Hypothesis Testing

A *statistical hypothesis* is a statement about the nature of the distribution of a random variable. How can such a hypothesis be tested?

Consider the genetic relationships among certain Indo-European Languages. Ross (1950) constructed a table as follows: Allot a column to each branch language and a row to each of a certain set of attested Indo-European roots; if a root appears in a particular branch language, put an 'x' in the appropriate cell of the table.

Root-number	L_1	L_2	L_3	...
1	x	x		
2		x	x	
3	x	x	x	
:				

Ross suggests that the question "Is L_i closely related to L_j?" is equivalent to the question 'Given the number of crosses in the *ith* and *jth* columns, what is the probability of obtaining the given number (or a greater number) of cases of a row with a cross in each of the two columns if the crosses were placed in the two columns at random?' If this probability is sufficiently small, we might be tempted to infer that the two languages have some casual (genetic) relationship.

Let N be the number of rows in the table (number of roots under consideration), let n_i be the number of rows which contain a cross in column *i,* let n_j be the number of rows which contain a cross in column *j,* and let r be the number of rows in which both the *ith* and *jth* columns are marked with a cross.

Our problem is to compute the probability that r agreeing rows occur by chance, given that n_i and n_j entries in column i and column j respectively are marked with a cross.

We proceed as follows. We first determine the number of ways of marking n_i entries in column i and n_j entries in column j with exactly r agreements. The number of ways that n_i crosses can be placed in the *ith* column is

$$\binom{N}{n_i} = \frac{N!}{n_i!(N-n_i)!}$$

which is the number of ways of selecting a subset of n_i objects from a set containing N objects.

Once these have been chosen, we put r crosses in the *jth* column on r of the n_i rows already marked in column *i.* This can be done in

$$\binom{ni}{r}$$

ways. Finally we mark at random the N-n_i rows which do not contain a cross in the *ith* column with the n_i-r remaining crosses allotted to the *jth* column. This can be accomplished in

$$\binom{N-n_i}{n_j-r}$$

ways. Thus the entire operation can be done in

$$\binom{N}{n_i}\cdot\binom{n_i}{r}\cdot\binom{N-n_i}{n_j-r}$$

different ways. Now we must find the number of ways n_i crosses can be placed in column i and n_i in column j without any restriction on agreements. Since n_i of the entries of column i can be marked with a cross in

$$\binom{N}{n_i}$$

ways and nj of the entries of column j can be marked with a cross in

$$\binom{N}{n_j}$$

ways, the total number of ways both columns can be marked is

$$\binom{N}{n_i}\cdot\binom{N}{n_j}$$

We let the set of outcomes be the set of all these possible ways of marking the two columns with crosses. Since each way is as likely to occur as any other, the probability of any one such marking is

$$\frac{1}{\binom{N}{n_i}\cdot\binom{N}{n_j}}$$

and since, as we have seen,

$$\binom{N}{n_i}\cdot\binom{n_i}{r}\cdot\binom{N-n_i}{n_j-r}$$

cases are favourable to the event of r agreeing markings, the probability p_r of r agreements is

$$p_r = \left[\binom{N}{n_i}\cdot\binom{n_i}{r}\cdot\binom{N-n_i}{n_j-r}\right] \Big/ \left[\binom{N}{n_i}\cdot\binom{N}{n_j}\right]$$

or

$$p_r = \left[\binom{n_i}{r} \cdot \binom{N-n_i}{n_j-r}\right] \Big/ \left[\binom{N}{n_j}\right]$$

Suppose that we fine s coincident crosses when we consider the actual case of Indo-European roots in L_i and L_j. The probability of at least that many occurring by chance is

$$P(m \geq s) = \sum_{q=s}^{k} p_q$$

where k is the smaller of the two numbers n_i and n_i. If P(mŠs) is small, say 0.05, this means that the probability of obtaining at least as many as s coincident roots by chance is 0.05. Since the probability of obtaining s or more common roots by chance is small indeed, one is tempted to reject the hypothesis of random root phenomenon. Thus it would appear that some causal factor is at work. If, on the other hand, P(mŠs) turned out to be, say 0.5, we would not be tempted to look for causal factors, for in this case there is a 50-50 chance that s or more roots coincide. This sort of argument is basic to *hypothesis testing*, which is where we turn our discussion to now.

