SANKHYĀ

THE INDIAN JOURNAL OF STATISTICS

Edited by : P. C. MAHALANOBIS

VOL. 21, PARTS 3 & 4 AUGUST 1959

THE ANALYSIS OF HETEROGENEITY. I.

By J. B. S. HALDANE

Indian Statistical Institute, Calcutta

SUMMARY. Estimators of the mean and variance of a frequency are given when this frequency varies through a series of samples.

INTRODUCTION

The following situation frequently arises in biological research, and doubtless in other branches of science. A number of experiments or observations are made under as nearly similar conditions as possible. Each of them leads to the production of a sample, whose members may be classed into two types which we may call successes and failures, though they may be females and males, fertile and sterile matings, survivals and deaths, and so on. If we have n samples, and the i-th consists of s_i members of which a_i are successes and b_i are failures, we can draw up a $(2\times n)$ fold table and apply the χ^2 or some other test of homogeneity. If the test is judged compatible with homogeneity, we can adopt the simple hypothesis that the probability of success was the same in each sample, and estimate it as $p = \sum\limits_{i=1}^{n} a_i / \sum\limits_{i=1}^{n} s_i$. If however the test is judged significant of heterogeneity, we must conclude that p has varied from one experiment to another, its value in the i-th experiment being p_i. We can then proceed to estimate the mean of p_i. Although the estimate given above is unbiassed, it will be shown that it is not efficient unless the sample number s_i is constant. We can also in general estimate the variance and higher moments of p_i. While χ^2 is a test for heterogeneity it is not a measure of it, but we shall see that the estimate of the variance of p is related to χ^2. In this paper I shall only deal with the estimation of the mean and variance.

In my experience this problem has arisen in two rather different contexts. On the one hand we may have to analyse a series of litters of mice or other small animals produced from parents as similar as possible, under standardised conditions. The mean value of s is of the order of 6, and for most values of s there is a fair number of samples. If heterogeneity has been detected, we can estimate the mean value of p for various sample numbers, and see whether they regress significantly on s. If they do not, we can weight each with the appropriate amount of information, and combine them. If our data are numerous enough, we can do the same for other moments.

On the other hand we may have to deal with a series of insect or plant families, in which the mean value of s is about 100, and two samples with the same value of s are unusual.

As Robertson (1951) pointed out, this problem is the inverse of the problem studied by Lexis (1877). Lexis considered the effect on the variance of a_i of a known variance of p_i.

In what follows I use κ_r to mean the r-th cumulant of the true distribution of p. k_r means an unbiassed estimate of κ_r and $\kappa(r^s)$ the expectation of the s-th cumulant of the distribution of k_r, while $k(r^s)$ is an unbiassed estimate of $\kappa(r^s)$. We have to consider expectations at two levels. I denote expectations for a given value of p_i, and thus within a single sample, with an asterisk. Thus $\mathcal{E}^*(a_i) = p_i s_i$, $\mathcal{E}^*(a_i^2) = p_i^2 s_i(s_i-1)+p_i s_i$, $\mathcal{E}^*(a_i\, b_i) = p_i(1-p_i)\ s_i(s_i-1)$ and so on. I denote expectations within the whole group of n samples without an asterisk. Thus $\mathcal{E}(p)=\mathcal{E}(p_i)=\kappa_1$. I use $\Sigma\, a$ or $\Sigma\, a_i$ to mean $\sum\limits_{i=1}^{n} a_i$, and so on. If s is constant $\mathcal{E}(\Sigma\, a) = \kappa_1 ns$. If s is variable I assume that s_i and p_i are uncorrelated, though this should be verified where possible. In any case I assume that p_i and p_j are uncorrelated, that is to say $\mathcal{E}(p_i\, p_j) = \kappa_1^2$, if $i \neq j$. Also

$$\mathcal{E}(p_i) = \kappa_1,\ \mathcal{E}(p_i^2) = \kappa_1^2+\kappa_2.$$

Samples of constant size

If every sample consists of s members, then since $\mathcal{E}^*(a_i) = p_i\, s$, so $\mathcal{E}(\Sigma a) = \kappa_1 ns$, whence

$$k_1 = (ns)^{-1}\Sigma\, a. \qquad \dots \quad (1)$$

This estimate is clearly unbiassed and efficient.

$$(\Sigma\, a)^2 = \Sigma\, a^2+2\sum_i \sum_{j\neq i} a_i a_j.$$

So
$$\begin{aligned}\mathcal{E}[(\Sigma\, a)^2] &= ns(s-1)\ \mathcal{E}(p^2)+ns\ \mathcal{E}(p)+n(n-1)s^2\ \mathcal{E}(p_i\, p_j)\\ &= ns(s-1)\ (\kappa_1^2+\kappa_2)+ns\ \kappa_1+n(n-1)s^2\kappa_1^2\\ &= ns(ns-1)\kappa_1^2+ns\ \kappa_1+ns(s-1)\kappa_2.\end{aligned}$$

