributions

- **Joint distribution**: the probabilities of each possible pair of categories for the First Variable and the Second Variable; these are the values $p_{11}, p_{12}, \ldots, p_{1c}$, $p_{21}, p_{22}, \ldots, p_{2c}, \cdots, p_{r1}, p_{r2}, \ldots, p_{rc}$.
- **Marginal distribution**: the probabilities of each possible category for one of the variables;

# 9.1 Inference for Two-Way Tables

## Joint, Marginal, and Conditional Distributions

- **Joint distribution**: the probabilities of each possible pair of categories for the First Variable and the Second Variable; these are the values $p_{11}, p_{12}, \ldots, p_{1c}$, $p_{21}, p_{22}, \ldots, p_{2c}, \cdots, p_{r1}, p_{r2}, \ldots, p_{rc}$.
- **Marginal distribution**: the probabilities of each possible category for one of the variables;
  - the probabilities for the marginal distribution of the First Variable are the row totals $\sum_{j=1}^{c} p_{1j}, \sum_{j=1}^{c} p_{2j}, \ldots, \sum_{j=1}^{c} p_{rj}$
  - the probabilities for the marginal distribution of the Second Variable are the column totals $\sum_{i=1}^{r} p_{i1}, \sum_{i=1}^{r} p_{i2}, \cdots, \sum_{i=1}^{r} p_{ic}$

# 9.1 Inference for Two-Way Tables

## Joint, Marginal, and Conditional Distributions

- **Conditional distribution**: the probabilities of each possible category for one of the variables given the other variable takes a particular value

# 9.1 Inference for Two-Way Tables

## Joint, Marginal, and Conditional Distributions

- **Conditional distribution**: the probabilities of each possible category for one of the variables given the other variable takes a particular value
  - the probabilities for the conditional distribution of the First Variable given that the Second Variable is in Category $j$ are

$$\frac{p_{1j}}{\sum_{i=1}^{r} p_{ij}}, \frac{p_{2j}}{\sum_{i=1}^{r} p_{ij}}, \ldots, \frac{p_{rj}}{\sum_{i=1}^{r} p_{ij}}$$

  - the probabilities for the conditional distribution of the Second Variable given that the First Variable is in Category $i$ are

$$\frac{p_{i1}}{\sum_{j=1}^{c} p_{ij}}, \frac{p_{i2}}{\sum_{j=1}^{c} p_{ij}}, \ldots, \frac{p_{ic}}{\sum_{j=1}^{c} p_{ij}}$$

# 9.1 Inference for Two-Way Tables

- **Example 9A**: To examine the effects of salt intake on risk for cardiovascular disease (CVD), a study collect the following data.

| | Low salt | High salt | Total |
|---|---|---|---|
| CVD | 88 | 112 | 200 |
| No CVD | 1081 | 1134 | 2215 |
| Total | 1169 | 1246 | 2415 |

What are the estimates for the joint probabilities for salt intake and CVD status? What is the estimates for the marginal probabilities for each variable?

# 9.1 Inference for Two-Way Tables

- *Answer*:

# 9.1 Inference for Two-Way Tables

- Now we study how to test if two categorical variables are independent.

# 9.1 Inference for Two-Way Tables

- Now we study how to test if two categorical variables are independent.
- Recall from Chapter 4 that if two events $A$ and $B$ are independent, then $P(A \text{ and } B) = P(A) \times P(B)$.

# 9.1 Inference for Two-Way Tables

- Now we study how to test if two categorical variables are independent.
- Recall from Chapter 4 that if two events $A$ and $B$ are independent, then $P(A \text{ and } B) = P(A) \times P(B)$.
- So, if we want to assess whether two variables are independent, we can compare the observed joint distribution with what the joint distribution would be based on the marginal distributions of each variable if the variables are independent.

