CS 565: Intelligent Systems and Interfaces

Lecture: Finding Collocations

21st Jan 2016

Semester: Jan - May 2016

Ashish Anand

IIT Guwahati

Issues with Mean & Variance Approach

- Unable to differentiate with chance cases
  - Example: *last year*, *next year*, *new companies*

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

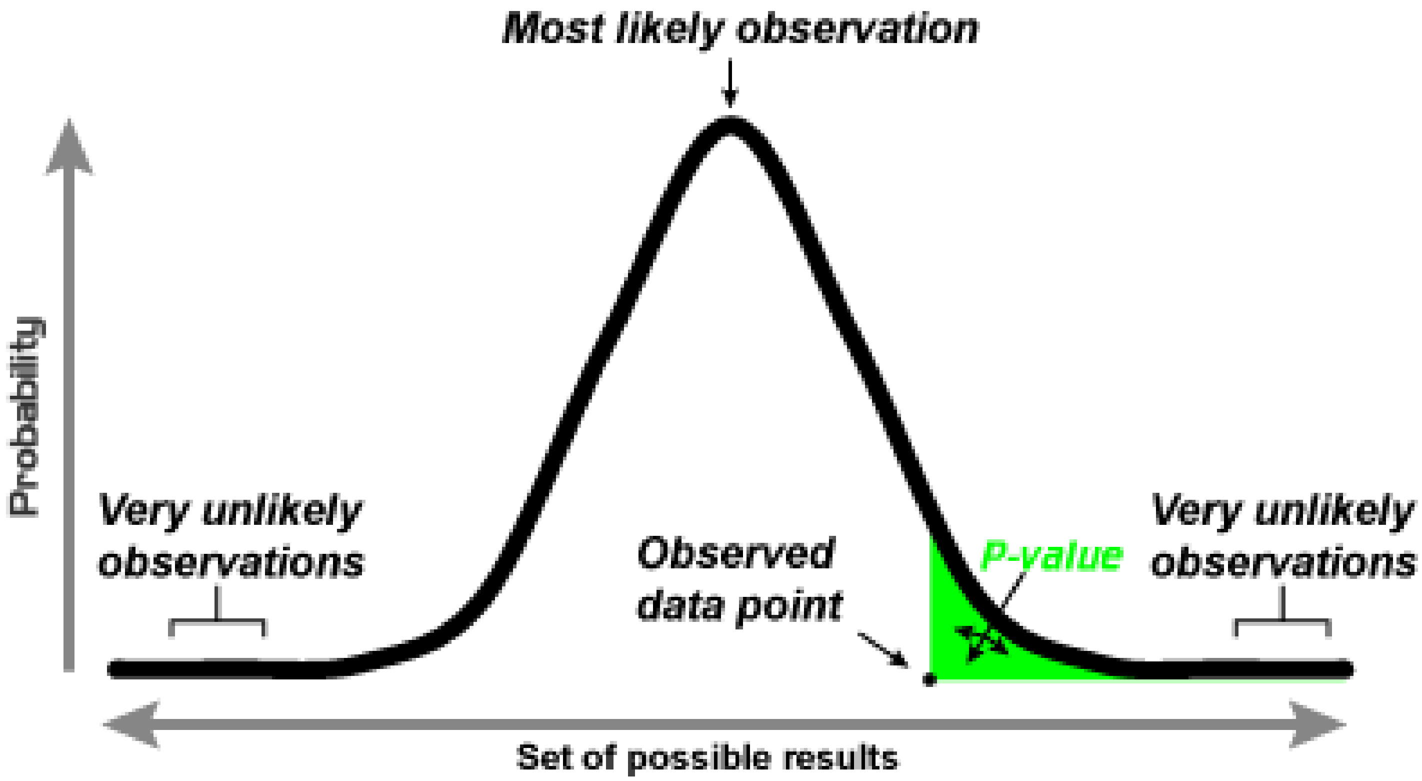
Hypothesis Testing: Mitigating the chance issue

- 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* 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*

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



A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

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,* $\mu = p$; *variance* $= p(1-p) \approx p$
- Calculate $\bar{x}$ and std. dev.

