1. Introduction[1]

In many cases where a large number of similar things are inspected, only a limited number of these items are examined. This is sometimes necessary because the article gets destroyed or becomes useless as a result of this examination, but as a rule this is done because one economizes. That portion of it which is examined is called the sample and the number of items present in the sample is called the sample size.

If each item in the aggregate or population has an equal chance of being selected the sample is called a random sample. A systematic study of random sampling is made in the manufacturing industry (e.g. the inspection of produced articles), in medical investigation (e.g. the comparison of medicines), in market analysis (e.g. official enquiries), for population estimates and in numerous other fields. The question arises whether random sampling which is so economical, may be useful in accountancy work.

Naturally we cannot compare an accountancy check with that of a lot of goods in all aspects. An important difference lies in the consequencies of drawing incorrect inferences from a sample. A batch of goods incorrectly approved may often be replaced without involving inordinate expenses as soon as the defective items have been discovered. If, however, a fraud is not discovered in time it may have very serious repercussions. In medical research similar cases sometimes occur when in comparing medicines by means of a sample an incorrect choice will have serious consequencies.

It is clear that sample inspection will yield less information than a complete examination.[2] The most important point is the nature of the deductions which definitely can be made on the basis of the information obtained from a sample. In mathematical statistics this question is examined and answered from two different points of

----

[1] Translated from Rapport S 274, "Toepassingen van aselecte steekproeven bij accountantscontroles" by J.A. Uijterlinden.

[2] Strictly speaking this only applies when at an examination of all items the inspection is carried out with the same amount of care as at an examination which restricts itself to a sample only. In cases where there is a risk that each item is subject to a less careful inspection as a result of the monotony of the work or of fatigue, sample inspection need not necessarily lead to less information than a complete investigation.

view. The principle of both methods of reasoning is briefly given below.

Assume that a long sequence of entries has to be examined for errors. A number of entries will be correct and some of them incorrect. A definite fraction p varying between nought and one will, therefore, exist indicating which portion of the entries is classed as "incorrect". This fraction (of the entries marked "incorrect") is called the fraction p and we want to examine this fraction on the basis of the results of the sample. Let us first assume that we have to estimate the value of p.

If we wish to ascribe a single number to the value of the estimate of p then it is obvious that we choose the fraction of incorrect items in the sample. If the sample size were 100 and the number of incorrect entries 10 then we would get as an estimate of p the number $f = \frac{10}{100} = 0.10$. If the sample does not contain all the entries f is not necessarily equal to p. It is also obvious that p will usually have a value close to that of f and not one deviating much from f. We can, therefore, give a range within which p presumably lies, a socalled confidence interval.

Therefore we do not ascribe a definite value to p but we give limits g_l and g_r within which the unknown fraction p probably lies (see fig. 1). We could for example give g_l the value 0.046 and g_r the value 0.186 on the basis of the value 0.10 found for f.

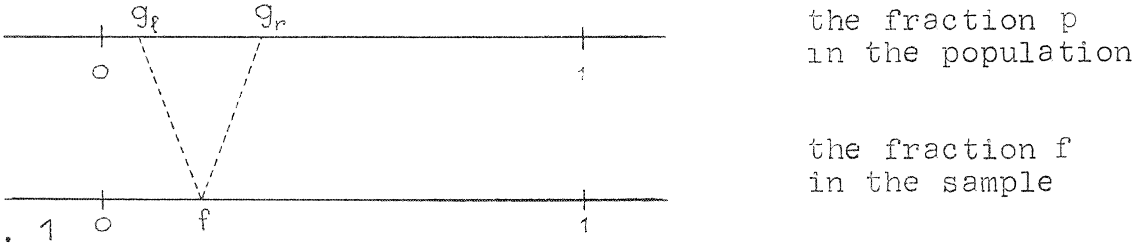
g_l
g_r
0
1
the fraction p in the population
the fraction f in the sample
0
f
1

fig. 1

The relationship between the fraction f in the sample and the fraction p in the population

A confidence interval need not necessarily have an upper limit and a lower limit. If, for instance we deduct from the sample that p is not greater than a certain value, there would only be one limit. In the given example the conclusion could be: the value of p is at most equal to 0.186.

Although p probably has a value not deviating much from f we may not exclude other values of p. It would actually be possible that our sample contains all incorrect entries of the list. In this case p can have a value which is much smaller than that of f. The conclusion we draw from these considerations is that even our statement in the form of an interval may be incorrect. The larger the interval we give, the less will be the number of inaccurate conclusions we make, but this advantage implies also the disadvantage of a less accurate prediction of p. We are now interested in the relationship between the fraction f found, the limits g_l and g_r of the confidence interval and the probability that p actually lies in this interval.

