Homework

- Read *simpleR* (John Verzani, CUNY), pages 1–24
  

- Try exercises 2.1–2.6 (pages 7–8), 3.1–3.7 (pages 18–19), 4.1 (page 31), 4.9 (page 32)

- Just try the exercises, you *don’t* need to write them up
Paired sign test

• Suppose we have a baseline grammar $G_1$ and an ‘improved’ grammar $G_2$.

• Evaluating each with a test suite of 1,200 sentences yields 52.5% coverage for $G_1$ and 54.1% coverage for $G_2$. Can we conclude that the $G_2$ is better than $G_1$?

• If $G_2$ is better than $G_1$, then there should be more sentences that $G_2$ gets right but $G_1$ gets wrong than the other way around. Our test statistic $r$ is the number of sentences on which, if the grammars did not agree, $G_2$ performed better than $G_1$.

• Of 1,200 sentences, $G_2$ performed better on 38 sentences, and $G_1$ performed better on 20.
Paired sign test

• Non-parametric, one-tailed, matched pair sign test, $\alpha = 0.05$:
  $H_0 : p = 0.5$
  $H_A : p > 0.5$

• If the probability $p$ of getting one sentence right is 0.5, then the probability of getting at least 38 out 58 sentences right is:

$$P(R \geq 38|H_0) = \sum_{i=38}^{58} \binom{58}{i}0.5^i(1-0.5)^{58-i} = 0.01237302$$

• Since $P(R \geq 38|H_0) < 0.05$, we can safely reject the null hypothesis and conclude with better than 95% confidence that $G_2$ is indeed an improvement over $G_1$. 
Paired sign test

```r
> binom.test(38, 38+20, alternative="greater")

  Exact binomial test

data: 38 and 38 + 20
number of successes = 38, number of trials = 58, p-value = 0.01237
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
  0.5395568 1.0000000
sample estimates:
probability of success
  0.6551724
```
Paired sign test
Paired sign test

- The paired sign test depends on few assumption beyond that the sample is *independent and identically distributed* (i.i.d.).

- Other standard tests have more power, but put stricter requirements on the kind of data that can be analyzed: Wilcoxon rank sum test (median), *t* test (mean)

- If a more powerful test applies, use it. But, always read the fine print.
The chi-square test is used to compare counts.

For example, we count the number of third person singular references of each type in an English corpus and a Japanese corpus.

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ellipsis</td>
<td>0</td>
<td>104</td>
</tr>
<tr>
<td>Central pronouns</td>
<td>314</td>
<td>73</td>
</tr>
<tr>
<td>Non-central pronouns</td>
<td>28</td>
<td>12</td>
</tr>
<tr>
<td>Names</td>
<td>291</td>
<td>314</td>
</tr>
<tr>
<td>Common NPs</td>
<td>174</td>
<td>205</td>
</tr>
</tbody>
</table>

Do English and Japanese differ in the way they express third person singular referents?
Nominal data

```r
> counts <- matrix(c(104,73,12,314,205,0,314,28,291,174),ncol=2)
> counts
   [,1] [,2]
[1,] 104  0
[2,]  73 314
[3,]  12  28
[4,] 314 291
[5,] 205 174
> outer(rowSums(counts),colSums(counts))/sum(counts)
   [,1]      [,2]
[1,] 48.60198 55.39802
[2,] 180.85545 206.14455
[3,]  18.69307  21.30693
[4,] 282.73267 322.26733
[5,] 177.11683 201.88317
> chisq.test(counts)

Pearson's Chi-squared test

data:  counts
X-squared = 258.5247, df = 4, p-value < 2.2e-16
```
Nominal data

- The chi-square test gives unpredictable results if expected values are too small (sparse data problems).

- For corpus data, chi-square test almost always indicates significance (non i.i.d. samples).

- Results are usually too coarse to interpret.

- Generalized linear models are a better way to analyze contingency tables.

- But, the statistics ($x, X^2$) can be useful descriptively.