But
$$[\mathcal{E}(\Sigma\ a)]^2 = n^2s^2\kappa_1^2,$$

so
$$\text{var}\ (\Sigma a) = ns(\kappa_1-\kappa_1^2)+ns(s-1)\kappa_2,$$

whence
$$\kappa(1^2) = \text{var}\ (k_1) = \frac{\kappa_1(1-\kappa_1)+(s-1)\kappa_2}{ns}. \qquad \dots \quad (2)$$

The first term is the component due to the small sample size, while the second, $n^{-1}(1-s^{-1})\kappa_2$, is due to the variance of p. The second term may greatly exceed the first.

$$\Sigma\ a_i\ \Sigma\ b_i = \sum_i\ a_i b_i + \sum_i \sum_{j\neq i}\ a_i b_j.$$

So $$\mathcal{E}[\Sigma\ a_i\ \Sigma\ b_i] = ns(s-1)\ (\kappa_1-\kappa_1^2-\kappa_2)+n(n-1)s^2(\kappa_1-\kappa_1^2)$$
$$= ns(ns-1)\kappa_1(1-\kappa_1)-ns(s-1)\kappa_2.$$

Also $$\mathcal{E}[\Sigma\ a_i\ b_i] = ns(s-1)\ (\kappa_1-\kappa_1^2-\kappa_2).$$

Hence $$\mathcal{E}[(s-1)\Sigma\ a\ \Sigma\ b-(ns-1)\ \Sigma\ ab] = n(n-1)s^2(s-1)\kappa_2.$$

So $$k_2 = \frac{(s-1)\ \Sigma\ a\ \Sigma\ b-(ns-1)\ \Sigma\ ab}{n(n-1)s^2(s-1)} \qquad \dots \quad (3)$$

is an unbiassed estimate of κ_2. Also

$$\mathcal{E}\left[\frac{\Sigma\ a\ \Sigma\ b-\Sigma\ ab}{n(n-1)s^2}\right] = \kappa_1(1-\kappa_1).$$

So we can put (2) in terms of observed quantities.

$$k(1^2) = \frac{\Sigma\ a\ \Sigma\ b-n\ \Sigma\ ab}{n^2(n-1)s^2} = \frac{\text{Cov}\ (a,\ b)}{(n-1)s^2}. \qquad \dots \quad (4)$$

Robertson (1951) gave an expression for the variance of p which in my symbolism becomes :

$$k_2 = [n^2\ s^2(s-1)]^{-1}\ [(s-1)\Sigma a\ \ \Sigma b-ns\ \Sigma ab].$$

Unless n is small, this is very near to my expression (3), the difference being the value (4) of $k(1^2)$. However when n is small the difference is not negligible. For example if $s = 100$, $n = 5$, and the values of a are 4, 8, 10, 14, 17, then $\Sigma a = 53$, $\Sigma b = 447$, $\Sigma\ ab = 4635$. $k_1 = 0.106$, and expression (3) gives $.00164\dot{3}\dot{6}$ for the variance or .04054 for the standard deviation of p, while Robertson's expression gives $.001127\dot{6}\dot{3}$ and .03358. If my own value is judged to be more accurate, it should be used.

χ^2_{n-1}, used as a test of homogeneity, may be written

$$\chi^2_{n-1} = \frac{ns(\Sigma a\ \Sigma\ b-n\ \Sigma ab)}{\Sigma a\ \Sigma b}$$

When $\kappa_2 = 0$, that is to say p is constant, its exact expectation is

$$\mathcal{E}(\chi^2_{n-1}) = (ns-1)^{-1}\ n(n-1)s.$$

So $$\chi^2_{n-1}-(ns-1)^{-1}n(n-1)s = [(ns-1)\ \Sigma a\ \Sigma b]^{-1}n^3(n-1)s^3(s-1)k_2,$$

or $$k_2 = \frac{\Sigma a\ \Sigma b\ [(ns-1)\chi^2_{n-1}-n(n-1)s]}{n^3(n-1)s^3(s-1)}. \qquad \dots \quad (5)$$

Since χ^2_{n-1} can be zero with a finite probability, but cannot be negative, it follows that k_2 can be negative, its minimum value, if p is constant, being $-p(1-p)(s-1)^{-1}$. The null sampling distribution of k_2 is given by that of χ^2_{n-1}, and the significance of a positive value is that of the corresponding value of χ^2. If k_2 is negative we must suppose that κ_2 is zero or too small to estimate, while drawing the appropriate conclusions from the small value of χ^2_{n-1}.

