HEADS

PhD PROGRAMME IN HEALTH DATA SCIENCE

Chi-squared test

1. Contingency tables
2. Chi-squared test
3. McNemar test

Basics of Health Intelligent Data Analysis

PhD Programme in Health Data Science

Cláudia Camila Dias

Pedro Pereira Rodrigues

Categorical X Categorical

- So far we have considered situations involving a continuous variable
- How can we build hypothesis tests that only involve categorical variables?
- To study the relationship between two categorical variables we can use contingency tables (double entry tables or cross tables)

Contingency Tables

Suppose we want to study the relationship between smoking during pregnancy (cigb4) and the mother's age group (mage) - both categorical variables

```
> with(alcohol, CrossTable(mage,cigb4,format="SPSS", digits=1, prop.r=F, prop.t=F, prop.chisq=F, prop.c=F, expected=F))
```

mage	cigb4 No	Yes	Row Total
13-20	13	8	21
21-30	216	70	286
31-35	79	11	90
36-55	49	6	55
Column Total	357	95	452

Contingency Tables

```
> with(alcohol, CrossTable(mage,cigb4,format="SPSS", digits=1, prop.r=T, prop.t=T, prop.chisq=F, prop.c=T, expected=F))
```

```
   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  452
```

```
             | cigb4
        mage |        No  |       Yes  | Row Total |
-------------|------------|------------|-----------|
       13-20 |        13  |         8  |       21  |
             |     61.9%  |     38.1%  |     4.6%  |
             |      3.6%  |      8.4%  |           |
             |      2.9%  |      1.8%  |           |
-------------|------------|------------|-----------|
       21-30 |       216  |        70  |      286  |
             |     75.5%  |     24.5%  |    63.3%  |
             |     60.5%  |     73.7%  |           |
             |     47.8%  |     15.5%  |           |
-------------|------------|------------|-----------|
       31-35 |        79  |        11  |       90  |
             |     87.8%  |     12.2%  |    19.9%  |
             |     22.1%  |     11.6%  |           |
             |     17.5%  |      2.4%  |           |
-------------|------------|------------|-----------|
       36-55 |        49  |         6  |       55  |
             |     89.1%  |     10.9%  |    12.2%  |
             |     13.7%  |      6.3%  |           |
             |     10.8%  |      1.3%  |           |
-------------|------------|------------|-----------|
Column Total |       357  |        95  |      452  |
             |     79.0%  |     21.0%  |           |
-------------|------------|------------|-----------|
```

- 38% of the mothers in age group of 13-20 years, smoked during pregnancy
- 74% of mothers who smoked during pregnancy are in age group of 21-30 years
- 1.3% of mothers smoked and are in age group of 36-55 years

Chi-squared test

```
> with(alcohol, CrossTable(mage,cigb4,format="SPSS", digits=1, prop.r=T, prop.t=F, prop.chisq=F, prop.c=F, expected=F))

   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|-------------------------|

Total Observations in Table:  452

             | cigb4
        mage |        No  |       Yes  | Row Total |
-------------|-----------|-----------|-----------|
       13-20 |        13  |         8  |        21  |
             |     61.9%  |     38.1%  |      4.6%  |
-------------|-----------|-----------|-----------|
       21-30 |       216  |        70  |       286  |
             |     75.5%  |     24.5%  |     63.3%  |
-------------|-----------|-----------|-----------|
       31-35 |        79  |        11  |        90  |
             |     87.8%  |     12.2%  |     19.9%  |
-------------|-----------|-----------|-----------|
       36-55 |        49  |         6  |        55  |
             |     89.1%  |     10.9%  |     12.2%  |
-------------|-----------|-----------|-----------|
Column Total |       357  |        95  |       452  |
-------------|-----------|-----------|-----------|
```

Looking for the % of smokers in each age group there seems to be a relationship between the variables:

- We find more smokers in younger mothers than in older mothers

Chi-squared test

Let's assume that the two variables are not associated:

- H_0: % of the smokers are identical in all age groups

or

- H_0: Tobacco use is independent of age group

Chi-squared test

- If H0 were true, which table would we expect to look at?
- In total 21% of mothers smoked during pregnancy, so if smoking is not dependent on age group we should observe the same % of smokers in all age groups.
- If H0 is true, the expected value for the number of smokers in the 21-30 year age group (E22) would be 21% of 286 mothers = 60.1

	Did not smoke	Smoked	Total
13-20			21
21-30		E_{22}=21%x286	286
31-35			90
36-55			55
Total	357 *(79%)*	95 *(21%)*	452

Chi-squared test

```
> with(alcohol, CrossTable(mage,cigb4,format="SPSS", digits=1, prop.r=F, prop.t=F, prop.chisq=F, prop.c=F, expected=T))

   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
|-------------------------|

Total Observations in Table:  452

             | cigb4
        mage |        No  |       Yes  | Row Total |
-------------|-----------|-----------|-----------|
       13-20 |        13  |         8  |       21  |
             |      16.6  |       4.4  |           |
-------------|-----------|-----------|-----------|
       21-30 |       216  |        70  |      286  |
             |     225.9  |      60.1  |           |
-------------|-----------|-----------|-----------|
       31-35 |        79  |        11  |       90  |
             |      71.1  |      18.9  |           |
-------------|-----------|-----------|-----------|
       36-55 |        49  |         6  |       55  |
             |      43.4  |      11.6  |           |
-------------|-----------|-----------|-----------|
Column Total |       357  |        95  |      452  |
-------------|-----------|-----------|-----------|
```

Chi-squared test

Chi squared statistic: sum of the differences (squared and relativized) between the expected values and observed values of each cell:

$$\chi^2 = \sum \frac{(E-O)^2}{E}$$

Chi-squared test

- If the statistic value is too large it means that the expected values (assuming true null hypothesis) are quite different from those observed.

