CS565: Intelligent Systems and Interfaces

Words: Finding Collocations

Semester: Jan – May 2019

Ashish Anand

Associate Professor, Dept of CSE

IIT Guwahati

Announcements

- Scribe for Next two lectures
  - Vaibhav Pandey, Dhananjay, Susrita: 23rd Lec
- Extra Class on this Thursday, 24th Jan at 2 PM

Recap

- Understand corpus data at word level
  - Uneven Distribution with long tail
  - Zipf’s and Mandelbrot’s Laws to describe this distribution

  - Collocation
    - What is it and its characteristics
    - How to Find them
      - Frequency based approach
      - Frequency with linguistic knowledge in form of syntactic patterns

Objective

- Continuing with ways to find collocation
  - Deal with collocation at a distance
  - Making sure observation is not random
    - Hypothesis Testing Methods

Finding Collocation

Pros and Cons of Frequency + Syntactic Pattern 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 *<u>knocked</u>* on his *<u>door</u>*
They *<u>knocked</u>* at the *<u>door</u>*
100 women *<u>knocked</u>* on Donaldson's *<u>door</u>*
a man *<u>knocked</u>* on the metal front *<u>door</u>*

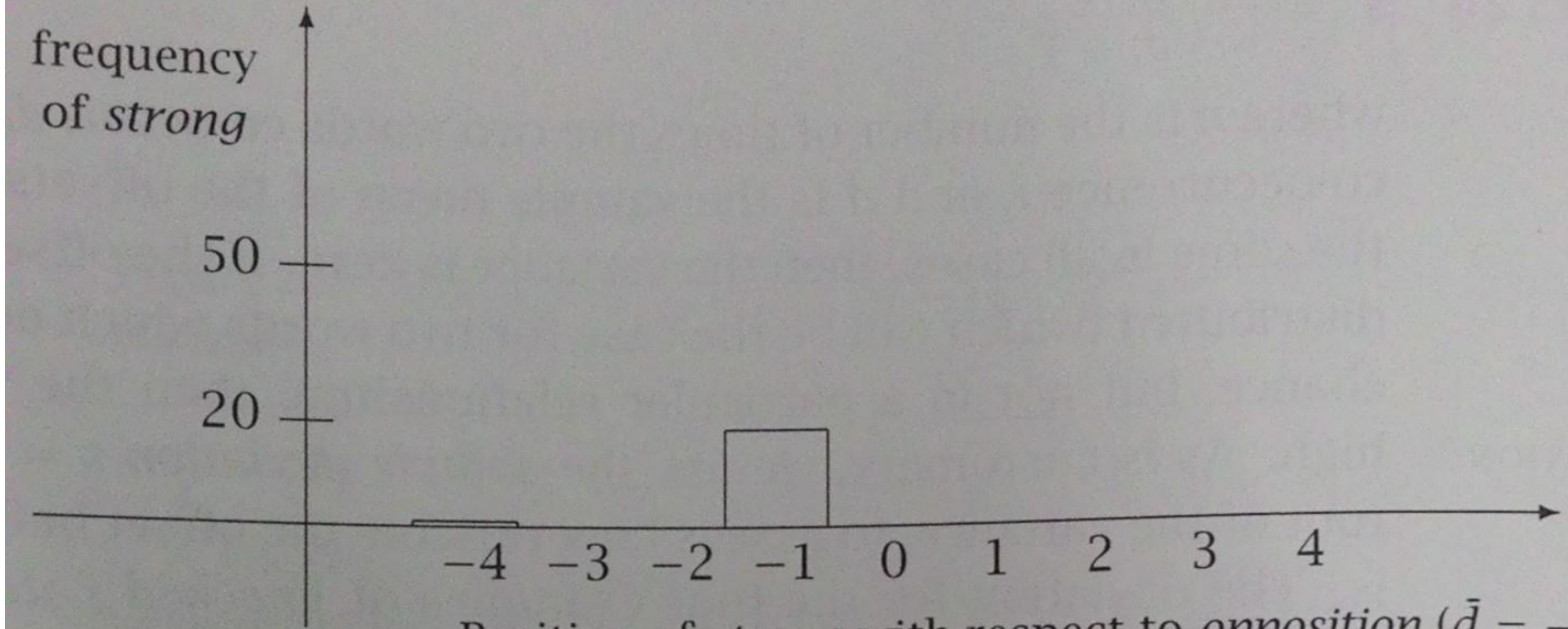
  - Mean, $\bar{d}$ = ¼(3 + 3 + 5 + 5)
    <u>[Donaldson's tokenized as : Donaldson, apostrophe, s]</u>