We now see that χ^2, as a test of homogeneity, has a triple function. Firstly its excess over its null value furnishes a test of whether the variance of p exceeds zero. Secondly it allows us, by means of (5), to estimate the variance of p. And thirdly it measures the uncertainty of the estimate of the mean of p. For (4) may be written as

$$\text{var }(k_1) = [n^3(n-1)s^2]^{-1}\chi^2_{n-1}\,\Sigma a\,\Sigma b. \qquad \ldots \quad (6)$$

Workers are rightly suspicious of a mean based on a heterogeneous set of samples. (4) or (6) tells them just how suspicious they should be. I may add that if χ^2_{n-1} is calculated by the method of Haldane (1955) which, it is claimed, saves a good deal of computation in some cases, (5) and (6) are more useful than (3) and (4).

(5) is analogous to the well-known relation between r and χ^2 for a (2×2)-fold table. I hope to give estimators of κ_3, κ_4, and of the sampling variance of k_2, in a later paper. The latter is not however of immediate importance, since we have an expression for the significance of a given value of k_2.

SAMPLES OF VARIABLE SIZE

If we have a large number of samples for each of a few small values of s, as with human families, we can treat each set for a given value of s separately, and combine the estimates of p, the amount of information for each value of s being given by (2). When however, the values of s are all or mostly different, we proceed as follows.

If w_i be any weighting factor, then,

$$\mathcal{E}(\Sigma w_i\, a_i) = \kappa_1 \Sigma w_i\, s_i.$$

So provided $\Sigma w_i s_i = 1$, $\Sigma w_i a_i$ is an unbiassed estimate of κ_1. Clearly w must be a one-valued function of s. Clearly also when $\kappa_2 = 0$, that is to say p is constant, w_i should be constant, and therefore equal to $(\Sigma s)^{-1}$. When however κ_2 is not zero, w should be an increasing function of s. One can derive the expression $w_i \propto [s_i - 1 + \kappa_2^{-1}\kappa_1(1-\kappa_1)]^{-1}$ which follows, directly from (2) by a somewhat intuitive argument. But it can be derived more rigorously as follows.

The most efficient form of w_i is that which minimizes the variance of $k_1 = \sum\limits_i w_i a_i$. Now

$$k_1^2 = \sum_i w_i^2 a_i^2 + 2\sum_i \sum_{j \neq i} w_i a_i w_j a_j.$$

Since $$\mathcal{E}[a_i^2] = (\kappa_1^2+\kappa_2)s_i(s_i-1)+\kappa_1 s_i$$

and $$\mathcal{E}[a_i a_j) = \kappa_1^2 s_i s_j,$$

it follows that $$\mathcal{E}[k_1^2] = \sum_i w_i^2 s_i[(s_i-1)(\kappa_1^2+\kappa_2)+\kappa_1]+2\sum_i\sum_{j\neq i} w_i w_j s_i s_j \kappa_1^2$$

$$= \kappa_1^2(\Sigma\, w_i s_i)^2+\kappa_1(1-\kappa_1)\Sigma w_i^2 s_i+\kappa_2\Sigma w_i^2 s_i(s_i-1).$$

But $\Sigma w_i s_i = 1$, so on subtracting κ_1^2 we find

$$\text{var}\,(k_1) = \sum_i [\{\kappa_2+(\kappa_1-\kappa_1^2-\kappa_2)s_i^{-1}\}w_i^2 s_i^2].$$

Since $\Sigma\, w_i s_i = 1$, this is minimal when $w_i s_i \propto [\kappa_2+(\kappa_1-\kappa_1^2-\kappa_2)s_i^{-1}]^{-1}$

or $$w_i = [\kappa_2 s_i+\kappa_1-\kappa_1^2-\kappa_2]^{-1}[\Sigma\{\kappa_2+(\kappa_1-\kappa_1^2-\kappa_2)s_i^{-1}\}]^{-1}.$$

So if $$c = \kappa_1(1-\kappa_1)\kappa_2^{-1}-1,\ w_i\propto(s_i+c)^{-1},$$

and $$k_1 = \frac{\sum\left(\frac{a}{s+c}\right)}{\sum\left(\frac{s}{s+c}\right)} \qquad \dots \quad (7)$$

$$\text{var}\,(k_1) = \kappa(1^2) = \frac{\kappa_2}{\sum\left(\frac{s}{s+c}\right)}. \qquad \dots \quad (8)$$

If κ_2 is small, that is to say p is nearly constant, c is very large, and k_1 approximates to $(\Sigma s)^{-1}\,\Sigma a$, as is obvious. If p can only assume the values of zero and unity, $c = 0$, and $k_1 = n^{-1}\Sigma\, a\, s^{-1}$. Otherwise the values of c are intermediate. Thus if all values of p from 0 to 1 are equally frequent, $c = 2$, and if all values from 0 to $\frac{1}{4}$ are equally frequent, $c = 20$, and so on. Usually therefore it will be necessary to estimate κ_2.

