QUALITY CONTROL BY SAMPLING

(A PLEA FOR BAYES' RULE)

BY

H. STEINHAUS (WROCŁAW)

1. The belief in inverse probabilities has been shattered by scientific criticism a generation ago. "The theory of inverse probabilities is founded upon an error, and must be wholly rejected" — writes a man who has contributed to the development of statistical science in our century more than anybody else [1]). New methods which are told to be independent of hypotheses essential for inverse probabilities were discovered by and by, and subsequently adapted to practical aims, the old ones thrown to rubbish. The resulting situation resembles one created by the Copernican discovery, which invests the average educated man of our epoch with the feeling of superiority over every Ancient, may he be Ptolemy himself. Nevertheless, nine out of ten modern lawyers, journalists or politicians give erroneous answers when asked about the causes of certain very simple celestial phenomena the true explanation of which was well known to the Ancients. For instance, the relative situation of Sun and Moon, as directly seen by a naïve observer who believes in the immobility of the Earth, is sufficient to explain the phases of the Moon; as soon as he tries to take into account the circling and spinning motion of the Earth, he ceases to believe his own eyes, and not to be deceived by apparences he abstains from looking at the sky, and confounds the phases with the eclipses.

The purpose of this communication is to try the old theory against the new one on the special problem of quality control by sampling. We are not concerned with the problem as a whole; a few examples suggested by the simplest kind of quality control are sufficient to make our point clear.

We have to accept or to reject a *lot* submitted to inspection, the *decision* being a result of a partial examination of the lot.

[1]) R. A. Fisher, *Statistical Methods for Research Workers*, 10th edition, London, 1948, p. 9.

Let us suppose that the producer of the lot has agreed previously with the consumer upon the *quality* of the lot, i. e. upon the fraction of good items in the lot qualifying it as fit for acceptance; let us call this quality α. As it is impossible in most practical cases to examine every item of the lot, a *plan*, i. e. a system of rules, is to be worked out, by which the decision is derived from the result of the examination of a part of the lot drawn at random from the lot, and called a *sample*. Every *item* of the sample is classified as *good* or *defective*; the plan is defined by the *size* n of the sample, by the maximum number m_1 of defectives in the sample allowed for *acceptance*, and by the minimum number m_2 of defectives in the sample required for *rejection* of the lot.

Let us suppose a particular case with exactly m_2 defectives found in the sample and rejection ensuing. The producer asks the mathematical expert responsible for the plan to explain this particular decision. The answer he gets depends on the principles serving as the theoretical basis for the plan.

If the expert belongs to the old school, he answers: "The result of the examination gives me a probability of $95^0/_0$ for your lot having a quality inferior to α".

If he is up to date, he says: "Were your lot of quality α, you would have a probability of $95^0/_0$ of showing a sample better (i. e. a sample with less defectives) than you have actually shown".

The first expert is fifty years behind the mathematical statistics of our times: his answer refers to the question proposed, but it is erroneous. His younger colleagues' statement is perfectly correct, but it is evidently not an answer to the producer's question.

The first answer can be formulated more cautiously: "If you submit lots of different qualities, taking care that every possible quality appears with the same frequency as any other (i. e. if qualities α belonging to any closed interval $\alpha' \leqslant \alpha \leqslant \alpha''$ do appear with the frequency $\alpha''-\alpha'$), and if you keep records of the qualities of the rejected lots, putting always in the same file the records showing m_2 defectives and in another the records with more than m_2 defectives, you will have in the long run $95^0/_0$ just rejections in the first file and more than $95^0/_0$ just rejections in the other".

The second answer can also be formulated as a statement about frequencies: "If you always submit lots of the same quality α, you will have, in the long run, only $5^0/_0$ rejections, and if you submit systematically lots of quality superior to α, you will suffer rejection even less often than five times in hundred".

Now both answers are clearly explained, but neither is satisfactory for the practical man, as both speak about "ifs"; the hypothetical occurrences mentioned are pretty far from the reality of the economical life. This objection [2]) applies as well to the new as to the old point of view.

2. To approach the subject more closely some definitions are necessary.

We call B the hypothesis that a lot L has been drawn at random from a collection C of lots, C being such that the probability of L having a quality between α' and α'' always equals $\alpha''-\alpha'$. We design with $H(\alpha)$ the hypothesis that the lot submitted to quality control has the quality α. A *single sampling plan* $S(m_1,n,m_2)$ is defined by the number n of items to be examined, called the *size of the sample*, by the *acceptance number* m_1, the highest number of defectives in the sample allowed for acceptance of the lot, and by the *rejection number* m_2, being the least number of defectives in the sample implying rejection of the lot. The condition $0\leqslant m_1<m_2\leqslant n$ makes the plan *consistent*, the supplementary condition $m_1=m_2-1=m$ makes it *categorical*: it guarantees the dichotomy "accept or reject" in every case; such a plan $S(m,n,m+1)$ may be designed shortly by $m\|n$.