Unless we limit ourselves to trivial information, such as: p has a value between 0 and 1, mistakes will usually occur in a long series of statements about unknown fractions. We now arrange our statements in such a way that not more than a certain fraction of statements is incorrect in this series. How large we choose the size of this fraction depends on the nature of the investigation. Therefore, we will choose a smaller fraction for tests on an important medicine than at a preliminary inquiry in the field of market analysis. The values usually chosen for the permissable fraction of incorrect statements are 0.01 and 0.05 respectively. This fraction, indicated by α_0 is called the significance level. Once this is fixed, mathematical statistics enable us to calculate a confidence interval for each sample size n and for each fraction f found, such that the average fraction of incorrect statements equals the selected significance level α_0. In this way the above mentioned interval $0.046 \leqslant p \leqslant 0.186$ is deduced for $\alpha = 0.02$. This interval can, therefore, be given when the sample size n is one hundred, the chosen significance level is 0.02 and the value found for f is 0.10. If we only fix an upper limit g_r then the prediction can only be incorrect if g_r is too small. If we, therefore, conclude from the sample that p is less than or at most equal to 0.186, then the probability of an incorrect prediction will become less than 0.02 and we can prove that this probability is not more than half of 0.02, and equal to 0.01.

Usually in the field of accountancy only the upper limits g_r of the unknown fractions p are of importance. In tables I and II these upper limits are given for the values $\alpha_0 = 0.01$ and $\alpha_0 = 0.05$ respectively. The values of the sample size chosen are n=50, n=100, n=200, n=300, n=400 and n=500. The first column contains the number of inaccuracies k found in the sample; in using these tables it is, therefore, not necessary to calculate the fraction $f = \frac{k}{n}$ first. It should be emphasized that the tables may only be used when the sample contains only a small portion of the whole population.

Table I[1]

Upper limits of the fraction of inaccuracies in the population for $\alpha_0 = 0.01$

Number of inaccuracies in sample k	Sample size n					
	50	100	200	300	400	500
0	0.053	0.027	0.014	0.009	0.007	0.006
1	0.106	0.054	0.028	0.018	0.014	0.011
2	0.141	0.073	0.037	0.025	0.019	0.015
3	0.172	0.089	0.045	0.031	0.023	0.018
4	0.201	0.104	0.053	0.036	0.027	0.021
5	0.228	0.119	0.061	0.041	0.031	0.024
6	0.254	0.133	0.068	0.046	0.034	0.027
7	0.279	0.147	0.075	0.050	0.038	0.030
8	0.304	0.160	0.082	0.055	0.042	0.033
9	0.328	0.173	0.089	0.060	0.045	0.036
10	0.351	0.186	0.095	0.064	0.048	0.039
11	0.374	0.198	0.101	0.068	0.051	0.042
12	0.396	0.210	0.108	0.073	0.054	0.044
13	0.418	0.222	0.114	0.078	0.058	0.046
14	0.440	0.234	0.121	0.082	0.061	0.049
15	0.462	0.246	0.127	0.086	0.064	0.052
16	0.483	0.258	0.133	0.090	0.067	0.054
17	0.504	0.270	0.140	0.094	0.070	0.056
18	0.524	0.281	0.146	0.098	0.073	0.059
19	0.544	0.292	0.151	0.102	0.076	0.062
20	0.564	0.303	0.157	0.106	0.079	0.064

[1] Tables I and II are reproduced from H.C. HAMAKER: "Average Confidence" limits for binomial probabilities, Review of the International Statistical Institute 21 (1953), 17-27.

Table II

Upper limits of the fraction of inaccuracies in the population for $\alpha_0 = 0.05$

Number of inaccuracies in sample k	Sample size n					
	50	100	200	300	400	500
0	0.037	0.019	0.010	0.006	0.005	0.004
1	0.074	0.038	0.020	0.013	0.010	0.008
2	0.106	0.054	0.028	0.018	0.014	0.010
3	0.134	0.068	0.034	0.024	0.018	0.014
4	0.161	0.082	0.042	0.028	0.021	0.016
5	0.186	0.096	0.048	0.032	0.024	0.020
6	0.211	0.108	0.055	0.037	0.028	0.022
7	0.235	0.122	0.062	0.041	0.030	0.025
8	0.258	0.134	0.068	0.046	0.034	0.028
9	0.282	0.146	0.074	0.050	0.038	0.030
10	0.304	0.158	0.080	0.054	0.040	0.032
11	0.326	0.170	0.086	0.058	0.044	0.035
12	0.348	0.182	0.093	0.062	0.046	0.038
13	0.370	0.192	0.099	0.066	0.050	0.040
14	0.392	0.204	0.104	0.070	0.052	0.042
15	0.414	0.216	0.110	0.074	0.056	0.044
16	0.434	0.227	0.116	0.078	0.058	0.047
17	0.455	0.238	0.122	0.082	0.062	0.049
18	0.476	0.250	0.128	0.086	0.064	0.052
19	0.496	0.262	0.134	0.090	0.068	0.054
20	0.516	0.273	0.139	0.094	0.070	0.056

In comparing table I and table II we note that, for a fixed sample size and for an equal number of inaccuracies, the interval is wider for a small level of significance than for a larger one. By means of simple calculations it can be shown that, for equal confidence coefficients and for equal fractions of inaccuracies in the sample, a shorter interval is given as the sample size n increases. In table I we see, that for n=50 and k=2 the upper limit for p=0.141, and for n=100 and k=4 the upper limit for p=0.104 and for n=500 and k=20 we find 0.064. Both conclusions agree with what one would intuitively expect.

