STAN ARONOFF*
TES Research and Consulting Ltd.
Calgary, Alberta T3A 2G6, Canada

Classification Accuracy: A User Approach

The importance of considering both the probability of incorrectly rejecting an acceptable map (producer's risk), as well as the probability of accepting an inaccurate map (consumer's risk) are emphasized.

ABSTRACT: *One of the important criteria used in selecting a remote sensing system is the accuracy of the information it can provide. This paper discusses the statistical basis of map or classification accuracy estimation with reference to the binomial distribution and its normal approximation. The importance of considering both the probability of incorrectly rejecting an acceptable map (producer's risk), as well as the probability of accepting an inaccurate map (consumer's risk) are emphasized. When remote sensing systems are used operationally, unnecessarily re-checking a map and the delay in transmitting the information to the user can be costly. This cost can be considerably reduced by designing a sampling system which provides for both low consumer's and producer's risks.*

INTRODUCTION

OVER THE PAST DECADE remote sensing applications have been developed which can meet a wide variety of mapping information needs. The user can generally choose from several remote sensing systems. One of the important criteria used in selecting a remote sensing system is the accuracy of the information it can provide. The accuracy is commonly assessed by selecting a sample of points from the map product and comparing the map classification with some verification data.

The objective of this paper is to review the theory and application of classification accuracy tests. It is addressed to both users and producers of what are termed "attribute maps," such as maps of land use, soils, and vegetation. Attention is focused on the user's need for an effective test to estimate the classification accuracy of specific map classes or remotely sensed image classes.

The selection of a sampling design is of more than academic interest. As an example, in northern Alberta the accurate identification and avoidance of organic soils produces considerable savings in road construction costs. Thus, inaccurate mapping of these soils can be very costly. However, there are also real costs associated with rejecting as inaccurate a map which actually is of acceptable standard. These costs include field and office time for re-checking and costs to users for the delay in obtaining the information. Similarly in comparing the classification accuracy of remote sensing systems, a suitable inexpensive system might be rejected in favor of a more expensive method because the sampling design was not powerful enough to identify the system as "sufficiently accurate."

The criteria for judging any proposed sampling design have been summarized by Ginevan (1979) as follows:

- There should be a low probability of accepting a map of low accuracy,
- There should be a high probability of accepting a map of high accuracy, and
- A minimum number of ground data sample points should be required.

[*] Presently at the Department of Forestry, University of California, Berkeley, CA 94720.

0099-1112/82/4808-1299$02.25/0

In order to decide whether a map is of acceptable accuracy, a sample of map points is checked against ground data and a probabilistic statement is made about the true accuracy of the map. This statement generally claims some minimum level of accuracy with some high level of confidence, e.g., a minimum of 85 percent accurate at the 95 percent confidence level. The sampling problem, therefore, is one of determining the optimal number (N) of map samples to be compared with ground data, and an allowable number (X) of misclassifications of these samples. After these values are determined, N map samples can be selected and their classifications compared against the true classification of the sample point (e.g., ground data). If X or fewer points are misclassified, then the map can be accepted as accurate at the specified level of precision.

Hypothesis Testing

Map or classification accuracy estimation can be viewed as a hypothesis test. The strategy of hypothesis testing is to state the problem in terms of two mutually exclusive choices, then to accept the conservative hypothesis (null hypothesis) unless there is a low probability of it being true. For map accuracy assessment the test may be stated as a null hypothesis (H_0) and alternate hypothesis (H_1) as follows:

H_0: The map is less accurate than required.
H_1: The map accuracy is equal to or greater than that required.

(In the appendix, a more formal statement of the null hypothesis and an alternate formulation often found in the literature are given. The way in which the alternate formulation can give misleading results is also discussed.)

This test can give two correct and two erroneous decisions. The two types of correct decisions are (1) to accept a sufficiently accurate map, and (2) to reject a substandard map. The two types of erroneous decisions are (1) to accept a substandard map, termed the consumer's risk, and (2) to reject a sufficiently accurate map, termed the producer's risk. (Type I and II errors are discussed in the appendix.)

The terms "consumer's risk" and "producer's risk" are taken from a branch of statistics known as acceptance sampling. Used extensively in the manufacturing industry for quality control, acceptance sampling theory considers problems essentially the same as the evaluation of map accuracy.

Evaluating Classification Accuracy

To illustrate the statistical theory of classification accuracy testing, consider a map of some unknown accuracy. A sample of N points can be randomly selected and the "true" classification of each point determined. (It is assumed that misclassification of a site can be unambiguously determined.) The proportion of correctly mapped points can then be calculated. If this process were repeated and another N points were selected from the same map, the proportion of correctly classified points would probably be different. Repeating this process a large number of times and tallying the frequency that each value of "proportion correct" occurs, would generate a sampling distribution which could be graphed as the proportion correct or accuracy *of the sample* against frequency or probability of obtaining the value. Instead of performing these iterations, a mathematical model can be selected (such as the normal or binomial distribution) which is considered to best represent the distribution of sample proportions which would have been obtained if N points had been randomly selected from a map with accuracy Q_L.