Prof. C. R. Rao has pointed out to me that the estimate (7) is not quite unbiassed, because in general estimates of κ_1 and κ_2 are correlated. However the bias will seldom be large. The most efficient estimator of κ_2 can only be given when higher moments are known, so that formally the problem is very complicated. But an infinite number of unbiassed estimators of κ_2 can be derived, according to the weights given to different samples. The weight to be attached to any sample will always increase with s, but the weight as a function of sample size will be somewhere between that appropriate when s is large and κ_2 small, and that appropriate when s is small and κ_2 large. We can write down any number of expectations, including the following :

$$\mathcal{E}[\Sigma s^{-1}\, ab] = (\kappa_1-\kappa_1^2-\kappa_2)(\Sigma s-n)$$

$$\mathcal{E}[\Sigma(s-1)^{-1}ab] = (\kappa_1-\kappa_1^2-\kappa_2)\Sigma s$$

$$\mathcal{E}[\Sigma s^{-1}(s-1)^{-1}ab] = (\kappa_1-\kappa_1^2-\kappa_2)n$$

$$\mathcal{E}[\Sigma a\ \ \Sigma b] = \kappa_1(1-\kappa_1)(\Sigma s-1)\Sigma s-\kappa_2(\Sigma s^2-\Sigma s)$$

$$\mathcal{E}[\Sigma s^{-1}a\Sigma s^{-1}b] = \kappa_1(1-\kappa_1)(n^2-\Sigma s^{-1})-\kappa_2(n-\Sigma s^{-1}).$$

From these we can at once derive a number of unbiassed estimates of κ_2, of which the most likely to be useful are :

$$k_{2\alpha} = \frac{(\Sigma s-n)\Sigma a\,\Sigma b-(\Sigma s-1)\Sigma s\;\Sigma(s^{-1}ab)}{(\Sigma s-n)[(\Sigma s)^2-\Sigma s^2]}$$

$$= \frac{\Sigma a\Sigma b[(\Sigma s-1)\chi^2_{n-1}-(n-1)\Sigma s]}{(\Sigma s-n)[(\Sigma s)^2-\Sigma s^2]} \qquad \ldots \quad (9)$$

$$k_{2\beta} = \frac{\Sigma a\Sigma b-(\Sigma s-1)\Sigma(s-1)^{-1}ab}{(\Sigma s)^2-\Sigma s^2} \qquad \ldots \quad (10)$$

$$k_{2\gamma} = \frac{n\Sigma s^{-1}a\Sigma s^{-1}b-(n^2-\Sigma s^{-1})\Sigma s^{-1}(s-1)^{-1}ab}{n^2(n-1)}. \qquad \ldots \quad (11)$$

Of these estimates k_{2n} and $k_{2\beta}$ should differ very little, and $k_{2\beta}$ is the easiest to compute unless χ^2_{n-1} has already been computed, which however will usually be the case. It will be seen that $k_{2\alpha}$ and $k_{2\beta}$ have about the same weighting as χ^2, while $k_{2\gamma}$ assigns approximately equal weight to each sample.

From the example which follows it will be seen that these estimates may be very close to one another. Indeed $k_{2\alpha}$ and $k_{2\beta}$ agree to four significant figures. They thus furnish a fairly precise estimate of κ_2, which, in turn, allows an accurate estimate of κ_1, and of the sampling variance of this estimate of κ_1.

TABLE 1. RECOMBINANTS IN 13 CULTURES OF *DROSOPHILA SUBOBSCURA*

s	a	b	$p' = s^{-1}a$
224	69	155	.30804
206	59	147	.28641
255	70	185	.27451
267	70	197	.26217
247	61	186	.24696
238	57	181	.23950
166	36	130	.21687
199	42	157	.21106
210	39	171	.18571
284	50	234	.17608
190	33	157	.17368
187	32	155	.17113
248	40	208	.16461
2916	658	2258	2.91673

A numerical example

In Table 1 the successive values of s are the numbers of imagines of *Drosophila subobscura* in 13 bottles each derived from a single pair mating. In each bottle the father was homozygous for a pair of autosomal recessive genes belonging to the same linkage group and therefore located on the same chromosome. The mother was heterozygous at these two loci. The values of a are the numbers of flies in which these loci had undergone recombination, commonly described as "cross-overs." I have to thank Mrs Trent, of the Department of Biometry, University College, London, for these figures. $b = s-a$, and $p'_i = a_i s_i^{-1}$. That is to say the value of p'_i is the estimate of recombination frequency from the i-th culture. The cultures are arranged in descending order of p'_i. $\chi^2_{12} = 37.059$, $P(\chi^2) = .00022$, so there is very strong evidence of heterogeneity.