Summarizing we may state the following: in order to determine the unknown fraction p we have done away with the complete analysis by taking a sample that enables us to find an interval within which p probably lies, and at the same time it appears to be possible to choose these intervals in such a way that the fraction of incorrect statements in a long range averages a prefixed value α_0. Tables I and II give the upper limits for p at levels of significance $\alpha_0=0.01$ and $\alpha_0=0.05$ for different sample sizes. They may be used provided that the sample size is small in relation to the whole population. The problem of finding an estimate for p from a sample has thus been solved. We give some examples in paragraph 3.

Sometimes another problem arises, related to the one of estimating an unknown fraction p. In many cases we have to examine whether the quality of a lot of goods or that of a list of entries is such, that they can be accepted. In these cases we are not so much interested in a confidence interval for the fraction p as in the risk of accepting a lot where it actually should have been rejected. The possibility of erroneously accepting a lot when not all items are examined, always exists,because the sample may not contain one single bad or incorrect item, although a great number of these are present in the whole lot or population.

Before accepting the lot or the list we can take a random sample of size n first and count the number of items which are defect. This number is indicated by k. When this number is equal to,or greater than k_0,the lot is rejected and when it is smaller than k_0 the lot is accepted. Suppose that a lot is unacceptable when the fraction of defects is p_0 or larger. The probability β that the lot is erroneously accepted will in this case depend on the sample size n, the rejection limit k_0 and the value of p_0. The relationship between these quantities is deduced in mathematical statistics. The following example will make this clear.

Suppose that a register with entries will be unacceptable if the fraction of inaccuracies is greater than or equal to $p_0=0.10$. Let us also assume that, if we should find $k_0=7$ or more incorrect entries in a sample of size n=100, the list will be rejected; the register will be accepted if there are less than 7 incorrect entries in the sample. The risk lies in the possibility that the sample contains less than 7 incorrect entries while the fraction of incorrect entries in the whole register is more than 0.10. The register is then wrongly accepted. The probability of being wrongly accepted can be calculated and amounts to β =0.12 at most.

One usually approaches the problem the other way round, and we set the condition that the probability of wrong approval may be $\beta = \beta_0$ (e.g. 0.01 again or 0.05) at most and we then determine what the limiting value k_0 should be in order to reject the list. If one wishes to accept lists at a fraction β =0.01 of the checks, when the fraction of wrong entries is more than 0.10, then we have to reject it as soon as the number of mistakes in a sample of size 100 is 4 or more. If we allow a fraction β_0=0.05 then rejection will take place when $k \geqslant 5$. Numerous tables exist in which the value of k_0 for different values of β_0, n and p_0 is given.

To resume: we can show that the complete check for finding out whether a lot of goods or a register of entries is acceptable, can be replaced by a sample inspection for which we can decide in advance the risk we want to take of erroneously accepting bad lots. By a suitable choice of the sample size n and the rejection limit k_0 the examination can, for example, be arranged in such a way that only a fraction 0.01 of the unacceptable lots or lists, may pass.

The mathematical approach to this problem is further explained in the appendix and the check for fraud as an application of the above discussion is given in paragraph 4.

Remarks

1. We should see clearly the difference between a statement formulated as a confidence interval and the kind of statements which are discussed in the second section of this paragraph. When a long series of confidence intervals with an average significance level of 0.01 is given, it means that an interval not containing the true p is given in 1% of all predictions. If we formulate the conclusion in the form of approving or rejecting lots of goods and if we set the condition that the probability of approving a bad lot is at most 0.01, it means that at most 1% of the cases in which a lot is bad will not be discovered. In the first case the total number of incorrect decisions is therefore 1% of all decisions and in the second case the total number

of incorrect dicisions is 1% of all decisions in which a bad lot is involved.

2. We can sometimes draw conclusions of both types from the same sample. We may then conclude that the lot with the one particular feature should be approved or rejected, and at the same time we can state a confidence interval for the fraction of items having another feature (see also remark 6 in paragraph 4).

2. Sampling in accountancy checks

In this paragraph we want to find out whether the methods described in paragraph 1 may be applied in the checks of an accountant.

The accountant often meets situations in which an opinion must be given about "lots of goods". A "lot of goods" may for instance exist of a balance sheet of debitors, the claims paid out during the course of one year by an insurance company or the list of weekly wages to be paid to the workmen of a factory.

When one has to check whether such a lot of goods may be accepted the question arises whether all entries should be checked separately or whether part of same will be sufficient. Strictly speaking it is in theory always possible to check everything, but whether this can be done in practice and whether it is justified from a financial point of view, is questionable in many cases. There are many cases where a large number of entries have to be checked and where the possible risks are not set off by the high cost of control. One has, therefore, often doubted whether the checking of all single items of a party is required under all circumstances. In this connection we refer the reader to different articles in the latest volumes of Maandblad voor Accountancy en Bedrijfshuishoudkunde. In practice it has unfortunately occurred that one has frequently been tempted to apply unsound methods in sampling which leaded to ambiguous deductions.

