Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

November 19, 2019

Topics

Objectives

Objectives:

- Carry out a chi-squared test for homogeneity.
- Carry out a chi-squared test for independence.

Chi-Squared Tests for Two-Way Contingency Tables

Introduction

- **Correlation** and the t **test** for a regression **slope** are used to decide if there's an **association** between **two numerical** variables.

Chi-Squared Tests for Two-Way Contingency Tables

Introduction

- **Correlation** and the t **test** for a regression **slope** are used to decide if there's an **association** between **two numerical** variables.

  One-factor ANOVA and the **two-sample** t **test** are used decide if there's an **association** between a **numerical** variable and a **categorical** one.

Chi-Squared Tests for Two-Way Contingency Tables

Introduction

- **Correlation** and the t **test** for a regression **slope** are used to decide if there's an **association** between **two numerical** variables.

  One-factor ANOVA and the **two-sample** t **test** are used decide if there's an **association** between a **numerical** variable and a **categorical** one.

  The ***chi-squared test*** is used to decide if there's an **association** between **two categorical** variables.

Example

To determine if there's any **association** between a person's **socio-economic status** and their cigarette **smoking status**, ***three*** **random samples** of sizes **200**, **300**, and **300**, respectively, are drawn from each of ***three*** **populations** defined by **socio-economic class** (**SEC** – *high*, *middle*, and *low*).

Example

To determine if there's any **association** between a person's **socio-economic status** and their cigarette **smoking status**, ***three*** **random samples** of sizes **200**, **300**, and **300**, respectively, are drawn from each of ***three*** **populations** defined by **socio-economic class** (**SEC** – *high*, *middle*, and *low*).

The ***contingency table*** on the next slide shows the results.

Example

To determine if there's any **association** between a person's **socio-economic status** and their cigarette **smoking status**, ***three*** **random samples** of sizes **200**, **300**, and **300**, respectively, are drawn from each of ***three*** **populations** defined by **socio-economic class** (**SEC** – *high*, *middle*, and *low*).

The ***contingency table*** on the next slide shows the results.

Note that the sum of each row gives the sample size from that population.

		Smoking Status			Sample
		Current	Former	Never	Size
Population	High SEC	40	20	140	200
	Middle SEC	75	45	180	300
	Low SEC	105	60	135	300

Example

A study was carried out to decide if there's any **association** between **hair color** (*dark* or *light*) and **eye color** (*dark* or *light*).

Example

A study was carried out to decide if there's any **association** between **hair color** (*dark* or *light*) and **eye color** (*dark* or *light*).

A ***single*** **random sample** of $n = 6,800$ men was taken (from a ***single*** **population**), and each man ***cross-classified*** according to his **hair color** and **eye color**.

Example

A study was carried out to decide if there's any **association** between **hair color** (*dark* or *light*) and **eye color** (*dark* or *light*).

A ***single*** **random sample** of $n = 6,800$ men was taken (from a ***single*** **population**), and each man ***cross-classified*** according to his **hair color** and **eye color**.

The ***contingency table*** on the next slide shows the results.

Example

A study was carried out to decide if there's any **association** between **hair color** (*dark* or *light*) and **eye color** (*dark* or *light*).

A ***single*** **random sample** of $n = 6,800$ men was taken (from a ***single*** **population**), and each man ***cross-classified*** according to his **hair color** and **eye color**.

The ***contingency table*** on the next slide shows the results.

Note that the sum of the four table entries gives the sample size.

		Hair Color	
		Dark	Light
Eye Color	Dark	726	131
	Light	3,129	2,814

- The data are summarized in a ***contingency table*** having the form below.

- The data are summarized in a ***contingency table*** having the form below.

		Column Category				
		1	2	$\cdots$	J	**Total**
Row Category	1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
	2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I\cdot}$
Total		$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot J}$	n

- **Notation**: For a given contingency table,

$$
\begin{aligned}
I &= \text{The number of row categories.} \\
J &= \text{The number of column categories.} \\
n_{ij} &= \text{The } i,j\text{th } \underline{\textbf{\textit{cell count}}}. \\
n_{.j} &= \text{The } j\text{th } \underline{\textbf{\textit{column total}}}. \\
n_{i.} &= \text{The } i\text{th } \underline{\textbf{\textit{row total}}}. \\
n &= \text{The } \underline{\textbf{\textit{overall sample size}}}.
\end{aligned}
$$

