PROVING A PROPOSITION IN R.A. FISHER'S PREFATORY NOTE TO *THEORY OF STATISTICAL ESTIMATION*

HENRY BOTTOMLEY

Abstract. This is a proof of a prefatory proposition assumed by R.A. Fisher in his 1925 paper *Theory of Statistical Estimation* to justify his use of an infinite hypothetical population distributed in a definite manner.

1. Introduction

In 1925 R.A. Fisher [1] justified the use of an infinite hypothetical population distributed in a definite manner with these words in a prefatory note to his paper *Theory of Statistical Estimation*:

> Imagine a population of N individuals belonging to s classes, the number in each class k being $p_k N$. This population can be arranged in order in $N!$ ways. Let it be so arranged and let us call the first n individuals in each arrangement a sample of n. Neglecting the order within the sample, these samples can be classified into the several possible types of sample according to the number of individuals of each class which appear. Let this be done, and denote the proportion of samples which belong to type j by q_j, the number of types being t. Consider the following proposition.
>
> Given any series of proper fractions P_1, P_2, ..., P_s, such that $S(P_k) = 1$, and any series of positive numbers η_1, η_2, ..., η_t, however small, it is possible to find a series of proper fractions Q_1, Q_2, ..., Q_t, and a series of positive numbers ϵ_1, ϵ_2, ..., ϵ_s, and an integer N_0, such that, if $N > N_0$ and $|p_k - P_k| < \epsilon_k$ for all values of k, then will $|q_j - Q_j| < \eta_j$ for all values of j.

He went on to comment:

> I imagine it possible to provide a rigorous proof of this proposition, but I do not propose to do so. If it be true, we may evidently speak without ambiguity or lack of precision of an infinite population characterised by the proper fractions, P, in relation to the random sampling distributions of samples of a finite size n.

Since Fisher did not prove this at the time, I intend to do so here. The proof is not difficult, since it only requires finding an N_0 more than large enough to ensure that the proposition is satisfied.

Some clarifications might be worth making, since from personal experience the proposition can sometimes appear confusing and abbreviated.

Date: 4 February, 2010.

1991 *Mathematics Subject Classification.* Primary 62D05; Secondary 62A01.

Key words and phrases. Fisher, sampling, infinite, population.

First, $S(P_k) = 1$ simply means that the sum over k is 1, i.e. $\sum_{k=1}^{s} P_k = 1$.

Second, there are more possible types of samples than there are classes (except in the degenerate case $s = 1$ or when the sample size $n = 1$) and, for large enough N to make every potential type possible, we are looking at the number of compositions of n into s non-negative parts. In other words t is simply a function of s and n namely $t = \binom{s+n-1}{s-1}$. So when we are given s and t at the start of the proposition, we are also given the sample size n.

Third, the values of p_k have to be decided once N is known since each $p_k N$ must be an integer, so after Q_1, Q_2, …, Q_t, ϵ_1, ϵ_2, …, ϵ_s, and N_0 are chosen; the values of q_j are determined by those of p_k. In order, we are given the values of P_k and η_j and so implicitly n, then we have to choose the values of Q_j, ϵ_k and N_0, so that all permitted values of p_k and N provide satisfactory values of q_j.

Fourth, P_k and Q_j being proper fractions seems to mean they are real numbers in the interval $[0, 1]$; unlike p_k and q_j, they do not have to rational. In fact all we need is $P_k \geq 0$ for all k together with their sum over k being 1; any $P_k = 0$ can be ignored by setting the corresponding $p_k = 0$, while the degenerate case where $P_1 = 1$ is trivially true because with $p_1 = 1$ there would only be one type of sample of size n. The restrictions on Q_j resulting from q_j and η_j mean that no additional restriction is required.

Fifth, by the definition, we need $\sum_{k=1}^{s} p_k N = N$ as every member of the population is in one of the classes, or equivalently $\sum_{k=1}^{s} p_k = 1$. This will lead to $\sum_{j=1}^{t} q_j = 1$. Although the positive values of η_j mean there is no necessary requirement for $\sum_{j=1}^{t} Q_j = 1$, it would be sensible to choose values of Q_j which would work for sufficiently large N whatever the values of η_j are, and so which do sum to 1.

