Corpus Linguistics

The Chi Square Test for Statistical Significance

Niko Schenk

Institut für England- und Amerikastudien
Goethe-Universität Frankfurt am Main
Winter Term 2015/2016

January 27, 2016

Motivation

Three common pitfalls when comparing *n*-gram frequencies to **draw conclusions**:

Motivation

Three common pitfalls when comparing *n*-gram frequencies to **draw conclusions**:

1. No **normalization** of the frequencies.

Motivation

Three common pitfalls when comparing *n*-gram frequencies to **draw conclusions**:

1. No **normalization** of the frequencies. (comparing equal population sizes?)

Motivation

Three common pitfalls when comparing *n*-gram frequencies to **draw conclusions**:

1. No **normalization** of the frequencies. (comparing equal population sizes?)
2. Comparing *n*-grams of different length.

Motivation

Three common pitfalls when comparing
n-gram frequencies to **draw conclusions**:

1. No **normalization** of the frequencies. (comparing equal population sizes?)
2. Comparing *n*-grams of different length.
3. Are the frequencies **"really" different**?

Motivation

Three common pitfalls when comparing
n-gram frequencies to **draw conclusions**:

1. No **normalization** of the frequencies. (comparing equal population sizes?)
2. Comparing *n*-grams of different length.
3. Are the frequencies **“really” different**? (chance?)

Motivation

Three common pitfalls when comparing
n-gram frequencies to **draw conclusions**:

1. No **normalization** of the frequencies. (comparing equal population sizes?)
2. Comparing *n*-grams of different length.
3. Are the frequencies **“really” different**? (chance?)
   - $\rightarrow$ Needs to be tested for statistical significance!

1. Introduction
2. Testing for Statistical Significance
   - Binomial Distribution
   - Chi Square Test

Motivation

- **Coin example**

Motivation

- **Coin example**
- A coin is a binomial random variable.

Motivation

- **Coin example**
- A coin is a binomial random variable.
  Two outcomes, **H/T** (heads, tails), usually with p=0.5, i.e. a **fair** coin.

12 Coin Tosses—Probability Distribution for n Times **H**

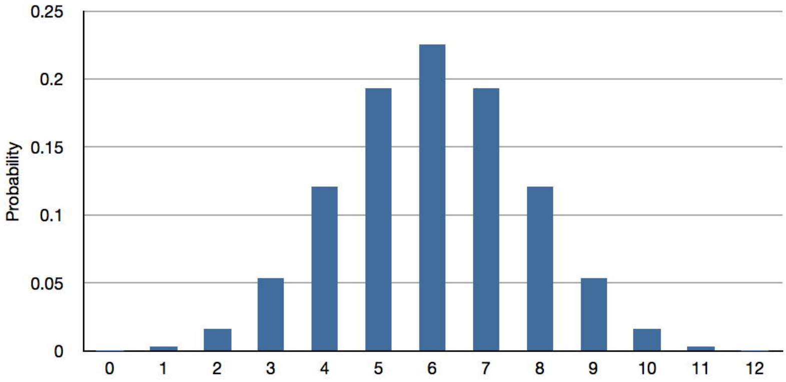
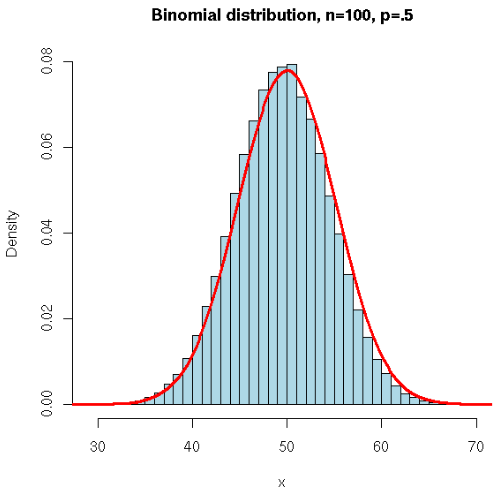


Figure: **12** coin tosses. X-axis: frequency of n times **H**. Y-axis: probability of n times **H**.

Motivation cont'd—100 coin tosses

Motivation cont'd—100 coin tosses



Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T?

Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

- Possible results:
    - 50 times H vs. 50 times T? (expected)

Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T? (expected)
  - 51 times H vs. 49 times T?

Assume you have a **fair** coin.
When would you consider the coin “unfair”?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T? (expected)
  - 51 times H vs. 49 times T? (still possible)

Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T? (expected)
  - 51 times H vs. 49 times T? (still possible)
  - 52 times H vs. 48 times T?

Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T? (expected)
  - 51 times H vs. 49 times T? (still possible)
  - 52 times H vs. 48 times T? (still possible)

Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T? (expected)
  - 51 times H vs. 49 times T? (still possible)
  - 52 times H vs. 48 times T? (still possible)
  - 53 times H vs. 47 times T?

Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T? (expected)
  - 51 times H vs. 49 times T? (still possible)
  - 52 times H vs. 48 times T? (still possible)
  - 53 times H vs. 47 times T? (still possible?)

Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T? (expected)
  - 51 times H vs. 49 times T? (still possible)
  - 52 times H vs. 48 times T? (still possible)
  - 53 times H vs. 47 times T? (still possible?)
  - 54 times H vs. 46 times T?

Assume you have a **fair** coin.
When would you consider the coin “unfair”?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T? (expected)
  - 51 times H vs. 49 times T? (still possible)
  - 52 times H vs. 48 times T? (still possible)
  - 53 times H vs. 47 times T? (still possible?)
  - 54 times H vs. 46 times T? (still possible?)

Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T? (expected)
  - 51 times H vs. 49 times T? (still possible)
  - 52 times H vs. 48 times T? (still possible)
  - 53 times H vs. 47 times T? (still possible?)
  - 54 times H vs. 46 times T? (still possible?)
  - 55 times H vs. 45 T?...

Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T? (expected)
  - 51 times H vs. 49 times T? (still possible)
  - 52 times H vs. 48 times T? (still possible)
  - 53 times H vs. 47 times T? (still possible?)
  - 54 times H vs. 46 times T? (still possible?)
  - 55 times H vs. 45 T?...
  - 90 times H vs. 10 times T?

Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T? (expected)
  - 51 times H vs. 49 times T? (still possible)
  - 52 times H vs. 48 times T? (still possible)
  - 53 times H vs. 47 times T? (still possible?)
  - 54 times H vs. 46 times T? (still possible?)
  - 55 times H vs. 45 T?...
  - 90 times H vs. 10 times T? (still possible?)

Assume you have a **fair** coin.
When would you consider the coin "unfair"?
(i.e. when would you be **surpised**)?

- Possible results:
  - 50 times H vs. 50 times T? (expected)
  - 51 times H vs. 49 times T? (still possible)
  - 52 times H vs. 48 times T? (still possible)
  - 53 times H vs. 47 times T? (still possible?)
  - 54 times H vs. 46 times T? (still possible?)
  - 55 times H vs. 45 T?...
  - 90 times H vs. 10 times T? (still possible?)
- Intuition/Formal Explanation:
  $\Rightarrow$ Usually, less than **5%** of the area under the curve in the tail of the distribution is an indicator of "surprise".

Motivation cont'd—100 coin tosses

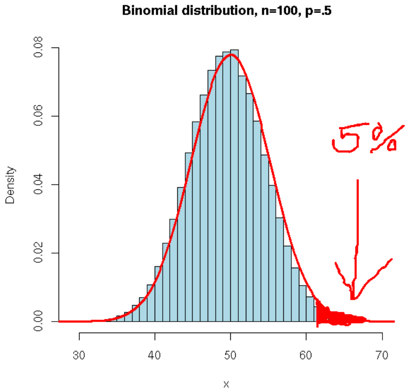


Figure: Number of **H** ≥ 63 = a reason to be surprised!

Surprise, surprise...

Summarizing the previous slides,

- Flipping a coin 100 times, and obtaining
  - 60 times **H** and 40 times **T**

is still **no reason to be surprised**, i.e. there is **no statistically significant difference** between the two frequencies given a fair coin. A result like this is perfectly fine and statistically probable.

Surprise, surprise...

Summarizing the previous slides,

- Flipping a coin 100 times, and obtaining
  - 60 times **H** and 40 times **T**

  is still **no reason to be surprised**, i.e. there is **no statistically significant difference** between the two frequencies given a fair coin. A result like this is perfectly fine and statistically probable.

- In a corpus linguistic scenario, researchers usually report and compare two (or more) frequencies. We need to find out whether the numbers **really differ** or whether they happen to be different **just by chance**.

1. Introduction

2. Testing for Statistical Significance
   - Binomial Distribution
   - Chi Square Test

χ^2—Motivation

- χ = chi
- Formal test for "surprise" given a random variable with n outcomes. (e.g., when tossing a coin, when rolling a die. . .).

χ^2—Computation

Chi square is computed as follows:

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i},$$

where
each i is a unique outcome of the random variable (e.g., **H** or **T**)
O_i = observed value at index i
E_i = expected value at index i

Coin Example

1. **Null hypothesis**:
   - H & T appear equally often.

Coin Example

1. **Null hypothesis**:
   - H & T appear equally often.
   - There is no significant difference between observed and expected values.

Example

2. Look at the data:

Example: 50 coin tosses.

	H	T
O	28	22

Example

2. Look at the data:

Example: 50 coin tosses.

	H	T
O	28	22
E	25	25

Example

2. Look at the data:

Example: 50 coin tosses.