# 9.1 Inference for Two-Way Tables

In Example 9A, if salt intake and CVD status are independent, then the joint probabilities could be computed based on the marginal probabilities as follows.

| | Low salt | High salt | Total |
|---|---|---|---|
| CVD | | | .0828 |
| No CVD | | | .9172 |
| Total | .4841 | .5159 | 1 |

# 9.1 Inference for Two-Way Tables

In Example 9A, if salt intake and CVD status are independent, then the joint probabilities could be computed based on the marginal probabilities as follows.

| | Low salt | High salt | Total |
|---|---|---|---|
| CVD | (.0828)(.4841)= .0401 | (.0828)(.5159)= .0427 | .0828 |
| No CVD | ( .9172)(.4841)= .4440 | ( .9172)(.5159)= .4732 | .9172 |
| Total | .4841 | .5159 | 1 |

# 9.1 Inference for Two-Way Tables

- Instead of comparing probabilities, it is typical to compare observed cell counts with expected cell counts.

# 9.1 Inference for Two-Way Tables

- Instead of comparing probabilities, it is typical to compare observed cell counts with expected cell counts.
- This is equivalent to comparing the observed joint distribution with what it would be based on the marginal distributions if the variables are independent. We merely multiply each table by the overall total number of observations in the table (denoted by $n$).

# 9.1 Inference for Two-Way Tables

- Instead of comparing probabilities, it is typical to compare observed cell counts with expected cell counts.
- This is equivalent to comparing the observed joint distribution with what it would be based on the marginal distributions if the variables are independent. We merely multiply each table by the overall total number of observations in the table (denoted by $n$).
- To compute an **expected cell count** under the assumption that the variables are independent, we use the following formula:

$$\text{expected cell count} = \frac{\text{row total} \times \text{column total}}{n}$$

# 9.1 Inference for Two-Way Tables

In Example 9A, if salt intake and CVD status are independent, then the expected cell counts are computed as follows.

| | Low salt | High salt | Total |
|---|---|---|---|
| CVD | | | 200 |
| No CVD | | | 2215 |
| Total | 1169 | 1246 | 2415 |

# 9.1 Inference for Two-Way Tables

In Example 9A, if salt intake and CVD status are independent, then the expected cell counts are computed as follows.

| | Low salt | High salt | Total |
|---|---|---|---|
| CVD | $\frac{(200)(1169)}{2415} = 96.81$ | $\frac{(200)(1246)}{2415} = 103.19$ | 200 |
| No CVD | $\frac{(2215)(1169)}{2415} = 1072.19$ | $\frac{(2215)(1246)}{2415} = 1142.81$ | 2215 |
| Total | 1169 | 1246 | 2415 |

# 9.1 Inference for Two-Way Tables

## Chi-square Statistic

- The **chi-square statistic**

$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

is a measure of how much the observed cell counts in a two-way table diverge from the expected cell counts where the sum is over all $r \times c$ cells in the table.

# 9.1 Inference for Two-Way Tables

## Chi-square Statistic

- The **chi-square statistic**

$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

  is a measure of how much the observed cell counts in a two-way table diverge from the expected cell counts where the sum is over all $r \times c$ cells in the table.

- When the variables are independent, this statistic approximately follows a chi-square distribution with $(r-1)(c-1)$ degrees of freedom when the sample size is large enough (all expected cell counts are at least 5).

# 9.1 Inference for Two-Way Tables

## Chi-square Distribution

- The density curve of the chi-square distribution with df degrees of freedom has the form

$$f(x) = \frac{1}{2^{df/2}\Gamma(\frac{df}{2})} x^{df/2-1} e^{-x/2}, x > 0.$$

# 9.1 Inference for Two-Way Tables

## Chi-square Distribution

- The density curve of the chi-square distribution with df degrees of freedom has the form

$$f(x) = \frac{1}{2^{df/2}\Gamma(\frac{df}{2})} x^{df/2-1} e^{-x/2}, x > 0.$$

- This distribution is skewed to the right.

