CS 565: Intelligent Systems and Interfaces

Lecture: Words - Collocations

17th Jan, 2017

Semester: Jan - May 2017

Ashish Anand

IIT Guwahati

Recap

- NLP is hard
    - Ambiguity at multiple levels
        - Word
        - Syntax
        - Semantic
        - Discourse

- Getting started with NLP
    - Word segmentation/Tokenization

Objective of lecture

- Sentence Segmentation

- Collocations
  - Definition
  - Characteristics
  - Finding them

Sentence Segmentation

Defining Sentence Boundary

- Something ending with a ‘.’, ‘?’, or ‘!’
  - Language specific
- Problem with ‘.’
  - Still 90% of periods are sentence boundary indicators [Riley 1989].
- Sub-sentence structure with the use of other punctuation
  - “The scene is written with a combination of unbridled passion and sure-handed control: In the exchanges …….. inexorability of separation”
- Other issues
  - “You remind me,” she remarked, “of your mother.”

Defining Sentence Boundary: A heuristic

- Put putative sentence boundaries after occurrences of ., ?, ! (and may be ;, :, -)
- Check presence of following quotation marks, if any move the boundary.
  - “You remind me,” she remarked, “of your mother.”
- Disqualify a period boundary if –
  - It is preceded by a known abbreviation that does not generally occur at the end of sentence such as Dr., Mr. or vs.
  - It is preceded by a know abbrev. that is generally not followed by an uppercase word such as etc. or Jr.
- Disqualify a boundary with a ? or ! If
  - It is followed by a lowercase letter (or name)

Issues with Heuristic or set of pre-defined rules

- Is it possible to define such rules without the help of experts?
- Will it work for all languages?

Machine Learning Methods: Sentence boundary as classification problem

- Riley (1989) used classification trees
  - Features: case & length of the words preceding and following a period; prior prob of words occurring before and after a sentence boundary etc.
- Palmer and Hearst (1997) used neural network model
  - Instead of prior probability, PoS distribution of the preceding and following words.
  - Language-independent model with accuracy of 98-99%
- Reynar and Ratnaparkhi (1997) and Mikheev (1998) used Max. Ent approach
  - Language independent model with accuracy of 99.25%

Collocation: Continuing with words

Collocations: Examples

Strong Tea, Stiff breeze, Take a risk, Start up, New Delhi, Fly High

Vs

Last class, Next lecture, New companies

Collocations: Definition

- [Choueka, 1988]: "*A sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact, unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components*“

- Limitation
  - We may do away with the requirement of words being consecutive.

- Example
  - <u>*Knocked*</u> on the <u>*door*</u>
  - <u>*Knocked*</u> on my <u>*door*</u>
  - <u>*Knocked*</u> at the class-room <u>*door*</u>

Characteristics: subtle and not easily explainable

- *“Strong tea”* but not “*Powerful tea”*
- *“Stiff breeze”* but not *“Stiff wind”*
- *“White wine”* but not *“Yellow wine”*
- *“Broad daylight”* but not *“Bright daylight”*

Characteristics

- Limited compositionality
  - Example: Strong Tea

An expression is ***compositional*** if its meaning can be predicted from the meaning of the parts.

- Non substitutability
  - Example: *yellow* cannot replace *white* in "*white wine*".
- Non-modifiability: can't be modified using additional lexical materials or through grammatical transformations.
  - Example: *people as poor as church <u>mice</u>; to get an <u>ugly</u> frog in one's throat.*

Why it is important?

- Computational lexicography
- Parsing
- Natural Language Generation
- Machine Translation
- Linguistic research

Finding Collocations

Frequency

- Assumption: More frequent occurrence of two words together may imply special function or property which can’t be simply explained

$C(w^1 w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a

Frequency based methods for finding collocations

Source: Table 5.1[FSNLP: Page 154]

Corpus: New York Times newswire-Aug to Nov 1990.
Statistics: 115 MB text with roughly 14 million words

Adding linguistic knowledge to Frequency

Tag Pattern
A N
N N
A A N
A N N
N A N
N N N
N P N

Source: Table 5.2 [FSNLP: 154]

1. Part of Speech (PoS) tag patterns for collocation filtering.
2. Patterns were proposed by *Justeson and Katz* (1995).
3. [*A*]djective; [*N*]oun; [*P*]reposition

$C(w^1\ w^2)$	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
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Table 5.3 Finding Collocations: Justeson and Katz' part-of-speech filter.

