χ^2 Test

Für den χ^2-Test verwenden wir die folgende 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 χ^2-Teststatistik wird folgendermaßen berechnet:

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

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

Das Signifikanzniveau, das der Statistik entspricht, schlägt man in einer χ^2-Statistiktabelle nach. (Beispiel: $\chi^2 \approx 1.55 < 3.84 \quad \Rightarrow$ nicht signif.)

A **chi-squared test**, also written as **χ^2 test**, is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the null hypothesis, specifically Pearson's chi-squared test and variants thereof. Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.

Main Idea: hypothesis testing
We want to test of “new companies” is a collocation or not.
Given observed information, the hypothesis testing result tells us whether “new companies” happens more than expected or not.

Our assumption is that, if the word pair happens more than expected then it could be because it is a collocation

-but we can not be so sure that the information we have(what we observe) is enough to make this claim.
-**hypothesis testing** can tell us if there is enough evidence (significant deviation from the expected value) that our hypothesis is correct or not

Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.

Beispiel: Ist das Wortpaar new companies signifikant häufiger als erwartet?

Daten: In einem Korpus mit n=14,307,668 Wörtern, taucht new f_{new}=15,828 Mal auf, companies $f_{companies}$=4,675 Mal und new companies 8 Mal.

χ^2 Test

Für den χ^2-Test verwenden wir die folgende 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 χ^2-Teststatistik wird folgendermaßen berechnet:

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

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

Das Signifikanzniveau, das der Statistik entspricht, schlägt man in einer χ^2-Statistiktabelle nach. (Beispiel: $\chi^2 \approx 1.55 < 3.84 \quad \Rightarrow$ nicht signif.)

Explain the notations

O_position1, position2
* pos1 represents the existence w1. It can take 2 values. Value 1 means the word exists, 2 means does not exist.
* pos2 represents w2.

11: f(w1 = new, w2 = company)
22: f(w1 != new, w2 != company)
12: f(w1 = new, w2 != company)
21: f(w1 != new, w2 = company)

1_ : f(w1 = new, w2 = any)
_1: f(w1 = any, w2 = company)
_2: f(w1 = any, w2 != company)
2 _ : f(w1 != new, w2 = any)
_ _ = : f(w1 = any, w2 = any)

Example

-create a contingency table for testing “new companies”
-In a corpus of 100 words, freq(new) = 20, freq(companies) =40, freq(new companies) = 10

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

Put the values from the table in the formula

Example

-create a contingency table for testing “polar bear”
-In a corpus of 100 words, freq(polar) = 20, freq(bear) =40, freq(polar bear) = 10, fill the table

	w2	not w2	
w1	O11 = 10 f(polar, bear) **(1)**	O12 = f(polar, not bear) (8) = f(polar, any) - f(polar, bear) O12 = O1_ - O11	O1_ = 20 f(polar, any) **(3)**
not w1	O21 = f(not polar, bear) (7) = f(any, bear) - f(polar, bear) O21 = O_1 - O11	O22 = f(not polar, not bear) = f(any, not bear)-f(polar, not bear) (9) O22 = O_2 - O12	O2_ = f(not polar, any) (6) = f(any, any) - f(polar, any) O2_ = N - O1_
	O_1 =40 f(any, bear) **(2)**	O_2 = f(any, not bear) (5) =f(any, any) - f(any,bear) O_2 = N - O_1	O _ _ = 100 **(4)** f(any, any)

Trick on how to calculate the score
- You will be given 4 values
- Start by writing the values in the correct cells (cells with number 1, 2, 3, 4)
- see the table as two parts: inner part (blue cells) and out part (light green cells)
- next, you will fill cell 5, 6, 7,8, 9
- to compute cell 5, you calculate cell4 - cell2, [4] [**5**] [2]
- then to compute cell 6, you calculate cell 4 - cell 3, [4] [**6**] [3]
- now the outer cells are computed, next you want to compute the inner cells
- you start by calculating cell 7 by taking cell 2- cell 1
-then to compute cell 8, you take cell 3 - cell1 , [1] [**8**] [3]
-for the last cell 9, you take cell 5 - cell 8, [8] [**9**] [5]
You will see a pattern here (if you have 3 cell [a] [ ] [c], you compute the middle cell by taking c - a
- [a]
- [ ]
- [c]
for this you take c- a as well (the outer cell c has higher value, so we subtract a from c)
- start from filling most outer cells, then going to the inner cells

Übung:

-create a contingency table for testing “blue eyes”

-In a corpus of 1000 words, freq(blue) = 25, freq(eyes) =30, freq(blue eyes) = 15