Every result of examination of a sample can be symbolized by (k,n), n being the size of the sample, and k — the number of defectives it contains. The result (m_1,n) may be called the *extreme acceptance*, and the result (m_2,n) — the *extreme rejectance* for the plan $S(m_1,n,m_2)$.

3. A plan $S(m_1,n,m_2)$ being given, two numbers, α_1 and α_2, where $0\leqslant\alpha_1\leqslant\alpha_2\leqslant 1$, can be defined, called respectively the *lower* and the *upper quality level* corresponding to the plan. To define these numbers we have to settle first by convention the so-called *probability levels*; to avoid a generality irrelevant for our purposes we assume once for all these levels to be $0{,}95(=95^0/_0)$.

Then we have to choose between two definitions, R and P:

Definition R. The *lower quality level* α_1 is such that B and the extreme acceptance imply with the probability 0,95 the inequality $\alpha>\alpha_1$ for the quality α of the lot submitted.

We can write the defining statement briefly as $(m_1,n)\supset(\alpha>\alpha_1)$, if we agree to read also B in the major and the probability clause, suppressed for brevity's sake.

The *upper quality level* α_2 is defined by $(m_2,n)\supset(\alpha<\alpha_2)$, m_2 being the rejection number; the same commentary as to the reading holds here too.

Definition P. The *lower quality level* α_1 is such that $H(\alpha_1)$ for the lot submitted implies with the probability 0,95 the inequality $m>m_1$ for the result (m,n) of the examination of the sample.

We can write the defining statement briefly as $H(\alpha_1)\supset(m>m_1)$, remembering to read the probability clause.

The *upper quality level* α_1 is defined by $H(\alpha_2)\supset(m<m_2)$, with the same commentary.

The definitions R and P are not equivalent. The plan $1\|20$, for instance, yields $79{,}3^0/_0$ as the lower and $96{,}1^0/_0$ as the upper quality level with definition R, whereas with definition P the levels in question are $78{,}4^0/_0$ and $98{,}2^0/_0$ respectively.

The definition R is not to be found in British or American standards for quality control; they deal exlusively with definition P. The reason for this predilection is the common opinion of hypothesis B being false or meaningless.

4. An important connection between the definitions R and P has been established by Oderfeld. Let us, namely, introduce the following notations:

$R_1(m_1,n)$ — the lower quality level corresponding to the size n of the sample and to the acceptance number m_1, if definition R is adopted,

$R_2(m_2,n)$ — the upper quality level corresponding to the size n of the sample and to the rejection number m_2, if definition R is adopted,

$P_1(m_1,n)$ — the lower quality level corresponding to the size n of the sample and to the acceptance number m_1, if definition P is adopted,

[2]) I am indebted for this remark to Mr J. Oderfeld from the Polish Standards Committee.

$P_2(m_2,n)$ — the upper quality level corresponding to the size n of the sample and to the rejection number m_2, if definition P is adopted.

We have then the following *rule of dualism* [3]):

$$(1)\qquad R_1(m_1,n)=P_1(m_1,n+1),\qquad R_2(m_2,n)=P_2(m_2+1,n+1).$$

This result must seriously invalidate the prejudice against the definition R. Suppose 79,3$^0/_0$ and 96,1$^0/_0$ were prescribed as the respective quality levels, and definition R adopted. The resulting plan is $S(1,20,2)$. With definition P the same levels would correspond to the plan $S(1,21,3)$, which is not categorical. As the procedure described by $S(1,20,2)$ is, roughly speaking, not very different from the procedure summarized by $S(1,21,3)$, somebody watching two inspectors working at quality control by sampling would hardly notice this difference. Both inspectors were given the levels 79,3$^0/_0$ and 96,1$^0/_0$, but one of them is aided by a mathematical adviser who believes in the obsolete rules of inverse probabilities, whereas his colleague has the advantage of having at his disposal printed instructions published by the Statistical Research Group of a famous university. The observer would eventually notice the difference, but he would rather explain it as something like two ways of reading 1^0 temperature: a Frenchman reads it as 1^0 centigrade, and a German — as 1^0 Réaumur; the difference is but slight. If the observer is a practical man, he will be greatly astonished if he is told that one of the inspectors uses a hypothesis which is false or meaningless! His astonishment will increase, if he learns how the blameful inspector could continue his work with the plan $S(1,20,2)$ without risking damnation: he has only to admit that the quality levels are respectively 78,4$^0/_0$ and 98,2$^0/_0$. The observer will consider, after all, his own comparison with the scales of temperature as not too far from truth.

