Homework

• For Wednesday 3/9:
  
  • Get bigram and unigram frequencies from ANC website:
    http://americannationalcorpus.org/frequency.html
  
  • Implement two of the collocation-finding methods described in
chapater 5
  
  • Find the top \( n \) bigrams by each measure
  
  • Write it up – a paragraph or two describing what's going on

Chi square

<table>
<thead>
<tr>
<th></th>
<th>male %</th>
<th>female %</th>
<th>( X^2 )</th>
<th></th>
<th>male %</th>
<th>female %</th>
<th>( X^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>fucking</td>
<td>0.08</td>
<td>0.01</td>
<td>1233.1</td>
<td>she</td>
<td>0.42</td>
<td>0.87</td>
<td>3109.7</td>
</tr>
<tr>
<td>er</td>
<td>0.56</td>
<td>0.36</td>
<td>945.4</td>
<td>her</td>
<td>0.14</td>
<td>0.28</td>
<td>965.4</td>
</tr>
<tr>
<td>the</td>
<td>2.60</td>
<td>2.20</td>
<td>698.0</td>
<td>said</td>
<td>0.29</td>
<td>0.47</td>
<td>872.0</td>
</tr>
<tr>
<td>yeah</td>
<td>1.29</td>
<td>1.10</td>
<td>310.3</td>
<td>n't</td>
<td>1.44</td>
<td>1.70</td>
<td>443.9</td>
</tr>
<tr>
<td>aye</td>
<td>0.07</td>
<td>0.03</td>
<td>291.8</td>
<td>l</td>
<td>3.24</td>
<td>3.58</td>
<td>357.9</td>
</tr>
<tr>
<td>right</td>
<td>0.36</td>
<td>0.27</td>
<td>276.0</td>
<td>and</td>
<td>1.73</td>
<td>1.94</td>
<td>245.3</td>
</tr>
<tr>
<td>hundred</td>
<td>0.09</td>
<td>0.05</td>
<td>251.1</td>
<td>to</td>
<td>1.37</td>
<td>1.54</td>
<td>198.6</td>
</tr>
<tr>
<td>fuck</td>
<td>0.02</td>
<td>0.00</td>
<td>239.0</td>
<td>cos</td>
<td>0.20</td>
<td>0.26</td>
<td>194.6</td>
</tr>
<tr>
<td>is</td>
<td>0.79</td>
<td>0.67</td>
<td>233.3</td>
<td>oh</td>
<td>0.78</td>
<td>0.90</td>
<td>170.2</td>
</tr>
<tr>
<td>of</td>
<td>0.81</td>
<td>0.69</td>
<td>203.6</td>
<td>Christmas</td>
<td>0.02</td>
<td>0.04</td>
<td>163.9</td>
</tr>
</tbody>
</table>

Bootstrap

• We have two samples of written Dutch, one collected from
newspapers and one from weekly news magazines:

<table>
<thead>
<tr>
<th>Type</th>
<th>Length (w)</th>
<th>Length (s)</th>
<th>Avg w/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>newspapers</td>
<td>140,869</td>
<td>7,155</td>
<td>19.69</td>
</tr>
<tr>
<td>magazines</td>
<td>145,423</td>
<td>7,507</td>
<td>19.37</td>
</tr>
</tbody>
</table>

• Question: is the mean sentence length different for the two text
types?
  
  \( H_0 \) : \( \mu_0 = \mu_1 \)

  \( H_A \) : \( \mu_0 \neq \mu_1 \)

• The standard test for a difference of means is the \( t \) test, which
assumes that the populations being compared are normally
distributed with equal variances.

Bootstrap

> hist(news,breaks="FD",main="Newspaper",xlim=c(0,120),xlab="Words",col=5)
Sentence lengths are skewed, so the results of a *t* test may or may not be accurate.

Non-parametric tests lack power, but misapplied parametric tests may be misleading.

*Bootstrap* methods provide an alternative. Rather than making assumptions about the sampling distribution of a statistic, we can estimate it directly by computational simulation.

Bootstrap techniques are related to permutation methods, which in turn are related to tests like the sign test.

Applying the bootstrap to the difference of means:

- For our sample, \( m = m_2 - m_1 = 0.3165 \). What would be the probability of getting a value of \( m \) at least this extreme if the means of the population were equal?

  - Combine the newspaper and magazine sentences into a single dataset.
  - Select (with replacement) 7,155 sentences for set 1 and 7,507 sentences for set 2.
  - Calculate the \( m \) for these two new samples, call it \( m^* \).
  - Repeat until we have 999 \( m^* \)’s.
  - The distribution of \( m^* \) reflects the sampling distribution of \( m \) under the null hypothesis:

\[
p = \frac{1 + \#\{|m^*| \geq |m|\}}{1000}
\]
Bootstrap

• In this case:
  \[ p = \frac{104}{1000} = 0.104 \]
so we cannot reject the null hypothesis.

• As it turns out, the sampling distribution of \( \mu_1 - \mu_0 \) is normal, so the
  \( t \) test would have been appropriate. If we apply it, we get \( p = 0.101 \)

• The bootstrap has power comparable to more standard ‘normal theory’ tests.

• The bootstrap can be applied to statistics that don’t have any
  ‘normal theory’ results (precision, recall, crossing brackets, etc.)

• Main requirement: samples must be i.i.d.

• In R: `library(boot), boot(), boot.ci()`

Hypothesis testing

• \( t \) test, Wilcoxon rank sum test, sign test, chi square test

• Non-parametric tests depends on few assumption beyond that the
  sample is independent and identically distributed (i.i.d.).

• Parametric tests have more power, but put stricter requirements on
  the kind of data that can be analyzed

• If a more powerful test applies, use it. But, always read the fine print.

• Simulation (bootstrap) can yield results even when analysis is
  impossible