Only sampling methods based on modern methods which are scientifically sound, may be applied. Random samples, for instance, as described in the first paragraph enable us to draw exact conclusions. We emphasize that these conclusions are possible in all cases where random samples are taken. The deductions made are not influenced in any way be the nature of the material checked.

It is, for instance, unimportant whether possible errors only occur in a certain section of a list or whether frauds are carried out in a systematic way.[1] We may, therefore, draw the conclusion that the taking of random samples in accountancy checks is perfectly justified from a statistical point of view. Whether it has sense to do this in a particular case depends on the circumstances. Some examples of applications are given in paragraphs 3 and 4.

Remark

We have already stated in the first paragraph that in a random sample each item of the lot must have an equal chance of being selected. In the case of a list of entries this condition may be interpreted in two different ways.

The most obvious would be to consider the entries in the list as elements and to see to it that each entry in the list has an equal chance of being selected. However, we can imagine that the large entries are of more importance than the smaller ones and that the former should be given a greater chance of being included in the sample. This can be done preserving the nature of the random sample by considering all guilders entered in the list as elements, and not the separate entries. In this case a random sample is one in which each guilder has an equal chance of being selected. Strictly speaking we then only have to check the indicated guilder; in practical accountancy we naturally do not confine ourselves to this procedure but we examine the whole entry to which the guilder concerned belongs. However, this involves that large entries have a greater chance of being checked than the smaller ones, and this probability is directly proportional to the size of the entry. In paragraphs 3 and 4 we will show how in some cases the conception of the list with entries and how in other cases the one with the list with guilders is more suitable.

3. Accuracy Checks

We can imagine that the method of confidence intervals as described in the 1st paragraph can be applied to checks in accuracy.

----

[1] The conclusions stated by S. Kleerekoper ("De steekproeven als middel van accountantscontrole") in M.A.B. 10 (1933), p. 54 are, therefore, incorrect.

We draw a random sample of n entries from the whole aggregate, we ascribe a range to the fraction p of inaccuracies in all entries on the strength of the sample result and we then investigate whether the limits found will give rise to a closer investigation or to intervention.

We now give a few examples of the application of confidence intervals to accuracy checks.

1. A factory employs 2000 workers payed weekly. The wages have to be calculated anew each week and in order to see whether this is done accurately enough we draw a random sample of 50 from the 2000 calculations for a certain week. We find two mistakes. In this case we thus have n=50 and k=2. If we wish to work with a significance level of 0.01 (or 1%) we see from table I that the fraction of mistakes made, amounts to a maximum of 0.141 or 14% approximately.

   If we draw samples of this type regularly in the same industry or in different industries and if we choose an upper limit from table I, the significance level of 0.01 thus means that, on an average, an incorrect limit in one of the 100 decisions is given. If we choose a significance level of 0.05 then according to table II we can give p an upper limit g_r=0.106 or 10.6% for the given sample. Regular application of this method leads to an average of 5 incorrect decisions out of 100.

2. A useful application of the second method of paragraph 1 can be made by a weekly control of the quality of a pay office in order to check whether the number of mistakes made is permissible. Let us assume that a sample of 100 wages is checked each week and that we regard a list with 5% faulty calculations as just permissable. The list with wage calculations is accepted or not, depending on the number of mistakes in the sample. Rejection could mean that all wages would have to be calculated again. If we reject a list on the strength of finding two or more mistakes, we will accept the list wrongly when the sample contains one or no mistakes in spite of the fact that the fraction of mistakes in the whole list is greater than 0.05, the probability of which is 0.037 or 3.7%. The probability of erroneous approval is 0.118 or approximately 12% should we also accept lists with two mistake.

Such calculations may also be carried out for other values of the sample size n and for the permissable limit p_0. We refer to an article by H.E.J. Botje[1] for a more detailed discussion of the whole problem.

3. After the 1st April the treasurer of a large club with over 100,000 members has to send reminders to those members who have not or who have not fully paid up their membership fees. A random sample of 300 was drawn towards the end of April from those members who had not yet paid at the 1st of April, in order to check whether the reminding takes place quickly enough and whether the correct accounts were entered on the reminders. The checks are performed by means of the copies of the reminders and five cases are found where no reminder was sent at all or where an incorrect sum of money was mentioned. From this we may conclude that, with a level of significance of 0.05, at most a fraction $p \leqslant 0.032$ of all members who should be reminded, had received no notice or a wrong notice.

   Besides the examples given above there are numerous other checks in which by sampling we can obtain an insight into the accuracy with which has been worked. In this connection we mention the checking of invoices and of store administration.

4. Fraud checks

As in the case of checks of accuracy we may, in principle, use confidence intervals in checks on fraud. We will naturally confine ourselves to giving an upper limit for the fraction p of frauded entries and besides that we will also choose the level of significance α_0 small. If we proceed in this way we may then conclud that an interval, not containing the true p, is given in at most a fraction α_0 of all cases of a large number of checks.