	H	T
O	28	22
E	25	25

3. Compute χ^2

Formula:

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i},$$

Example

2. Look at the data:

Example: 50 coin tosses.

	H	T
O	28	22
E	25	25

3. Compute χ^2

Formula:

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i},$$

$(28\text{-}25)^2 + (22\text{-}25)^2 =$

Example

2. Look at the data:

Example: 50 coin tosses.

	H	T
O	28	22
E	25	25

3. Compute χ^2

Formula:

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i},$$

$(28\text{-}25)^2 + (22\text{-}25)^2 = 0.72$

Example

2. Look at the data:

Example: 50 coin tosses.

	H	T
O	28	22
E	25	25

3. Compute χ^2

Formula:

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i},$$

(28-25)^2 + (22-25)^2 = 0.72
$\chi^2 = 0.72$

Next steps...

4. Compute the **degrees of freedom** (**df**).

Next steps...

4. Compute the **degrees of freedom (df)**.
   - Usually, df = number of distinct outcomes -1

Next steps...

4. Compute the **degrees of freedom (df)**.
    - Usually, df = number of distinct outcomes -1
5. Use **df** and **lookup** χ^2 **value** for a predefined level of significance (usually 5%).

Next steps...

4. Compute the **degrees of freedom** (**df**).
   - Usually, df = number of distinct outcomes -1
5. Use **df** and **lookup** χ^2 **value** for a predefined level of significance (usually 5%).
6. If your χ^2 value is smaller than the precomputed one, **you cannot reject the null hypothesis**, i.e. you have a non-significant result.
   - This means: The differences in frequencies **not** statistically significantly different and are **only due to chance**.

Next steps...

4. Compute the **degrees of freedom** (**df**).
   - Usually, *df* = number of distinct outcomes -1
5. Use **df** and **lookup** χ^2 **value** for a predefined level of significance (usually 5%).
6. If your χ^2 value is smaller than the precomputed one, **you cannot reject the null hypothesis**, i.e. you have a non-significant result.
   - This means: The differences in frequencies **not** statistically significantly different and are **only due to chance**.

A chi square table:
`http://whichbobareyou.com/uploads/2/9/4/6/2946053/9419235_orig.png?288`

A Chi Square Table

Upper critical values of chi-square distribution with ν degrees of freedom

ν	Probability of exceeding the critical value 0.10	0.05	0.025	0.01	0.001
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264
12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123
15	22.307	24.996	27.488	30.578	37.697
16	23.542	26.296	28.845	32.000	39.252
17	24.769	27.587	30.191	33.409	40.790
18	25.989	28.869	31.526	34.805	42.312
19	27.204	30.144	32.852	36.191	43.820
20	28.412	31.410	34.170	37.566	45.315
21	29.615	32.671	35.479	38.932	46.797
22	30.813	33.924	36.781	40.289	48.268

Figure: Chi square distribution table. *df* / critical values and *p*-values.

A Chi Square Table

Upper critical values of chi-square distribution with ν degrees of freedom

	Probability of exceeding the critical value				
ν	0.10	0.05	0.025	0.01	0.001
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264
12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123
15	22.307	24.996	27.488	30.578	37.697
16	23.542	26.296	28.845	32.000	39.252
17	24.769	27.587	30.191	33.409	40.790
18	25.989	28.869	31.526	34.805	42.312
19	27.204	30.144	32.852	36.191	43.820
20	28.412	31.410	34.170	37.566	45.315
21	29.615	32.671	35.479	38.932	46.797
22	30.813	33.924	36.781	40.289	48.268

Figure: Only inspect the critical values for $p = 5\%$.

A Chi Square Table

Upper critical values of chi-square distribution with ν degrees of freedom

	Probability of exceeding the critical value				
ν	0.10	0.05	0.025	0.01	0.001
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264
12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123
15	22.307	24.996	27.488	30.578	37.697
16	23.542	26.296	28.845	32.000	39.252
17	24.769	27.587	30.191	33.409	40.790
18	25.989	28.869	31.526	34.805	42.312
19	27.204	30.144	32.852	36.191	43.820
20	28.412	31.410	34.170	37.566	45.315
21	29.615	32.671	35.479	38.932	46.797
22	30.813	33.924	36.781	40.289	48.268

Figure: Degrees of freedom (*df*) should be 1. (we have two outcomes H/T)

A Chi Square Table

Upper critical values of chi-square distribution with ν degrees of freedom

	Probability of exceeding the critical value				
ν	0.10	0.05	0.025	0.01	0.001
1	2.706	$\boxed{3.841}$	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264
12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123
15	22.307	24.996	27.488	30.578	37.697
16	23.542	26.296	28.845	32.000	39.252
17	24.769	27.587	30.191	33.409	40.790
18	25.989	28.869	31.526	34.805	42.312
19	27.204	30.144	32.852	36.191	43.820
20	28.412	31.410	34.170	37.566	45.315
21	29.615	32.671	35.479	38.932	46.797
22	30.813	33.924	36.781	40.289	48.268

Figure: Our χ^2 (0.74) is **smaller** than the precomputed one! (3.841)
No statistically significant difference between 28 times H and 22 times T!