  - 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

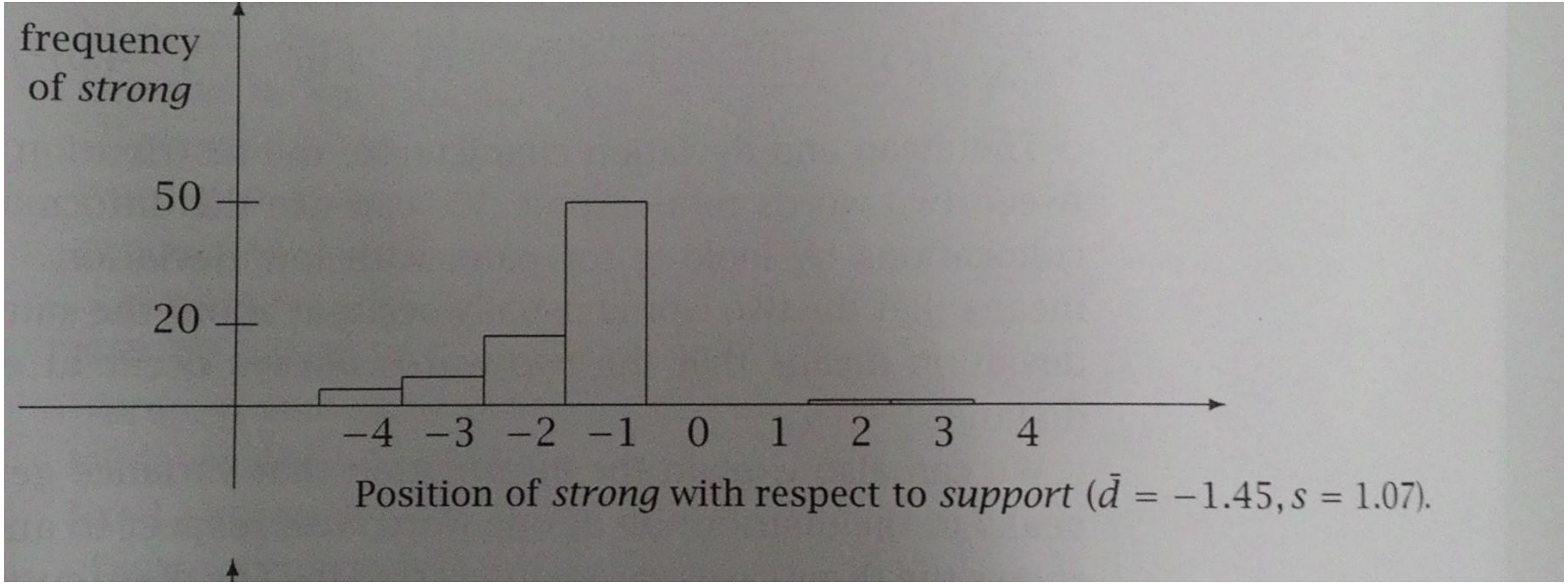
Table 5.5 Finding collocations based on mean and variance. Sample deviation s and sample mean $\bar{d}$ of the distances between 12 word pairs.

Source: Table 5.5 [FSNLP: page 161]