Although this method of thinking is correct from a statistical point of view, there are two important objections against it. In the first place we are not interested in the fraction of frauded entries but mainly in the total sum of money frauded. Secondly, we will not limit ourselves to the checking of a sample of entries when a fraud

----

[1] H.E.J. BOTJE, Steekproefcontrole in de loonadministratie, Sigma 4 nr 5, 1958, 93-98.

is found, but we would rather proceed to the checking of all entries Both difficulties can be obviated.

Let us assume that frauds can only be achieved by entering sums of money that are too high or by adding entries which don't exist, or by entering higher totals than the actual ones or by carrying forward totals incorrectly. Two methods have been developed in order to overcome these difficulties.

P. de Wolff[1] suggested that all large posts be checked and the smaller ones only partially. If a case of fraud has been found,all entries are checked and we then only make probability statements when not a single fraud has been found in the sample.

An other method of overcoming the difficulties was developed by A. VAN HEERDEN. Certain difficulties encountered in DE WOLFF's method are not present in VAN HEERDEN's method, and we will, therefore, limit ourselves to an extensive discussion of the former. The procedure is as follows.

Let the register which has to be checked for fraud,consist of N entries and let the sum of all the entries be B guilders. We do now not consider the entries as single items, but the single guilders and we, therefore, act as if a list with B guilders, about which a deduction has to be made on the basis of a sample, is given. n Guilders are selected at random from the list and the examination may thus be reduced to checking whether these n guilders are actually present. However in most cases these n guilders will belong to entries. We now not only check the indicated guilders but the entire entries of which these guilders are part. Larger entries will have a greater possibility of being indicated in this way than the smaller ones and may even be indicated several times. We thus only check n entries at most. As soon as a case of fraud has been detected all entries are checked. The question now arises, how large n should be chosen in order to keep the risks sufficiently small. Here the risk for the accountant exists in not detecting a

----

[1] P. DE WOLFF, Steekproeven bij administratieve controle, Statistica Neerlandica 10 (1956), 35-44.

do , On the application stratified sampling to an auditing problem, Istituto di Statistica, Roma (vol. in honor of C. GINI, 21 pp.).

fraud while in reality there is a fraud and we can, therefore, set the following condition: if the fraction of fraudulated guilders is more than a fraction p_0 of the total number of guilders then the probability of not finding one fraudulated guilder in the check may be, at most, β_0.

This method of reasoning links up with the one at the end of paragraph 1. For here we are also interested in the probability β, that the list will be approved wrongly, i.e. the probability that the fraud will not be discovered, while in reality there has been a fraud. We can see, on intuitive grounds, that this probability decreases as there are more frauds and as the sample size increases. The probability β that frauds are not discovered is given as a function of the fraction p of fraudulated guilders for different values of the sample size n in fig. 2. As this probability depends on p we write $\beta(p)$ instead of β to make this clear. It appears that $\beta(p)$ definitely decreases for increasing p and for increasing n.

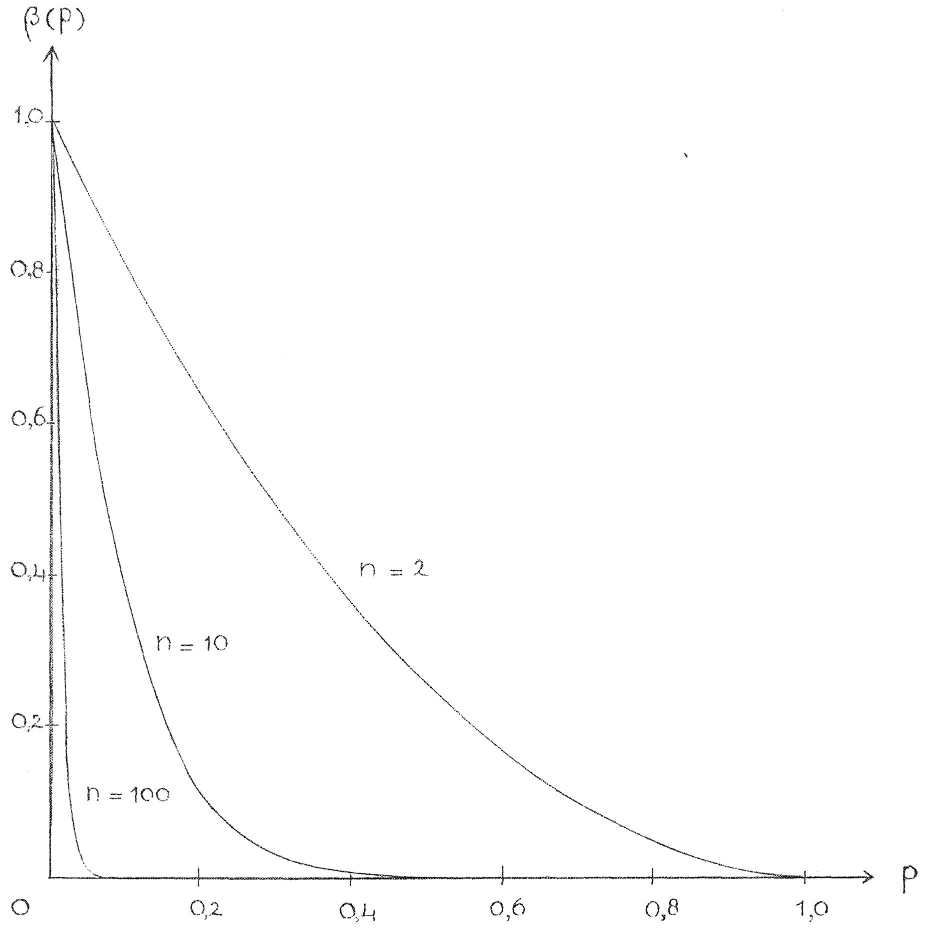

fig. 2

The probability $\beta(p)$ of not detecting a fraud as a function of the fraction p of frauded guilders and of the sample size n