Example II—Rolling a Die

Figure: http://www.youtube.com/watch?v=WXPBoFDqNVkPossible outcome for n=36, p=$\frac{1}{6}$.

Example II—Rolling a Die

Figure: http://www.youtube.com/watch?v=WXPBoFDqNVkPossible outcome for $n=36$, $p=\frac{1}{6}$.

$\chi^2 = 9.6$

Example II—Rolling a Die

Figure: `http://www.youtube.com/watch?v=WXPBoFDqNVk`Possible outcome for $n=36$, $p=\frac{1}{6}$.

$\chi^2 = 9.6 \rightarrow$ again, no significant difference between observed and expected values!

Individual Steps Summarized

1. You start with the null hypothesis, e.g., there is **no** correlation / significant difference between my two variables (O/E).

Individual Steps Summarized

1. You start with the null hypothesis, e.g., there is **no** correlation / significant difference between my two variables (O/E).
2. Create contingency table with observed values.

Individual Steps Summarized

1. You start with the null hypothesis, e.g., there is **no** correlation / significant difference between my two variables (O/E).
2. Create contingency table with observed values.
3. Create contingency table with expected values.

Individual Steps Summarized

1. You start with the null hypothesis, e.g., there is **no** correlation / significant difference between my two variables (O/E).
2. Create contingency table with observed values.
3. Create contingency table with expected values.
4. Compute χ^2

Individual Steps Summarized

1. You start with the null hypothesis, e.g., there is **no** correlation / significant difference between my two variables (O/E).
2. Create contingency table with observed values.
3. Create contingency table with expected values.
4. Compute χ^2
5. (Compute degrees of freedom and) lookup value for a predefined level of significance.

Individual Steps Summarized

1. You start with the null hypothesis, e.g., there is **no** correlation / significant difference between my two variables (O/E).
2. Create contingency table with observed values.
3. Create contingency table with expected values.
4. Compute χ^2
5. (Compute degrees of freedom and) lookup value for a predefined level of significance.
6. Accept or reject your (null) hypothesis.

Individual Steps Summarized

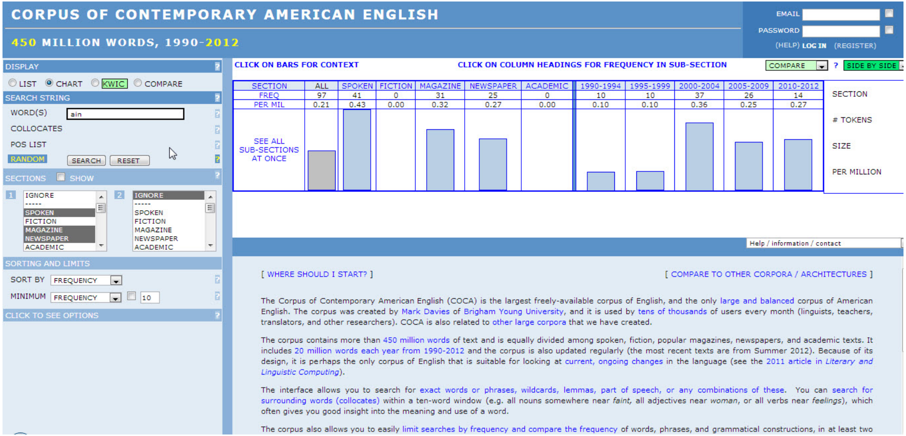
1. You start with the null hypothesis, e.g., there is **no** correlation / significant difference between my two variables (O/E).
2. Create contingency table with observed values.
3. Create contingency table with expected values.
4. Compute χ^2
5. (Compute degrees of freedom and) lookup value for a predefined level of significance.
6. Accept or reject your (null) hypothesis.

Online calculator:

- `http://www.quantpsy.org/chisq/chisq.htm`

Example III—COCA

Consider this chart for the frequency for “ain’t” (ain) in the COCA corpus.



A linguist reports in his scientific work:
"Focusing only on a subset of all sections in the COCA corpus (spoken, magazine, newspaper), we found that frequencies for **ain** *(negation) differ with respect to the specific genres: the usage of* **ain't** *is much more frequent in spoken language compared to standard newspaper texts, for instance."*

Homework:
Verify this claim formally by means of the χ^2 test.

Example IV—BNC

Assume you have the following frequency distribution for the word "funny" in the BNC:

- Spoken: 100
- News: 520
- Academic: 120

Is there a statistically significant difference between the frequencies?