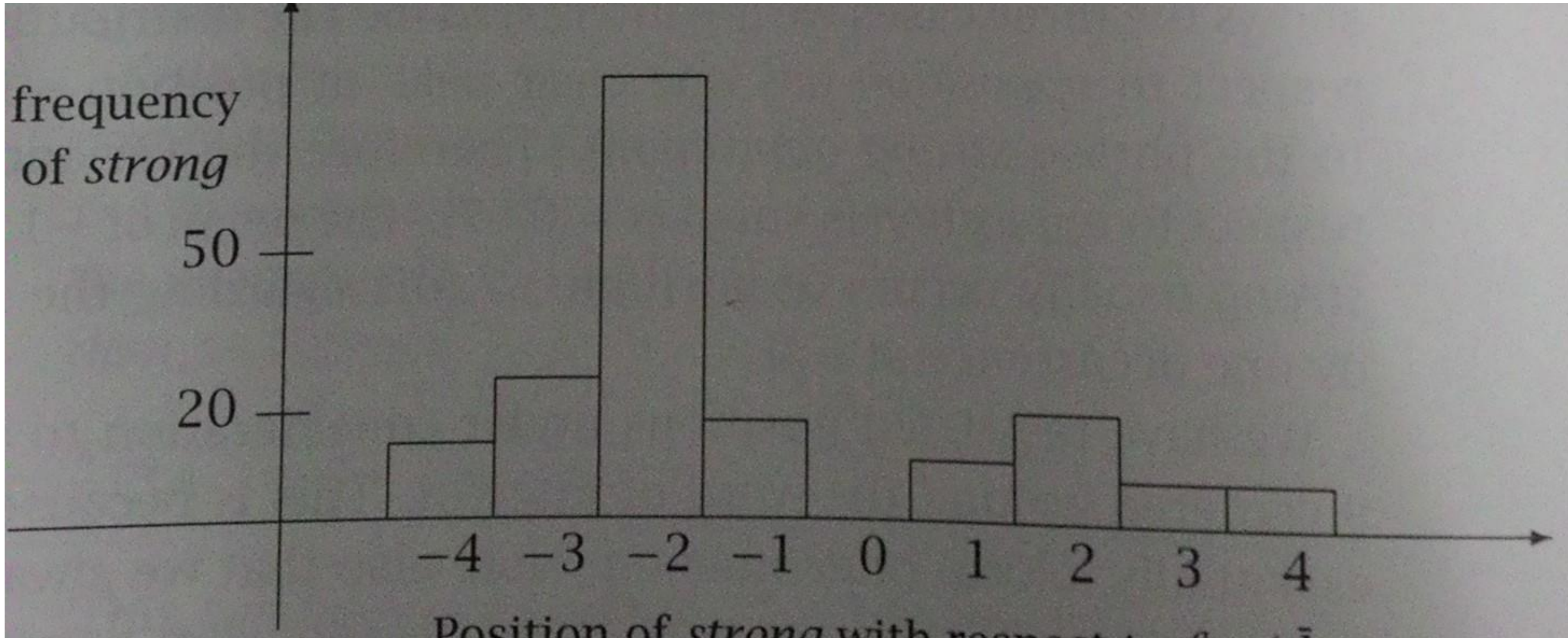


Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).

Source: Figure 5.2 [FSNLP: page 160]



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: Whether the observation is significantly different than just being a random event

- Objective in our case: whether words occur together more frequently than they would have occurred together by chance

- Steps are
  - Formulate *Null Hypothesis, H_0 :* model random event appropriately
  - Decide Significance Level: Probability of rejecting H_0 when it is true
  - Compute the probability p that the *event (corresponding statistics)* occurs if H_0 is true.
  - Reject null hypothesis if p is less than the significance level

Statistical Test: t-test

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

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

Example: Study of men heights

<u>Null Hypothesis, H_0</u> : Sample is drawn from general population of men with mean heights = 158 cm

Sample size, *N* = 200; Observed/sample mean = 169 cm; sample variance = 2600

$t \approx 3.05$

Critical value of t-statistics = ±2.83

Give your verdict

Question: How to use t-test in this problem?

- What are my samples?
- What is sample size?
- What is sample mean?
- What is expected mean?

Deciding sample answers all questions

- Consider corpus : collection of n-grams
- Samples: Indicator random variable corresponds to the target n-gram.
- Sample size: # of n-grams
- x_i ~ Bernoulli (p)

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 [MLE]
- 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 <u>*Bernoulli Trial*</u> with $p = P(w_i, w_j) = P(w_i) * P(w_j)$
- <u>*Mean,* μ</u> $= p$; <u>*variance*</u> $= p(1-p) \approx p$
- Calculate $\bar{x}$ and std. dev.

References

- Chapter 5 [FSNLP]