5. Let us call, for brevity's sake, the old method R (retrospective) and the new one P (prospective). The adherent of P may object against all that has been said above that it neither touches the mathematical correctness of the P-method nor its applicability. "Even if the R-method were applicable in some cases, it is scientifically suspect and, consequently, inferior to P" — says the P-man.

[3]) J. Oderfeld, *On the dual aspect of sampling plans*, this fascicle, p. 89.

We must therefore proceed to show the advantages of R *not* shared by P.

A problem of practical importance is the distinction between the *good* and the *bad* producer. The first is known as being, as a rule, up to the standards, the second — as being behind them. The customer who has to pay for the goods feels it only fair to treat them differently. The probability a priori does not conform any longer with hypothesis B, this hypothesis assuming all qualities of lots as equally probable. We simplify our problem by asking: "How do we know that a producer is good or bad?" The answer is obvious: our knowledge is based on the experience of previous sampling. The R-method faces the problem this way: if a producer presents his ware for the first time, we may as well assume the hypothesis B, and choose a suitable sampling plan according to definition R; if (m,n) was the result of the first examination, we may take this experience into account at the next shipment; there are many devices to do it. One of them, which seems simple enough for practical use, is the following: m_1 being the acceptance number, and $m<m_1$, the decision relative to the first lot was favourable, but as the same decision would have been taken if the result were (m_1,n), we have employed for it only a part of the information contained in the effective result (m,n). The rest is free and has to be put on the balance of the producer as the positive quantity m_1-m. At the next delivery this record has to be employed to modify the result of the new sampling: the result (m',n) is changed into $(m'-m_1+m,n)$. We proceed with the modified result as with the first: a part of it yields the decision and the rest remains on the balance. The free rest can be sometimes a negative quantity; by subtracting it from m' we increase the number of defectives, and the procedure remains unaltered. We can settle limits for the amounts to be kept on the balance; if they are reached, the sampling plan is changed, becoming respectively more or less exacting as the negative or the positive limit has been attained. This decision is accompanied by the cancelling out of all previous records.

It is not our purpose to give here a system of rules covering all possible situations. For our discussion R *versus* P it is important to show that the idea of informations contained in the results of sampling gives a clue to the problem of good and bad producers. Now, this idea is natural only to somebody accustomed to think in terms of the R-method. It is incompatible with the P-method. The P-method does not care what the a priori distribution is like. This is exactly what the P-method is boasting of. Ignoring the a priori distribution it never makes inferences from the sample on the lot. As it denies the possibility of learning by experience, the P-method is deprived of a natural approach to the problem of good and bad producers. It would be perhaps not impossible to solve it by an artificial device which would present in P-terms the results of R-thinking, just as there are people acting in terms of this world and speaking in terms of the other. One could hardly find a better description for this situation than the statement written recently by Wiener[4]:

"The development of our theory" (the time series problem) "beyond this point, as a practical statistical theory, involves an extension of existing methods of sampling... It involves all the complexities of the use, either of Bayes' law on the one hand, or of those terminological tricks in the theory of likelihood, on the other, which seem to avoid the necessity for the use of Bayes' law, but which in reality transfer the responsibility for its use to the working statistician, or the person who ultimately employs his results. Meanwhile the statistical theorist is quite honestly able to say that he has said nothing which is not perfectly rigorous and unimpeachable..."

6. Suppose we have to face the following practical problem, to be formulated best on an example. The plan $1\|20$ has been agreed upon. A lot of 10 000 items has been submitted to inspections with the result $(2,20)$. The lot is rejected. The producer proposes to improve the lot by discarding a certain number of defectives under control of the purchaser. He asks how many defectives will he have to eliminate to have the lot accepted without further sampling. Only P-plans are available.