The choice of n is still free and we may utilise this in order to set a certain condition for $\beta(p)$. It is obvious that we demand that a fraud larger than a fraction p_0, only has a small probability of not being discovered. We can, for instance, set the condition that this probability may at most be 0.01 for $p = 0.01$. We can calculate that n should be at least 459 for the values of β_0 and p_0 chosen (see appendix). Values of n for $p_0=0.05$, 0.01 and 0.001 and for $\beta_0=0.05$, 0.01 and 0.001 are given in table III as other values may also be chosen for p_0 and β_0.

Table III

Values of n for different values of p_0 and β_0

p_0 \ β_0	0.05	0.01	0.001
0.05	59	90	135
0.01	299	459	688
0.001	2995	4603	6905

In table III we see that β_0 may be reduced considerably without n increasing too strongly, but that we have to take voluminous samples for more stringent conditions relating to p_0.

Conclusion

The probability that a fraud larger than a fraction 0.01 of the total amount will not be detected is at most 0.01 if we take a random sample of 459 guilders from a total of B guilders, then check the entries to which these guilders belong and check all entries as soon as a case of fraud has been found. Formulating it in an other way: in a long sequence of checks at most 1% of the cases, where more than 1% is frauded, are not detected when we apply the check method by means of random samples as explained above. In addition to this we notice that the risk of not detecting a fraud may be considerably less (compare remark 1).

Remarks

1. We check more than the mere n guilders, on which the calculations are based as we not only check the indicated guilders, but the whole entry to which they belong. The probability of not detecting a fraud of more than 1% is, therefore, only exactly 0.01 in those cases in which each entry is either completely incorrect or not frauded at all. The probability can be considerably less than 0.01 in those cases where entries are partly frauded (e.g. by entering a too large amount). The exact values of these probabilities cannot be calculated without making further suppositions.

2. The fraud is always expressed as a fraction of the total amount B given. The actual amount comes to B' which is less than B in cases where fraud has taken place. The fraud should then be expressed as a fraction of B'. This is quite easy because a fraud fraction p in B' is equivalent to a fraud fraction $\frac{p}{1+p}$ in B and we could, therefore, replace p_0 by $\frac{p_0}{1+p_0}$ in what has been said previously. This substitution is of little influence for small values of p_0; we would thus have found n=463 instead of n=459 for $p_0=0.01$ and $\beta_0=0.01$. This correction is omitted in the light of what is said in remarks 1 and 5.

3. The samples as discussed above are random samples, i.e. samples in which each of the B guilders has an equal chance of being selected. This can be achieved by working with lists of random numbers. If, for instance, we have a total amount of B=30,000 guilders, we use a list with n random numbers between 0 and 30,000. Though, in principle, it is possible to carry out all checks with one long series of random numbers (which then should contain more than n random numbers) it is easier to make separate lists of n numbers between 0 and 10,000, between 0 and 15,000, etc.

   Not only the total amount B of the list with entries is known, but all sorts of partial sums can also be found easily in many cases. The search for the guilders chosen at random,can then be simplified by making lists of n random numbers which are arranged according to magnitude. We, therefore, must have lists of n ordered random numbers between 0 and 10,000, 0 and 15,000, etc., up to 0 and 95,000 at our disposal. These lists may also be used when the total amounts to more than 100,000 guilders. If for instance, B=200,000, we use the list of 0 to 20,000 and we interpret guilders as being ten-guilder pieces.

4. When searching for an indicated guilder we need not count from the beginning if all kinds of partial sums of the list are known. Let us assume that these partial totals are the amounts carried forward to the next page. If we now have to check the 6530th guilder, we look up the page with a number smaller than 6530 at the top and larger than 6530 at the bottom and we start counting from the top of the page.

   A fraud in the addition on the page will be found when it appears that the page with 6000 at the top and 6800 at the bottom contains only 500 guilders. The indicated guilder then is a guilder from a completely frauded "entry of 300 guilders". A fraud committed by carrying forward incorrectly is detected when in searching for the 6530th guilder, it appears that the one page finishes at 6500, while the next one starts at 7000. Frauds in the adding and in the carrying forward of sums are, therefore, also checked by means of the check described above.

5. In connection with what we said in remark 3, we can further reduce the risk of not detecting a fraud by inserting a new random number in those cases where the jump between two consecutive random numbers is more than 1% of the total amount B. In general we need only a few new random numbers and we thereby exclude the possibility that a fraud of more than 1% of B in a single entry is not noticed.

