Collocations

VL Embeddings

Uni Heidelberg

SS 2019

Word cooccurrences

Cooccurrences, associations, collocations

J.R. Firth (1890-1960): *Contextual Theory of Meaning*

You shall know a word by the company it keeps. (Firth 1957)

- **Collocations**
  - lexical units that often occur together in a corpus
  - conventionalised ("ein starker Wind", "eine steife Brise")

strong tea	powerful tea
powerful drug	mächtiger Tee

  - Collocations can tell us something about
    - the language
    - the world

Collocations

- **Non-compositionality**
  - meaning not predictable based on meaning of individual components, e.g.: *den Löffel abgeben,*

Collocations

- **Non-compositionality**
  - meaning not predictable based on meaning of individual components, e.g.: *den Löffel abgeben,*
- **Non-substitutability**
  - can not be replaced by words with similar semantics, e.g.: *Zähne putzen* versus *Zähne bürsten*; *weißer Wein* versus *gelber/heller Wein*

Collocations

- **Non-compositionality**
  - meaning not predictable based on meaning of individual components, e.g.: *den Löffel abgeben,*
- **Non-substitutability**
  - can not be replaced by words with similar semantics, e.g.: *Zähne putzen* versus *Zähne bürsten*; *weißer Wein* versus *gelber/heller Wein*
- **Non-modifiability**
  - e.g.: *arm wie eine Kirchenmaus* versus *arm wie Kirchenmäuse*; *ins Gras beißen* versus *ins grüne Gras beißen*

(also see Manning & Schütze:172/173)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- ohne Blätter vor den Mund zu nehmen (Pluralisierung)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- ohne Blätter vor den Mund zu nehmen (Pluralisierung)
- Hier nahm er manches Blatt vor den Mund (Quantifizierung)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- ohne Blätter vor den Mund zu nehmen (Pluralisierung)
- Hier nahm er manches Blatt vor den Mund (Quantifizierung)
- mit einem postmodernen Blatt vor dem Munde / kein Blatt vor seinen republikfeindlichen Mund (adjektivische Modifikation)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- ohne Blätter vor den Mund zu nehmen (Pluralisierung)
- Hier nahm er manches Blatt vor den Mund (Quantifizierung)
- mit einem postmodernen Blatt vor dem Munde / kein Blatt vor seinen republikfeindlichen Mund (adjektivische Modifikation)
- ohne das geringste (Klee-)Blatt vor den vorlauten Mund zu nehmen (nominale Modifikation)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- ohne Blätter vor den Mund zu nehmen (Pluralisierung)
- Hier nahm er manches Blatt vor den Mund (Quantifizierung)
- mit einem postmodernen Blatt vor dem Munde / kein Blatt vor seinen republikfeindlichen Mund (adjektivische Modifikation)
- ohne das geringste (Klee-)Blatt vor den vorlauten Mund zu nehmen (nominale Modifikation)
- Hier wird kein Blatt vor den Mund genommen (Passivierung)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- ohne Blätter vor den Mund zu nehmen (Pluralisierung)
- Hier nahm er manches Blatt vor den Mund (Quantifizierung)
- mit einem postmodernen Blatt vor dem Munde / kein Blatt vor seinen republikfeindlichen Mund (adjektivische Modifikation)
- ohne das geringste (Klee-)Blatt vor den vorlauten Mund zu nehmen (nominale Modifikation)
- Hier wird kein Blatt vor den Mund genommen (Passivierung)
- mit so einem kecken Mund, vor den kein Blatt genommen wird / auf dem Blatt, das sie nicht vor den Mund nimmt (Relativierung)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- Die Leipziger nahmen damals im Guten den Mund noch nicht so voll, und im Schlechten hielten sie sich noch kein Blatt davor (Pronominalisierung)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- Die Leipziger nahmen damals im Guten den Mund noch nicht so voll, und im Schlechten hielten sie sich noch kein Blatt davor
  (Pronominalisierung)
- Und ein Blatt nimmt keiner vor den Mund an diesem Wahlabend
  Kein Blatt vor den Mund nahm dafür einer von Müllers Vorgängern
  (Topikalisierung)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- Die Leipziger nahmen damals im Guten den Mund noch nicht so voll, und im Schlechten hielten sie sich noch kein Blatt davor (Pronominalisierung)
- Und ein Blatt nimmt keiner vor den Mund an diesem Wahlabend
  Kein Blatt vor den Mund nahm dafür einer von Müllers Vorgängern (Topikalisierung)
- Ich nehme mir eben kein Blatt vor den Mund (Reflexivierung)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- Die Leipziger nahmen damals im Guten den Mund noch nicht so voll, und im Schlechten hielten sie sich noch kein Blatt davor (Pronominalisierung)
- Und ein Blatt nimmt keiner vor den Mund an diesem Wahlabend
  Kein Blatt vor den Mund nahm dafür einer von Müllers Vorgängern (Topikalisierung)
- Ich nehme mir eben kein Blatt vor den Mund (Reflexivierung)
- das Blatt vom Mund nimmt (Variation der Präposition und der assoziierten Bedeutung)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- Die Leipziger nahmen damals im Guten den Mund noch nicht so voll, und im Schlechten hielten sie sich noch kein Blatt davor (Pronominalisierung)
- Und ein Blatt nimmt keiner vor den Mund an diesem Wahlabend
  Kein Blatt vor den Mund nahm dafür einer von Müllers Vorgängern (Topikalisierung)
- Ich nehme mir eben kein Blatt vor den Mund (Reflexivierung)
- das Blatt vom Mund nimmt (Variation der Präposition und der assoziierten Bedeutung)
- Ich habe ihr kein Blatt vor den Mund gehalten (Valenz und Verbvariation)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- Die Leipziger nahmen damals im Guten den Mund noch nicht so voll, und im Schlechten hielten sie sich noch kein Blatt davor (Pronominalisierung)
- Und ein Blatt nimmt keiner vor den Mund an diesem Wahlabend Kein Blatt vor den Mund nahm dafür einer von Müllers Vorgängern (Topikalisierung)
- Ich nehme mir eben kein Blatt vor den Mund (Reflexivierung)
- das Blatt vom Mund nimmt (Variation der Präposition und der assoziierten Bedeutung)
- Ich habe ihr kein Blatt vor den Mund gehalten (Valenz und Verbvariation)
- Gilbert Ziebura tut es ohne Blatt vor den Mund (Verbauslassung)

