ELEMENTARY INEQUALITIES BETWEEN THE EXPECTED VALUES OF CURRENT ESTIMATES OF VARIANCE

BY

H. STEINHAUS (WROCŁAW)

The inequalities I refer to in the title are statements about the so-called Bernoullian, Lexian and Poissonian schemes, which are commonly employed to explain differences between empirical and theoretical variance. They are to be found in most text-books on probabilities and statistics.

Nevertheless, when faced with a particular problem about the spacial distribution of leucocytes in human blood, I was compelled to answer questions of the haematologists to whom certain rule-of-thumb simplifications in the usual computations appeared suspect. As I have found nowhere a fairly complete exposition of these important inequalities I consider their elementary character rather as an argument in favour of the publication.

Preliminary remarks, definitions and statements. Let x be any random variable and $E(x)$ its expected value; if y is a random variable too, and a and b constants, we have

$$E(ax+by)=a\cdot E(x)+b\cdot E(y). \tag{1}$$

If x and y are independent, we have

$$E(xy)=E(x)\cdot E(y). \tag{2}$$

The expression

$$v(x)=E(x^2)-(Ex)^2 \tag{3}$$

is called the *true variance* of x.

The true variance has obviously the following properties:

$$v(cx)=c^2v(x) \quad \text{for any constant } c, \tag{4}$$

$$v(x+y)=v(x)+v(y) \quad \text{for independent } x, y, \tag{5}$$

$$E(x-c)^2\geqslant v(x) \quad \text{for any constant } c, \tag{6}$$

the sign of equality being valid in (6) only for $c=E(x)$. It follows

$$v(x)=E(x-Ex)^2. \tag{7}$$

All these properties can be got immediately from (1), (2) and (3).

If $a_1, a_2, \ldots, a_m$ are constants, we put

$$v(a)=\frac{1}{m}\sum_{j=1}^{m}a_j^2-\left(\frac{1}{m}\sum_{j=1}^{m}a_j\right)^2 \tag{8}$$

and we get

$$v(a)\geqslant 0, \tag{9}$$

the sign of equality holding only if $a_1=a_2=\ldots=a_m$.

We call $v(a)$ the *arithmetical variance* of the finite set $\{a_j\}$.

To simplify the notations we will designate by $\bar{a}$ the mean of a_j:

$$\bar{a}=\frac{1}{m}\sum_{j=1}^{m}a_j, \tag{10}$$

and we will extend this notation on random variables; as to the natural number m it will be n in some cases and N in others, the subscript j becoming respectively k and i.

Let us denote, for instance, by x, the number of segmented leucocytes encountered among n leucocytes chosen at random on a blood preparation; we can put

$$x=u_1+u_2+\ldots+u_n, \tag{11}$$

the u being independent random variables assuming the values 1 and 0 only.

We call the procedure yielding x an *experiment*; it consists of n *trials*.

We can imagine this experiment being repeated N times and giving the values $x_1, x_2, \ldots, x_N$ respectively for x. The set x_i enables us to compute the *experimental variance* and to compare it with the true variance; such comparison is a means deciding between different hypotheses underlying the computation of the true variance.

We shall need the identity

$$\sum_{i=1}^{N}(x_i-\bar{x})^2=\sum_{i=1}^{N}x_i^2-N\bar{x}^2 \tag{12}$$

resulting directly from $\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$.

Hypotheses. We have to examine the three most important: the *Bernoulli scheme* (B), every u_k in (11) having the same probability p of being 1 (and consequently the same probability $q=1-p$ of being 0); the *Lexis scheme* (L), when every experiment is performed with a constant probability p_i, but the value of p_i changes from experiment to experiment; the *Poisson scheme* (P), where the probabilities change from trial to trial, so that p_k is the probability of u_k being 1, but the same conditions prevail in each experiment, so that the probabilities do not change with i.