Example

The **contingency table** below shows the **marginal row** and **marginal column totals** and the **overall sample size.**

		Smoking Status			Sample
		Current	Former	Never	Size
Population	High SEC	40	20	140	200
	Middle SEC	75	45	180	300
	Low SEC	105	60	135	300
	Total	220	125	455	800

Example

The **contingency table** below shows the **marginal row** and **marginal column totals** and the **overall sample size.**

		Hair Color		
		Dark	Light	**Total**
Eye Color	Dark	726	131	857
	Light	3,129	2,814	5,943
	Total	3,855	2,945	6,800

Chi-Squared Tests

- The ***chi-squared test*** is used in **two contexts**:

Chi-Squared Tests

- The ***chi-squared test*** is used in **two contexts**:
  - ***Separate*** **random samples** of sizes $n_{1\cdot}, n_{2\cdot}, ..., n_{I\cdot}$ from I **populations** defined by a **categorical variable**, where each individual is **classified** according to **one *other* categorical variable**.

Chi-Squared Tests

- The ***chi-squared test*** is used in **two contexts**:
    - ***Separate* random samples** of sizes $n_1.$, $n_2.$, ..., $n_I.$ from I **populations** defined by a **categorical variable**, where each individual is **classified** according to **one *other* categorical variable**.

      A **chi-squared test *test for homogeneity*** of the populations is used in this case.

Chi-Squared Tests

- The ***chi-squared test*** is used in **two contexts**:
  - ***Separate*** **random samples** of sizes $n_{1\cdot}$, $n_{2\cdot}$, ..., $n_{I\cdot}$ from I **populations** defined by a **categorical variable**, where each individual is **classified** according to **one *other* categorical variable.**

    A **chi-squared test *test for homogeneity*** of the populations is used in this case.
  - A ***single*** **random sample** of n individuals from a ***single*** **population** of individuals ***cross-classified*** according to **two categorical variables**.

Chi-Squared Tests

- The ***chi-squared test*** is used in **two contexts**:
  - ***Separate* random samples** of sizes $n_{1\cdot}$, $n_{2\cdot}$, ..., $n_{I\cdot}$ from I **populations** defined by a **categorical variable**, where each individual is **classified** according to **one *other* categorical variable**.

    A **chi-squared test *test for homogeneity*** of the populations is used in this case.
  - A ***single* random sample** of n individuals from a ***single* population** of individuals ***cross-classified*** according to **two categorical variables**.

    A **chi-squared test *test for independence*** between the two categorical variables is used in this case.

Test for Homogeneity

Test for Homogeneity

- **Notation**:

$$p_{ij} = \text{The } i\text{th population proportion in the } j\text{th category.}$$

		Column Category				
		1	2	$\cdots$	J	
Row Category	1	p_{11}	p_{12}	$\cdots$	p_{1J}	1.0
	2	p_{21}	p_{22}	$\cdots$	p_{2J}	1.0
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	p_{I1}	p_{I2}	$\cdots$	p_{IJ}	1.0

		Column Category				
		1	2	$\cdots$	J	
Row Category	1	p_{11}	p_{12}	$\cdots$	p_{1J}	1.0
	2	p_{21}	p_{22}	$\cdots$	p_{2J}	1.0
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	p_{I1}	p_{I2}	$\cdots$	p_{IJ}	1.0

(Note: The proportions sum to one in each row.)

- The **null hypothesis** in the ***test for homogeneity*** is:

Null Hypothesis:

H_0 : The population proportion for each column category is the same across the populations (rows).

We write this as

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{Ij} \qquad \text{for each } j = 1, 2, \ldots, J$$

- The **alternative hypothesis** will be

Alternative Hypothesis:

H_a : The population proportion for at least one column category is different across the populations (rows).

We can write this as

H_a : $p_{1j}, p_{2j}, \cdots, p_{Ij}$ Aren't all equal for at least one $j = 1, 2, \ldots, J$

- ***When*** H_0 is **true**, we can use $p_1, p_2, \ldots, p_J$ to denote the **common proportions** for the J categories.

- ***When*** $\boldsymbol{H_0}$ is **true**, we can use $\boldsymbol{p_1, p_2, \ldots, p_J}$ to denote the **common proportions** for the J categories.