The solution is not difficult, if we do not fear to think in R-terms. The plan $1\|20$ read as a R-plan gives as the lower quality level the same number as the plan $1\|21$ read in P-terms. This number is $79{,}3^0/_0$ and is to be found in the list of P-plans. Read in R-terms it means that the agreement stipulates a $95^0/_0$ probability for the lot's quality being above $79{,}3^0/_0$, as a condition for acceptance. To give a meaning to the result $(2,20)$ we must find the lower quality level for the plan $2\|20$ in R-terms. Equation (1) says that this level is the same as for the plan $2\|21$ read in P-terms. The list gives $72{,}9^0/_0$. In R-terms it means that the actual lot has with $95^0/_0$ probability a quality superior to $72{,}9^0/_0$.

To improve the lot we must discard x defectives, x being given by the equation

$$\frac{10\,000-7290-x}{10\,000-x}=1-0{,}793.$$

$x=807$ is the number sought for. The producer has to examine the lot item by item until 807 defectives are found and discarded.

The solution given above could be simplified, if R-plans were available. The whole reasoning would run exclusively on R-lines and the translation by means of the rule of dualism (1) would be superfluous. Our example shows what happens if the agreement is written in P-terms, and only P-plans are available. The interesting feature of such a situation is that the inspector is compelled to ask for help somebody who is not ashamed to think in R-terms. This friend in need makes the translation, reads the P-plans after his fashion and gives an unambiguous advice of discarding 807 defectives. There is one thing he does not: he abstains from explaining the correctness of his answer in P-terms; the inspector is, however, humiliated enough not to insist on this question. There is, doubtlessly, a solution to this riddle — it would sound probably very queer even for the sworn adherents of the P-method.

7. P-methodists have the habit of putting the sequential analysis on their side of the balance. Sequential analysis is easily explained in R-terminology. The single plan with an invariable size of the sample is replaced by a sequential plan, where the size of the sample depends of what happens during the examination. We examine item after item getting the results $(r_1,1)$, $(r_2,2)$, $(r_3,3)$, ..., and stop examination at the result (r_k,k) which is the first to give a sufficient basis for a categorical decision conformably to the definition R. More precisely: we stop examination as soon as $(r_k,k)\supset(a>a_1)$ or $(r_k,k)\supset(a<a_2)$ (with abbreviations of the definition R in force). The first case implies acceptance, the latter — rejection of the lot; in the case of neither of the two implications being true we have to ex-

[4] N. Wiener, *Cybernetics*, New York, 1948, p. 109-110.

amine the $(k+1)$th item and so on, until a categorical decision is reached. To construct single R-plans, tables of Incomplete Beta-function may be employed; the same tables are sufficient for the sequential plans defined above.

It is quite natural to stop examination as soon as the partial result of sampling yields an information about the quality of the lot, as strong as that which we considered a sufficient basis for a decision in single sampling. Thus, sequential analysis is a refinement of the R-method, which employs the same concepts and the same means of computation as the R-theory of single sampling plans.

It would be far more difficult to explain sequential analysis in the P-language. The rule of stopping examination as soon as the necessary information is obtained means nothing without inverse probabilities. This is the reason why this very useful and natural device of reducing inspection to the necessary minimum has appeared comparatively late: quality control by sampling and the defeat of R-troops belong to the same epoch and thus it happened that the problem of minimum inspection was put before mathematicians who voluntarily deprived themselves of the primitive weapon invented by Bayes. They had to wait until the difficulties of combining the sequential principle with the new methods were vanquished by Wald[5]). He proceeds as follows:

Every function $F(r,k)$, defined for all natural numbers k and all natural numbers $r \leqslant k$, taking the values *accept, reject,* and *continue,* gives a system of rules, if we interpret (r,k) as a result of sampling (r defectives among k examined items). This system may be called the *sequential plan F.*

The postulate for F to give to a lot of quality α_1 the probability of 0,05 for being eventually accepted, and the same probability to a lot of quality α_2 for being eventually rejected, restricts the choice of F to a certain set $\{F\}$. Every plan F defines, for a given quality α, a random variable: the least number k which happens to give to F one of the two categorical values; if a lot of quality α is examined conformably to F. The expected value of k is the expected size of the sample to be examined.

[5]) A. Wald, *Sequential tests of statistical hypotheses,* The Annals of Mathematical Statistics 16 (1945), p. 118-186.

This expected size has to be as small as possible. This restriction reduces the set $\{F\}$ to a subset $\{F\}^*$. As the expected size depends also on α, the postulate of smallest size leads to as many solutions as there are different qualities α. Wald considers only the values α_1 and α_2, and minimizes the sum of the expected sizes; the exact proof of the minimum property for Wald's solution is still outstanding. Nevertheless, it was a remarkable achievement to find Wald's formula.