6. We can apply the suggestion made in remark 2 of paragraph 1 to use one and the same sample for the examination of two different features by treating a check both as a check on fraud and as a check on accuracy. When working with a sample of size n=459 we can conclude with the risks stated above whether a fraud was carried out or not and, at the same time, we can give an upper limit for the fraction of inaccuracies in the whole list by means of table I or table II.

Appendix

Let us assume that we have an aggregate (a batch of goods, a list with entries, the population of a city), the elements of which possess either the feature F or the feature G. There is, therefore, a fraction p possessing the feature F and in many cases it is of importance to check whether this fraction is equal to a certain value p_0. We limit ourselves to cases where this can be done by means of a sample.

Before taking the sample we make the supposition that p has a value p_0 and we call this the null hypothesis (H_0). All other values that p can assume, i.e. all values between 0 and 1 and unequal to p_0, together form the alternative hypothesis (H_1) and we say that the null hypothesis H_0 is tested against the alternative hypothesis H_1.

The result of the sample will not agree with the hypothesis made when the fraction f of the elements in the sample with the feature F is either much larger or much smaller than p_0, and the null hypothesis $p=p_0$ will then be rejected. All values of f which lead to rejection, form the critical region Z, a region which is split in two sections Z_l and Z_r. The test in this case is called a two-sided test (see fig. 3).

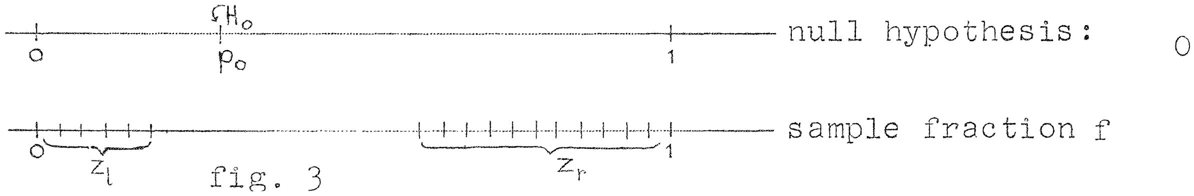

fig. 3

Two-sided test of the hypothesis $p=p_0$

Except for cases, where we want to examine whether p has a value p_0, there are also cases in which we want to check whether p has a value $\leqslant p_0$. We then test the null hypothesis $p \leqslant p_0$ against the alternative hypothesis $p > p_0$ and we will reject the null hypothesis when the sample fraction possesses a high value. The critical region, therefore, only consists of high values in the interval 0-1 so that there is only a right critical region Z_r and we, therefore, call it a right one-sided test (see fig. 4).

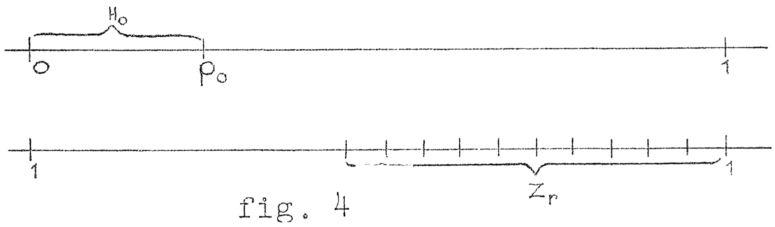

fig. 4

Right one-sided test of the hypothesis $p \leqslant p_0$

The reverse case may also occur, namely the testing of the null hypothesis $p \geqslant p_0$ against the alternative hypothesis $p < p_0$. There is a left-critical region consisting of low values of f and we call it a left one-sided test (see fig. 5)

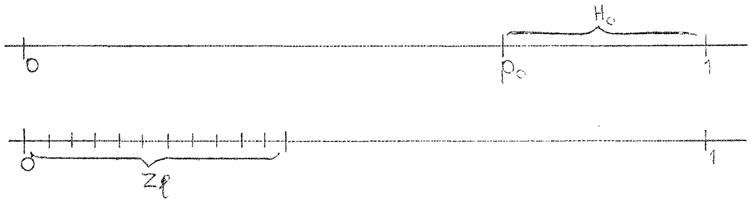

fig. 5

Left one-sided test of the null-hypothesis $p \geqslant p_0$

In all three cases the following holds: in general sample results can also be found in the critical region when the null hypothesis is correct. Thus rejecting the null hypothesis wrongly, cannot be excluded; we call this an error of the first kind. Its probability is usually indicated by α and is called the level of significance of the test, or the probability of an error of the first kind.

On the other hand it cannot be excluded that a value is found in the sample, outside the critical region, while one of the alternative hypothesis indicates the correct value of p and thus the null hypothesis is incorrect. The null hypothesis is then not rejected wrongly. This is known as an error of the second kind, the probability of which is indicated by β. The different possibilities in testing are given in table IV.