Non-modifiability? Evidence from large corpora

kein Blatt vor den Mund nehmen

- Die Leipziger nahmen damals im Guten den Mund noch nicht so voll, und im Schlechten hielten sie sich noch kein Blatt davor (Pronominalisierung)
- Und ein Blatt nimmt keiner vor den Mund an diesem Wahlabend
  Kein Blatt vor den Mund nahm dafür einer von Müllers Vorgängern (Topikalisierung)
- Ich nehme mir eben kein Blatt vor den Mund (Reflexivierung)
- das Blatt vom Mund nimmt (Variation der Präposition und der assoziierten Bedeutung)
- Ich habe ihr kein Blatt vor den Mund gehalten (Valenz und Verbvariation)
- Gilbert Ziebura tut es ohne Blatt vor den Mund (Verbauslassung)
- Mahneke nimmt keine Rücksichten mehr und kein Blatt vor den Mund (Zeugma)

Collocations

- **Non-compositionality**
  - meaning not predictable based on meaning of individual components, e.g.: *den Löffel abgeben,*
- **Non-substitutability**
  - can not be replaced by words with similar semantics, e.g.: *Zähne putzen* versus *Zähne bürsten*; *weißer Wein* versus *gelber/heller Wein*
- **Non-modifiability**
  - e.g.: *arm wie eine Kirchenmaus* versus *arm wie Kirchenmäuse*; *ins Gras beißen* versus *ins grüne Gras beißen*

- **Better:** limited compositionality/substitutability/modifiability

(also see Manning & Schütze:172/173)

Word cooccurrences

Cooccurrences, associations, collocations

- Words that often cooccur together trigger certain associations
  - **syntagmatic** relations (drink, coffee)
  - **paradigmatic** relations (coffee, tea)
- Evidence from psycholinguistics (priming experiments)

Word cooccurrences

Cooccurrences vs collocations

Different views on collocations

- **Distributional semantics**
  - directly observable quantity, descriptive
  - adjacent words in text, no additional (linguistic) information

Word cooccurrences

Cooccurrences vs collocations

Different views on collocations

- **Distributional semantics**
  - directly observable quantity, descriptive
  - adjacent words in text, no additional (linguistic) information
- **Linguistic approach**
  - Collocations are somewhere between fixed idioms *(kick the bucket)* and loose word sequences (read a book)
  - semi-compositional word pairs with free element (basis) and lexically determined element (collocator)
    - z.B.: starker Raucher, eine Herde Zebras, Widerstand leisten
    - free element keeps original meaning
    - collocator adds meaning component: *ein Rudel Wölfe* vs *eine Herde Zebras* vs *ein Schwarm Bienen*

(also see Evert, 2005:15 ff.)

Word cooccurrences

Cooccurrences vs collocations

- Distributional approach looks at cooccurrences
  - based on frequency information
  - indicator for statistical associations

  "collocation is the occurrence of two or more words within a short space of each other in a text"
  (Neo-Firthian school, Sinclair 1991:170)

- Linguistic approach looks at collocations
  - linguistic definition
  - not dependent on frequency information

  "a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components."
  (Choueka 1988)

Collocations in NLP

- Broad definition of *collocations:*
  - word compounds *(black box)*
  - idioms *(kick the bucket, ins Gras beißen)*
  - ⇒ limited compositionality
    (handsome man vs. *beautiful woman)*

- Terms sometimes used as synonyms:
  - Multi-word expressions (MWE)
  - Multi-word units (MWU)
  - Bigramms/Ngrams
  - Idioms

Collocations in NLP

- Broad definition of *collocations:*
  - word compounds *(black box)*
  - idioms *(kick the bucket, ins Gras beißen)*

  ⇒ limited compositionality
  (handsome man vs. *beautiful woman)*

- Terms sometimes used as synonyms:
  - Multi-word expressions (MWE)
  - Multi-word units (MWU)
  - Bigramms/Ngrams
  - Idioms

In sum:

No clear, agreed upon definition for collocations!