Source: Table 5.3 [FSNLP: Page 155]

Pros and Cons of Frequency+PoS Filter

- Advantages
  - Simple method

- Disadvantages
  - Too much dependency on hand-designed filter
  - High frequency can be random without any specific meaning
  - Works well for fixed phrases but will not work for cases where variable number of words may exist between two words
  - Example
    - She *knocked* on his *door*
    - They *knocked* at the *door*
    - 100 women *knocked* on Donaldson’s *door*
    - a man *knocked* on the metal front *door*

Sliding window could be savior

Sentence:

man knocked on the front door

Bigrams:

man knocked	*man on*	*man the*	*man front*	
	knocked on	*knocked the*	*knocked front*	*knocked door*
		on the	*on front*	*on door*
			the front	*the door*
				front door

Four word collocational window to capture bigrams at a distance

Mean and Variance

- Can implicitly take care of varying distance issue
- Method
  - Calculate mean of *offsets* (signed distance) between the two words.

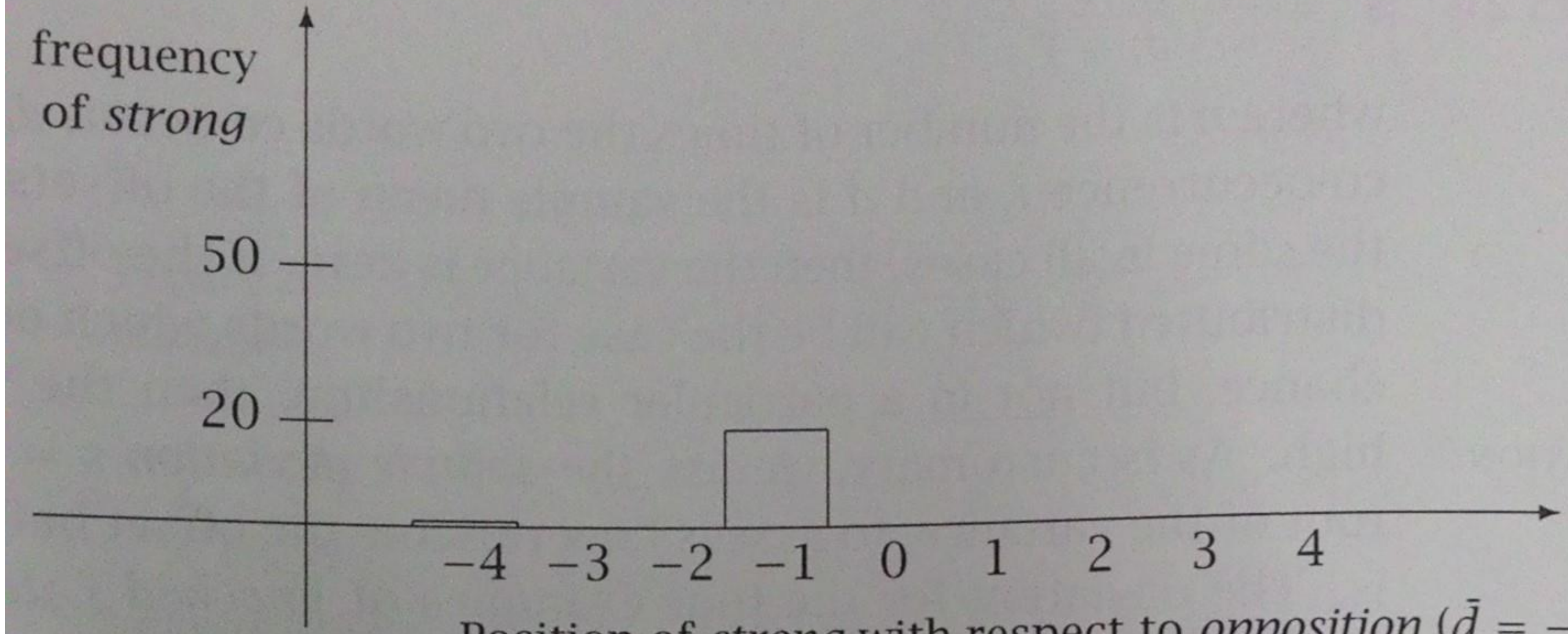
She *knocked* on his *door*
They *knocked* at the *door*
100 women *knocked* on Donaldson's *door*
a man *knocked* on the metal front *door*