  Then regardless of the population $\boldsymbol{i}$, the **expected count** for the jth category is

$$\text{Expected Count } = n_{i\cdot}\, p_j$$

- Replacing the unknown true proportions $p_1, p_2, \ldots, p_J$ by their **estimates** $\boldsymbol{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_J}$, where

$$\hat{p}_j = \frac{n_{\cdot j}}{n},$$

we get the ***estimated expected count*** (under H_0), denoted $\boldsymbol{\hat{e}_{ij}}$:

$$\hat{e}_{ij} = n_{i\cdot}\,\hat{p}_j = \frac{n_{i\cdot}n_{\cdot j}}{n} = \frac{(i\text{th row total})(j\text{th column total})}{n},$$

Chi-Squared Test Statistic for Homogeneity:

$$\chi^2 = \sum_{\text{all cells}} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

$$= \sum_{\text{all cells}} \frac{(\text{observed} - \text{estimated expected})^2}{\text{estimated expected}}$$

- The numerator of χ^2 will be large if the **observed** counts differ substantially from the **estimated expected** counts under H_0, so ...

- The numerator of χ^2 will be large if the **observed** counts differ substantially from the **estimated expected** counts under H_0, so ...

Large values of χ^2 provide **evidence against H_0 in favor of H_a : p_{1j} , p_{2j} , $\cdots$, p_{Ij} aren't all equal for at least one $j = 1, 2, \ldots, J$.**

Sampling Distribution of the Test Statistic Under H_0:
If χ^2 is the test statistic for homogeneity, and the sample sizes $n_{1\cdot}$, $n_{2\cdot}$, ..., $n_{I\cdot}$ are ***large***, then when H_0 is true,

$$\chi^2 \sim \chi^2((I-1)(J-1)).$$

- The **sample sizes** $n_{1\cdot}$**,** $n_{2\cdot}$**, ...,** $n_{I\cdot}$ are considered **large enough** as long as each of the **estimated expected counts** is **five** (or higher).

- **Comment**: The **df** are $(I-1)(J-1)$ because the deviations $n_{ij} - \hat{e}_{ij}$ sum to zero in each row and in each column, leaving only $(I-1)(J-1)$ **"free to vary"**.

- The $\chi^2((I-1)(J-1))$ curve gives us:

- The $\chi^2((I-1)(J-1))$ curve gives us:
    - The ***rejection region*** as the **extreme largest 100α% of χ^2 values**.

- The $\chi^2((I-1)(J-1))$ curve gives us:
  - The ***rejection region*** as the **extreme largest 100α% of χ^2 values**.
  - The ***p-value*** as the **tail area to the right of the observed χ^2 value**.

Exercise

The table below shows the **estimated expected** counts $\hat{e}_{ij}$ in parentheses:

		Smoking Status			Sample Size
		Current	Former	Never	
Population	High SEC	40 (55.0)	20 (31.3)	140 (113.8)	$n_{1\cdot} = 200$
	Middle SEC	75 (82.5)	45 (46.9)	180 (170.6)	$n_{2\cdot} = 300$
	Low SEC	105 (82.5)	60 (46.9)	135 (170.6)	$n_{3\cdot} = 300$
	Total	$n_{\cdot 1} = 220$	$n_{\cdot 2} = 125$	$n_{\cdot 3} = 455$	$n = 800$

We'll carry out a **chi-squared test for homogeneity** to decide if there's an **association** between **socio-economic class** and **smoking** status.

We'll carry out a **chi-squared test for homogeneity** to decide if there's an **association** between **socio-economic class** and **smoking** status.

a) Verify that the **sample sizes** $n_{1\cdot}$, $n_{2\cdot}$, and $n_{3\cdot}$ are **large enough** to justify the use of the **chi-squared test.**

We'll carry out a **chi-squared test for homogeneity** to decide if there's an **association** between **socio-economic class** and **smoking** status.

a) Verify that the **sample sizes** $n_{1\cdot}$, $n_{2\cdot}$, and $n_{3\cdot}$ are **large enough** to justify the use of the **chi-squared test.**

b) Compute the chi-squared **test statistic.**

We'll carry out a **chi-squared test for homogeneity** to decide if there's an **association** between **socio-economic class** and **smoking** status.

a) Verify that the **sample sizes** $n_{1\cdot}$, $n_{2\cdot}$, and $n_{3\cdot}$ are **large enough** to justify the use of the **chi-squared test.**

b) Compute the chi-squared **test statistic.**

Hint: You should get $\chi^2 = 32.7$.

