Lecture 1: motivating examples

25 September 2008

1 Introduction

The theory and methods of probability and statistics are becoming more and more useful in the study of language. Today we'll cover two home-grown examples of how probabilistic methods are being used here at UCSD Linguistics to study various types of linguistic phenomena.

2 Analysis of variance for a self-paced reading study

This is the outcome of a self-paced reading experiment conducted by Hannah Rohde, in collaboration with me and Andy Kehler.

The question under investigation is whether certain kinds of verbs (implicit causality (IC) verbs) such as “detest”, which intuitively demand some sort of explanation, can affect readers’ online syntactic attachment preferences.

(1) John detests the children of the musician who is generally arrogant and rude (IC,LOW)
(2) John detests the children of the musician who are generally arrogant and rude (IC,HIGH)
(3) John babysits the children of the musician who is generally arrogant and rude (NONIC,LOW)
(4) John babysits the children of the musician who are generally arrogant and rude (NONIC,HIGH)

Hannah hypothesized that the use of an IC verb should facilitate reading of high-attached RCs, which are generally found in English to be harder to read than low-attached RCs (Cuetos & Mitchell, 1988). The reasoning here is that the IC verbs demand an explanation, and one way of encoding that explanation linguistically is through a relative clause. In these cases, the most plausible type of explanation will involve a clause in which the object of the IC verb plays a role, so an RC modifying the IC verb’s object should become more expected. This stronger expectation may facilitate processing when such an RC is seen (Levy, 2008).
The stimuli for the experiment consist of 20 quadruplets of sentences of the sort above. Such a quadruplet is called an EXPERIMENTAL ITEM in the language of experimental psychology. The four different variants of each item are called the CONDITIONS. Since a participant who sees one of the sentences in a given item is liable to be strongly influenced in her reading of another sentence in the item, the convention is only to show each item once to a given participant. To achieve balance, each participant will be shown five items in each condition.

<table>
<thead>
<tr>
<th>Participant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IC,HIGH</td>
<td>NONIC,HIGH</td>
<td>IC,LOW</td>
<td>NONIC,LOW</td>
<td>IC,HIGH</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>NONIC,LOW</td>
<td>IC,HIGH</td>
<td>NONIC,HIGH</td>
<td>IC,LOW</td>
<td>NONIC,LOW</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>IC,LOW</td>
<td>NONIC,LOW</td>
<td>IC,HIGH</td>
<td>NONIC,HIGH</td>
<td>IC,LOW</td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td>NONIC,HIGH</td>
<td>IC,LOW</td>
<td>NONIC,LOW</td>
<td>IC,HIGH</td>
<td>NONIC,HIGH</td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>IC,HIGH</td>
<td>NONIC,HIGH</td>
<td>IC,LOW</td>
<td>NONIC,LOW</td>
<td>IC,HIGH</td>
<td>...</td>
</tr>
</tbody>
</table>

The experimental data will be analyzed for effects of verb type and attachment level.

In self-paced reading, the observable effect of difficulty at a given word often shows up a word or two downstream, so in this case we will focus on the second word after the disambiguator—i.e., “generally”. The words in this part of the sentence were presented one word at a time, so the second word is in the second region after the critical region—this is called the SECOND SPILLOVER REGION. First, let us look at reading times in each of the four conditions of this experiment:

```r
data <- read.table("rohde-levy-kehler-expt2-results.txt", header=T, sep="\t", quote="")
with(subset(data, crit=="RC_VERB+2"), tapply(rt, list(verb, attachment), mean))
```

<table>
<thead>
<tr>
<th></th>
<th>high</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC</td>
<td>437.0605</td>
<td>451.2396</td>
</tr>
<tr>
<td>NONIC</td>
<td>473.0942</td>
<td>436.2296</td>
</tr>
</tbody>
</table>

You’ll notice that the reading times (RTs) are numerically fastest in the IC, HIGH) and NONIC, LOW) conditions, and numerically slowest in the IC, LOW) and NONIC, HIGH) conditions. This pattern is an INTERACTIVE pattern—going from LOW to HIGH speeds reading in the IC conditions but slows reading in the NONIC conditions. However, we need to establish how unlikely it would be to obtain such a numerical pattern of results as a mere chance occurrence if there were in fact no interaction between attachment and verb type.