2. An example

To take a relatively simple example, suppose there are three classes so $s = 3$ and we are given $P_1 = 0.6$, $P_2 = 0.3$ and $P_3 = 0.1$; furthermore the sample size is to be $n = 2$ and the required accuracy is to be 0.001.

There are clearly six potential types of samples of size two: they are two from class 1, two from class 2, two from class 3, one from class 1 and one from class 2, one from class 1 and one from class 3, or one from class 2 and one from class 3. Since we aim to have the values of Q_j equal to the limiting values as N_0 and N increase and the η_j reduce, we will choose them as if we were drawing twice *with replacement* from an urn with ten balls (6 of class 1, 3 of class 2 and 1 of class 3) where the values corresponding to the respective types would be

$$\begin{aligned} Q_1 &= P_1^2 &&= 0.36 \\ Q_2 &= P_2^2 &&= 0.09 \\ Q_3 &= P_3^2 &&= 0.01 \\ Q_4 &= 2P_1P_2 &&= 0.36 \\ Q_5 &= 2P_1P_3 &&= 0.12 \\ Q_6 &= 2P_2P_3 &&= 0.06. \end{aligned}$$

These are not the proportions of the types of samples resulting from counting the first two elements of Fisher's permutations, or equivalently sampling *without replacement* from the same urn.

If in terms of the proposition we had $N = 10$ and $p_1N = 6$, $p_2N = 3$ and $p_3N = 1$ then

$$\begin{aligned} q_1 &= \tfrac{6}{10}\tfrac{5}{9} &&= 0.3333\ldots \\ q_2 &= \tfrac{3}{10}\tfrac{2}{9} &&= 0.0666\ldots \\ q_3 &= \tfrac{1}{10}\tfrac{0}{9} &&= 0 \\ q_4 &= \tfrac{6}{10}\tfrac{3}{9} + \tfrac{3}{10}\tfrac{6}{9} &&= 0.4 \\ q_5 &= \tfrac{6}{10}\tfrac{1}{9} + \tfrac{1}{10}\tfrac{6}{9} &&= 0.1333\ldots \\ q_6 &= \tfrac{3}{10}\tfrac{1}{9} + \tfrac{1}{10}\tfrac{3}{9} &&= 0.0666\ldots, \end{aligned}$$

giving in this case $\max_j |q_j - Q_j| = q_4 - Q_4 = 0.04$.

The position becomes more complicated for some other values of N, since there may not then a natural set of choices for p_k, but we could always choose them so that $P_kN - 1 < p_kN < P_kN + 1$ for all k or equivalently $|p_k - P_k| < \frac{1}{N}$, and this would give us one or two choices for each p_k while enabling us to make their sum 1. So for example if $N = 23$, then we can have p_1 as $\frac{13}{23}$ or $\frac{14}{23}$, p_2 as $\frac{6}{23}$ or $\frac{7}{23}$, and p_3 as $\frac{2}{23}$ or $\frac{3}{23}$; since they must sum to 1, we have the choice between $\frac{13}{23} + \frac{7}{23} + \frac{3}{23}$, $\frac{14}{23} + \frac{6}{23} + \frac{3}{23}$, or $\frac{17}{23} + \frac{7}{23} + \frac{2}{23}$. But we must decide each ϵ_k before knowing N, so if instead we set each ϵ_k equal to $\frac{1}{N_0}$, we have $P_kN - \frac{N}{N_0} < p_kN < P_kN + \frac{N}{N_0}$, which for $N \geq N_0$ is a wider interval and so we can still make suitable choices for various p_k.

If we do then we will have

$$\begin{aligned} q_4 - Q_4 &= 2p_1p_2\tfrac{N}{N-1} - 2P_1P_2 \\ &< 2\left(P_1 + \tfrac{1}{N_0}\right)\left(P_2 + \tfrac{1}{N_0}\right)\tfrac{N}{N-1} - 2P_1P_2 \\ &= \tfrac{1.8}{N_0} + \tfrac{2}{N_0^2} + \tfrac{0.36}{N-1} + \tfrac{1.8}{N_0(N-1)} + \tfrac{2}{N_0^2(N-1)} \end{aligned}$$

which is a decreasing function of both N_0 and N and which tends towards 0 as N_0 and so N increase. The other bounds, positive or negative, will involve similar calculations, but this one is potentially largest in absolute terms. It is less than 0.001 if $N \geq N_0 \geq 2162$.