# 9.1 Inference for Two-Way Tables

## Chi-square Distribution

- The density curve of the chi-square distribution with df degrees of freedom has the form

$$f(x) = \frac{1}{2^{df/2}\Gamma(\frac{df}{2})} x^{df/2-1} e^{-x/2}, x > 0.$$

- This distribution is skewed to the right.
- Sometimes, we abbreviate the chi-square distribution with df degrees of freedom by $\chi^2(df)$.

# 9.1 Inference for Two-Way Tables

## Chi-square Distribution

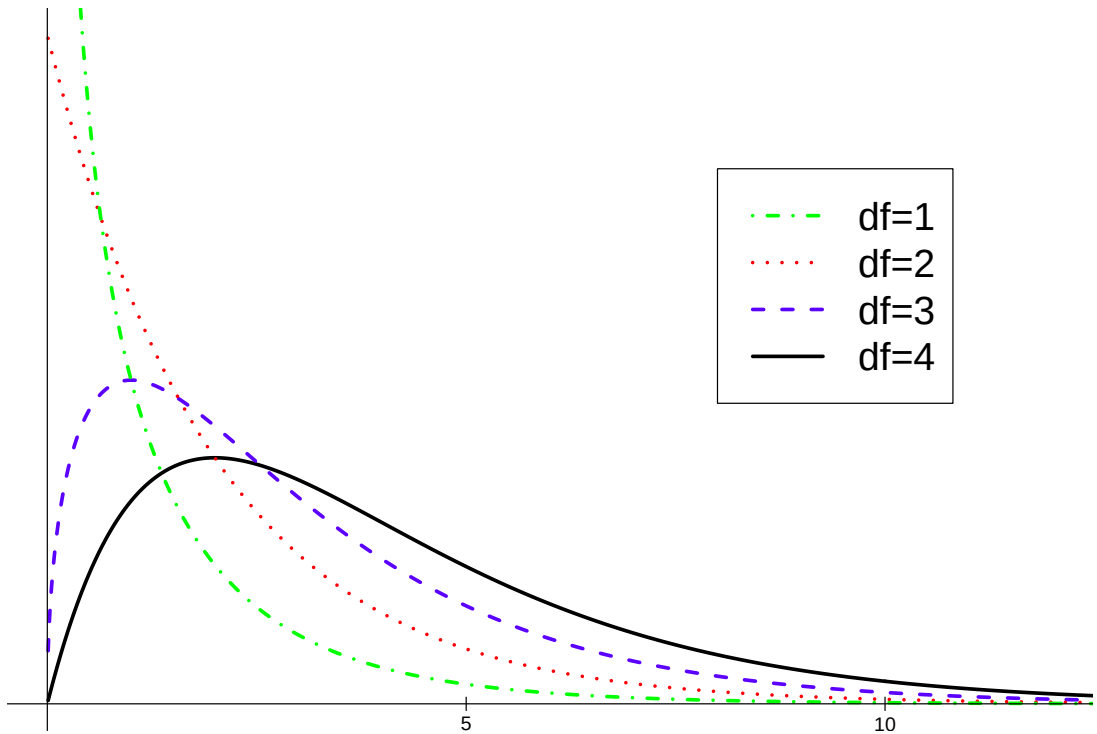
- The density curve of the chi-square distribution with df degrees of freedom has the form

$$f(x) = \frac{1}{2^{df/2}\Gamma(\frac{df}{2})} x^{df/2-1} e^{-x/2}, x > 0.$$

- This distribution is skewed to the right.
- Sometimes, we abbreviate the chi-square distribution with df degrees of freedom by $\chi^2(df)$.
- If $df$ is an integer, then the sum of $df$ squared independent standard normal random variables follows a $\chi^2(df)$-distribution.