This is illustrated in Figure 1. (The normal distribution is used here for ease of illustration though, as discussed below, it is often not appropriate for map accuracy estimation.) For a test at the 95 percent confidence level, Q_L is set at the minimum accuracy and H_0 will be accepted unless there is a 5 percent (0.05 probability) or less chance that the map's accuracy is less than Q_L. (Since the probability of exceeding a maximum level of accuracy is not generally of concern here, only one tail of the probability distribution is used for calculations.) The value Q_T is calculated such that the probability of obtaining a sample value as high or higher than Q_T is 5 percent, assuming that the map has an accuracy of Q_L (see Figure 1). If the proportion of correct points in the sample is equal to or greater than Q_T, then the map's accuracy is considered to be equal to or greater than Q_L and the map is accepted. If the sample accuracy is less than Q_T, the map is rejected because it is considered that there is too great a probability that the actual accuracy of the map being tested is less than the accuracy required.

To illustrate the determination of producer's risk consider, as example, the probability dis-

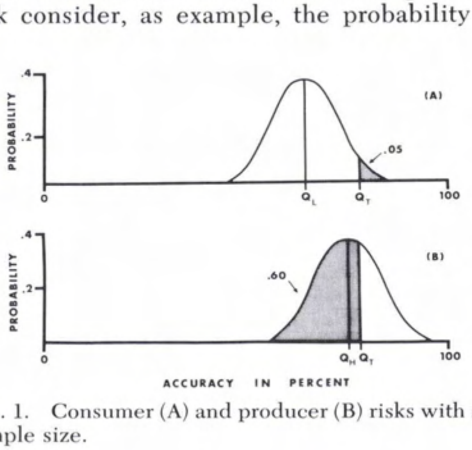


FIG. 1. Consumer (A) and producer (B) risks with small sample size.

tribution, shown in Figure 1B, of the same N test points randomly selected now from a map which actually has an accuracy Q_H, where Q_H is higher than Q_L. The producer's risk B (shaded area) is the probability that the proportion of correct points is less than Q_T; in this case, $B = 0.60$ or 60 percent. This means that, in *designing this sampling test with emphasis on ensuring that low accuracy maps are rejected, we have made it difficult for a map which actually meets these requirements (i.e., a map with a true accuracy Q_H) to pass the test and be accepted.* In fact, there is a 60 percent chance that this map would be rejected.

The producer's risk could be reduced by increasing the consumer's risk (i.e., reducing Q_T) or by increasing the sample size. Figure 2 illustrates the effect of increased sample size which reduces the sample variance, thereby "narrowing" the sampling distribution. Note that the value of Q_T can now be lower and still maintain the same consumer's risk of 5 percent while giving a lower producer's risk, in this example, of 10 percent.

Map accuracy assessment tests discussed in the literature have usually considered only the consumer's risk, generally set at 5 percent. However, for small sample sizes this may result in producer's risks on the order of 60 to 70 percent. Thus, while there is a low probability of accepting an inaccurate map, there is a high probability that an accurate map will fail the test and incur unnecessary costs such as re-checking and the delayed availability of information to users.

Accuracy Estimation Methods

Accuracy estimation methods presented in the literature differ primarily in the way they resolve two questions. The first is which mathematical model to use. The second involves how to deal with the fact that for a given map different classes will be mapped with different levels of accuracy.

THE MATHEMATICAL MODEL

The binomial distribution is considered to be the appropriate model when sampling is conducted under the following conditions:

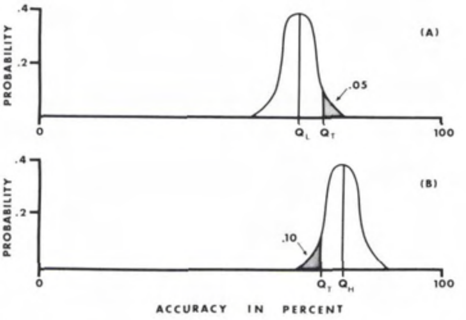


FIG. 2. Consumer (A) and producer (B) risks with large sample size.

- Each trial or each item selected can be assigned to one of two categories (e.g., correct versus incorrect),
- The probability of obtaining a "correct" result is the same for each trial,
- Each trial is conducted independently of any other, and
- A fixed number of trials are performed (Guenther, 1977).