So in this example, given $P_1 = 0.6$, $P_2 = 0.3$ and $P_3 = 0.1$ and $\eta_1 = \eta_2 = \eta_3 = \eta_4 = \eta_5 = \eta_6 = 0.001$, a solution is (with the types in the order described earlier) $Q_1 = 0.36$, $Q_2 = 0.09$, $Q_3 = 0.01$, $Q_4 = 0.36$, $Q_5 = 0.12$ and $Q_6 = 0.06$, with $\epsilon_1 = \epsilon_2 = \epsilon_3 = \frac{1}{2162} = 0.0004625\ldots$, and $N_0 = 2162$.

In fact 2162 is more than enough given that each p_kN must be an integer and the inequalities of $|p_k - P_k| < \frac{1}{N_0}$ have to be satisfied for all k, restricting further the choice of values for p_1, p_2 and p_3. As far as I can tell, the largest counter-example comes when $N_0 = 1559$, $N = 1560$, $p_1N = 936$, $p_2N = 469$ and $p_3N = 155$, leading to $|q_4 - Q_4| = 0.0010006\ldots$, meaning we could say $N \geq N_0 \geq 1560$ is sufficient in this example. If we reduced the ϵ_k to say $\frac{1}{2N_0}$ then we could find an even smaller N_0.

3. Proof of the proposition

In the degenerate case where $s = 1$ and all the individuals are in the same class (so $P_1 = 1$ and $t = 1$), we simply take $Q_1 = 1$, $\epsilon_1 = \eta_1$ and $N_0 = 1$ to satisfy the proposition, no matter what the sample size n is. The rest of the proof deals with the more interesting case where $s > 1$.

Given the number of classes s and the number of potential types of samples t, we have $t = \binom{s+n-1}{s-1}$, where n is the size of the sample. This is $\frac{s+n-1}{s-1}\frac{s+n-2}{s-2}\cdots\frac{n+1}{1}$, which is an increasing function of n (for $s > 1$), so there is a unique value of n which will give the given value of t for the given value of s.

Let $\eta = \min_j \eta_j$. If we can ensure $|q_j - Q_j| < \eta$ for all values of j then we will also have $|q_j - Q_j| < \eta_j$ for all values of j.

If a particular type j of sample of size n has $n_{j,1}$ members of class 1, $n_{j,2}$ members of class 2, and so on up to $n_{j,s}$ members of class s, then the number of different ways of distinctly ordering that type of sample is $o_j = \frac{n!}{n_{j,1}!n_{j,2}!\cdots n_{j,s}!}$. This cannot be greater than $n!$ (with equality only when $n \leq s$ and each of the $n_{j,k}$ are 1 or 0). We also have $\sum_{k=1}^{s} n_{j,k} = n$.

We will take the obvious choice for values for each of the Q_j, namely

$$Q_j = o_j \prod_{k=1}^{s} P_k^{n_{j,k}}$$

but the formula for q_j in terms of the p_k is more complicated, namely

$$q_j = o_j \frac{(N-n)!}{N!} \prod_{k=1}^{s} \frac{(p_k N)!}{(p_k N - n_{j,k})!}$$

and we need to consider the positive and negative extremes of the differences between these namely

$$q_j - Q_j = o_j \left(\frac{(N-n)!}{N!} \prod_{k=1}^{s} \frac{(p_k N)!}{(p_k N - n_{j,k})!} - \prod_{k=1}^{s} P_k^{n_{j,k}} \right).$$

But

$$\frac{1}{N^n} \leq \frac{(N-n)!}{N!} \leq \frac{1}{(N-n)^n}$$

and

$$(p_k N - n)^{n_{j,k}} \leq \frac{(p_k N)!}{(p_k N - n_{j,k})!} \leq (p_k N)^{n_{j,k}}$$

so there are loose bounds

$$o_j \left(\prod_{k=1}^{s} \left(\frac{p_k N - n_{j,k}}{N} \right)^{n_{j,k}} - \prod_{k=1}^{s} P_k^{n_{j,k}} \right) \leq q_j - Q_j$$
$$\leq o_j \left(\prod_{k=1}^{s} \left(\frac{p_k N}{N-n} \right)^{n_{j,k}} - \prod_{k=1}^{s} P_k^{n_{j,k}} \right).$$

