# Chi Square test and contingency table

# $\chi^2$ Test

Another test that can be used for collocation detection.

### 5.3.3 Pearson's chi-square test

Use of the $t$ test has been criticized because it assumes that probabilities are approximately normally distributed, which is not true in general (Church and Mercer 1993: 20). An alternative test for dependence which does not assume normally distributed probabilities is the $\chi^2$ test (pronounced 'chi-square test'). In the simplest case, the $\chi^2$ test is.applied to 2-by-2 tables like table 5.8. The essence of the test is to compare the observed frequencies in the table with the frequencies expected for independence. If the difference between observed and expected frequencies is large, then we can reject the null hypothesis of independence.

Table 5.8 shows the distribution of *new* and *companies* in the reference corpus that we introduced earlier. Recall that $C(new) = 15{,}828$, $C(companies) = 4{,}675$, $C(new\ companies) = 8$, and that there are 14,307,668 tokens in the corpus. That means that the number of bigrams $w_i w_{i+1}$ with the first token not being *new* and the second token being *companies* is $4667 = 4675 - 8$. The two cells in the bottom row are computed in a similar way.

The $\chi^2$ statistic sums the differences between observed and expected values in all squares of the table, scaled by the magnitude of the expected values, as follows:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{5.6}$$

Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

# $\chi^2$ Test

Für den $\chi^2$-Test brauchen wir die Kontingenztabelle:

| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} = 8$ | $O_{12} = 15820$ | $O_{1\_} = 15828$ |
| $w_1 \neq new$ | $O_{21} = 4667$ | $O_{22} = 14287173$ | $O_{2\_} = 14291840$ |
| | $O_{\_1} = 4675$ | $O_{\_2} = 14302993$ | $O_{\_\_} = 14307668$ |

Die $\chi^2$-Teststatistik wird folgendermaßen berechnet:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{wobei } E_{ij} = p_{i\_}p_{\_j}O_{\_\_} = \frac{O_{i\_}O_{\_j}}{O_{\_\_}}$$

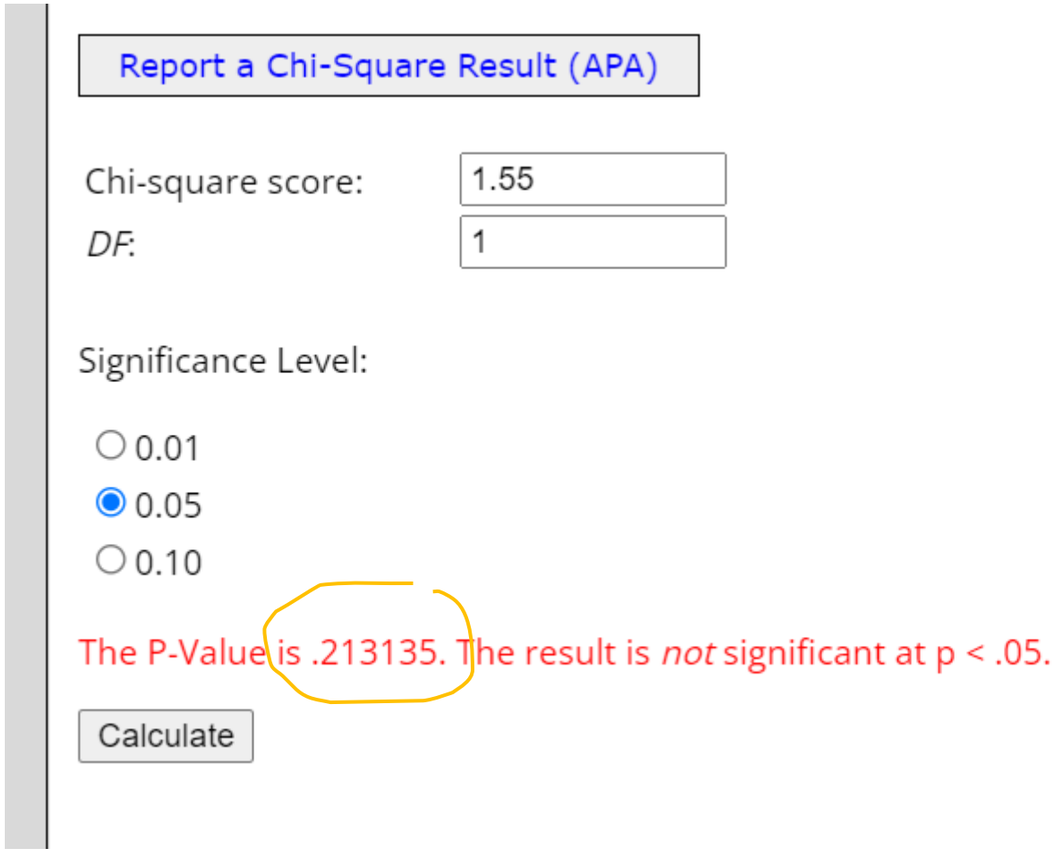
Der $\chi^2$-Wert misst die Abweichung von den erwarteten Werten in der Kontingenz-Tabelle. Je größer er ist, desto kleiner ist der p-Wert.