Again: What are collocations?

- Cooccurrences of two or more words
- Conventionalised (Zähne putzen vs. brush teeth)

Again: What are collocations?

- Cooccurrences of two or more words
- Conventionalised (Zähne putzen vs. brush teeth)

Test: Can be translated literally?

- Collocations need own entry in dictionary
  - brush hair – Haare bürsten
  - brush teeth – Zähne putzen

Again: What are collocations?

- Cooccurrences of two or more words
- Conventionalised (Zähne putzen vs. brush teeth)

Test: Can be translated literally?

- Collocations need own entry in dictionary
  - brush hair – Haare bürsten
  - brush teeth – Zähne putzen

- Identification of collocations important for
  - language learning, (machine) translation, lexicography, ...

Finding collocations in corpora

- **Task:** list all collocations that occur in the corpus
- **Methods:**
  - **Frequency**
  - **Mean and variance** of the distance between basis and collocator
  - Testing of hypotheses: **t-test, Chi-square (χ^2)**
  - **Mutual Information** (MI)

Frequency

- Hypothesis:
  - Two words that cooccur with high frequency ⇒ collocation
- Approach:
  - extract bigrams with highest frequency in the corpus

Frequency	**Bigram**
4441	, die
2854	, daß
1437	in den
...	
508	Millionen Mark
...	
53	Oskar Lafontaine

Frequency

- Hypothesis:
  - Two words that cooccur with high frequency $\Rightarrow$ collocation
- Approach:
  - extract bigrams with highest frequency in the corpus

Frequency	Bigram
4441	, die
2854	, daß
1437	in den
...	
508	Millionen Mark
...	
53	Oskar Lafontaine

- Problem: many bigrams with function words (high frequency!)

Frequency

- **Use POS filter**
  (Justeson & Katz 1995, Ross & Tukey 1975, Kupiec et al. 1995)

  - Look for patterns that are likely to be phrases

Tag Pattern	Example
A N	linear function
N N	regression coefficients
A A N	Gaussian random variable
A N N	cumulative distribution function
N A N	mean squared error
N N N	class probability function
N P N	degrees of freedom

Frequency

- **Use POS filter**
  (Justeson & Katz 1995, Ross & Tukey 1975, Kupiec et al. 1995)

  - Combine with frequency information

$C(w_1, w_2)$	w_1	w_2	**Tag Pattern**
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
...	...	...	...
1073	real	estate	A N

Frequency

- **Use POS filter**
  (Justeson & Katz 1995, Ross & Tukey 1975, Kupiec et al. 1995)

  - Combine with frequency information

$C(w_1, w_2)$	w_1	w_2	**Tag Pattern**
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
...	...	...	...
1073	real	estate	A N

Simple quantitative method + linguistic information $\rightarrow$ good results

Frequency

- What collocations we find depends on
  - the corpus (what type of corpus is the best for this task?)
  - the corpus size (there's no data like more data!)

Frequency

- What collocations we find depends on
    - the corpus (what type of corpus is the best for this task?)
    - the corpus size (there's no data like more data!)

- Sum-up: frequency method works well for fixed word combinations (with fixed position)

  But: how can we find free(r) constructions?

  Sie putzt die Zähne.
  versus
  Ihre Zähne hat Sie noch nie oft geputzt.
  versus
  Putzt sie ihre Zähne auch regelmäßig?

Mean and variance

Descriptive statistics

Given a population of students with the following marks (from 1–6)

- **Sample 1**

student	1	2	3	4	5	6	7	8	9	10	total
mark	3	3	3	3	3	3	3	3	3	3	**30**

- **Sample 2**

student	1	2	3	4	5	6	7	8	9	10	total
mark	4	2	1	1	5	6	2	3	5	1	**30**

- What is the sample mean for each of these samples?

Mean and variance

Descriptive statistics

Given a population of students with the following marks (from 1–6)

- **Sample 1**

student	1	2	3	4	5	6	7	8	9	10	total
mark	3	3	3	3	3	3	3	3	3	3	**30**

- **Sample 2**

student	1	2	3	4	5	6	7	8	9	10	total
mark	4	2	1	1	5	6	2	3	5	1	**30**

- What is the sample mean for each of these samples?
- What is the difference between the two samples?