- That is, if H0 is true, the probability of obtaining the difference between the expected and observed values we observed (or more extreme), i.e. the p-value, is very small and we should reject H0.

- On the other hand, a chi-squared value close to 0 (expected values identical to those observed) will have a very high associated p-value.

Chi-squared test

The chi-squared test is then used for independent samples. As in the other tests:

We defined the hypothesis:

H_0: There is no association between the categories of one factor and the categories of the other

We set the level of significance (alpha) - usually 0.05

We get the test statistic with the sample data

$$\chi^2 = \frac{(O-E)^2}{E}$$ Follows a chi squared distribution with n-1 degree of freedom

O – observed values
E – expected value if H_0 true

We get the p-value

We interpret the value of p

Chi-squared test

```
> with(alcohol, CrossTable(mage,cigb4,format="SPSS", digits=1, prop.r=F, prop.t=F, prop.chisq=F, prop.c=F, expected=T, chisq=T))

   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
|-------------------------|

Total Observations in Table:  452


 Statistics for All Table Factors


 Pearson's Chi-squared test
 ------------------------------------------------------------
 Chi^2 =  13.32908     d.f. =  3     p =  0.003976397



       Minimum expected frequency: 4.413717
 Cells with Expected Frequency < 5: 1 of 8 (12.5%)
```

```
             | cigb4
        mage |       No  |       Yes  | Row Total |
-------------|-----------|-----------|-----------|
       13-20 |       13  |         8  |        21 |
             |     16.6  |       4.4  |           |
-------------|-----------|-----------|-----------|
       21-30 |      216  |        70  |       286 |
             |    225.9  |      60.1  |           |
-------------|-----------|-----------|-----------|
       31-35 |       79  |        11  |        90 |
             |     71.1  |      18.9  |           |
-------------|-----------|-----------|-----------|
       36-55 |       49  |         6  |        55 |
             |     43.4  |      11.6  |           |
-------------|-----------|-----------|-----------|
Column Total |      357  |        95  |       452 |
-------------|-----------|-----------|-----------|
```

Chi-squared test

Assumptions:

At most 20% of expected values are less than 5.

If the assumption is not fulfilled, then

```
Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  13.32908     d.f. =  3     p =  0.003976397

Fisher's Exact Test for Count Data
------------------------------------------------------------
Alternative hypothesis: two.sided
p =  0.003244598

       Minimum expected frequency: 4.413717
Cells with Expected Frequency < 5: 1 of 8 (12.5%)
```

Fisher's exact test

Based on hypergeometric distribution, the probability of obtaining the found set of values is computed using factorials and binomial coefficients, hence increasing the computational effort.

For hand calculations, the test is only feasible in the case of a 2 × 2 contingency table. However the principle of the test can be extended to the general case of an $m \times n$ table.

McNemar test

Paired Sample

One hundred patients with frequent headaches were evaluated.

The same 100 patients took one drug A for one month, and drug B the next month.

Patients were asked to record whether or not they had headaches each month.

	A – Without headaches	A – With headaches
B – Without headaches	45	4
B – With headaches	17	34
Total	62	38

McNemar test

Paired Sample

- In this situation, the values in the main diagonal support the null hypothesis (45+34), so we ignore them for the particular test statistic...

- But, if the drugs were identical, found diferences should be random; therefore, V should be identical to U.

	A – Without headaches	A – With headaches
B – Without headaches	45	4 (U)
B – With headaches	17 (V)	34
Total	62	38

McNemar test

Mcnemar’s test is based on the statistic

$$\chi^2 = \frac{(V-U)^2}{V+U} = \frac{(17-4)^2}{17+4} \approx 8$$

	A – Without headaches	A – With headaches
B – Without headaches	45	U=4
B – With headaches	V=17	34
Total	62	38

McNemar test

```
> CrossTable(b,a,format="SPSS", digits=1, prop.t=F, mcnemar = T)

   Cell Contents
|-------------------------|
|                   Count |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|-------------------------|

Total Observations in Table:  100

             | a
           b |      com  |      sem  | Row Total |
-------------|-----------|-----------|-----------|
         com |       34  |       17  |       51  |
             |     11.0  |      6.8  |           |
             |    66.7%  |    33.3%  |    51.0%  |
             |    89.5%  |    27.4%  |           |
-------------|-----------|-----------|-----------|
         sem |        4  |       45  |       49  |
             |     11.5  |      7.0  |           |
             |     8.2%  |    91.8%  |    49.0%  |
             |    10.5%  |    72.6%  |           |
-------------|-----------|-----------|-----------|
Column Total |       38  |       62  |      100  |
             |    38.0%  |    62.0%  |           |
-------------|-----------|-----------|-----------|
```

```
McNemar's Chi-squared test
------------------------------------------------------------
Chi^2 =  8.047619     d.f. =  1     p =  0.00455635

McNemar's Chi-squared test with continuity correction
------------------------------------------------------------
Chi^2 =  6.857143     d.f. =  1     p =  0.008828761


        Minimum expected frequency: 18.62
```