

Contingency Tables

Test of Independence

The background of the slide features several thick, light gray wavy lines that flow from the bottom left towards the right side, creating a sense of movement and depth.

Learning Outcomes

After studying the material, you should be able to:

- *Set up a contingency analysis table and perform a chi-square test of independence.*

Contingency Tables

- A *contingency table* or *cross-tabulation table* is a statistical table in which row entries classify data according to one variable and column entries classify data according to another variable. When there are r rows and c columns in the table, it is called an $r \times c$ contingency table.
 - The frequencies in the cells are called **cell frequencies**.
 - The total of the frequencies in each row or each column is called the **marginal frequency**.

$r \times c$ Contingency Tables

Categorical Variable 2

Categorical Variable 1

O_{11}	O_{12}	O_{1j}	O_{1c}
O_{21}	O_{22}	O_{2j}	O_{2c}
....
O_{i1}	O_{i2}	O_{ij}	O_{ic}
....
O_{r1}	O_{r2}	O_{rj}	O_{rc}

Test of Independence

The hypotheses for testing independence are

- The null hypothesis H_0 is simply that ***there is no association*** between the row and column variable.
- The alternative hypothesis H_a is that ***there is an association*** between the two variables. It doesn't specify a particular direction and can't really be described as one-sided or two-sided.

Example: 2x2 Contingency Table

- The table shows the data from a study of 91 patients who had a myocardial infarction (Snow 1965). One variable is treatment (propranolol versus a placebo), and the other is outcome (survival for at least 28 days versus death within 28 days).

		<u>OUTCOME</u>		Total
		Survival for at least 28 days	Death	
<u>Treatment</u>	Propranolol	38	7	45
	Placebo	29	17	46
Total		67	24	91

Hypothesis statement in Our Example

- Null hypothesis: the method of treating the myocardial infarction patients did not influence the proportion of patients who survived for at least 28 days.
- The alternative hypothesis is that the outcome (survival or death) depended on the treatment, meaning that the outcomes was the dependent variable and the treatment was the independent variable.

Calculation of Expected Cell Count

- To test the null hypothesis, we compare the observed cell frequencies to the expected cell frequencies if H_0 is true (i.e., no association).
 - If difference is large, then there is evidence of association
 - If difference is not large, then insufficient evidence to conclude an association
- The process of comparing observed counts with expected counts is called a **goodness-of-fit test.**

Test Statistic

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} = observed frequency in cell i

E_{ij} = expected or theoretical frequency in cell i

r = number of row categories

c = number of column categories

The number of degrees of freedom of an $r \times c$ contingency table is $(r - 1)(c - 1)$.

Contingency Analysis

EXPECTED CELL FREQUENCIES

$$E_{ij} = \frac{(\text{ith Row total})(\text{jth Column total})}{\text{Total sample size}}$$

Observed cell counts

		<u>OUTCOME</u>		Total
		Survival for at least 28 days	Death	
<u>Treatment</u>	Propranolol	38	7	45
	Placebo	29	17	46
	Total	67	24	91

Expected cell counts

		<u>OUTCOME</u>		Total
		Survival for at least 28 days	Death	
<u>Treatment</u>	Propranolol	33.13	11.87	45
	Placebo	33.87	12.13	46
	Total	67	24	91

Chi-Square Random Variable for Contingency Tables

It can be shown that under the null hypothesis the random variable associated with

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

has, to a good approximation, a chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom. The approximation works well if each of the estimated expected numbers E_{ij} is at least 5 and the smallest expected cell count is > 1 . Sometimes adjacent classes can be combined in order to meet these assumptions.

Chi-Square Test

- Some limitations:
 - Does not describe the magnitude or the direction of the association
 - Relies on “large sample theory”. Thus avoid use under these conditions (**Cochran’s guidelines**)
 - None of the expected cell counts less than 1.
 - No more than 20% of the expected cell frequencies are less than 5.

Example

A randomized, double blind, placebo-controlled experiment was conducted in which patients with Alzheimer's disease were given either extract of Ginkgo biloba (EGb) or a placebo for one year. The change in each patient's Alzheimer's disease Assessment Scale-Cognitive subscale (ADAS-Cog) score was measure. The results are given in the next table.

At 0.05 level, is the change in Assessment Scale-Cognitive subscale (ADAS-Cog) independent of treatment?

Example (continued)

Change in ADAS-Cog Score

	<u>-4 or better</u>	<u>-2 to +1</u>	<u>+1 or worth</u>	<u>Totals</u>
EGb	54	25	32	111
Placebo	19	22	23	64
Totals	73	47	55	175

- First, compute the expected values and contributions to χ^2 for each of the six cells.
- Then to the hypothesis test....