Mean and variance

Descriptive statistics

Given a population of students with the following marks (from 1–6)

- **Sample 1**

student	1	2	3	4	5	6	7	8	9	10	total
mark	3	3	3	3	3	3	3	3	3	3	**30**

- **Sample 2**

student	1	2	3	4	5	6	7	8	9	10	total
mark	4	2	1	1	5	6	2	3	5	1	**30**

- What is the sample mean for each of these samples?
- What is the difference between the two samples?

- Variance σ^2 : average of the squared differences from the mean

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N} \qquad (1)$$

- x individual data points
- μ mean of the population
- N number of data points

Mean and variance

Descriptive statistics

- Compute average distance between two words in the corpus
- Compute variance between those distances
  - no collocation $\rightarrow$ random distribution: *high variance* of distance between w_1 and w_2
  - collocation $\rightarrow$ non-random distribution: fixed word order or frequent occurrence with exactly n words in between w_1 and w_2: *low variance*

Sie putzt die Zähne.	2
Hat sie ihre Zähne geputzt?	-1
Putzt sie ihre Zähne auch regelmäßig?	3

Mean: $\mu =$

Mean and variance

Descriptive statistics

- Compute average distance between two words in the corpus
- Compute variance between those distances
  - no collocation $\rightarrow$ random distribution:
    high variance of distance between w_1 and w_2
  - collocation $\rightarrow$ non-random distribution:
    fixed word order or frequent occurrence with exactly n words in between w_1 and w_2: *low variance*

Sie putzt die Zähne.	2
Hat sie ihre Zähne geputzt?	-1
Putzt sie ihre Zähne auch regelmäßig?	3

Mean: $\mu = \frac{1}{3}(2 + -1 + 3) = 1.3$

Mean and variance

Descriptive statistics

- Variance: deviation between distances from mean between w_1 and w_2

Sie putzt die Zähne.	2
Hat sie ihre Zähne geputzt?	-1
Putzt sie ihre Zähne auch regelmäßig?	3

Mean and variance

Descriptive statistics

- Variance: deviation between distances from mean between w_1 and w_2

Sie putzt die Zähne.	2
Hat sie ihre Zähne geputzt?	-1
Putzt sie ihre Zähne auch regelmäßig?	3

Mean: $\mu = 1.3$

Variance: $\sigma^2 =$

Mean and variance

Descriptive statistics

- Variance: deviation between distances from mean between w_1 and w_2

Sie putzt die Zähne.	2
Hat sie ihre Zähne geputzt?	-1
Putzt sie ihre Zähne auch regelmäßig?	3

Mean: $\mu = 1.3$

Variance: $\sigma^2 = \frac{\sum_{i=1}^{n}(d_i-\mu)^2}{n-1}$

Mean and variance

Descriptive statistics

- Variance: deviation between distances from mean between w_1 and w_2

Sie putzt die Zähne.	2
Hat sie ihre Zähne geputzt?	-1
Putzt sie ihre Zähne auch regelmäßig?	3

Mean: $\mu = 1.3$

Variance: $\sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}$

$$= \frac{(2-1.3)^2 + (-1-1.3)^2 + (3-1.3)^2}{3-1}$$

Mean and variance

Descriptive statistics

- Variance: deviation between distances from mean between w_1 and w_2

Sie putzt die Zähne.	2
Hat sie ihre Zähne geputzt?	-1
Putzt sie ihre Zähne auch regelmäßig?	3

Mean: $\mu = 1.3$

Variance:
$$\sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}$$
$$= \frac{(2-1.3)^2 + (-1-1.3)^2 + (3-1.3)^2}{3-1}$$
$$= \frac{0.49+5.29+2.89}{2} = \frac{8.67}{2} = 4.335$$

Mean and variance

Descriptive statistics

- Variance: deviation between distances from mean between w_1 and w_2

Sie putzt die Zähne.	2
Hat sie ihre Zähne geputzt?	-1
Putzt sie ihre Zähne auch regelmäßig?	3

Mean: $\mu = 1.3$

Variance:
$$\sigma^2 = \frac{\sum_{i=1}^{n}(d_i-\mu)^2}{n-1} = \frac{(2-1.3)^2+(-1-1.3)^2+(3-1.3)^2}{3-1} = \frac{0.49+5.29+2.89}{2} = \frac{8.67}{2} = 4.335$$

Standard deviation: $\sigma = \sqrt{\sigma^2}$

Mean and variance

Descriptive statistics

- Variance: deviation between distances from mean between w_1 and w_2

Sie putzt die Zähne.	2
Hat sie ihre Zähne geputzt?	-1
Putzt sie ihre Zähne auch regelmäßig?	3

$$\text{Mean:} \quad \mu = 1.3$$

$$\text{Variance:} \quad \sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}$$

$$= \frac{(2-1.3)^2 + (-1-1.3)^2 + (3-1.3)^2}{3-1}$$

$$= \frac{0.49+5.29+2.89}{2} = \frac{8.67}{2} = 4.335$$

$$\text{Standard deviation:} \quad \sigma = \sqrt{\sigma^2}$$

$$= \sqrt{4.335} = 2.08$$

Mean and variance

Collocations

- Mean and variance can describe the distribution of distances between two words in the corpus
- Collocations: word pairs with low standard deviation
  $\rightarrow$ w_1 and w_2 frequently cooccur within same distance
- If standard deviation $= 0$
  $\rightarrow$ distance between w_1 and w_2 always the same

Mean and variance

Example

σ	μ	Frequency	w_1	w_2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

Mean and variance

Example

σ	μ	Frequency	w_1	w_2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

Sum-up: Mean and variance

- good for identifying collocations with non-fixed order
- but: high frequency and low variance can also be random!