Introduction

Ross' argument to the effect that the hypothesis that two languages L_1 and L_2, say, are not genetically related is equivalent to the following hypothesis:

(**H**) If, in a set of N Indo-European roots, n_1 have a cognate in L_1 and n_2 a cognate in L_2, then the distribution of the random variable R equal to the number of roots with cognates in both languages is governed by chance alone.

We showed that under this hypothesis R has the hyper-geometric distribution, that is

$$P(R = r) = \left[\binom{n_2}{r} \cdot \binom{N-n_2}{n_1-r}\right] \Big/ \left[\binom{N}{n_1}\right]$$

To test this hypothesis, we can actually list the N Indo-European roots, find n_1, n_2, and r, and finally compute *P(RŠr),* the probability of obtaining at least r cognates by chance. If this probability is small, we would tend to reject **H** and assume that L_1 and L_2 are related. If *P(RŠr),* were not small, then it might be possible that this number of cognates could have occurred by chance, and no case could be made on probabilistic grounds for the two languages to be any more closely related than any other pair of Indo-European languages. [This is a special case of Fisher's exact case (test for independence) and is a very commonly used test].

A few observations about the situation above may prove helpful. Two kinds of error can be made by a researcher: (1) he may reject the hypothesis when it is in fact true; and (2) he may accept the hypothesis as being true when it is indeed false. Let us see what we can do to control errors of the first kind.[1]

For our example, we may reject the hypothesis **H** that there is no genetic relationship between L_1 and L_2 when there is indeed no such relationship. This is an error of type 1. On the other hand we may accept the hypothesis of no genetic relationship when it is false; this is an error of the second kind or a type 2 error. Of course accepting **H** when it is true and rejecting **H** when it is false present no problem.

In our example we tend to reject the hypothesis of chance occurrence if the number of shared cognate roots is high.If we were to choose a *critical value c* and reject the hypothesis whenever the observed value *r* of *R* is greater than or equal to *c,* then the probability of making a type 1 error is

$$P(R \geq c) = \sum_{i \geq c} P(R = i) \qquad [1]$$

= probability of obtaining a value of *R* as high as *c* under the given hypothesis

This probability $P(R \check{S} c) = \alpha$ of rejecting hypothesis **H** when it is true can be controlled by first choosing α, the *critical level,* and then finding *c* so that $P(R \check{S} c) = \alpha$. Generally α is chosen to be small when type 1 errors are especially hazardous; critical levels like 0.05, 0.01, and 0.005 are common, so that for example, if

$$P(R \geq c_{0.01}) = 0.01 \qquad [2]$$

then rejecting the hypothesis at $r \check{S} c_{0.01}$ renders the probability of a type 1 error at 0.01. However, if the observed value *r* of *R* were $r < c_{0.01}$, this would not necessarily mean that we could automatically assume **H** to be valid. If indeed $P(R \check{S} r) = 0.05$ this would mean that the probability of obtaining an observed number of cognates at least as high as *r* is quite small, 0.05, but not as small as the critical level, 0.01.

There is a distinct asymmetry to the situation just described. The hypothesis **H** is an assertion that there is nothing to the claim that L_1 and L_2 are related. Thus in employing $c_{0.01}$ in expression [2], if we obtain an observed $r \check{S} c_{0.01}$ and reject the hypothesis, we are claiming that L_1 and L_2 are related, with a probability of 0.01 that we are rejecting the hypothesis when it is true. On the other hand, if $r < c_{0.01}$ we are not in a position to accept **H** without incurring the possibility of an error of type 2.