Example (Continued)

Change in ADAS-Cog Score

		<u>-4 or better</u>	<u>-2 to +1</u>	<u>+1 or</u>
<u>worth</u>				
EGb:	O -	54	25	32
	E -	46.3029	29.8114	34.8857
	χ^2 contribution -	1.2795	0.7765	0.2387
Placebo:	O -	19	22	23
	E -	26.6971	17.1886	20.1143
	χ^2 contribution -	2.2192	1.3468	0.4140

$\Sigma \chi^2$ contributions = 6.2747

Example (Continued)

- H_0 : The change in ADAS-Cog score independent of treatment

H_1 : The change in ADAS-Cog score and treatment are not independent.

- $\alpha = 0.05$

$$df = (r - 1)(k - 1) = (2 - 1) \cdot (3 - 1) = 1 \cdot 2 = 2$$

- Test Statistic : $\chi^2 = 6.2747$

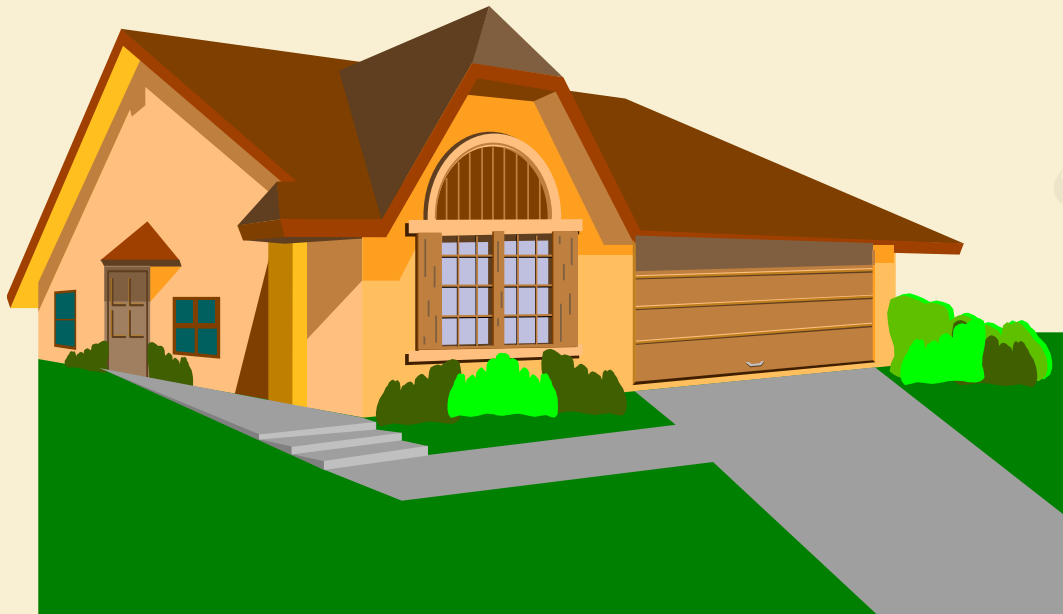
- **p -Value:** $0.025 < p\text{-Value} < 0.05$

- Conclusion: we reject the null with 95% confidence.

- Implications: There is enough evidence to show that the change in ADAS-Cog score is not independent from treatment.

χ^2 Test of Independence: Example

A Survey was conducted to determine whether there is a relationship between heart disease (yes or no) and treatment (placebo or aspirin).



Given the survey data, test at the 99% level to determine whether there is a relationship between heart disease and treatment.

χ^2 Test of Independence: Example

(continued)

1. Set hypothesis:

H_0 : the two categorical variables (treatment and heart disease) are independent

H_1 : the two categorical variables are related

2. Contingency table:

		Treatment		Total
		Aspirin	Placebo	
Levels of Variable 1	No	63	49	112
	Yes	15	33	48
Total		78	82	160

Levels of Variable 2

χ^2 Test of Independence: Example

(continued)

3. Computing expected frequencies

	Treatment				Total
	Aspirin		Placebo		
Heart disease	Obs.	Exp.	Obs.	Exp.	
No	63	54.6	49	57.4	112
Yes	15	23.4	33	24.6	48
Total	78	78	82	82	160

$\frac{78 \cdot 112}{160}$ $\frac{82 \cdot 112}{160}$

χ^2 Test of Independence: Example

(continued)

4. Calculate Test Statistic:

$$\chi^2 = \sum_{\text{All Cells}} \frac{(f_o - f_e)^2}{f_e}$$

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
63	54.6	8.4	70.56	1.292
49	57.4	-8.4	70.56	1.229
15	23.4	-8.4	70.56	3.015
33	24.6	8.4	70.56	2.868

$$\chi^2 \text{ Test Statistic} = 8.404$$

χ^2 Test of Independence: Example Solution

H_0 : The two categorical variables (treatment and heart disease) are independent

H_1 : The two categorical variables are related

$$df = (r - 1)(c - 1) = 1$$

Decision:

Reject H_0 if p-value < 1%

Conclusion:

Since p-value < 0.01, there is evidence that the choice of treatment and heart disease are related.