We'll carry out a **chi-squared test for homogeneity** to decide if there's an **association** between **socio-economic class** and **smoking** status.

a) Verify that the **sample sizes** $n_{1\cdot}$, $n_{2\cdot}$, and $n_{3\cdot}$ are **large enough** to justify the use of the **chi-squared test.**

b) Compute the chi-squared **test statistic.**

Hint: You should get $\chi^2 = 32.7$.

c) Find the **p-value** and state the **conclusion** using $\alpha = 0.05$.

We'll carry out a **chi-squared test for homogeneity** to decide if there's an **association** between **socio-economic class** and **smoking** status.

a) Verify that the **sample sizes** $n_{1\cdot}$, $n_{2\cdot}$, and $n_{3\cdot}$ are **large enough** to justify the use of the **chi-squared test.**

b) Compute the chi-squared **test statistic.**

Hint: You should get $\boldsymbol{\chi^2 = 32.7}$.

c) Find the **p-value** and state the **conclusion** using $\alpha = 0.05$.

Hint: You should get **p-value** $<$ **0.001**.

Test for Independenc

Test for Independence

- **Notation**:

p_{ij} = The ***joint probability*** (or ***population proportion***) of the ith row and jth column cross-classification.

$p_{\cdot j} = \sum_i p_{ij}$ = The ***marginal probability*** of the ith row category.

$p_{i\cdot} = \sum_j p_{ij}$ = The ***marginal probability*** of the jth column category.

		Column Category				
		1	2	$\cdots$	J	
Row Category	1	p_{11}	p_{12}	$\cdots$	p_{1J}	$p_{1\cdot}$
	2	p_{21}	p_{22}	$\cdots$	p_{2J}	$p_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	p_{I1}	p_{I2}	$\cdots$	p_{IJ}	$p_{I\cdot}$
		$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot J}$	1.0

		Column Category				
		1	2	$\cdots$	J	
Row Category	1	p_{11}	p_{12}	$\cdots$	p_{1J}	$p_{1\cdot}$
	2	p_{21}	p_{22}	$\cdots$	p_{2J}	$p_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	p_{I1}	p_{I2}	$\cdots$	p_{IJ}	$p_{I\cdot}$
		$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot J}$	1.0

(Note: The probabilities in the IJ cells sum to one.)

Example

In **hair** and **eye color** study,

$$p_{1\cdot} = \text{The } \textbf{probability} \text{ that a randomly selected man has } \textbf{dark eyes.}$$

$$p_{\cdot 1} = \text{The } \textbf{probability} \text{ that he has } \textbf{dark hair.}$$

- Recall that two events A and B are ***independent*** if

$$P(A \mathbin{\&} B) = P(A)P(B).$$

- The **null hypothesis** in the ***test for independence*** is:

Null Hypothesis:

H_0 : An individual's row category is independent of that individual's column category.

- The **null hypothesis** in the ***test for independence*** is:

Null Hypothesis:

H_0 : An individual's row category is independent of that individual's column category.

We write this as

$$H_0 : p_{ij} = p_{i\cdot}\, p_{\cdot j} \qquad \text{for all pairs } i \text{ and } j$$

- The **alternative hypothesis** will be

Alternative Hypothesis:

H_a : An individual's row category is dependent on the individual's column category.

We can write this as

$H_0 : p_{ij} \neq p_{i.}\, p_{.j}$ for at least one pair i and j

- The **expected count** for the i, jth cell is

$$\text{Expected Count} = n\, p_{ij}$$

which, ***when*** $\boldsymbol{H_0}$ is **true**, is

$$\text{Expected Count} = n\, p_{i\cdot}\, p_{\cdot j}\ .$$

- Replacing the unknown true marginal probabilities $p_{i\cdot}$ and $p_{\cdot j}$ by their **estimates $\hat{p}_{i\cdot}$** and **$\hat{p}_{\cdot j}$**, with

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \qquad \text{and} \qquad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n},$$

- Replacing the unknown true marginal probabilities $p_{i\cdot}$ and $p_{\cdot j}$ by their **estimates** $\hat{\boldsymbol{p}}_{\boldsymbol{i}\cdot}$ and $\hat{\boldsymbol{p}}_{\cdot \boldsymbol{j}}$, with

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \qquad \text{and} \qquad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n},$$

  we get the ***estimated expected count*** (under H_0), denoted $\hat{\boldsymbol{e}}_{\boldsymbol{ij}}$:

$$\hat{e}_{ij} = n\,\hat{p}_{i\cdot}\,\hat{p}_{\cdot j} = \frac{n_{i\cdot}n_{\cdot j}}{n} = \frac{(i\text{th row total})(j\text{th column total})}{n}.$$

- Replacing the unknown true marginal probabilities $p_{i\cdot}$ and $p_{\cdot j}$ by their **estimates** $\hat{\boldsymbol{p}}_{\boldsymbol{i}\cdot}$ and $\hat{\boldsymbol{p}}_{\cdot\boldsymbol{j}}$, with

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \qquad \text{and} \qquad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n},$$

we get the ***estimated expected count*** (under H_0), denoted $\hat{\boldsymbol{e}}_{\boldsymbol{ij}}$:

$$\hat{e}_{ij} = n\,\hat{p}_{i\cdot}\,\hat{p}_{\cdot j} = \frac{n_{i\cdot}n_{\cdot j}}{n} = \frac{(i\text{th row total})(j\text{th column total})}{n}.$$

(Note: It's the **same** as the **estimated expected count** for the **test for homogeneity**).

Chi-Squared Test Statistic for Independence:

$$\chi^2 = \sum_{\text{all cells}} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

$$= \sum_{\text{all cells}} \frac{(\text{observed} - \text{estimated expected})^2}{\text{estimated expected}}$$

Note: It's the **same** as the **test statistic** for the **test for homogeneity.**

Sampling Distribution of the Test Statistic Under H_0:
If χ^2 is the test statistic for independence, and the sample size n is ***large***, then when H_0 is true,

$$\chi^2 \sim \chi^2((I-1)(J-1)).$$

- The **sample size n** is considered **large enough** as long as each of the **estimated expected counts** is **five** (or higher)**.**

Exercise

The table below shows the **estimated expected** counts $\hat{e}_{ij}$ in parentheses:

		Hair Color		
		Dark	Light	**Total**
Eye Color	Dark	726 (485.8)	131 (371.2)	$n_{1.} = 857$
	Light	3,129 (3,369.2)	2,814 (2,573.8)	$n_{2.} = 5,943$
	Total	$n_{.1} = 3,855$	$n_{.2} = 2,945$	$n = 6,800$

We'll carry out a **chi-squared test for independence** to decide if there's an **association** between **hair** and **eye color**.

We'll carry out a **chi-squared test for independence** to decide if there's an **association** between **hair** and **eye color**.

a) Verify that the **sample size** n is **large enough** to justify the use of the **chi-squared test**.

We'll carry out a **chi-squared test for independence** to decide if there's an **association** between **hair** and **eye color.**

a) Verify that the **sample size** n is **large enough** to justify the use of the **chi-squared test.**

b) Compute the chi-squared **test statistic.**

We'll carry out a **chi-squared test for independence** to decide if there's an **association** between **hair** and **eye color**.

a) Verify that the **sample size** n is **large enough** to justify the use of the **chi-squared test**.

b) Compute the chi-squared **test statistic**.

Hint: You should get $\boldsymbol{\chi^2 = 313.7}$.

We'll carry out a **chi-squared test for independence** to decide if there's an **association** between **hair** and **eye color**.

a) Verify that the **sample size** n is **large enough** to justify the use of the **chi-squared test**.

b) Compute the chi-squared **test statistic**.

Hint: You should get $\boldsymbol{\chi^2 = 313.7}$.

c) Find the **p-value** and state the **conclusion** using a level of significance $\alpha = 0.05$.

We'll carry out a **chi-squared test for independence** to decide if there's an **association** between **hair** and **eye color.**

a) Verify that the **sample size** n is **large enough** to justify the use of the **chi-squared test.**

b) Compute the chi-squared **test statistic.**

Hint: You should get $\boldsymbol{\chi^2 = 313.7}$.

c) Find the **p-value** and state the **conclusion** using a level of significance $\alpha = 0.05$.

Hint: You should get **p-value** $<$ **0.001**.

We'll carry out a **chi-squared test for independence** to decide if there's an **association** between **hair** and **eye color.**

a) Verify that the **sample size** n is **large enough** to justify the use of the **chi-squared test**.

b) Compute the chi-squared **test statistic.**

Hint: You should get $\boldsymbol{\chi^2 = 313.7}$.

c) Find the **p-value** and state the **conclusion** using a level of significance $\alpha = 0.05$.

Hint: You should get **p-value** $<$ **0.001**.

d) If there's an association, describe the **nature** of the **association**.