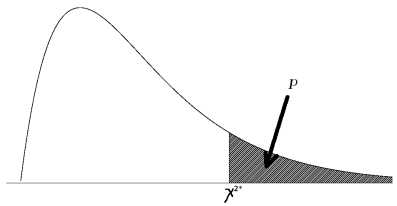
# 9.1 Inference for Two-Way Tables

## Chi-square Distribution



# 9.1 Inference for Two-Way Tables

## Table for $\chi^2$ distribution




**$\chi^2$ distribution critical values**

| | Upper-tail probability $p$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| df | .25 | .20 | .15 | .10 | .05 | .025 | .02 | .01 |
| 1 | 1.323 | 1.642 | 2.072 | 2.706 | 3.841 | 5.024 | 5.412 | 6.635 |
| 2 | 2.773 | 3.219 | 3.794 | 4.605 | 5.991 | 7.378 | 7.824 | 9.210 |
| 3 | 4.108 | 4.642 | 5.317 | 6.251 | 7.815 | 9.348 | 9.837 | 11.345 |
| | | | | ⋮ | | | | |
| 60 | 66.981 | 68.972 | 71.341 | 74.397 | 79.082 | 83.298 | 84.580 | 88.379 |
| 80 | 88.130 | 90.405 | 93.106 | 96.578 | 101.879 | 106.629 | 108.069 | 112.329 |
| 100 | 109.141 | 111.667 | 114.659 | 118.498 | 124.342 | 129.561 | 131.142 | 135.807 |

# 9.1 Inference for Two-Way Tables

Chi-square test for two-way tables
when all expected cell counts are at least 5

1. Test
   $H_0$ : there is no association between the row and column variables vs.
   $H_a$ : there is an association between the variables
2. Test statistic:
   $$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$
3. Use the $\chi^2$-table to find the value $\chi^{2*}$ such that $P(\chi^2 > \chi^{2*}) = \alpha$ where $\chi^2$ follows a $\chi^2$ distribution with $(r-1)(c-1)$ degrees of freedom
4. Reject $H_0$ if $X^2 > \chi^{2*}$

# 9.1 Inference for Two-Way Tables

- **Example 9B**: To examine the effects of salt intake on risk for CVD, a study collect the following data.

| | Low salt | High salt | Total |
|---|---|---|---|
| CVD | 88 | 112 | 200 |
| No CVD | 1081 | 1134 | 2215 |
| Total | 1169 | 1246 | 2415 |

Test the null hypothesis that salt intake and CVD status are not associated against the alternative that there is an association between these two variables at level .05.

# 9.1 Inference for Two-Way Tables

- *Answer*:

# 9.1 Inference for Two-Way Tables

- This test is equivalent to a test comparing whether two or more populations follow the same conditional distribution when response variable has two or more categories.

# 9.1 Inference for Two-Way Tables

- This test is equivalent to a test comparing whether two or more populations follow the same conditional distribution when response variable has two or more categories.
- **Example 9C**: To examine the effects of salt intake on risk for CVD, a study collect the following data.

| | Low salt | High salt | Total |
|---|---|---|---|
| CVD | 88 | 112 | 200 |
| No CVD | 1081 | 1134 | 2215 |
| Total | 1169 | 1246 | 2415 |

What are the estimates for the conditional probabilities of CVD status for each level of salt intake?

# 9.1 Inference for Two-Way Tables

- *Answer*:

# 9.1 Inference for Two-Way Tables

- **Example 9D**: To examine the effects of salt intake on risk for CVD, a study collect the following data.

| | Low salt | High salt | Total |
|---|---|---|---|
| CVD | 88 | 112 | 200 |
| No CVD | 1081 | 1134 | 2215 |
| Total | 1169 | 1246 | 2415 |

Perform a two-sided significance test of the hypothesis at level .05 of the statement that conditional distributions for CVD status are the same for each level of salt intake.

# 9.1 Inference for Two-Way Tables

- *Answer*:

# 9.1 Inference for Two-Way Tables

- Examples 9B and 9D illustrates the equivalence of the chi-square test for no association and the $z$-test of equality of population proportions for two populations.

