Statistics and Linguistic Applications
Hale
April 5, 2010

Proportion test

In Chapter 3 of his book, Baayen reports word frequencies from CELEX. These are based on a corpus sample of 18580121 tokens.

word	frequency	relative frequency
the	1093547	0.5885575
president	2469	0.000132884
hare	153	0.00000823
harpsichord	15	0.00000086

Even though we know better, lets for the moment conceptualize the observation of a word like "president" as a SUCCESS and the observation of any other word as a FAILURE. Our probability model is thus a Binomial with parameter $p = 0.000133$.

We have seen (e.g. page 30 of Vasishth's notes) that when the corpus size n and the success probability p are not too close to zero, the Binomial distribution is closely approximated[1] by a Normal distribution with mean np and variance npq where the failure probability $q = (1 - p)$.

We can now view other corpus samples as results from a kind of language-production experiment. From such a sample we can compute a statistic, the *sample proportion*, and look up how probable this statistic's value is under the assumed parameterization. Is the 1-million word Brown corpus, with 382 attestations of "president" a wacky or run-of-the-mill sample? Across many many corpora, what fraction would attest "president" that many times if the parameter were really $p = 0.000133$? Let us compute the standardized score and make a judgment. A standardized score looks like this

$$Z = \frac{x - \mu}{\sigma}$$

where x is the value to be standardized, σ is the (population) standard deviation and μ the (population) mean. In this case x is our observed proportion, $\hat{p}$, and we know σ, μ in virtue of approximating the Binomial with the Normal. The *proportional* standard deviation σ_p of the Binomial is $\sqrt{pq/n}$. The kind of decreasing proportional variability as sample size goes up is suggested in Figure 2.6 on page 19 of the Vasishth notes. Page 152 of Johnson writes out this formula and graphs the interrelationship of p, N and the denominator σ_p.

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

Multiplying by a form of 1 translates the proportion-based Z score into one based on an absolute number of successes. Define the success-count x in terms of the proportion of success such that $\hat{p} = x/n$.

$$\begin{aligned} Z &= \frac{\hat{p} - p}{\sqrt{pq/n}} \\ &= \frac{\hat{p} - p}{\sqrt{pq/n}} \times \frac{n}{n} \\ &= \frac{x - np}{\sqrt{pq/n} \times n} \\ &= \frac{x - np}{\sqrt{pq}/\sqrt{n} \times (\sqrt{n} \cdot \sqrt{n})} \\ &= \frac{x - np}{\sqrt{pq} \cdot \sqrt{n}} \\ &= \frac{x - np}{\sqrt{npq}} \end{aligned} \tag{1}$$

[1] A derivation of the Binomial approximation to the Normal is given on the Mathworld Binomial page

The denominator now shows exactly the standard deviation of the Normal approximation to the Binomial. In our Brown corpus example,

$$Z = \frac{382 - (0.000133 \times 1,000,000)}{\sqrt{1,000,000 \times 0.000133 \times (1 - 0.000133)}}$$

```
> success <- 0.000133
> failure <- 1 - success
> n <- 1e+06
> (382 - (success * n))/sqrt(n * success * failure)

[1] 21.59247
```

Wow! a Z-score of 21.59! That's twenty one standard deviations above the mean. What's the probability of getting a sample *that* extreme or more? Lets "look it up in our table."

```
> 1 - pnorm(21.59)

[1] 0
```

382 attestations in the Brown corpus is highly unlikely under the null hypothesis that "president" appears 0.000133 of the time in all corpora. Baayen dryly remarks, "The resulting probability is indistinguishable from zero given machine precision and provides ample reason for surprise." He gets to this same conclusion via a different route, calculating the Binomial probabilities directly with `pbinom`. In fact, the Normal was originally introduced by de Moivre as a way of approximating the Binomial (computers were expensive in 1733). Moreover, the Normal approximation leads us far beyond mere success/failure proportions as we shall see.