Hence in order to be able to make a certain claim **C** about the values of some random variable *X,* we try to arrange to make the hypothesis **H** that there is nothing to **C** and then proceed to test **H** with an eye to rejecting it, and hence expousing **C**, if the value observed for *X* is sufficiently improbable. For this reason **H** is often called a *null hypothesis* because, as indicated above, it usually amounts to a denial of some particular claim **C**. A null hypothesis **H** concerning a random vari-

[1] Type 2 errors are especially problematic for linguistics researchers for two reasons. First, s/he has less control over them than s/he does over type 1 errors and secondly, s/he is often in the position where being able to accept the statistical hypothesis yields positive scientific results. Humanities researchers guard against this situation by forming hypothesis in such a way that rejection yields positive results, i.e., in this case indicates a genetic relationship between the two languages.

able X is usually of the form: X has some particular distribution F_X, the *null distribution*, from which such probabilities as

$$P(X \geq c) \quad [3]$$

$$P(|X| \geq c) \quad [4]$$

$$P(|X - E(X)| \geq c) \quad [5]$$

can be computed. Then a critical or significance level is chosen and, depending on the requirements of the problem, a critical value c is obtained from [3], [4], and [5], etc., so that the probability in question equals α, c is the *critical value* and α is the corresponding *level of significance* or *critical level*.

Thus a *hypothesis test* consists of making the null hypothesis about the distribution of X, choosing a level of significance α, choosing a probability configuration like [3], [4], or [5], or something else appropriate for a *critical region,* obtaining a critical value c corresponding to α for that particular critical region, and rejecting or accepting the hypothesis according as the observed value of X lies in the critical region or not. If the critical region is $X \check{S}\, c$ (as in [3]) or $X \eth c$, the test is called a *one-tailed test*. If, on the other hand, one wishes to use a critical region like $|X| \check{S} c$ or $|X\text{-}E(X)| \check{S} c$, the test is called a *two-tailed test*. The problem at hand determines which test is appropriate.

Example 1 Ross' scheme to test for significant relationships between languages is an example of an (upper) one-tailed hypothesis test. It is often complicated by the amount of computation required to obtain $P(R=r)$. For example, Ross cites in his table 6, that in N=1860 Indo-European roots there are n1=1184 Italo-Celtic cognates and n2=1165 Greek cognates, and of these r=783 are common cognates in both languages. This means that to find $P(R \check{S} r)$ in this case we must compute

$$P(R \geq 783) = \sum_{i \geq 783} \binom{1165}{i} \cdot \binom{1860 - 1165}{1184 - i} \Big/ \binom{1860}{1184}$$

which is a prodigious task.

Since the hyper-geometric distribution has been tabulated, in Lieberman and Owen (1961), for $N \eth 50$ and for certain selected values of $N > 50$, a careful choice of N may render Ross' test easier to use.

Example 2 In Hayden's study of American English (1950), there were 325 occurrences of |j| in 65,122 phonemes. If we assume that Robert's survey (1965) is over a sufficiently large corpus that his relative frequency 0.0036 of |j| is a good approximation of the probability of |j| and that both samples are random, then we can make the null hypothesis that both samples come from the same population for which the probability of selecting |j| is p=0.0036.

Hayden's sample can be taken as 65,122 Brownell trials with probability *p=0.0036* of obtaining |*j*| in each trial. Thus, if the random variable X_i (i=1,2,...,65,122) is defined to take the value 1 if the *ith* phoneme is |*j*| and 0 otherwise, then with

$$\bar{X} = \sum_{i=1}^{65,122} X_i = number \cdot of \cdot occurrences \cdot of \cdot |j| \cdot in \cdot a \cdot 65,122 \cdot phoneme \cdot running \cdot text$$

the null hypothesis can be stated as follows: X has a binomial distribution with n=65,122 and p=0.0036, that is

$$P(\bar{X} = k) = \binom{65,122}{k} \cdot 0.0036^k \cdot 0.9964^{65,122-k}$$

Since *npŠ5*, the approximation of this binomial distribution by the normal distribution is adequate. Using this normal approximation, we can write

$$P(\bar{X} \le k) = N(k, np, \sqrt{npq}) = N(k, 234.4, \sqrt{233.6})$$

which according to

$$N(x, \mu, \sigma) = N\left(\frac{x-\mu}{\sigma}, 0, 1\right) \qquad [A]$$

can be written

$$P(\bar{X} \le k) = N\left(\frac{k-234.4}{\sqrt{233.6}}, 0, 1\right)$$

Let us perform a two-tailed test by choosing the critical value c for some deviation of *X* from *E(X̄)* corresponding to some significance level α, that is, let us choose c such that