Hypothesis B. Supposing p to be known let us compare first the result of N experiments, with the true variance, by computing the *statistical variance*

$$(13)\qquad s=\frac{1}{N}\{(x_1-np)^2+(x_2-np)^2+\ldots+(x_N-np)^2\}.$$

We have to answer the question whether s is an estimate for the true variance $v(x)$, i. e. whether $E(s)=v(x)$. By (1) and (13) we get

$$(14)\qquad E(s)=E(x_1-np)^2,$$

as B implies the same distribution for every x_i. By (1) and (11) we get

$$(15)\qquad E(x)=E(u_1)+E(u_2)+\ldots+E(u_k),$$

and by (5) and (11)

$$(16)\qquad v(x)=v(u_1)+v(u_2)+\ldots+v(u_k),$$

the subscript i on the left of (15) and (16) being omitted purposely. Now it results obviously from (3)

$$(17)\qquad E(u_k)=p,\qquad E(u_k^2)=p,\qquad v(u_k)=p-p^2=pq$$

($q=1-p$; $k=1,2,\ldots n$), and from (15), (16) and (17)

$$(18)\qquad E(x)=np,\qquad v(x)=npq.$$

(18) and (3) give

$$(19)\qquad E(x^2)=n^2p^2+npq,$$

and (14) gives

$$(20)\qquad E(s)=E(x^2)-2np\cdot E(x)+n^2p^2.$$

(18), (19) and (20) imply

$$(21)\qquad E(s)=npq=v(x).$$

Thus s is an estimate for the true variance.

There are only a few examples (in genetics, for instance) where the theoretical p is known; in such cases we can compute s by means of (13) putting in the formula the experimental set x_i and compare s with $v(x)$ as given by (18); a sensible difference between the two would be considered as an argument against the hypothesis B. In most cases, however, p is not known *a priori* and we are compelled to use instead the fraction defined by the result of N experiments, i. e. the fraction of favorable trials. This amounts to use $\overline{x}$, the mean of all x, instead of np, and $1-\overline{x}/n$ instead of q. The question arises, how does this procedure affect the variances; we must note that $v(x)$ cannot be computed by the exact formula (18) and the introduction of quantities depending of the result of experiments, like $\overline{x}$, turns the expression npq into a random variable.

Let us denote by s_e the purely empirical expression

$$(22)\qquad s_e=\frac{1}{N-1}[(x_1-\overline{x})^2+(x_2-\overline{x})^2+\ldots+(x_N-\overline{x})^2],$$

and let us compute the expected value of the parenthesis [ ]; the identity (12) shows it to be equal to

$$(23)\qquad E\left(\sum_{i=1}^{N}x_i^2\right)-E(N\overline{x}^2)=N\cdot(Ex_1^2)-N\cdot E(\overline{x}^2),$$

and we have only to work out $E(\overline{x}^2)$, as $E(x_1^2)$ is already given by (19). We get from (3) (4) and (5)

$$(24)\qquad E(\overline{x}^2)=v(\overline{x})+[E(\overline{x})]^2=\frac{1}{N^2}v\left(\sum_{i=1}^{N}x_i\right)+[E(x)]^2=\frac{1}{N}v(x_1)+[E(x_1)]^2,$$

and from (23) and (24)

$$(25)\qquad \begin{aligned}E[\quad]&=N\cdot E(x_1^2)-v(x_1)-N\cdot[E(x_1)]^2\\&=N\cdot v(x_1)-v(x_1)=(N-1)\cdot v(x_1).\end{aligned}$$

(22) and (25) show immediately

$$(26)\qquad E(s_e)=v(x)=npq,$$

i. e. the empirical s_e being an estimate for the true variance.

The usual manner to test the identity (26) is to replace $v(x)=npq$ by

$$(27)\qquad v_e=\overline{x}(1-\overline{x}/n);$$

v_e is a random variable resulting from N experiments and it is customary to compare s_e with v_e as if v_e were the true variance.

It results from (27), (18) and (24) however

$$E(v_e) = E(\bar{x}) - E(\bar{x}^2)/n = np - [E(x)]^2/n - v(x)/Nn$$
$$= np - np^2 - npq/Nn,$$

and

(28) $$E(v_e) = npq - pq/N.$$

Thus we get from (26) and (28)

(29) $$E(s_e) = E(v_e) + pq/N,$$

which formula proves that the customary comparison could create an appearance of *hypernormal* variance, as the expected value of $s_e - v_e$ is positive. Nevertheless the difference is in most case too small to be taken care of.

Let us count segmented leucocytes in ten experiments, each of them comprising 100 leucocytes of all kinds; let us assume $p=0.6$, i. e. a theoretical probability of 0.6 for a leucocyte to be segmented. Hence

$$n=100, \quad N=10, \quad q=0.4, \quad npq=24, \quad pq/N=0.024.$$

Thus the difference is only $1\,^0/_{00}$ of the quantity measured; as the standard deviation is the square root of the variance, the difference between the two standard deviations which could be ascribed to the last term of (29), could alter only statements in which the third decimal of the ratio of both deviations is significant; in problems connected with the repartition of leucocytes even the second decimal is beyond the attainable accuracy.

