Co-financed by the Connecting Europe Facility of the European Union

Contingency Table Analysis

Data Driven Healthcare
Module B
Prof. Paola Cerchiello – University of Pavia

Goals (1 of 2)

you should be able to:

- Set up a contingency analysis table and perform a chi-square test of association
- Recognize when and how to use the Wilcoxon signed rank test for paired or matched samples

Introduction

- Nonparametric Statistics
  - Fewer restrictive assumptions about data levels and underlying probability distributions
    - Population distributions may be skewed
    - The level of data measurement may only be ordinal or nominal

Contingency Tables

- Used to classify sample observations according to a pair of attributes
- Also called a cross-classification or cross-tabulation table
- Assume *r* categories for attribute A and *c* categories for attribute B
- Then there are $(r \times c)$ possible cross-classifications

r Times *c* Contingency Table

	Attribute B				
Attribute A	**1**	**2**	**. . .**	***c***	**Totals**
1	O_{11}	O_{12}	…	O_{1c}	R_1
2	O_{21}	O_{22}	…	O_{2c}	R_2
.	.	.	…	.	.
.	.	.	…	.	.
.	.	.	…	.	.
r	O_{r1}	O_{r2}	…	O_{rc}	R_r
Totals	C_1	C_2	…	C_c	n

Test for Association (1 of 2)

- Consider n observations tabulated in an $r \times c$ contingency table
- Denote by O_{ij} the number of observations in the cell that is in the i^{th} row and the j^{th} column
- The null hypothesis is

H_0 : No association exists between the two attributes in the population

- The appropriate test is a chi-square test with degrees of freedom

$$(r-1)(c-1)$$

Test for Association (2 of 2)

- Let R_i and C_j be the row and column totals
- The expected number of observations in cell row *i* and column *j*, given that H_0 is true, is

$$E_{ij} = \frac{R_i C_j}{n}$$

- A test of association at a significance level *α* is based on the chi-square distribution and the following decision rule

$$\text{Reject } H_0 \text{ if } \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} > \chi^2_{(r-1)(c-1),\alpha}$$

Contingency Table Example (1 of 2)

Left-Handed vs. Gender

- Dominant Hand: Left vs. Right
- Gender: Male vs. Female

H_0 : There is no association between hand preference and gender

H_1 : Hand preference is not independent of gender

Contingency Table Example (2 of 2)

Sample results organized in a contingency table:

sample size = n = 300:

120 Females, 12 were left handed

180 Males, 24 were left handed

Gender	Hand Preference: Left	Hand Preference: Right	
Female	12	108	120
Male	24	156	180
	36	264	300

Logic of the Test

H_0 : There is no association between hand preference and gender

H_1 : Hand preference is not independent of gender

- If H_0 is true, then the proportion of left-handed females should be the same as the proportion of left-handed males
  - The two proportions above should be the same as the proportion of left-handed people overall

Finding Expected Frequencies

120 Females, 12 were left handed

180 Males, 24 were left handed

➡

Overall:

$$P(\text{Left Handed}) = \frac{36}{300} = .12$$

If no association, then

$$P(\text{Left Handed} \mid \text{Female}) = P(\text{Left Handed} \mid \text{Male}) = .12$$

So we would expect 12% of the 120 females and 12% of the 180 males to be left handed…

i.e., we would expect

$(120)(.12) = 14.4$ **females to be left handed**

$(180)(.12) = 21.6$ **males to be left handed**

Expected Cell Frequencies

- Expected cell frequencies:

$$E_{ij} = \frac{R_i C_j}{n} = \frac{\left(i^{\text{th}} \text{ Row total}\right)\left(j^{\text{th}} \text{ Column total}\right)}{\text{Total sample size}}$$

Example:

$$E_{11} = \frac{(120)(36)}{300} = 14.4$$

Observed vs. Expected Frequencies (3 of 4)

Observed frequencies vs. expected frequencies:

Gender	Hand Preference		
	Left	Right	
Female	Observed = 12 Expected = 14.4	Observed = 108 Expected = 105.6	120
Male	Observed = 24 Expected = 21.6	Observed = 156 Expected = 158.4	180
	36	264	300

The Chi-Square Test Statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \quad \text{with } \text{d.f.} = (r-1)(c-1)$$

- where:

  $O_{ij} =$ observed frequency in cell (*i, j*)

  $E_{ij} =$ expected frequency in cell (*i, j*)

  r = number of rows

  c = number of columns

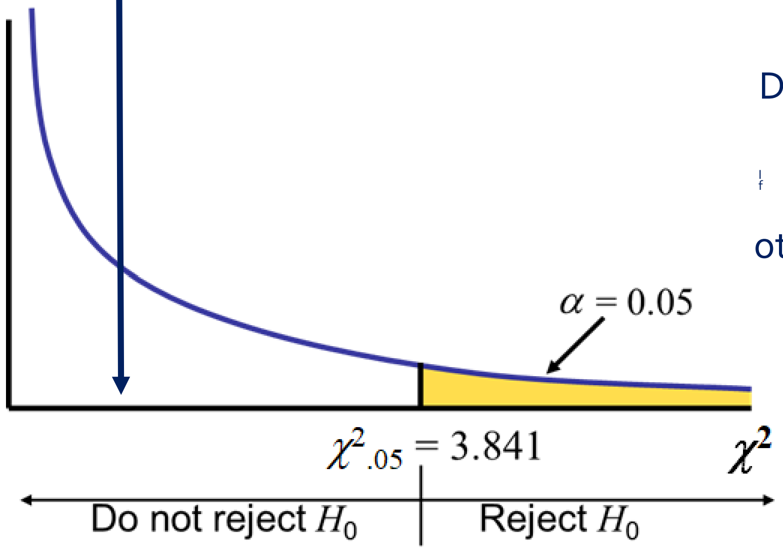
Observed vs. Expected Frequencies (4 of 4)

Gender	Hand Preference		
	Left	Right	
Female	Observed = 12 Expected = 14.4	Observed = 108 Expected = 105.6	120
Male	Observed = 24 Expected = 21.6	Observed = 156 Expected = 158.4	180
	36	264	300

$$\chi^2 = \frac{(12-14.4)^2}{14.4} + \frac{(108-105.6)^2}{105.6} + \frac{(24-21.6)^2}{21.6} + \frac{(156-158.4)^2}{158.4} = 0.7576$$

Contingency Analysis

$\chi^2 = 0.7576$ with $\text{d.f.} = (r-1)(c-1) = (1)(1) = 1$



Decision Rule:

If $\chi^2 > 3.841$, reject H_0,

otherwise, do not reject H_0

Here, $\chi^2 = 0.7576 < 3.841$,

so we do not reject H_0 and conclude that gender and hand preference are not associated