  - Mean, $\bar{d}$ = ¼(3 + 3 + 5 + 5)
    [Donaldson's tokenized as : Donaldson, apostrophe, s]
  - Variance, $s^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}$

s	$\bar{d}$	Count	Word 1	Word 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

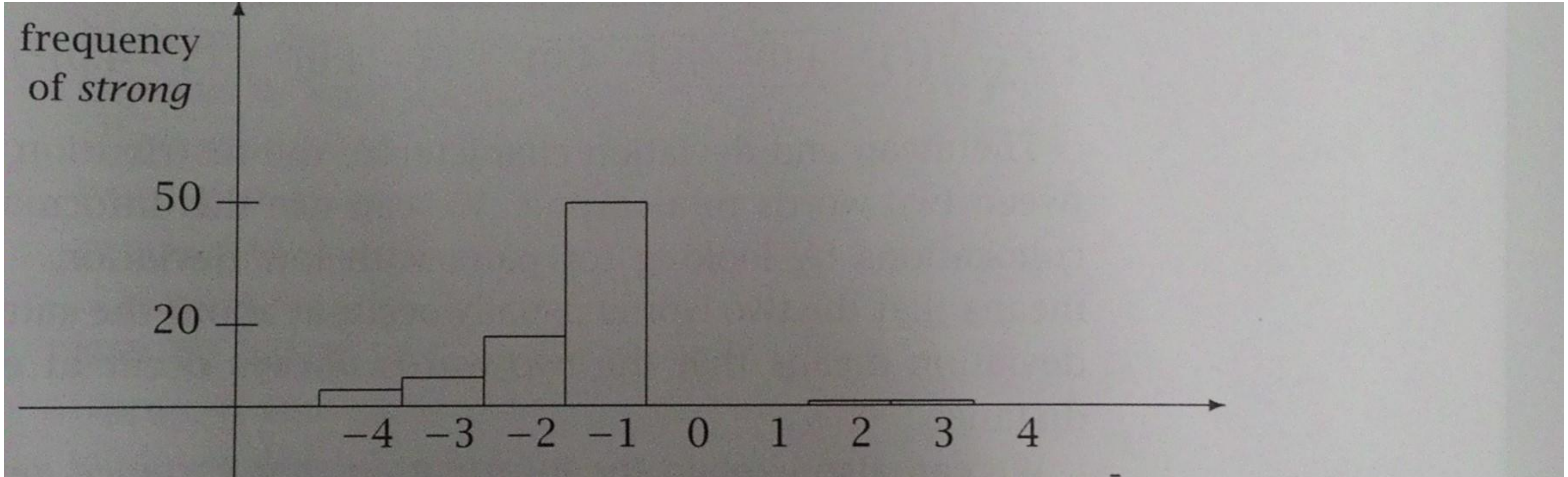
Table 5.5 Finding collocations based on mean and variance. Sample deviation s and sample mean $\bar{d}$ of the distances between 12 word pairs.

Source: Table 5.5 [FSNLP: page 161]