Now if p were constant, its estimate would be $658/2916 = .2257 \pm .0073$. In fact all estimates known to me have been made in this way.

The mean value of p'_i is .2244, its median .2169. var $(p'_i) = .002380$, so $\sigma_{p'} = .0487$. This is much too high as an estimate of the variance of p. The formulae (9) and (11) give $k_{2\alpha} = k_{2\beta} = .001636$, $k_{2\gamma} = .001682$. This is a very satisfactory agreement and we may estimate κ_2 as .00166, giving $\sigma_p = .0407$ which is considerably below the crude estimate of .0487.

Adopting the provisional value of .225 for κ_1, we find $c = 105.03$. Putting $c = 105$ in (7), $k_1 = .2273$. If we repeat the process we find $c = 106.3$, which does not alter the value of k_1. From (8) we find $\kappa_2(1^2) = .0001908$. So

$$k_1 = .2273 \pm .0138.$$

Thus the estimate of the mean of p is only changed from its "classical" value by 12% of its standard sampling error. The change could be much greater if the values of s had a larger coefficient of variation. On the other hand its standard sampling error is nearly doubled. And we have at least an estimate of the variance of p.

Discussion

This paper is a preliminary attempt to develop a field of statistics opened up by Robertson (1951). If the sample size s is constant it is merely a matter of algebraical accuracy to obtain unbiassed and efficient estimates of all the moments or cumulants of the distribution of p, upto and including the s-th. On the other hand, at least with the approach here adopted, when s is not constant, one requires statistics of order $2r$ to obtain efficient estimates of the r-th moment or cumulant. Formally this involves an infinite regress. It may be that the problem will be soluble in finite terms by Robertson's or related methods.

The example shows that second order statistics may suffice for practical purposes when all values of s are of the order of 100 and not very variable. On the other hand had the same number of individuals occurred in some hundred samples in which s ranged from 1 to about 10, as in human families, fourth order statistics would have been desirable to obtain the correct weightings in evaluating k_2.

The numerical example shows that most of the published data on linkage are probably inaccurate. The mean recombination values found may only require slight revision. Their sampling errors are consistently larger than those published. The variances of recombination values will require estimation. A sufficiently variable value leads to spurious "interference" of the frequencies of crossing over in adjacent segments. In fact the whole theory of linkage will require revision when sufficient data are available. An attempt is being made to collect such data in this Institute.

I believe that the approach here outlined may be of a certain value in the design of sample surveys. We have seen that in the very simple case here considered, it is desirable, when k_2 is large, to divide up the total population sampled into a large number of small samples even if the total cost or effort is thereby increased. This diminishes the sampling variance of the estimate k_1. However for a given total effort or cost the optimal design is not known till we have at least a rough estimate of κ_2. So ideally the procedure would be sequential. Further the optimal design for the estimation of κ_2 is quite different from that optimal for the estimation of κ_1.

Sample surveys are seldom matters of mere counting. So the results of this investigation have no immediate relevance to them. But analogous problems will arise in sample surveys when variances as well as means are to be estimated not merely to determine the precision of means or differences between them, but for their own sake.

Some of the expressions here found can be derived, as limiting cases, from the theory of the analysis of variance. For example (1) to (6) can be derived by considering n sets of s samples, each of one member, the value of p_i being constant in each set. However I have not been able to obtain all the results of the paper by discussing such limiting cases, and the more direct approach here used can be applied to the evaluation of higher moments.

I have to thank Mrs Trent (Miss J. M. Clarke) for kindly putting her unpublished data at my disposal.

References

Haldane, J. B. S. (1955): The rapid calculation of χ^2 as a test of homogeneity from a $2 \times n$ table. *Biometrika*, 42, 519–520.

Lexis, W. (1877): *Zur Theorie der Massenerscheinungen in der menschlichen Gesellschaft*, Freiburg.

Robertson, A. (1951): The analysis of heterogeneity in the binomial distribution. *Ann. Eugen.*, 16, 1–14.

Paper received : February, 1959.