Testing hypotheses

- Variance method biased towards high-frequency words (e.g. new companies)
  - if two words are extremely frequent in the corpus $\rightarrow$ might cooccur by chance
- How can we distinguish those from "real" collocations?

Testing hypotheses

- Variance method biased towards high-frequency words (e.g. new companies)
  - if two words are extremely frequent in the corpus $\rightarrow$ might cooccur by chance
- How can we distinguish those from "real" collocations?
- **Question:**
  Can the cooccurrence of w_1 and w_2 be due to chance?
- **Methodological approach:**
  1. Formulate **null hypothesis** (H0):
     cooccurrance of w_1 and w_2 is due to chance (no collocation)
  2. If we can reject the H0, we can take that as evidence for the **alternative hypothesis** (w_1 and w_2 are collocations)

Testing hypotheses

Collocations

How can we know if two words w_1 and w_2 cooccur more often than chance?

- Null hypothesis:
  - w_1 and w_2 are independent from each other (no collocation!)
  - chance of w_1 and w_2 occurring together is

    $$P(w_1 w_2) = P(w_1)P(w_2)$$

- We can compute the probability P of w_1 and w_2 occurring together
- But how do we know whether P is larger than chance?

Testing hypotheses: t-Test

Given a sample of measurements

- H0: our sample is drawn from a distribution with mean μ
- t-Test looks at difference between expected and observed mean, scaled by the variance of the data

  $\rightarrow$ How likely is it that a sample with the observed values for mean and variance was drawn from a distribution with mean μ

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}} \tag{2}$$

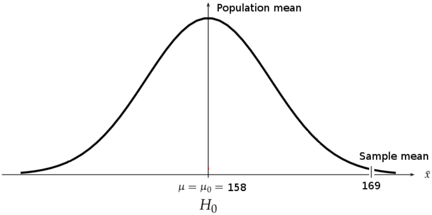
  - $\bar{x}$: sample mean
  - σ^2: sample variance (avg. of squared differences from mean)
  - N: sample size
  - μ: mean of the expected distribution

- If the t statistics is large enough, we reject the null hypothesis

Testing hypotheses: t-Test

Example

- **Null hypothesis**:
  the average size in a population of women is 158 cm
- We have a sample of 200 women with average size of $\bar{x}$ =169 cm and σ^2=2600



How likely is it that the sample comes from our population with mean $\mu = 158$ cm? → null hypothesis true

How likely is it that the sample comes from a different population? → null hypothesis false

Testing hypotheses: t-Test

Example

- **Null hypothesis**:
  average size in a population of women is 158 cm
- We have a sample of 200 women with average size of $\bar{x}$ =169 cm and σ^2=2600

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}} =$$

Testing hypotheses: t-Test

Example

- **Null hypothesis**:
  average size in a population of women is 158 cm
- We have a sample of 200 women with average size of $\bar{x}$ =169 cm and σ^2=2600

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}} = \frac{169 - 158}{\sqrt{\frac{2600}{200}}} = 3.05 \qquad (3)$$

- Look up t statistics:
  Confidence level $\alpha = 0.005$ determined by us
  Degrees of freedom = 199 (Freiheitsgrade) (Sample size -1)
  Sample size of 200 → **t=2.576**

Testing hypotheses: t-Test

Example

- **Null hypothesis**:
  average size in a population of women is 158 cm
- We have a sample of 200 women with average size of $\bar{x}$ =169 cm and σ^2=2600

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}} = \frac{169 - 158}{\sqrt{\frac{2600}{200}}} = 3.05 \quad (3)$$

- Look up t statistics:
  Confidence level $\alpha = 0.005$ determined by us
  Degrees of freedom = 199 (Freiheitsgrade) (Sample size -1)
  Sample size of 200 $\rightarrow$ **t=2.576**

Observed value $t = 3.05$ is larger than value in table ($t = 2.576$)
$\Rightarrow$ we can reject H0 with a likelihood of 99.5% ($\alpha = 0.005$)

Testing hypotheses: t-Test lookup table

Degrees of Freedom	Probability, p			
	0.1	0.05	0.01	0.001
1	6.31	12.71	63.66	636.62
2	2.92	4.30	9.93	31.60
3	2.35	3.18	5.84	12.92
4	2.13	2.78	4.60	8.61
5	2.02	2.57	4.03	6.87
6	1.94	2.45	3.71	5.96
7	1.89	2.37	3.50	5.41
8	1.86	2.31	3.36	5.04

- Confidence (significance) level α: probability of the study rejecting the null hypothesis, given that H0 were true
- $p-$value: probability of obtaining a result at least as extreme, given that H0 were true

Testing hypotheses: t-Test

Collocation example

new companies – a collocation?

