Statistics in Linguistics Tutorial
Just a sip...

Mike Hammond
Linguistics, U. of Arizona
Overview
Overview

- Are our data categorical?
Overview

- Are our data categorical?
- Typological claims
Overview

- Are our data categorical?
- Typological claims
- Claims about corpora
Overview

- Are our data categorical?
- Typological claims
- Claims about corpora
- An easy appropriate test: $\chi^2$ (Chi-square)
Why do statistics?

Some linguistic facts are categorical:
Why do statistics?

Some linguistic facts are categorical:

- ‘John loves Mary’ is grammatical in English.
Why do statistics?

Some linguistic facts are categorical:

- ‘John loves Mary’ is grammatical in English.
- The past tense of look is looked.
Why do statistics?

Some linguistic facts are categorical:

- ‘John loves Mary’ is grammatical in English.
- The past tense of look is looked.
- The English word for cat is [kæt].
Typological claims
Typological claims

- Subject agreement is more common than object agreement.
Typological claims

- Subject agreement is more common than object agreement.
- Syntactic ergativity is rare, e.g. Dyirbal.
Typological claims

- Subject agreement is more common than object agreement.
- Syntactic ergativity is rare, e.g. Dyirbal.
- The vowel [a] is more frequent than [ü].
Claims about corpora
Claims about corpora

- English disprefers words like [spVp] and [skVk].
Claims about corpora

- English disprefers words like [spVp] and [skVk].
- Active sentences are more common than passive sentences.
Claims about corpora

- English disfavors words like [spVp] and [skVk].
- Active sentences are more common than passive sentences.
- Item $x$ is an exception to generalization $y$. 
How do we know if these are true?

Can we as linguists really make good judgments about what is more or less common?
For example
For example

Is [ü] under-represented in the languages of the world? Imagine we have a sample of 100 languages, and we find this:
For example

Is [ü] under-represented in the languages of the world? Imagine we have a sample of 100 languages, and we find this:

<table>
<thead>
<tr>
<th>with [ü]</th>
<th>without [ü]</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>35</td>
<td>65</td>
</tr>
</tbody>
</table>
For example

Is [ü] under-represented in the languages of the world? Imagine we have a sample of 100 languages, and we find this:

<table>
<thead>
<tr>
<th>with [ü]</th>
<th>without [ü]</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>35</td>
<td>65</td>
</tr>
</tbody>
</table>

Nothing
For example

Is [ü] under-represented in the languages of the world? Imagine we have a sample of 100 languages, and we find this:

<table>
<thead>
<tr>
<th>with [ü]</th>
<th>without [ü]</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>35</td>
<td>65</td>
</tr>
</tbody>
</table>

Nothing

Something
For example

Is \([\ddot{u}]\) under-represented in the languages of the world? Imagine we have a sample of 100 languages, and we find this:

<table>
<thead>
<tr>
<th>with [\ddot{u}]</th>
<th>without [\ddot{u}]</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>35</td>
<td>65</td>
</tr>
</tbody>
</table>
How a chi-square works

- Intuitively: how likely is it that the observed distribution would occur by chance?

- More formally: \( \chi^2 = \sum \frac{(O-E)^2}{E} \), where
  - \( O = \) observed frequency and
  - \( E = \) expected frequency

- More practically: Perlman ustats, Free R stats program, SPSS on the DASL machines and on the u-cluster, etc.
The moral

- Even orthodox syntacticians, morphologists, and phonologists can make use of statistics.
- Sometimes the required statistical tool can be really simple.