$$P(|\bar{X} - E(\bar{X})| \ge c) = \alpha \qquad [B]$$

The expression

$$(|\bar{X} - E(\bar{X})| \ge c) \qquad [C]$$

is equivalent to the disjunction of

$$E(\bar{X}) + c \le \bar{X} \qquad [D]$$

and

$$\bar{X} \le E(\bar{X}) - c \qquad [E]$$

that is, [C] holds if and only if either [D] or [E] holds. In the present case, the events [D] and [E] are disjoint events for c>0, so

$$P(|\bar{X} - E(\bar{X})| \ge c) = P(\bar{X} - E(\bar{X}) \ge c) + P(\bar{X} - (E(\bar{X}) \le -c)) \qquad [F]$$

$$= \quad 1 - (P(\overline{X} < E(\overline{X}) + c) + P(\overline{X} \le E(\overline{X}) - c))$$

$$= \quad 1 - (N(234.4 + c,\ 234.4,\ \sqrt{233.6}) + N(234.4 - c,\ 234.4,\ \sqrt{233.6}))$$

since $E(\overline{X}) = np = 234.4$ and $\sigma_{\overline{X}}^2 = 233.6$.

Using [A], we can write $P(|\overline{X}\text{-}E(\overline{X})|Šc)$ as follows:

$$P(|\overline{X} - E(\overline{X})| \ge c) \ = \ 1 - N\left(\frac{c}{\sqrt{233.6}},\ 0,\ 1\right) + N\left(\frac{-c}{\sqrt{233.6}},\ 0,\ 1\right) \qquad [G]$$

Because the distribution is symmetric with respect to the vertical coordinate axis, then N(-*x*,0,1)=1-N(*x*,0,1). Thus

$$P(|\overline{X} - E(\overline{X})| \ge c) \ = \ 2 \cdot \left[1 - N\left(\frac{c}{15.28},\ 0,\ 1\right)\right] \qquad [H]$$

If we set the significance level at 0.01, we must choose a *c* so that $P(|\overline{X}\text{-}E(\overline{X})|Šc)$=0.01. From a table of he normal distribution, we can obtain the result N(2.58,0,1)=0.995, so that the right side of [H] equals 0.01. Thus if *c*/15.28 = 2.58 or *c* = 15.28*2.58 = 39.42 in expression [H], we have $P(|\overline{X}\text{-}234.4|Š39.42) = 0.01$.

Hayden's sample yields the value 325 for $\overline{X}$ and hence $\overline{X}$-234.4=90.6, which is well beyond the critical value. Therefore in rejecting the hypothesis that $\overline{X}$ has binomial distribution with p=0.0036, we have a probability of less than 0.01 of being wrong, that is, of committing an error of type 1. Thus there is overwhelming evidence that $\overline{X}$ is not binomially distributed with probability $p=0.0036$. This is probably because, on the one hand, the X_i's may not be independent (that is, Hayden's sample may not be random) and, on the other, Hayden's sample is a very specialized one with perhaps a different value of *p*.

In general, when a null hypothesis takes the form the random variable $\overline{X}$ has normal distribution with mean μ and variance σ^2, then in terms of expression [5] the equation relating the critical value *c* to the level of significance α is

$$P(|X - E(X)| \ge c) \ = \ 2 \cdot \left[1 - N\left(\frac{c}{\sigma},\ 0,\ 1\right)\right] \ = \alpha$$

References

Ross, A.S.C. (1950). Philological probability problems. *J. Roy. Stat. Soc, ser. B, 12,* 19-59.

Lieberman, G.J., and Owen, D.B. (1961). *Tables of hypergeometric probability distribution* (Palo Alto: Stanford University Press).

Hayden, R.E. (1950). The relative frequency of phonemes in general-American English. *Word 6,* 217-223.

Roberts, A.H. (1965). *A Statistical Linguistic Analysis of American English* (The Hague: Mouton).