How likely is it that "new" and "companies" cooccurr by chance?

- high probability: "new companies" not a collocation
- low probability: "new companies" is a collocation

- Compute probability of w_1 and w_2 cooccurring by chance

$$P(w_1, w_2) = P(w_1)P(w_2)$$

occurrence of w_1 does not dependent on w_2

Testing hypotheses: t-Test

Collocation example

- Is *new companies* a collocation?
- Compute t value for *new companies* in a given corpus:

word	frequency
new	15828
companies	4675
new companies	8
all tokens in corpus	14307668

$$P(new) =$$

Testing hypotheses: t-Test

Collocation example

- Is *new companies* a collocation?
- Compute t value for *new companies* in a given corpus:

word	**frequency**
new	15828
companies	4675
new companies	8
all tokens in corpus	14307668

$$P(new) = \frac{15828}{14307668}$$

$$P(\text{companies}) =$$

Testing hypotheses: t-Test

Collocation example

- Is *new companies* a collocation?
- Compute t value for *new companies* in a given corpus:

word	frequency
new	15828
companies	4675
new companies	8
all tokens in corpus	14307668

$$P(new) = \frac{15828}{14307668}$$

$$P(\text{companies}) = \frac{4675}{14307668}$$

$$H_0 : P(\text{new companies}) = P(\text{new})\ P(\text{companies})$$

Testing hypotheses: t-Test

Collocation example

- Is *new companies* a collocation?
- Compute t value for *new companies* in a given corpus:

word	frequency
new	15828
companies	4675
new companies	8
all tokens in corpus	14307668

$$
\begin{aligned}
P(new) &= \frac{15828}{14307668} \\
P(\text{companies}) &= \frac{4675}{14307668} \\
H_0 : P(\text{new companies}) &= P(\text{new})\ P(\text{companies}) \\
&= \frac{15828}{14307668} x \frac{4675}{14307668} = 3.615x10^{-7}
\end{aligned}
$$

Testing hypotheses: t-Test

Collocation example

- Is *new companies* a collocation?

$H_0 : P(\text{new companies}) =$

Testing hypotheses: t-Test

Collocation example

- Is *new companies* a collocation?

$$H_0 : P(\text{new companies}) = P(\text{new})\ P(\text{companies}) = 3.615x10^{-7}$$

$\Rightarrow$ If H0 is true, the probability of finding cooccurrences of *new companies* in the corpus is $\approx 3.615x10^{-7}$

Testing hypotheses: t-Test

Collocation example

- Is *new companies* a collocation?

$H_0 : P(\text{new companies}) = P(\text{new})\ P(\text{companies}) = 3.615x10^{-7}$

$\Rightarrow$ If H0 is true, the probability of finding cooccurrences of *new companies* in the corpus is $\approx 3.615x10^{-7}$

- Randomly select bigrams from our sample
  - assign 1 if $w_1, w_2 =$ *new companies*
  - assign 0 if $w_1, w_2 \neq$ *new companies*

$\Rightarrow$ Bernoulli trial with $p = 3.615x10^{-7}$ for *new companies*

$$\text{mean } \mu = 3.615x10^{-7}$$
$$\text{variance } \sigma^2 = p(1-p) \approx p$$

Bernoulli distribution

- Discrete probability distribution of a random variable X
- Takes the value 1 with probability p and the value 0 with probability $q = 1 - p$
- Example: coin toss (heads or tails)

Bernoulli distribution

- Discrete probability distribution of a random variable X
- Takes the value 1 with probability p and the value 0 with probability $q = 1 - p$
- Example: coin toss (heads or tails)

- Parameters:
  - $0 \leq p \leq 1$
  - $q = 1$ - p
  - Mean: p $\quad \mu = \mathbb{E}[X] = \sum_{x \in X} xP(X = x)$
    $= 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = 1 \cdot p + 0 \cdot q = p$
  - Variance: $p(1 - p) = pq$

Testing hypotheses: t-Test

Collocation example

- Cooccurrences of *new* and *companies* in the corpus: 8
- Sample mean:

$$\bar{x} = \frac{8}{14307668} = 5.591x10^{-7}$$

- Sample variance:

$$\sigma^2 = p(1-p) \approx p$$

- Use t-Test:

$$t = \frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{N}}} = \frac{5.59110^{-7} - 3.61510^{-7}}{\sqrt{\frac{5.59110^{-7}}{14307668}}} \approx 0.999932$$