For many years dating back to Clark (1973), the gold standard in this situation has been to construct two separate analyses of variance (ANOVA, executed in R with `aov()`): one for subjects, and one for items. The question of interest is whether the interaction between RC attachment level and verb type is STATISTICALLY SIGNIFICANT. In the analysis over subjects, we take as our individual data points the mean value of all the observations in each cell of Subject × Verb × Attachment—that is, we AGGREGATE, or average, across items. Correspondingly, in the analysis over items, we aggregate across subjects.
> dat <- read.table("rohde-levy-kehler-expt2-results.txt", 
+   header=T,sep="\t",quote=""")
> # by-subjects analysis
> sp.1.subj <- with(subset(dat,crit=="RC_VERB+2"), aggregate(list(rt=rt), 
+   list(subj=subj,verb=verb,attachment=attachment),mean))
> summary(aov(rt ~ verb * attachment + 
+   Error(subj/(verb*attachment)), sp.1.subj))

Error: subj
  Df  Sum Sq  Mean Sq F value Pr(>F)
Residuals 54 2570366  47599

Error: subj:verb
  Df  Sum Sq  Mean Sq F value  Pr(>F)
verb  1   4307   4307  0.5912  0.4453
Residuals 54  393380  7285

Error: subj:attachment
  Df  Sum Sq  Mean Sq F value  Pr(>F)
attachment 1   4664   4664  0.5072  0.4794
Residuals 54  496607  9196

Error: subj:verb:attachment
  Df Sum Sq Mean Sq F value  Pr(>F)
verb:attachment 1  44255  44255  5.9171 0.01834 *
Residuals 54  403877  7479
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> # by-items analysis
> sp.1.item <- with(subset(dat,crit=="RC_VERB+2"), aggregate(list(rt=rt), 
+   list(item=item,verb=verb,attachment=attachment),mean))
> summary(aov(rt ~ verb * attachment + 
+   Error(item/(verb*attachment)), sp.1.item))

Error: item
  Df  Sum Sq  Mean Sq F value  Pr(>F)
Residuals 19  268455  14129

Error: item:verb
  Df  Sum Sq  Mean Sq F value  Pr(>F)
verb  1   1395   1395  0.4190  0.5252
Residuals 19  63264  3330
We see that the interaction between verb and attachment is significant \((p < 0.05)\) in both the by-subjects and by-items ANOVAs. Thus we conclude with some degree of confidence that IC verbs do indeed facilitate processing of high-attaching RCs. Read Rohde, Levy, and Kehler (2008) for more details.

## 3 Speaker choice in lexicosyntactic variation

Doyle and Levy (2008) conducted a corpus study of the lexicosyntactic alternation between the forms, for any verb \(V\),

\[
\text{needs } V\text{\text{-}ing} \sim \text{needs to be } V\text{e}.
\]

Here are some examples from the British National Corpus (BNC):

(5) The wine needs re-corking every twenty years.
(6) I’m not sure I need reminding of this trip.
(7) A special license needs to be obtained before animals can be brought into the United Kingdom.
(8) This question would need to be asked in a more specific way.

It is not clear that there are clear-cut, categorical syntactic or semantic conditions that determine when one form or the other can be used. Instead, we proceeded under the hypothesis that there are many factors that non-categorically favor one usage or the other, and that these factors, distributed together in natural use, lead to a gradient of preferences between the to be form and the -ing form. We attempted to draw inferences about what some of these factors are by constructing a probabilistic model of speaker choice in this alternation, based on 1,004 annotated examples from the BNC. Two specific hypotheses we explored are (a) that individual verbs have idiosyncratic biases favoring one variant or the other; and (b) that larger amounts of postverbal material dependent on \(V\) would more strongly favor the to be
form. We construct a **Mixed-Effects Logistic Regression** model of the choices made by speakers/writers in this dataset, with individual verb preferences treated as a **random effect** (to be discussed in detail later on), and, for pedagogical purposes, including the animacy of the subject (*wine* in (9), for example) as a control factor:

```r
> dat <- read.table("needs.txt", header = T, quote = "", sep = "\t")
> head(dat)

<table>
<thead>
<tr>
<th>Verb</th>
<th>Anim</th>
<th>Dep.Length</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>do</td>
<td>abst</td>
<td>2</td>
<td>ing</td>
</tr>
<tr>
<td>pull</td>
<td>inan</td>
<td>2</td>
<td>ing</td>
</tr>
<tr>
<td>thin</td>
<td>inan</td>
<td>0</td>
<td>ing</td>
</tr>
<tr>
<td>cook</td>
<td>inan</td>
<td>0</td>
<td>ing</td>
</tr>
<tr>
<td>sort</td>
<td>abst</td>
<td>0</td>
<td>ing</td>
</tr>
<tr>
<td>lock</td>
<td>anim</td>
<td>0</td>
<td>ing</td>
</tr>
</tbody>
</table>

> summary(dat)

<table>
<thead>
<tr>
<th>Verb</th>
<th>Anim</th>
<th>Dep.Length</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>do</td>
<td>abst</td>
<td>427</td>
<td>ing</td>
</tr>
<tr>
<td>replace</td>
<td>inan</td>
<td>309</td>
<td>tobe</td>
</tr>
<tr>
<td>remind</td>
<td>anim</td>
<td>197</td>
<td></td>
</tr>
<tr>
<td>take</td>
<td>inan?</td>
<td>26</td>
<td>Mean</td>
</tr>
<tr>
<td>make</td>
<td>anim+met</td>
<td>20</td>
<td>Max.</td>
</tr>
<tr>
<td>look</td>
<td>abst?</td>
<td>13</td>
<td>39.000</td>
</tr>
<tr>
<td>(Other):816</td>
<td>(Other):</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

> library(lme4)
> dat.lmer <- lmer(Response == "tobe" ~ log(1 + Dep.Length) + Anim +  + (1 | Verb), dat, family = "binomial")
> print(dat.lmer, corr = FALSE)

**Generalized linear mixed model fit by the Laplace approximation**

**Formula:** Response == "tobe" ~ log(1 + Dep.Length) + Anim + (1 | Verb)

**Data:** dat

**AIC** 1165  **BIC** 1219  **logLik** -571.5  **deviance** 1143

**Random effects:**

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>(Intercept)</td>
<td>1.0453</td>
<td>1.0224</td>
</tr>
</tbody>
</table>

**Number of obs:** 1004, **groups:** Verb, 393

**Fixed effects:**

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|---------|

---

*Linguistics 251 lecture 1 notes, page 5*  
Roger Levy, Fall 2008
The output of the model shows that more postnominal material is significantly associated with a preference for the *to be* form. In addition, we can get estimates of the effects of individual verbs (larger positive values indicate stronger preference for *to be*); the following displays the verbs with strongest preferences, and the directions of those preferences:

```r
> verb.effects <- ranef(dat.lmer)[[1]]
> verbs <- row.names(verb.effects)
> idx <- order(abs(verb.effects[, 1]), decreasing = T)[1:10]
> eff <- verb.effects[idx, 1]
> names(eff) <- verbs[idx]
> eff

make see give cut consider explain clean examine  
1.654616 1.354957 1.345043 -1.342962 1.287825 -1.264927 -1.236851 1.223335  
inform remind
1.200134 -1.148297
```

In this model, *make, see,* and *give* have idiosyncratic preferences for the *to be* form, whereas *cut, clean,* and *remind* have preferences for the *ing* form. It’s an open question, of course, whether these apparent preferences reflect correlations with other linguistic variables that ultimate drive speaker choice.

### 4 Getting from here to there

There was quite a bit of detail and technical know-how—not only theoretical and mathematical, but also R programming hackery—in the model-building and interpretation exercises in the preceding two sections. In order to get you to the point where you can execute these kinds of analyses with confidence, we’re going to start out with the basic foundations of probability theory and statistics, and move step by step toward practical model building and analysis for the kinds of linguistic data sets that you are likely to encounter. The first part of the course, starting in the remainder of the lecture today, is therefore a brief introduction to probability theory.
5 What are probabilities?

There are two basic schools of thought as to the philosophical status of probabilities. One school of thought, the frequentist school, considers the probability of an event to be its asymptotic frequency over an arbitrarily large number of repeated trials. For example, to say that the probability of a toss of a fair coin landing as Heads is 0.5 (ignoring the possibility that the coin lands on its edge) means to a frequentist that if you tossed the coin many, many times, the proportion of Heads outcomes would approach 50%.

The second, Bayesian school of thought considers the probability of an event \( E \) to be a principled measure of the strength of one’s belief that \( E \) will result. For a Bayesian, to say that \( P(\text{Heads}) \) for a fair coin is 0.5 (and thus equal to \( P(\text{Tails}) \)) is to say that you believe that Heads and Tails are equally likely outcomes if you flip the coin.

The debate between these interpretations of probability rages, and we’re not going to try and resolve it in this class. Fortunately, for the cases in which it makes sense to talk about both reasonable belief and asymptotic frequency, it’s been proven that the two schools of thought lead to the same rules of probability. If you’re further interested in this, I encourage you to read Cox (1946), a beautiful, short paper.

6 Sample Spaces

The underlying foundation of any probability distribution is the sample space—a set of possible outcomes, conventionally denoted \( \Omega \). For example, if you toss two coins, the event space is

\[
\Omega = \{ hh, ht, th, hh \}
\]

where \( h \) is Heads and \( t \) is Tails. Sample spaces can be finite, countably infinite (e.g., the set of integers), or uncountably infinite (e.g., the set of real numbers).

7 Events and probability spaces

An event is simply a subset of a sample space.

What is the sample space corresponding to the roll of a single six-sided die? What is the event that the die roll comes up even?

It follows that the negation of an event \( E \) (that is, \( E \) not happening) is simply \( \Omega - E \). A probability space \( P \) on \( \Omega \) is a function from events in \( \Omega \) to real numbers such that the following three properties hold:

1. \( P(\Omega) = 1 \).
2. \( P(E) \geq 0 \) for all \( E \subset \Omega \).
3. If \( E_1 \) and \( E_2 \) are disjoint, then \( P(E_1 \cup E_2) = P(E_1) + P(E_2) \).
8 Conditional Probability and Independence

We’ll use an example to illustrate conditional independence. In Old English, the object in a transitive sentence could appear either preverbally or postverbally. Suppose that among transitive sentences in a corpus, the frequency distribution of object position and pronominality is as follows:

\[
\begin{array}{c|cc}
\text{Object} & \text{Pronoun} & \text{Not Pronoun} \\
\hline
\text{Preverbal} & 0.224 & 0.655 \\
\text{Postverbal} & 0.014 & 0.107 \\
\end{array}
\]

Let’s interpret these frequencies as probabilities. What is the conditional probability of pronominality given that an object is postverbal?

The conditional probability of event \(B\) given that \(A\) has occurred/is known is defined as follows:

\[
P(B | A) \overset{\text{def}}{=} \frac{P(A \cap B)}{P(A)}
\]

In our case, event \(A\) is \textbf{Postverbal}, and \(B\) is \textbf{Pronoun}. The quantity \(P(A \cap B)\) is already listed explicitly in the lower-right cell of table (9): 0.014. We now need the quantity \(P(A)\). For this we need to calculate the MARGINAL TOTAL of row 2 of Table (9): 0.014 + 0.107 = 0.121. We can then calculate:

\[
P(\text{Pronoun} | \text{Postverbal}) = \frac{0.014}{0.014 + 0.107} = 0.116
\]

8.1 (Conditional) Independence

Events \(A\) and \(B\) are said to be CONDITIONALLY INDEPENDENT GIVEN \(C\) if

\[
P(A \cap B | C) = P(A | C)P(B | C)
\]
A more philosophical way of interpreting conditional independence is that if we are in the state of knowledge denoted by $C$, then conditional independence of $A$ and $B$ means that knowing $A$ tells us nothing more about the probability of $B$, and vice versa. You’ll also see the term we are in the state of “not knowing anything at all” ($C = \emptyset$) then we would simply say in this case that $A$ and $B$ are CONDITIONALLY INDEPENDENT.

It’s crucial to keep in mind that if $A$ and $B$ are conditionally independent given $C$, that does not guarantee they will be conditionally independent given some other set of knowledge $C'$.

9 Random Variables

Technically, a random variable $X$ is a function from $\Omega$ to the set of real numbers ($\mathbb{R}$). You can think of a random variable as an “experiment” whose outcome is not known in advance. In fact, the outcome of a random variable is a technical term simply meaning which number resulted from the “experiment”.

The relationship between the sample space $\Omega$, a probability space $P$ on $\Omega$, and a random variable $X$ on $\Omega$ can be a bit subtle so I’ll explain it intuitively, and also with an example. In many cases you can think of a random variable as a “partitioning” of the sample space into the distinct classes of events that you (as a researcher, or as a person in everyday life) care about. For example, suppose you are trying to determine whether a particular coin is fair. A natural thing to do is to flip it many times and see how many times it comes out heads. Suppose you decide to flip it eight times. The sample space $\Omega$ is then all possible sequences of length eight whose members are either H or T. The coin being fair corresponds to the probability space $P$ in which each point in the sample space has equal probability $\frac{1}{2^8}$. Now suppose you go ahead and flip the coin eight times, and the outcome is

$$\text{T T T T T T H T}$$

Intuitively, this is a surprising result. But under $P$, all points in $\Omega$ are equiprobable, so there is nothing about the result that is particularly surprising.

The key here is that you as an investigator of the coin’s fairness are not interested in the particular H/T sequence that resulted. You are interested in how many of the tosses came up heads. This quantity is your random variable of interest—let’s call it $X$. The logically possible outcomes of $X$ are the integers $\{0, 1, \ldots, 8\}$. The actual outcome was $X = 1$—and there were seven other possible points in $\Omega$ for which $X = 1$ would be the outcome! We can use this to calculate the probability of this outcome of our random variable under the hypothesis that the coin is fair:

$$P(X = 1) = \frac{1}{2^8} \times 8 = \frac{1}{2^5} = 0.03125$$

So seven tails out of eight is a pretty surprising result. Incidentally, there’s a very interesting recent paper (Griffiths & Tenenbaum, 2007) that deals with how humans actually do ascribe surprise to a “rare” event like a long sequence of heads in a coin flip.
References