# 9.1 Inference for Two-Way Tables

- Examples 9B and 9D illustrates the equivalence of the chi-square test for no association and the $z$-test of equality of population proportions for two populations.
- The advantage of the $z$-test for two populations is that we can test either one-sided or two-sided alternatives.

# 9.1 Inference for Two-Way Tables

- Examples 9B and 9D illustrates the equivalence of the chi-square test for no association and the $z$-test of equality of population proportions for two populations.
- The advantage of the $z$-test for two populations is that we can test either one-sided or two-sided alternatives.
- The advantage of the chi-square test is that we can compare more than two populations.

# 9.1 Inference for Two-Way Tables

The chi-square test provides a way of testing hypotheses from two models.

1. When comparing several populations with a single categorical response variable, we test

   $H_0$ : the distribution of the response variable is the same in all $c$ populations vs.

   $H_a$ : the distributions are not all the same.

# 9.1 Inference for Two-Way Tables

The chi-square test provides a way of testing hypotheses from two models.

1. When comparing several populations with a single categorical response variable, we test

   $H_0$ : the distribution of the response variable is the same in all $c$ populations vs.

   $H_a$ : the distributions are not all the same.

2. When testing independence for the variables in a two-way table, we test

   $H_0$ : there is no association between the row and column variables vs.

   $H_a$ : there is an association between the variables.

# 9.1 Inference for Two-Way Tables

- **Example 9E**: Here is some data on the relationship between pet ownership and gender.

| | Pet ownership status | | |
|---|---|---|---|
| Gender | Non-pet owners | Dog owners | Cat owners |
| Female | 1024 | 157 | 85 |
| Male | 915 | 171 | 82 |

Test the null hypothesis that gender and pet ownership status are independent against the alternative that they are dependent at level .10. Compute an appropriate test statistic, find the critical value of the distribution, and state your conclusion.

# 9.1 Inference for Two-Way Tables

- *Answer*:

## 9.2 Goodness of Fit

After completing this section, students should be able to:

- perform *goodness of fit tests* to determine if a categorical variable follows a hypothesized distribution.

# 9.2 Goodness of Fit

- In this section, we consider a different application of the chi-square test.

# 9.2 Goodness of Fit

- In this section, we consider a different application of the chi-square test.
- Here, we compare a sample of observed counts for a categorical variable from one population with a hypothesized distribution.

# 9.2 Goodness of Fit

- In this section, we consider a different application of the chi-square test.
- Here, we compare a sample of observed counts for a categorical variable from one population with a hypothesized distribution.
- In this setting, the expected cell counts equal the total sample size times $p_i$, the hypothesized probability for category $i$:

$$\text{expected cell count} = np_i.$$

# 9.2 Goodness of Fit

Chi-square goodness of fit test when the sample size is large

1. Test
   $H_0$ : the categorical variable follows the hypothesized distribution vs.
   $H_a$ : it does not follow the hypothesized distribution
2. Test statistic:
   $$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$
3. Use the $\chi^2$-table to find the value $\chi^{2*}$ such that $P(\chi^2 > \chi^{2*}) = \alpha$ where $\chi^2$ follows a $\chi^2$ distribution with $k-1$ degrees of freedom ($k$ is the number of categories)
4. Reject $H_0$ if $X^2 > \chi^{2*}$

# 9.2 Goodness of Fit

- **Example 9F**: M&M Mars Company recently reported that the color distribution for its M&M's is 13% brown, 14% yellow, 13% red, 20% orange, 24% blue, and 16% green. You open a bag of M&M's and find 61 brown, 59 yellow, 49 red, 77 orange, 141 blue, and 88 green. Use a goodness of fit test at level .01 to examine how well this bag fits the percents stated by the M&M's Mars Company.

# 9.2 Goodness of Fit

- *Answer*: