

STAT 100 Lecture 26:  
Analysis of Categorical Data  
Part 2:  
Contingency Table with One Margin Fixed. Test of  
Homogeneity

Nate Strawn

December 5 and 8

- 1 Categorical data. Examples.
- 2 Pearson's  $\chi^2$  test for goodness of fit

# Today's Agenda

- 1  $\chi^2$  test of homogeneity in a contingency table.
- 2  $Z$  test to compare two proportions.
- 3 Examples.

# Developing a $\chi^2$ test to Compare Two Diets

## Example

	<i>Excellent</i>	<i>Average</i>	<i>Poor</i>	<i>Total</i>
<i>Diet A</i>	37	24	19	80
<i>Diet B</i>	17	33	20	70
<i>Total</i>	54	57	39	150

The null hypothesis of 'homogeneity' or 'no difference between the diets' is  $H_0: p_{A1} = p_{B1}, p_{A2} = p_{B2}, p_{A3} = p_{B3}$

Under  $H_0$ , the estimated probabilities are

$$\hat{p}_1 = 54/150, \hat{p}_2 = 57/150, \hat{p}_3 = 39/150$$

Thus the expected frequencies in the first row are

$$80 \times \frac{54}{150} = \frac{80 \times 54}{150}, \quad \frac{80 \times 57}{150}, \quad \frac{80 \times 39}{150}$$

Notice the pattern in these calculations:

$$\text{Expected cell frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}.$$

# Developing a $\chi^2$ test to Compare Two Diets

## Example

*Observed and Expected Frequencies of the Diet Data*

	<i>Excellent</i>	<i>Average</i>	<i>Poor</i>
<i>Diet A (O)</i>	37	24	19
<i>Diet A (E)</i>	28.8	30.4	20.8
<i>Diet B (O)</i>	17	33	20
<i>Diet B (E)</i>	25.2	26.6	18.2

*The Values of  $(O - E)^2 / E$*

	<i>Excellent</i>	<i>Average</i>	<i>Poor</i>
<i>Diet A</i>	2.335	1.347	.156
<i>Diet B</i>	2.668	1.540	.178

$$\chi^2 = \sum_{\text{cells}} \frac{(O - E)^2}{E} = 8.224.$$

# Developing a $\chi^2$ test to Compare Two Diets

## Example

$\chi^2$  statistic has been computed from two independent samples; each contributes  $3-1 = 2$  d.f. because there are three categories. The added d.f.  $= 2 + 2 = 4$  must now be reduced by the number of parameters we have estimated. Since  $p_1$ ,  $p_2$ , and  $p_3$  satisfy the relation  $p_1 + p_2 + p_3 = 1$ , we have two undetermined parameters among them. Therefore, our  $\chi^2$  statistic has d.f.  $= 4-2 = 2$ .

With d.f.  $= 2$ , the tabulated upper 5% point of  $\chi^2$  is 5.99 (Appendix B, Table 5). Since the observed  $\chi^2 = 8.224 > 5.99$ , the null hypothesis is rejected at  $\alpha = .05$ . Therefore, a significant difference between the quality of the two diets is indicated by the data.

# Calculating degrees of freedom for $r \times c$ contingency table

## Fact

$$\begin{aligned}d.f \text{ of } \chi^2 &= r(c - 1) - (c - 1) \\ &= (r - 1)(c - 1) \\ &= (\text{No. of rows} - 1) \times (\text{No. of columns} - 1)\end{aligned}$$

# The $\chi^2$ Test of Homogeneity in a Contingency Table

## Rule

### Null hypothesis

*In each response category, the probabilities are equal for all the populations.*

### Test statistic

$$\chi^2 = \sum_{\text{cells}} \frac{(O - E)^2}{E}$$

$$d.f = (\text{No. of rows} - 1) \times (\text{No. of columns} - 1)$$

### Rejection region

$$\chi^2 \geq \chi_{\alpha}^2.$$

# Example. Germination of Seeds

## Example

	<i>Germinated</i>	<i>Not Germinated</i>	<i>Total</i>
<i>Treated (O)</i>	84	16	100
<i>Treated (E)</i>	86.40	13.60	
<i>Untreated (O)</i>	132	18	150
<i>Untreated (E)</i>	129.60	20.40	
<i>Total</i>	216	34	250

①  $H_0 : p_1 = p_2$  vs  $H_1 : p_1 \neq p_2$ .

②  $\chi^2 = \sum_{\text{cells}} (O - E)^2 / E$

③  $\chi^2 = \frac{(84-86.40)^2}{86.40} + \frac{(16-13.60)^2}{13.60} + \frac{(132-129.60)^2}{129.60} + \frac{(18-20.40)^2}{20.40}$   
 $= .067 + .424 + .044 + .282 = .817$

$d.f = (2 - 1)(2 - 1) = 1$

④ For  $\alpha = 0.05$   $\chi_{0.05}^2 = 3.84$ . **Rejection region**  $\chi^2 \geq 3.84$ .

⑤ Since  $.817 < 3.84$  we **do not** reject  $H_0$

# Z test to Compare Two Proportions

Independent Samples from Two Populations with Two Categories

	No. of Successes	No. of Failures	Sample Size
Population 1	$X$	$n_1 - X$	$n_1$
Population 2	$Y$	$n_2 - Y$	$n_2$

Let  $p_1$  and  $p_2$  be the probabilities of success for these populations.

$H_0 : p_1 = p_2$ .

$$\hat{p}_1 = X/n_1, \quad \hat{p}_2 = Y/n_2$$

When the sample sizes are LARGE we can use the Z test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } \hat{p} = \frac{X + Y}{n_1 + n_2}.$$

Level  $\alpha$  **rejection regions**  $H_1 : p_1 \neq p_2 \quad |Z| \geq z_{\alpha/2}$

$H_1 : p_1 < p_2 \quad |Z| \leq -z_{\alpha}$

$H_1 : p_1 > p_2 \quad |Z| \geq z_{\alpha}$

# Comparing $\chi^2$ and Z tests

## Fact

Although the test statistics  $Z$  and  $\chi^2$  appear to have quite different forms, there is an exact relation between them:  $Z^2 = \chi^2$  (for a  $2 \times 2$  contingency table **ONLY**)

$$Z^2 = \chi^2 \text{ (for a } 2 \times 2 \text{ contingency table \textbf{ONLY})}$$

Also  $z_{\alpha/2}^2 = \chi_{\alpha}^2$ , with d.f. = 1. For instance, for  $\alpha = .05$ ,

$$z_{.025}^2 = (1.96)^2 = 3.8416 = \chi_{0.5}^2 \text{ with d.f.}=1$$

Thus, the two test procedures are equivalent when the alternative hypothesis is **two-sided**.

However, if the alternative hypothesis is **one-sided**, such as  $H_1 : p_1 < p_2$ , **ONLY the Z test is appropriate**.

# Example. Germination of Seeds

## Example

	<i>Germinated</i>	<i>Not Germinated</i>	<i>Total</i>
<i>Treated</i>	84	16	100
<i>Untreated</i>	132	18	150
<i>Total</i>	216	34	250

$$H_0 : p_1 = p_2, H_1 : p_1 \neq p_2$$

$$\hat{p}_1 = 84/100 = .84, \quad \hat{p}_2 = 132/150 = .88,$$

$$\hat{p} = (84 + 132)/(100 + 150) = .864.$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})} \sqrt{(1/n_1) + (1/n_2)}} = -.904$$

Because the observed  $|Z|$  is smaller than  $z_{.025} = 1.96$ ,  $H_0$  is **not rejected** at  $\alpha = .05$ . Note that  $Z^2 = (-.904)^2 = .817$  agrees with the result  $\chi^2 = .817$  found before.

# Example. Germination of Seeds

## Example

**95% confidence interval.**

$$\hat{p}_1 = .84, \quad \hat{p}_2 = .88, \quad n_1 = 100, \quad n_2 = 150,$$

Therefore

$$\hat{p}_1 - \hat{p}_2 = .84 - .88 = -.04$$

$$\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} = .045$$

Since  $z_{0.05} = 1.96$  the 95% confidence interval is

$$-.04 \pm 1.96(0.045) = -.04 \pm .09 = (-.13, .05)$$

Since 0 is **inside** this confidence interval  $H_0$  was **not rejected**.

# Next time

- Read Section 13.4 from Johnson and Bhattacharyya
- FINAL REVIEW: Thursday, 11 December, from 5 to 7 pm in room 0126 ARM.
- Group Problems:

Group	1	2	3	4	5
Problem	13.11	13.13	13.15	13.21	13.23
Group	6	7	8	9	10
Problem	13.11	13.13	13.15	13.21	13.23