Testing hypotheses: t-Test

Collocation example

- Cooccurrences of *new* and *companies* in the corpus: 8
- Sample mean:

$$\bar{x} = \frac{8}{14307668} = 5.591x10^{-7}$$

- Sample variance:

$$\sigma^2 = p(1-p) \approx p$$

- Use t-Test:

$$t = \frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{N}}} = \frac{5.59110^{-7} - 3.61510^{-7}}{\sqrt{\frac{5.59110^{-7}}{14307668}}} \approx 0.999932$$

- $t = 0.999932$ is not larger than table score (2.576)

  $\Rightarrow$ we cannot reject the H0

Testing hypotheses: t-Test

Comparison with Frequency

t-Test applied to 10 bigrams with frequence $C(w_1, w_2) = 20$

t	$C(w_1)$	$C(w_2)$	$C(w_1, w_2)$	w_1	w_2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

Testing hypotheses: t-Test

Comparison with Frequency

t-Test applied to 10 bigrams with frequence $C(w_1, w_2) = 20$

t	$C(w_1)$	$C(w_2)$	$C(w_1, w_2)$	w_1	w_2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

- t-test takes into account number of cooccurrences of w_1, w_2 relative to frequency of individual components

When do we use $N-1$ for computing variance?

Population variance versus sample variance

Bessel's correction

- corrects bias in the estimation of the population variance
  - We want to estimate the variance of the population, based on a smaller sample
  - Problem: the estimate of the population variance based on the sample mean is always smaller than what we would get if we used the population mean
  - Exception: the sample mean happens to be the same as the population mean

When do we use $N - 1$ for computing variance?

Population variance versus sample variance

Bessel's correction

- corrects bias in the estimation of the population variance
  - We want to estimate the variance of the population, based on a smaller sample
  - Problem: the estimate of the population variance based on the sample mean is always smaller than what we would get if we used the population mean
  - Exception: the sample mean happens to be the same as the population mean

- To correct for this bias, we divide by $N - 1$ instead of N (*unbiased estimate* of the population variance)

Pearsons Chi-Square-Test χ^2

- Problem with t-Test: assumes that probabilities are normally distributed (never ever true for natural language!)
- χ^2 Test: does *not* assume normal distribution
- Approach:
  - compare **observed** events with **expected** events under the assumption that observed events are independent of each other

  ⇒ if difference between observed and expected values is large, reject H0

  Independence assumption
  - e.g. rolling two dice: outcome of first die does not depend on second die
- **Natural language**: independence assumption does not hold

  E.g. given a determinier, probability of next word being either a noun or an adjective is much higher than probability of seeing another determiner

Pearsons Chi-Square-Test χ^2

Collocation example

- Is “new companies” a collocation?

Pearsons Chi-Square-Test χ^2

Collocation example

- Is “new companies” a collocation?

word	frequency
new	15828
companies	4675
new companies	8
all bigrams in corpus	14307676

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

Pearsons Chi-Square-Test χ^2

Collocation example

- Is "new companies" a collocation?

word	frequency
new	15828
companies	4675
new companies	8
all bigrams in corpus	14307676

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

Observed frequencies: 2 x 2 table

	w_j =new	$w_j \neq$new
w_i = companies	(new companies)	(z.B. old companies)
$w_i \neq$ companies	(e.g. new machines)	(e.g. old machines)

Pearsons Chi-Square-Test χ^2

Collocation example

- Is "new companies" a collocation?

word	frequency
new	15828
companies	4675
new companies	8
all bigrams in corpus	14307676

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

Observed frequencies: 2 x 2 table

	w_j =new	$w_j \neq$new
w_i = companies	8 (new companies)	(z.B. old companies)
$w_i \neq$ companies	(e.g. new machines)	(e.g. old machines)

Pearsons Chi-Square-Test χ^2

Collocation example

- Is "new companies" a collocation?

word	**frequency**
new	15828
companies	4675
new companies	8
all bigrams in corpus	14307676

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

Observed frequencies: 2 x 2 table

	w_j =new	$w_j \neq$new
w_i = companies	8 (new companies)	4667 (z.B. old companies)
$w_i \neq$ companies	(e.g. new machines)	(e.g. old machines)

Pearsons Chi-Square-Test χ^2

Collocation example

- Is "new companies" a collocation?

word	frequency
new	15828
companies	4675
new companies	8
all bigrams in corpus	14307676

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

Observed frequencies: 2 x 2 table

	w_j =new	$w_j \neq$new
w_i = companies	8 (new companies)	4667 (z.B. old companies)
$w_i \neq$ companies	15820 (e.g. new machines)	(e.g. old machines)

Pearsons Chi-Square-Test χ^2

Collocation example

- Is "new companies" a collocation?

word	frequency
new	15828
companies	4675
new companies	8
all bigrams in corpus	14307676

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

Observed frequencies: 2 x 2 table

	w_j =new	$w_j \neq$new
w_i = companies	8 (new companies)	4667 (z.B. old companies)
$w_i \neq$ companies	15820 (e.g. new machines)	14287181 (e.g. old machines)