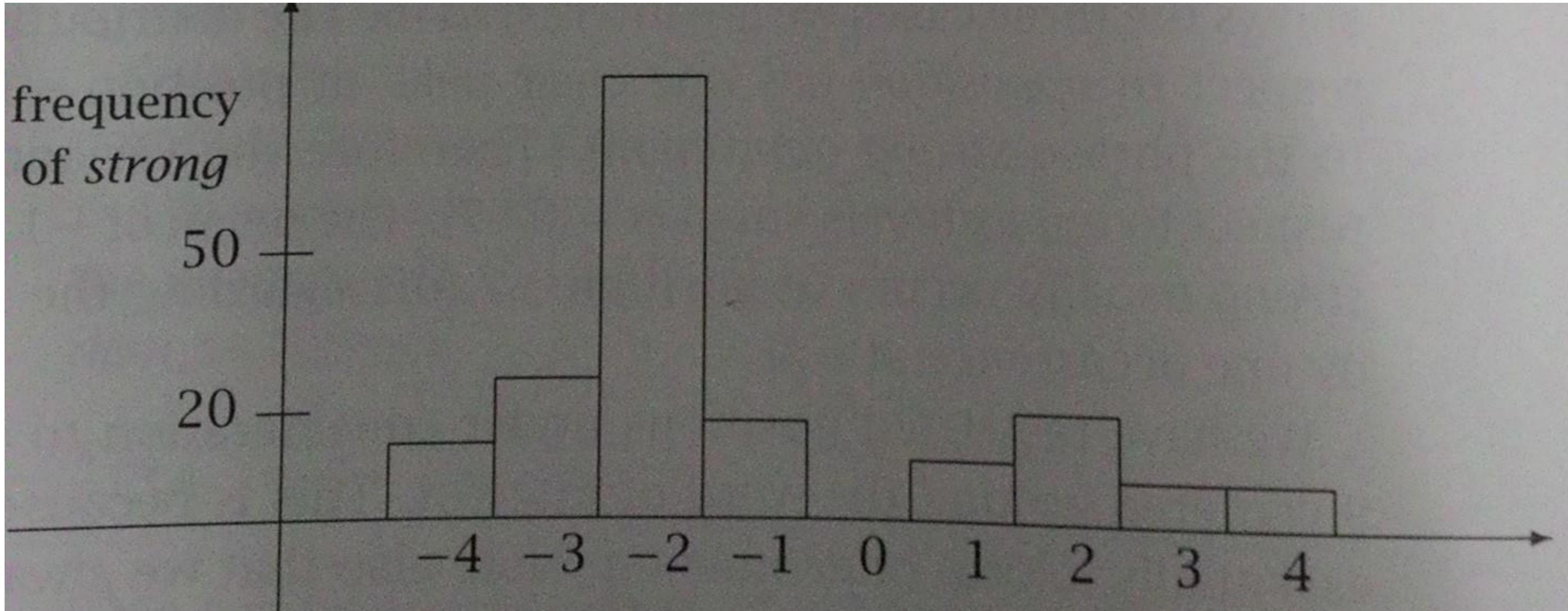
Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).

Source: Figure 5.2 [FSNLP: page 160]



Position of *strong* with respect to *support* ($\bar{d} = -1.45, s = 1.07$).

Source: Figure 5.2 [FSNLP: page 160]



Position of *strong* with respect to *for* ($\bar{d} = -1.12, s = 2.15$).

Figure 5.2 Histograms of the position of *strong* relative to three words.

Source: Figure 5.2 [FSNLP: page 160]

Issues with Mean & Variance Approach

- Unable to differentiate with chance cases

- Why this is happening?
  - High frequency of individual words, hence likely to co-occur together quite often