Hypothesis L. The probability is constant in every experiment but it changes from experiment to experiment, p_i being the probability of success in every trial of the i-th experiment.

Let us call $\bar{p}$ the mean of all p_i and let us evaluate first the expected value of the statistical variance:

(30) $$s = \frac{1}{N}[(x_1 - n\bar{p})_2 + (x_2 - n\bar{p})^2 + \ldots + (x_N - np)^2],$$

$$E(s) = \frac{1}{N}\sum_{i=1}^{N} E(x_i - np)^2,$$

$$E(x_i - n\bar{p})^2 = E(x_i^2) - 2n\bar{p}\,E(x_i) + n^2\bar{p}^2 = n^2p_i^2 + np_iq_i - 2n^2p_i\bar{p} + n^2\bar{p}^2,$$

where q_i signifies $1-p_i$ and the expected values have been taken from (18) and (19). Thus we get

(31) $$E(s) = \frac{1}{N}\sum_{i=1}^{N} np_iq_i + n^2\left(\frac{1}{N}\sum_{i=1}^{N} p_i^2 - \frac{2\bar{p}}{N}\sum_{i=1}^{N} p_i + \bar{p}^2\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N} np_iq_i + n^2 v(p),$$

$v(p)$ being defined by

(32) $$v(p) = \frac{1}{N}\sum_{i=1}^{N} p_i^2 - \bar{p}^2,$$

accordingly to (8). The case where all the p_i are equal being excepted by the hypothesis L we know by (9 that $v(p)$ is positive. As the true variance $v(x_i)$ of x_i is np_iq_i by (18), we see on the right side of (31) the mean variance $\bar{v}$ defined by

(33) $$\bar{v} = \frac{1}{N}\sum_{i=1}^{N} v(x_i) = \frac{1}{N}\sum_{i=1}^{N} np_iq_i$$

augmented by n^2 times the numerical variance of the set $\{p_i\}$:

(34) $$E(s) = \bar{v} + n^2 v(p).$$

The knowledge of $\bar{p}$ is sufficient to draw from a set of experiments the value of s as defined by (30), but it gives no means to compute $\bar{v}$. Let us therefore proceed as if we were in the Bernoulli case and had to evaluate the true variance corresponding to a constant probability $\bar{p}$; let us call it $v(x)$; writing $\bar{q}$ for $1-\bar{p}$ we get

(35) $$v(x) = n\bar{p}\bar{q}.$$

(32), (33) and (35) give

$$v(x) - \bar{v} = n\left(\bar{p}(1-\bar{p}) - \frac{1}{N}\sum_{i=1}^{N} p_i(1-p_i)\right) = n\left(\frac{1}{N}\sum_{i=1}^{N} p_i^2 - \bar{p}^2\right) = nv(p),$$

i. e.

(36) $$v(x) = \bar{v} + nv(p),$$

which with (34) leads to

(37) $$E(s) = v(x) + (n^2 - n)\,v(p).$$

Thus we can expect for the statistical variance s (putting apart the trivial case $n=1$) a value exceeding the Bernoulli variance $v(x)$.

In many manuals of statistics the equality (37) is considered as a proof of the hypernormal behaviour of the empirical variance under Lexis hypothesis. However, an experimental comparison of s with $v(x)$ is impossible in most cases, as the definitions of both quantities (30) and (35) do contain the unknown theoretic $\bar{p}$. Let us therefore take instead of s and v the quantities s_e and v_e, already defined by (22) and (27). We get as before for the expected value of the parenthesis in (22)

$$(38)\qquad E[\quad]=E\left(\sum_{i=1}^{N}\bar{x}_i^2\right)-E(N\bar{x}^2),$$

and from (19)

$$(39)\qquad E(x_i^2)=n^2p_i^2+np_iq_i.$$

(38), (39) and (24) give

$$(40)\qquad E[\quad]=\sum_{i=1}^{N}n^2p_i^2+\sum_{i=1}^{N}np_iq_i-\frac{1}{N}\sum_{i=1}^{N}v(x_i)-N\cdot[E(\bar{x})]^2$$
$$=n^2\sum_{i=1}^{N}p_i^2-N\left(\sum_{i=1}^{N}np_i/N\right)^2+[1-1/N]\cdot\sum_{i=1}^{N}np_iq_i$$
$$=n^2N\cdot v(p)+(N-1)\bar{v},$$