The multinomial

In the "president" case there were only two possible outcomes, identified with SUCCESS and FAILURE to produce "president" respectively. In a particular corpus sample $x = n\hat{p}$ is often called the observed frequency of success as opposed to failure. The expected frequency (of success) is np. The more general Multinomial distribution describes k different categories of events $A_1, A_2, \ldots, A_k$ with probabilities $p_1, p_2, \ldots p_k$. But the notions of "observed" and "expected" are the same.

If we draw a sample of size n from a Multinomial population, the observed frequencies for the events $A_1, \ldots, A_k$ can be described by random variables $X_1, \ldots, X_k$ whose specific values $x_1, x_2, \ldots, x_k$ would be the observed frequencies in the sample. The expected frequencies would just be $np_1, np_2, \ldots, np_k$.

Event	A_1	A_2	$\cdots$	A_k
Observed Frequency	x_1	x_2	$\cdots$	x_k
Expected Frequency	np_1	np_2	$\cdots$	np_k

Table 1: Multinomial assigns probability to k kinds of events

As an example of a Multinomial, consider bags of M&Ms. How many of each color (blue, brown, green, ...) are there in a bag of 30? The count of one affects the others, if 28 are red then none of the other colors can claim more than 2 of the candies for their own color.

$$P(X_1 = x_1, \ldots, X_k = x_k) = \binom{n}{x_1}\binom{n - x_1}{x_2} \cdots \binom{n - x_1 - x_2 - \cdots - x_{k-1}}{x_k} p_1 p_2 \cdots p_k$$

In the one-proportion Z test statistic (equation 1), there are exactly two outcomes. Viewed as a special case of the Multinomial, we might think of them as just the first two of potentially many more outcomes.

The square of the Z-score

What if we wanted to generalized beyond SUCCESS and FAILURE? Consider the square of the Z score

$$Z^2 = \frac{(x - np)^2}{npq}$$

To prepare notationally for a larger set of k event categories, rename the success count X_1 and the failure count X_2. In the two-event case, we have $X_1 + X_2 = n$ analogous to $p + q = 1$.

$$\begin{aligned} X_1 - np &= X_1 - np \\ &= (n - X_2) - np \\ &= n - X_2 - n(1 - q) \\ &= (-1)(X_2 - nq) \end{aligned}$$

and so the squares of these quantities should stand in the relationship

$$\begin{aligned} (X_1 - np)^2 &= ((-1)(X_2 - nq))^2 \\ (X_1 - np)^2 &= (X_2 - nq)^2 \end{aligned}$$

These equalities make it possible to rewrite the Z score as a mix of two addends.

$$Z^2 = \frac{(x - np)^2}{npq} = \frac{(X_1 - np)^2}{np} + \frac{(X_2 - nq)^2}{nq} \tag{2}$$

The denominators in equation 2 are the expected frequency of success and expected frequency of failure, respectively. The numerators represent the discrepancy between observed and expected counts. Generalizing this to the Multinomial

$$\frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} + \cdots + \frac{(X_k - np_k)^2}{np_k} = \sum_{j=1}^{k} \frac{(X_j - np_j)^2}{np_j}$$

gives a test statistic, χ^2 whose distribution, the sum of squares of Normal deviates, is known.

$$\chi^2 = \sum_{j=1}^{k} \frac{(X_j - np_j)^2}{np_j} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \tag{3}$$

The chi-square statistic has $k - 1$ degrees of freedom if the expected frequencies can be computed without having to estimate the population parameters from statistics. The degrees of freedom in this case is one less than the number of probabilities, reflecting the constraint that they must add up to 1. If it is necessary to estimate m population parameters to specify the null hypothesis, then the statistic follows a χ^2 distribution with $k - 1 - m$ degrees of freedom.