When reading the argument one sees the recipe appearing as a *deus ex machina* and one is tempted to guess behind it Bayes' rule pulling the threads. The machinery is now too complicated to be explained to a layman. It works, but it works no better than the sequential analysis based on the R-method. The greater the mathematical skill needed for the inoculation of the sequential analysis on the P-tree, the stronger the evidence that the sequential analysis is no valid argument for the P-method. We would rather consider the advantages of the sequential principle as an argument in favour of the R-method, in which it is a direct consequence of the principles involved, and no divination is needed to guess the solution.

It would be interesting to compare a sequential plan constructed conformably to Wald's method with the corresponding sequential R-plan. To do this we have to choose first the levels α_1, α_2, and to construct the Wald's plan for these levels. Now, a single P-plan for the levels α_1, α_2 can be read as an R-plan for certain levels α_1', α_2'. The procedure employed to define a sequential plan in R-terms, and explained at the beginning of this section, leads to a sequential R-plan for the levels α_1', α_2'. This plan is to be compared with Wald's plan for α_1, α_2. The plans would be different but the difference would not amount to very much. The advantage of the R-plan, besides its theoretical simplicity, is the limitation of the sample's size, while the random variable k in Wald's plans is not bounded.

8. There is an objection against the R-method which is generally considered as very serious. To apply inverse probabilities we admit the hypothesis B, which postulates uniform distribution a priori for a certain parameter α, whose true value is unknown. Now, there are other parameters connected with α; α^2, for instance, is such a parameter; let us call it β. We could as well suppose β to be equally distributed, which is not equivalent with the uniform distribution of α. How are we to know to which parameter, α or β, should we to apply the hypothesis of uniformity?

For problems of quality control the answer is simple: inspection by sampling is a game in which good items appear with probability α, and we have to determine this probability; β is not a probability. Thus in all problems of quality control a conventional rule settles the question: the unknown quality of the lot is a priori uniformly distributed.

The same situation prevails in experiments where the toxicity of a poison is to be determined by tests, in which the poison is sprayed on insects. There are different methods to define the unknown parameter; only one, the probability that an insect will be killed, corresponds to our conventional rule. The toxicity changes with the concentration of the poison, and for every concentration the hypothesis has to be made separately. This remark shows that hypothesis B has nothing to do with the real distribution of toxicities in nature.

Państwowy Instytut Matematyczny
The State Institute of Mathematics

THÉORÈMES ERGODIQUES ET LEURS APPLICATIONS

PAR
S. HARTMAN, E. MARCZEWSKI ET C. RYLL-NARDZEWSKI
(WROCŁAW)

Nous nous proposons de donner dans cette communication un aperçu raisonné des théorèmes ergodiques dans leur variante discrète, de quelques unes de leurs généralisations et de leurs applications. Nous y rappelons les théorèmes connus, envisagés en particulier dans les beaux travaux de F. Riesz [20, 21] et signalons les résultats récents, surtout ceux contenus dans les travaux de Ryll-Nardzewski qui se trouvent en préparation pour Studia Mathematica. Les démonstrations seront ici pour la plupart omises ou seulement esquissées; celles des théorèmes connus sont à trouver dans les travaux cités (voir p. 122-123), dont les numéros sont indiqués en crochets.

Cette communication est en même temps un rapport sur les études et recherches concernant les théorèmes ergodiques, faites à Wrocław par le Groupe des Fonctions Réelles de l'Institut Mathématique de l'État. Ce sont justement les travaux précités de F. Riesz qui ont été le point de départ de ces recherches.

1. Transformations mesurables. Nous entenderons par l'*espace*, et désignerons d'ordinaire par X, un ensemble abstrait fixé, par *ensembles mesurables* — les ensembles appartenant à un σ-corps fixé $\boldsymbol{M}$ de sous-ensembles de X, et par *mesure* — une fonction fixée $\mu(E)$ d'ensemble, définie pour tout $E \epsilon \boldsymbol{M}$, non-négative, finie[1]) et σ-additive.

Pour simplifier les énoncés, nous supposons une fois pour toutes que $\mu(X)=1$.

Les *fonctions* seront entendues partout comme fonctions réelles définies dans l'espace tout entier. Nous adoptons pour elles les définitions habituelles de la *mesurabilité*, de l'*intégrale*

[1]) Si l'on admet les mesures *σ-finies*, c'est-à-dire pour lesquelles X est somme d'une suite d'ensembles de mesure finie, il faut modifier convenablement les énoncés et les démonstrations.