if use is made of (32) and (33). So (22) and (40) lead to

$$(41)\qquad E(s_e)=\bar{v}+n^2v(p)\cdot\frac{N}{N-1}\qquad (N>1),$$

and — if we recall (36) — to

$$(42)\qquad E(s_e)=v(x)+v(p)\cdot\left(\frac{n^2N}{N-1}-n\right).$$

Thus it results from (42), (37) and (36)

$$(43)\qquad E(s_e)>E(s)>v(x)>\bar{v}.$$

Let us now consider v_e; (27) and (24) give

$$E(v_e)=E(\bar{x})-\frac{1}{n}E(\bar{x}^2)=n\bar{p}-\frac{1}{n}\left(\sum_{i=1}^{N}v(x_i)/N^2+E(\bar{x}^2)/N^2\right)$$
$$=n\bar{p}-\frac{n}{N}\left(\sum_{i=1}^{N}p_i\right)^2-\sum_{i=1}^{N}p_iq_i/N^2=n\bar{p}\bar{q}-\sum_{i=1}^{N}p_iq_i/N^2,$$

as it results from the identity

$$\bar{p}\bar{q}=\bar{p}(1-\bar{p})=\frac{1}{N}\sum_{i=1}^{N}p_i\cdot\left(1-\frac{1}{N}\sum_{i=1}^{N}p_i\right)=\sum_{i=1}^{N}p_i/N-\left(\sum_{i=1}^{N}p_i\right)^2/N^2,$$

and, with (35) and (36),

$$(44)\qquad E(v_e)=v(x)-\sum_{i=1}^{N}p_iq_i/N^2=\bar{v}+nv(p)-\sum_{i=1}^{N}p_iq_i/N^2.$$

Hence the difference $E(s_e)-E(v_e)$ is positive, as shown by (41) and (44):

$$(45)\qquad E(s_e)-E(v_e)=v(p)\cdot\left(\frac{Nn^2}{N-1}-n\right)+\sum_{i=1}^{N}p_iq_i/N^2.$$

The difference

$$E(s)-E(v_e)=v(p)\cdot(n^2-n)+\sum_{i=1}^{N}p_iq_i/N^2$$

resulting from (37) and (44) is positive too, but it is the inequality $E(s_e)>E(v_e)$ which is usually tested by experiments as both terms can be experimentally determined; (45) proves that the hypothesis of Lexis is sufficient to explain the empirical variance to be systematically greater than the so-called theoretic variance v_e computed in the same manner as in the Bernoulli case. To summarize, we have got the inequalities

$$(46)\qquad E(s_e)>E(s)>v>E(v_e),\qquad v>\bar{v},$$

contained in (43), (44) and (45); as to $E(v_e)>\bar{v}$ it is not true in general, but (44) proves it to be valid if n or N is sufficiently great.

Hypothesis P. The probability p_k changes from trial to trial but the same set p_k underlies every experiment.

We have now obviously by (15)

$$(47)\qquad E(x_i)=\sum_{k=1}^{n}E(u_k)=\sum_{k=1}^{n}p_k=n\bar{p}\qquad (i=1,2,\ldots,N),$$

denoting by $\bar{p}$ the mean $\sum_{k=1}^{n}p_k/n$. It follows immediately that

$$(48)\qquad E(\bar{x})=n\bar{p}$$

denoting by $\bar{x}$ the mean $\sum_{i=1}^{N} x_i/N$. It follows from (16) and (17) by replacing p by p_k in (17)

$$\mathfrak{v}(x_i)=\sum_{k=1}^{n}\mathfrak{v}(u_k)=\sum_{k=1}^{n}p_k q_k \qquad (q_k=1-p_k), \tag{49}$$

and from (4) and (49)

$$\mathfrak{v}(\bar{x})=\frac{1}{N}\sum_{k=1}^{n}p_k q_k. \tag{50}$$

From (47) and (49) we get

$$E(x_i^2)=\sum_{k=1}^{n}p_k q_k+n^2\bar{p}^2, \qquad (i=1,2,\ldots,N), \tag{51}$$

and from (48) and (50)