Im Beispiel erhalten wir einen $\chi^2$-Wert von 1.55, der einem p-Wert von 0.21 entspricht. Das Ergebnis ist also auch bei diesem Test **nicht signifikant**.

Wie die Ergebnisse zeigen, können sich die p-Werte verschiedener statistischer Tests deutlich unterscheiden.

Method:
- fill the Contingency table
- compute X^2
- compute p-value of X^2
- see if the p-value is significant (it is significant if it is lower than the significant level that we picked)
- decide if we will reject the null hypothesis or not

https://www.socscistatistics.com/pvalues/chidistribution.aspx



# $\chi^2$ Test

Für den $\chi^2$-Test brauchen wir die Kontingenztabelle:

| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} = 8$ | $O_{12} = 15820$ | $O_{1-} = 15828$ |
| $w_1 \neq new$ | $O_{21} = 4667$ | $O_{22} = 14287173$ | $O_{2-} = 14291840$ |
| | $O_{-1} = 4675$ | $O_{-2} = 14302993$ | $O_{--} = 14307668$ |

Die $\chi^2$-Teststatistik wird folgendermaßen berechnet:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{wobei } E_{ij} = p_{i-}p_{-j}O_{--} = \frac{O_{i-}O_{-j}}{O_{--}}$$

$O_{ij}$ ist der beobachtete Wert aus der Kontingenztabelle.
$E_{ij}$ sind die erwarteten Werte unter Annahme der Nullhypothese.

$p_{1-} = \frac{O_{1-}}{O_{--}}$ ist die Wahrscheinlichkeit, dass *new* das 1. Wort ist.

$p_{2-} = \frac{O_{2-}}{O_{--}}$ ist die Wahrscheinlichkeit, dass *new* **nicht** das 1. Wort ist.

$p_{-1} = \frac{O_{-1}}{O_{--}}$ ist die Wahrscheinlichkeit, dass *companies* das 2. Wort ist.

$p_{-2} = \frac{O_{-2}}{O_{--}}$ ist die Wahrscheinlichkeit, dass *companies* **nicht** das 2. Wort ist.

## How to fill the table

set of words that are not „companies“

| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} =$ | $O_{12} =$ | $O_{1-} =$ |
| $w_1 \neq new$ | $O_{21} =$ | $O_{22} =$ | $O_{2-} =$ |
| | $O_{-1} =$ | $O_{-2} =$ | $O_{--} =$ |

the number of (W1=new, W2= companies)

the number of
(W1=any word that is **not „**new“, W2= companies)
e.g. (W1=the, W2= companies)

the number of
(W1= any word, W2= companies)
e.g. (W1=new, W2= companies)
(W2=the, W2= companies)

- Fill what we know
  - f(new) = 15828
  - f(companies) = 4675
  - f(new, company) = 8
  - n = 14307668

new: 15828
n=14307668
company: 4675

| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} =$ 8 | $O_{12} =$ | $O_{1-} =$ 15828 |
| $w_1 \neq new$ | $O_{21} =$ | $O_{22} =$ | $O_{2-} =$ |
| | $O_{-1} =$ 4675 | $O_{-2} =$ | $O_{--} =$ 14307668 |

Fill the rest (in the correct order)