Example

For the bigram *new companies*

P(new) = 15828 / 14307668

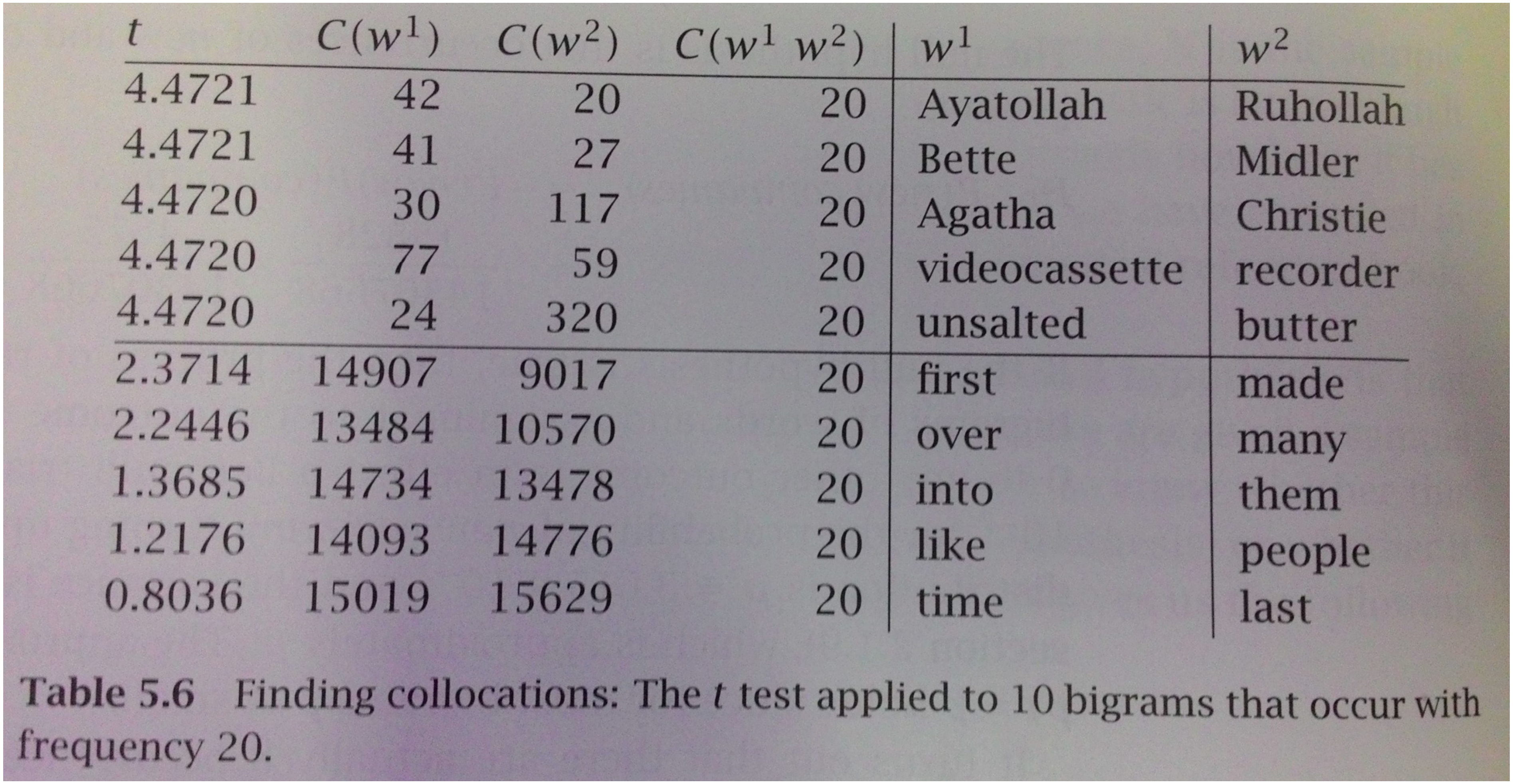
P(companies) = 4675 / 14307668

μ = P(new companies) = 3.615 x 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

- Reference
  - FSNLP: 5.3

- Next Lecture
  - Alternative tests