$$E(\bar{x}^2)=\sum_{k=1}^{n}p_k q_k/N+n^2\bar{p}^2. \tag{52}$$

The statistical variance s being defined by (30) we write

$$\begin{aligned}E(s)&=\sum_{i=1}^{N}E[(x_i-n\bar{p})^2]/N=\sum_{i=1}^{N}E\{[x_i-E(x_i)]^2\}/N\\&=\sum_{i=1}^{N}\mathfrak{v}(x_i)/N=\sum_{k=1}^{n}p_k q_k\end{aligned} \tag{53}$$

utilising (47), (7) and (49); the quantity $\bar{\mathfrak{v}}$ being defined by the first equality (33) we get

$$E(s)=\bar{\mathfrak{v}}. \tag{54}$$

To compute $E(s_e)$, s_e being given by (22), we can utilise (23), (24) and (25):

$$E(s_e)=\mathfrak{v}(x_i)=\bar{\mathfrak{v}}, \tag{55}$$

the last equality resulting from all $\mathfrak{v}(x_i)$ being equal. Thus we have $E(s)=E(s_e)$ as in the Bernoulli case.

Let us define $\mathfrak{v}(x)$ by (35) and compare it with $\bar{\mathfrak{v}}$:

$$\begin{aligned}\mathfrak{v}(x)-\bar{\mathfrak{v}}&=n\bar{p}\bar{q}-\sum_{k=1}^{n}p_k q_k=\sum_{k=1}^{n}p_k\cdot\left(1-\sum_{k=1}^{n}p_k/n\right)-\sum_{k=1}^{n}p_k+\sum_{k=1}^{n}p_k^2\\&=\sum_{k=1}^{n}p_k^2-\left(\sum_{k=1}^{n}p_k\right)^2\Big/\,n=n\mathfrak{v}(p),\end{aligned} \tag{56}$$

where $\mathfrak{v}(p)$ is an analogon to (32): i has been replaced by k, and N by n. We get from (54), (55) and (56)

$$E(s_e)=E(s)=\bar{\mathfrak{v}}=\mathfrak{v}(x)-n\cdot\mathfrak{v}(p). \tag{57}$$

The inequality $E(s_e)<\mathfrak{v}(x)$ is usually brought forward in textbooks as a reason why the hypothesis of Poisson is suited to explain a systematic surplus of the so-called theoretic variance over the empirical one. This explanation, however, does not apply to most examples, as they are worked out not with the variance $\mathfrak{v}(x)$ which presupposes the knowledge of $\bar{p}$, but with $\mathfrak{v}_e$ defined by (27). Now we have by (27), (48) and (52)

$$\begin{aligned}E(\mathfrak{v}_e)&=E(\bar{x})-E(\bar{x}^2)/n=n\bar{p}-\sum_{k=1}^{n}p_k q_k/Nn-n\bar{p}^2\\&=n\bar{p}\bar{q}-\sum_{k=1}^{n}p_k q_k/Nn,\end{aligned} \tag{58}$$

and by (35) and (33)

$$E(\mathfrak{v}_e)=\mathfrak{v}(x)-\bar{\mathfrak{v}}/Nn; \tag{59}$$

comparing with (57) we get therefrom

$$E(s_e)=E(\mathfrak{v}_e)-n\cdot\mathfrak{v}(p)+\bar{\mathfrak{v}}/Nn. \tag{60}$$

Thus the inequality

$$E(s_e)<E(\mathfrak{v}_e), \tag{61}$$

characteristic for the hyponormal dispersion, is not valid in the Poisson case without supplementary assumptions. Nevertheless it can be seen from (60) that it is valid for n or N sufficiently large. The condition for the validity of the inequality (61) in the Poisson scheme is

$$\mathfrak{v}(p)>\bar{\mathfrak{v}}/Nn^2; \tag{62}$$

as $p_k q_k\leqslant 1/4$ this condition is certainly fulfilled if

$$\mathfrak{v}(p)>1/4Nn. \tag{63}$$

In the example of leucocytes with $N=10$, $n=100$ let us assume for instance

$$p_1=p_2=\ldots=p_8=0.6;\qquad p_9=0.58,\qquad p_{10}=0.62;$$

this small arithmetical variance of the p_k would already satisfy the condition (63) and guarantee (61).

To summarize the case P we have got the inequalities

$$E(s_e)=E(s)=\bar{\mathfrak{v}}<\mathfrak{v}(x),\qquad E(\mathfrak{v}_e)<\mathfrak{v}(x) \tag{64}$$

and conditionally $E(s_e)<E(\mathfrak{v}_e)$.