Example: letter probabilities

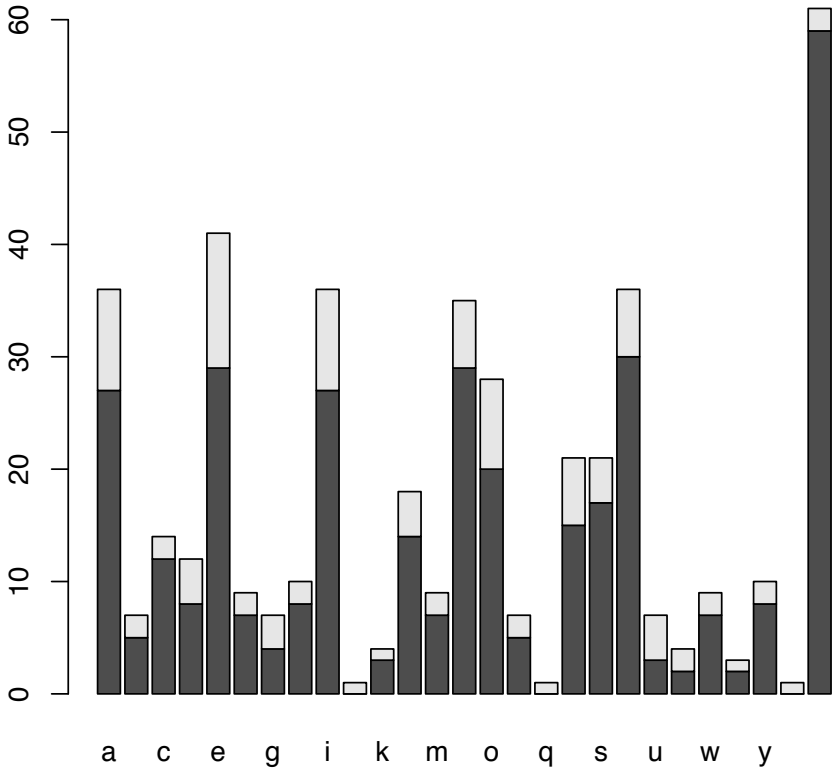
If you are a computer at Fort Meade in Maryland, your job might well be to guess which language a sample of text is from. A simple way to do this is to ask whether the letter-distribution is very different from what we would expect about language $L_1, L_2, \ldots$. Verzani's book includes official frequencies for English as specified in the Scrabble board game.

Lets take a look at the overall letter distribution in this quote from the New York Times.

> The Democratic presidential contest between Senator Hillary Rodham Clinton and Senator Barack Obama in Pennsylvania today will offer a new test of what exactly a win is. There are many potential different outcomes, and you can be sure the campaigns will be pointing to all kinds of things in trying to claim victory in this first contest in six weeks.

```
> quotevector <- unlist(strsplit(tolower(quote), ""))
> letter.dist <- sapply(c(letters, " "), function(x) sum(quotevector == x))

> barplot(rbind(letter.dist, scrabble$freq))
```



And in particular, the signature vowel distribution that Scrabble predicts.

```
> scrabblevowels <- scrabble$freq[scrabble$piece == "A" | scrabble$piece == "E" | scrabble$piece ==
+     "I" | scrabble$piece == "O" | scrabble$piece == "U"]
> quotevowels <- sapply(c("a", "e", "i", "o", "u"), function(x) sum(quotevector == x))
> chisq.test(x = quotevowels, p = scrabblevowels/sum(scrabblevowels))

        Chi-squared test for given probabilities

data:  quotevowels
X-squared = 6.6604, df = 4, p-value = 0.1550
```

The null hypothesis cannot be rejected on the basis of a sample that would be expected fifteen percent of the time. The Multinomial Scrabble model fits just fine.

The chi-square doesn't care which distribution you think generated the data; it is *non-parameteric*. If you can work out the expected frequencies, you can calculate the goodness of fit and then see how far out you are on the χ^2 distribution. In this way one may ascertain how well the entire sample (not just its mean) looks like it came from a particular distribution.