We can set $\epsilon_k = \frac{1}{N_0}$ for all k so we have $P_k N - \frac{N}{N_0} < p_k N < P_k N + \frac{N}{N_0}$. This enables $p_k N$ to be an integer and the sum over k to be N; in general there will be several possibilities. This gives

$$o_j \left(\prod_{k=1}^{s} \left(P_k - \frac{1}{N_0} - \frac{n_{j,k}}{N} \right)^{n_{j,k}} - \prod_{k=1}^{s} P_k^{n_{j,k}} \right) \leq q_j - Q_j$$
$$\leq o_j \left(\prod_{k=1}^{s} \left(P_k + \frac{P_k n}{N-n} + \frac{1}{N_0} + \frac{n}{N_0(N-n)} \right)^{n_{j,k}} - \prod_{k=1}^{s} P_k^{n_{j,k}} \right).$$

The lower bound is negative (certainly if $N_0 \geq \frac{n_{j,s}+1}{P_k}$) but is an increasing function of N_0 and N and tends towards 0 as N_0 and so N increase. Similarly the

upper bound is positive and tends towards 0 as N_0 and so N increase. So for each j there is some minimum value where any greater N_0 ensures the absolute value of both the lower and upper bounds are less than η. Since there a finite number t of types of sample giving different values of N_0, we take the largest such N_0 and $\epsilon_k = \frac{1}{N_0}$ so that for all j the absolute value of both the lower and upper bounds are less than η. □

We could go slightly further than this and give an explicit though loose expression for N_0 with $\epsilon_k = \frac{1}{N_0}$. The lower bound is a decreasing function of P_k (i.e. becomes more negative as it increases) and the upper bound an increasing function (more positive), so we will have looser bounds if we replace P_k by 1 and then take the product; we can also replace $\frac{n_{j,k}}{N}$ by $\frac{n}{N}$. We know $o_j \leq n!$ and that the bounds would also loosen if we replace N by N_0. So we have

$$n!\left(\left(1-\frac{n+1}{N_0}\right)^n-1\right) \leq q_j - Q_j \leq n!\left(\left(1+\frac{n+1}{N_0-n}\right)^n-1\right).$$

So the upper bound is larger in absolute terms than the lower bound, and we can concentrate on it. If we then use $\left(1+\frac{c}{n}\right)^n \leq \exp(c)$ then we have

$$|q_j - Q_j| \leq n!\left(\exp\left(\frac{n(n+1)}{N_0-n}\right)-1\right)$$

and we want this to be be less than η so requiring

$$N_0 > \frac{n(n+1)}{\log_e\left(1+\frac{\eta}{n!}\right)} + n.$$

We can make this simpler though looser, first by using $\log_e(1+c) > \frac{c}{1+c}$ and then $n(n+1)n! + n(n+2) < (n+2)!$ and $\eta \leq 1$ to give a loose requirement for N_0 (with $\epsilon_k = \frac{1}{N_0}$ and the natural values of Q_j based on P_k) so that $|q_j - Q_j| \leq \eta$ for all j, namely

$$N_0 > \frac{(n+2)!}{\eta}.$$

It is possible to tighten this, but Fisher only required proof of existence for his purpose of assuming that permutations and limits could be used to justify sampling theory. Just how loose this has become can be seen from the example: using $n = 2$ we find here that $N > N_0 > 24000$ and $\epsilon_k = \frac{1}{24000}$ was sufficient to satisfy the proposition, compared with the 2162 calculated earlier.

References

[1] Fisher, R. A.: *Theory of Statistical Estimation*, Proceedings of the Cambridge Philosophical Society, **22**, (1925), 700-725.
http://digital.library.adelaide.edu.au/coll/special//fisher/42.pdf

5 Leydon Close
London SE16 5PF
E-mail address, Henry Bottomley: se16@btinternet.com
URL: http://www.btinternet.com/~se16/hgb