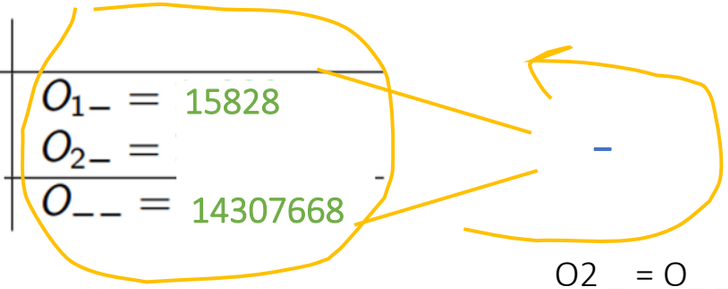
| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} =$ 8 | $O_{12} =$ | $O_{1-} =$ 15828 |
| $w_1 \neq new$ | $O_{21} =$ | $O_{22} =$ | $O_{2-} =$ |
| | $O_{-1} =$ 4675 | $O_{-2} =$ | $O_{--} =$ 14307668 |

-

O _2 = O_ _ - O _ 1
= 14307668 - 4675
= 14,302,993

| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} =$ 8 | $O_{12} =$ | $O_{1-} =$ 15828 |
| $w_1 \neq new$ | $O_{21} =$ | $O_{22} =$ | $O_{2-} =$ |
| | $O_{-1} =$ 4675 | $O_{-2} =$ 14,302,993 | $O_{--} =$ 14307668 |

| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} =$ 8 | $O_{12} =$ | $O_{1\_} =$ 15828 |
| $w_1 \neq new$ | $O_{21} =$ | $O_{22} =$ | $O_{2\_} =$ |
| | $O_{\_1} =$ 4675 | $O_{\_2} =$ 14,302,993 | $O_{\_\_} =$ 14307668 |



O2 _ = O_ _ - O _ 1
= 14307668 – 15828
= 14,291,840

| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} =$ 8 | $O_{12} =$ | $O_{1\_} =$ 15828 |
| $w_1 \neq new$ | $O_{21} =$ | $O_{22} =$ | $O_{2\_} =$ 14,291,840 |
| | $O_{\_1} =$ 4675 | $O_{\_2} =$ 14,302,993 | $O_{\_\_} =$ 14307668 |

| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} =$ 8 | $O_{12} =$ | $O_{1-} =$ 15828 |
| $w_1 \neq new$ | $O_{21} =$ | $O_{22} =$ | $O_{2-} =$ 14,291,840 |
| | $O_{-1} =$ 4675 | $O_{-2} =$ 14,302,993 | $O_{--} =$ 14307668 |

O12 = O1_ - O 11
= 15828 – 8
= 15,820

| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} =$ 8 | $O_{12} =$ 15,820 | $O_{1-} =$ 15828 |
| $w_1 \neq new$ | $O_{21} =$ | $O_{22} =$ | $O_{2-} =$ 14,291,840 |
| | $O_{-1} =$ 4675 | $O_{-2} =$ 14,302,993 | $O_{--} =$ 14307668 |

| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} =$ 8 | $O_{12} =$ 15,820 | $O_{1-} =$ 15828 |
| $w_1 \neq new$ | $O_{21} =$ | - $O_{22} =$ | $O_{2-} =$ 14,291,840 |
| | $O_{-1} =$ 4675 | $O_{-2} =$ 14,302,993 | $O_{--} =$ 14307668 |

4675 – 8 = 4667

| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} =$ 8 | $O_{12} =$ 15,820 | $O_{1-} =$ 15828 |
| $w_1 \neq new$ | $O_{21} =$ 4667 | $O_{22} =$ | $O_{2-} =$ 14,291,840 |
| | $O_{-1} =$ 4675 | $O_{-2} =$ 14,302,993 | $O_{--} =$ 14307668 |

| | $w_2 = companies$ | $w_2 \neq companies$ | |
|---|---|---|---|
| $w_1 = new$ | $O_{11} =$ 8 | $O_{12} =$ 15,820 | $O_{1-} =$ 15828 |
| $w_1 \neq new$ | $O_{21} =$ 4667 | $O_{22} =$ 14287173 | $O_{2-} =$ 14,291,840 |
| | $O_{-1} =$ 4675 | $O_{-2} =$ 14,302,993 | $O_{--} =$ 14307668 |

14,302,993 – 15820 = 14287173

finished