Contingency Tables

From the perspective of a Multinomial over k different event types, everything is a $1 \times k$ ("one-by-k)") table. If we extend into the second dimension we have an $n_r \times n_c$ *contingency* table. Frequently in linguistics we cross-classify attestations of a certain sound, word etc in two or more ways — these are contingency tables. Even though these data are arranged in a square-shaped table, we can still ask whether the table as a whole has a large discrepancy as compared to some expected values. To do this, compute equation 3 over the $n_r n_c$ cells in the table and compare the obtained χ^2 statistic to a chi-square distribution with particular degrees of freedom:

$(n_r - 1)(n_c - 1)$	if the expected frequencies can be computed without having to estimate population parameters from sample statistics
$(n_r - 1)(n_c - 1) - m$	if the expected frequencies can be computed only by estimating m population parameters from sample statistics

One fascinating hypothesis is that the *column* variables are probabilistically independent from the *row* variables. Remember, if two random variables are independent, then their joint distribution is the product of their individual distributions.

$$\mathcal{H}_0 : p_{ij} = p_i^r p_j^c \qquad \mathcal{H}_1\text{: the } p_{ij} \text{ are not independent}$$

For example, Cooper and Hale (2004) examined ah, you know, disfluencies in the Switchboard corpus, a sample of spoken English. Looking at pairs of conjoined constituents affectionately known as "lobes", they tabulated whether or not each one contains any disfluency.

		Lobe 2	
		Disfluent	Fluent
Lobe 1	Disfluent	150	126
	(Expected)	(99.2)	(176.8)
	% of total	18.3%	15.3%
	Fluent	145	400
	(Expected)	(195.8)	(349.2)
	% of total	17.7%	48.7%

N=821, df=1
χ^2=61.25
p<.001

Table 2: Disfluency status of conjoined lobes, obtained with `tgrep2`. A significant distribution.

To work out what we expect under the null hypothesis, consider the marginals. Ignoring Lobe2 for a moment, there are 150+126/N observations where Lobe1 is disfluent. Call $p_{1disf} = 0.336$. Of course, avoiding disfluency is all there is to fluency, so $p_{1fluent} = 1 - p_{1disf}$. Likewise we have $p_{2disf} = (145+150)/812 = 0.359$. Under $\mathcal{H}_0$, we should see $N p_{1disf} p_{2disf} = 99.17$ in the upper-left cell of the contingency table (these expected values are pre-parenthesized for you). But in fact all the squared deviations divided by the expected number add up over sixty. It is highly improbably that Cooper and Hale would have observed this pattern if disfluency the first constituent had no influence on disfluency in the second.

Arranging for R to calculate your chi-squared test of independence

Lets borrow an example from D.G. Altman 'Practical Statistics for Medical Research'. As quoted in Dalgaard, this data concerns caffeine consumption among women giving birth. The women are classified by marital status.

```
> caff.marital <- matrix(c(652, 1537, 598, 242, 36, 46, 38, 21, 218, 327, 106, 67), nrow = 3,
+     byrow = T)
> colnames(caff.marital) <- c("0", "1-150", "151-300", ">300")
> rownames(caff.marital) <- c("Married", "Prev.married", "Single")
> caff.marital

              0 1-150 151-300 >300
Married     652  1537     598  242
```

```
Prev.married  36    46      38   21
Single       218   327     106   67
```

```
> chisq.test(caff.marital)

        Pearson's Chi-squared test

data:  caff.marital
X-squared = 51.6556, df = 6, p-value = 2.187e-09
```

Marital status and caffeine consumption are not independent! But in what ways do they deviate from independence? We can work this out using some handy information that `chisq.test` hands back.

```
> cm <- chisq.test(caff.marital)
> E <- cm$expected
> O <- cm$observed
> (O - E)^2/E
```

```
                     0    1-150   151-300      >300
Married       4.1055981 1.612783 0.6874502 0.8858331
Prev.married  0.3007537 7.815444 4.5713926 6.8171090
Single       15.3563704 1.875645 7.0249243 0.6023355
```

The result shows the contribution of each cell to the overall χ^2. A lot of single mothers just don't consume any caffeine. The previously-married are shifted in the direction of greater consumption.