Table IV

Possible combinations when testing a null hypothesis

	H_0 correct	H_0 not correct
no rejection of H_0	rightly	error of the 2nd kind
rejection of H_0	error of the 1st kind	rightly

The probabilities α and β depend on the choice of the critical region and on the sample size. In general the probability of an error of the first kind is small when we select a small critical region, but in this case the probability of an error of the second kind is large. The reverse generally holds when we have a large critical region. Usually we arrange the test in such a way that the level of significance α is at most equal to a prefixed value α_0 (e.g. 0.01 or 0.05). We, therefore, first select α_0 and we then fix the critical region in such a way that the probability of an error of the first kind is $\leqslant \alpha_0$. Let there be different critical regions satisfying this condition. We may then use the freedom left to choose that region for which the probability of an error of the second kind is as favourable as possible. We can usually also minimize β by enlarging the sample size.

We will now apply these ideas to the problem of checking frauds by means of samples.

Consider a population of B guilders which may have two features: frauded or not frauded. We indicate the fraction of fraudulent guilders by p. The check whether a fraud has taken place means, from a statistical point of view, that we want to test whether $p=0$. The null hypothesis, therefore, is $p=0$ and this is tested against the alternative hypothesis $p>0$. The null hypothesis is rejected when one or more cases of fraud are detected. There is, therefore, a right critical range and we apply a right one-sided test. If we indicate the total number of fraudulent guilders in the sample by k, then the critical range consists of all values of k which are equal to or greater than 1.

The probability of an error of the first kind, i.e. the probability α of an incorrect rejection of the null hypothesis is nought, as the probability of one or more frauds in the sample equals nought when fraud has not been practised at all.

The probability of making an error of the second kind, i.e. drawing the conclusion that fraud has not been practised while it actually has been, is found by bearing in mind that this error can only be made when there are no cases of fraud in the sample, i.e. $k=0$. Furthermore, a fraud of fraction p of the total sum B means that there are p B frauded guilders and (1-p) B non-frauded guilders. If we now select n guilders at random, we can calculate the probability that none of them is frauded. This probability is

$$\frac{\binom{(1-p)B}{n}}{\binom{B}{n}},$$

which is closely approximated by $(1-p)^n$ for a value of n much smaller than B. The probability of an error of the second kind

$$\beta(p) = (1-p)^n \qquad (1)$$

is, therefore, a function of both p and n. This is the probability which was plotted as a function of p for different values of n in fig. 2.

In mathematical terms, the conclusion of page 14 is therefore:

If we select 459 guilders at random and if we take all sample results with one or more frauds as critical region, we apply a right one-sided test with significance level nought and with a probability of at most 0.01 that a fraud bigger than 0.01 of the total amount will not be detected.

The size of the sample is, therefore, determined by means of formule (1), where we only make use of the fact that n guilders are selected at random and where was proved in checking that none of these were frauded. In reality not only these n guilders are checked but also the complete entries to which they belong. We can now wonder whether the probability of <u>not</u> detecting a fraud becomes smaller as a result of checking far more than n guilders.

We can prove that this can actually be the case when sums of money which are too high are written down for some entries, but that the probability does not become smaller if all entries are either correct or completely frauded. We illustrate both statements by means of the following example.

Suppose we have a list with a total sum of B guilders, which are divided over $\frac{B}{10}$ entries of f 10.-. If a total of F guilders has been frauded, the fraction of frauded guilders is $p = \frac{F}{B}$. If we now select n guilders and if we check only these n guilders then the probability of not detecting the fraud is $(1-p)^n$ again.

If we select and check entries the probability of not detecting a fraud depends on the way in which fraud has taken place. Let us first consider the case where an entry has not been frauded at all or has been completely frauded. This means that $\frac{F}{10}$ of the $\frac{B}{10}$ entries have been frauded, i.e. a fraction $\frac{F}{B} = p$. If n is also small in relation to the total number of entries, then the probability that the fraud of a fraction p is not detected is again $(1-p)^n$.

The number of frauded entries is F in the case where the fraud amounts to only one guilder for each frauded entry. The fraction of frauded entries then amounts to $\frac{F}{B/10} = \frac{10F}{B} = 10p$. The probability of not detecting a fraud is now $(1-10p)^n$ and this probability is much smaller than $(1-p)^n$. A similar reasoning holds if the fraud is not *f* 1.- for each frauded entry, but an other amount which differs from entry to entry.

In the method applied guilders are selected and entries are checked. If an entry is not selected twice then n entries are checked and it follows from the above reasoning that the probability of not detecting a fraud can be considerably less than $(1-p)^n$. If in our example not n but n_1 entries are indicated by the n guilders (n_1 therefore being smaller than n), then the probability of not detecting a fraud becomes $(1-10p)^{n_1}$ in the case of frauds of *f* 1.- for each frauded entry and this probability is larger than $(1-10p)^n$. On the other hand the original reasoning that in each case n single guilders are checked still applies, so that the probability of not detecting a fraud is definitely smaller or equal to $(1-p)^n$.

The conclusion is, therefore, that in a check, where guilders are selected and entries are checked, the probability of not detecting a committed fraud may actually be smaller than the fixed limit: for a sample of size $n=459$ it is therefore smaller than 0.01. In some cases this probability can even be considerably smaller.