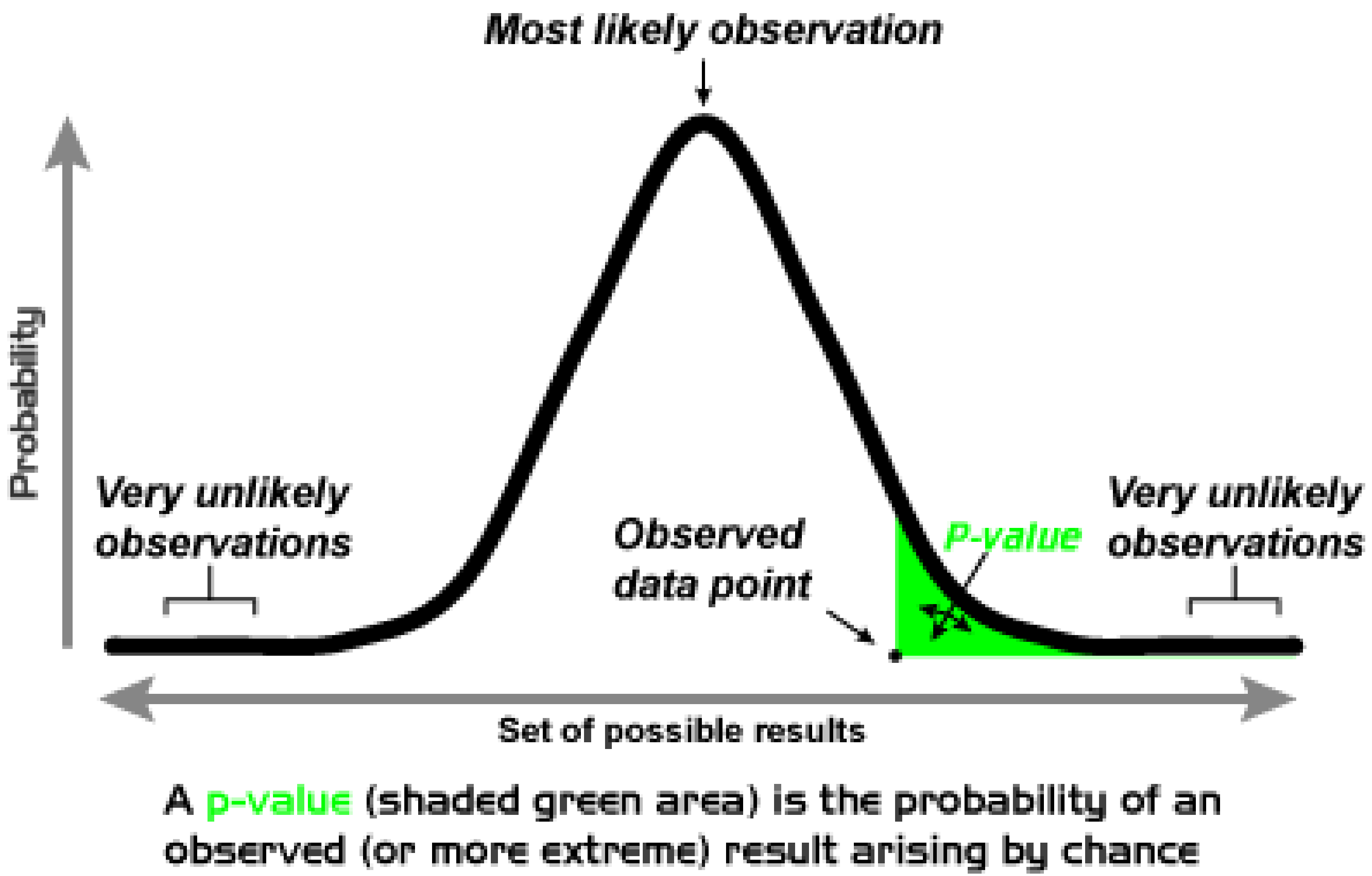
Hypothesis Testing: Mitigating the chance issue

- Objective: Able to make distinction whether two words are co-occurring more frequently *just by chance*.
- Method: Hypothesis Testing
- Steps are
  - Formulate *Null Hypothesis, H_0 :* There is no association between the words beyond chance occurrences.
  - Compute the probability *p* that the *event (corresponding statistics)* occurs if H_0 is true.
  - Reject null hypothesis if *p* is too low

Statistical Test: t-test

- Null Hypothesis: *Sample is drawn from a distribution with mean μ*

- $t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$



Source: https://en.wikipedia.org/wiki/One-_and_two-tailed_tests

Finding collocations: Formulating Hypothesis

- Formulation of Null Hypothesis, H_0 :
  - $P(w_i)$: Probability of occurrence of individual word
  - $P(w_i, w_j)$: Probability of co-occurrence of the two words
  - Under H_0 : $P(w_i, w_j) = P(w_i) * P(w_j)$

Using *t-test* for finding collocations

- Text corpus as a sequence of *N* bigrams
- $P(w_i)$ = # of occurrences of word w_i / total # of words
- H_0 : $P(w_i, w_j) = P(w_i) * P(w_j)$ [occurrence of the two words are independent]

- Under null hypothesis, process of random occurrence of the bigram is a *Bernoulli Trial* with $p = P(w_i, w_j) = P(w_i) * P(w_j)$
- *Mean, μ* = *p*; *variance* = $p(1-p) \approx p$
- Calculate $\bar{x}$ and std. dev.

Example

For the bigram *new companies*

P(new) = 15828 / 14307668

P(companies) = 4675 / 14307668

μ = P(new companies) = 3.615×10^{-7}

Actual occurrence of *new companies* = 8

t = 0.999932 < t_critical at 0.005 = 2.576

Give your verdict

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

Table 5.6 Finding collocations: The t test applied to 10 bigrams that occur with frequency 20.

Reference

- Chapter 5 FSNLP
- FSNLP: Foundations of Statistical Natural Language Processing, Manning & Schütze