Pearsons Chi-Square-Test χ^2

Collocation example

- Compute expected frequencies E_{ij} from marginal probabilities (totals of rows and columns converted into proportions)

Pearsons Chi-Square-Test χ^2

Collocation example

- Compute expected frequencies E_{ij} from marginal probabilities (totals of rows and columns converted into proportions)

Observed frequencies: 2 x 2 table

	w_1 =new	$w_1 \neq$new	total
w_2 = companies	8 (new companies)	4667 (z.B. old companies)	4675
$w_2 \neq$ companies	15820 (e.g. new machines)	14287181 (e.g. old machines)	14303001
total	15828	14291848	**14307676**

- e.g. for cell $c_{1,1}$ (*new companies*):

Pearsons Chi-Square-Test χ^2

Collocation example

- Compute expected frequencies E_{ij} from marginal probabilities (totals of rows and columns converted into proportions)

Observed frequencies: 2 x 2 table

	w_1 =new	w_1 ≠new	total
w_2 = companies	8 (new companies)	4667 (z.B. old companies)	4675
$w_2 \neq$ companies	15820 (e.g. new machines)	14287181 (e.g. old machines)	14303001
total	15828	14291848	**14307676**

- e.g. for cell $c_{1,1}$ (*new companies*):

$$E_{1,1} = \frac{8+4667}{N} \times \frac{8+15820}{N} \times N = 0.00033 \times 0.00111 \times 14307676 \approx 5.2$$

Pearsons Chi-Square-Test χ^2

Collocation example

- Compute expected frequencies E_{ij} from marginal probabilities (totals of rows and columns converted into proportions)

Observed frequencies: 2 x 2 table

	w_1 =new	$w_1 \neq$new	total
w_2 = companies	8 (new companies)	4667 (z.B. old companies)	4675
$w_2 \neq$ companies	15820 (e.g. new machines)	14287181 (e.g. old machines)	14303001
total	15828	14291848	**14307676**

- e.g. for cell $c_{1,1}$ (*new companies*):

$$E_{1,1} = \frac{8+4667}{N} \times \frac{8+15820}{N} \times N = 0.00033 \times 0.00111 \times 14307676 \approx 5.2$$

$\Rightarrow$ If *new* and *companies* are independent of each other, we expect to find 5.2 cooccurrences on average for a text of the size of our sample.

Pearsons Chi-Square-Test χ^2

Collocation example

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4}$$

- i all rows in table
- j alle columns in table
- O_{ij} observed value for cell (i, j)
- E_{ij} expected value for cell (i, j)

Pearsons Chi-Square-Test χ^2

Collocation example

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4}$$

- i all rows in table
- j alle columns in table
- O_{ij} observed value for cell (i, j)
- E_{ij} expected value for cell (i, j)

Simpler form for 2 x 2 tables:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$$

Pearsons Chi-Square-Test χ^2

Collocation example

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (4)$$

- i all rows in table
- j alle columns in table
- O_{ij} observed value for cell (i,j)
- E_{ij} expected value for cell (i,j)

Simpler form for 2 x 2 tables:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$$

$$= \frac{14307676(8x14287181-4667x15820)^2}{(8+4667)(8+15820)(4667+14287181)+(15820+14287181)} \approx 1.55$$

Pearsons Chi-Square-Test χ^2

Collocation example

Is *new companies* a collocation?

- We computed a χ^2 value of 1.55
- Look up in χ^2 table ($\alpha = 0.05$ and $df = 1$) $\Rightarrow \chi^2 =$ **3.841**

Pearsons Chi-Square-Test χ^2

Collocation example

Is *new companies* a collocation?

- We computed a χ^2 value of 1.55
- Look up in χ^2 table ($\alpha = 0.05$ and $df = 1$) $\Rightarrow \chi^2 = \mathbf{3.841}$

$\Rightarrow$ We cannot reject the H0 that *new* and *companies* are independent of each other

Pearsons Chi-Square-Test χ^2

Problems

- Not adequate for rare events:
  - Sample size $N \leq 20$ or
  - Sample size between 20 and 40 and table cells with expected frequencies ≤ 5
- for rare events: *Likelihood Ratio*
- for collocations: usually no huge differences between t-Test and χ^2


I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.
I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT *YEARS* AGO.

References

- Christopher Manning & Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA. (Kapitel 5: Collocations)
- Kenneth Church & Patrick Hanks (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16:1, 22–29.
- Stefan Evert (2004, published 2005). The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.
- Sinclair, J. 1991. Corpus, concordance, collocation: Describing English language. Oxford: Oxford University Press.
- Yaacov Choueka (1988). Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases. In RIAO: 609–624.
- Firth, J.R. (1957). Papers in Linguistics 1934-1951. London: Oxford University Press.
- Table: t-Test: `http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf`
- Table: X^2: `http://www.sjsu.edu/faculty/gerstman/StatPrimer/chisq-table.pdf`