In the field of acceptance sampling and in the recent literature on map accuracy estimation, the binomial distribution is considered the most appropriate mathematical model. Some researchers have calculated binomial probabilities from the binomial probability function itself, while others have used the normal approximation.

Single sample plan based on the binomial. Ginevan (1979) has approached the problem of map accuracy estimation as a problem in acceptance sampling and has calculated exact binomial probabilities by computer. The binomial probability is calculated as follows:

$$P(s) = \frac{N!}{(N-s)!\,s!} Q^s\,(1-Q)^{N-s} \qquad (1)$$

where $P(s)$ = the probability of obtaining s correct points in the sample,
N = sample size, and
Q = accuracy of the map, i.e., proportion of correctly classified points.

This formula can be more conveniently written to express the probability of a given number of misclassifications (X); thus, $s = N - X$ and, by substitution,

$$P(X) = \frac{N!}{X!(N-X)!} Q^{N-X}\,(1-Q)^X. \qquad (2)$$

To design the test, a minimum acceptable map accuracy (Q_L), a sample size (N), and a consumer's risk (A) (usually 0.05) are selected. Then the largest number of allowable misclassifications (X') is found such that the cumulative probability of having X' or fewer misclassifications is less than or equal to A. That is, find X' such that

$$A \geqslant \sum_{X=0}^{X'} \frac{N!}{X!(N-X)!} Q_L{}^{N-X}\,(1-Q_L)^X. \qquad (3)$$

The producer's risk B, the probability that a map of some high accuracy Q_H will have more than X' misclassifications and thus be rejected, can be calculated as follows:

$$B = \sum_{X=X'+1}^{N} \frac{N!}{X!(N-X)!} Q_H{}^{N-X}\,(1-Q_H)^X. \qquad (4)$$

Because the binomial distribution is discrete, several values of N have the same value of X. Increasing the sample size within the range that X

takes on a certain value, increases the producer's risk while exceeding the required consumer's risk. Thus, given Q_L, Q_H, and the consumer's risk A, it is the lowest value of N for a given X which will give the lowest producer's risk.

Ginevan (1979) has produced tables which allow sample sizes and allowable number of misclassifications to be conveniently evaluated for values of $Q_L = 0.85, 0.90$; $Q_H = 0.90, 0.95, 0.99$; and $A = 0.01, 0.05$. (Note that consumer's and producer's risks are designated A and B, respectively, in this paper and as β and α in Ginevan's (1979) paper.)

Suppose a minimum acceptable accuracy of 75 percent were chosen for Q_L, and the same value were chosen for Q_H. Then a map accuracy test which gives a 5 percent chance of accepting an inaccurate map (e.g., a map slightly less than 75 percent accurate) would give a 95 percent chance of rejecting a map with a true accuracy of 75 percent. When Q_H is chosen equal to Q_L, then the consumer's risk and producer's risk sum to 100 percent. However, the chance of rejecting a map of 85 percent accuracy (Q_H = 85 percent) might only be 60 percent. Thus, selection of the value Q_H determines the value of the producer's risk that will be calculated.

The sample size N, or values of the other variables, should be adjusted until a suitable design is obtained which reflects the costs of potential errors. If the consequences of a misclassification are not costly, then a higher consumer's risk or lower accuracy level might be acceptable. This would allow fewer samples to be checked, thereby reducing the cost of the map accuracy test as well. It should be recognized that, as required accuracy levels become higher and as consumer and producer risks are specified lower, the cost of performing the test increases considerably faster than improvements in the test parameter values. The extra bit of accuracy or risk improvement comes at a disproportionately high cost. Ultimately, a judgement must be made as to which parameters of the test are appropriate for the situation.

Curtailed sampling with the binomial. Having selected a sampling design to test classification accuracy and randomly selected all the points to be checked, it is usually not necessary to check each point. As soon as too many misclassified points are identified, the classification has failed the test and no further points need be checked. Similarly, if the number of points remaining to be checked is less than the remaining number of allowable errors, the classification has passed. When these "stopping rules" are used the sampling method is termed "curtailed." Curtailed sampling does not change consumer or producer risks; however, the sample size N becomes a random variable. The average sample size (ASN), using a binomial accuracy test for a map of accuracy Q, is calculated as follows (where $p = 1 - Q$:

$$ASN = \frac{X+1}{p}\,\mathrm{E}(X+2; N+1, p) + \frac{N-X}{1-p}\,(1 - \mathrm{E}(X+1; N+1, p)) \tag{5}$$

$$\text{where } \mathrm{E}(s; N,p) = \sum_{r=s}^{N} \frac{N!}{r!(N-r)!}\,p^r\,(1-p)^{N-r}, \tag{6}$$

r = number of misclassifications, and
p = proportion of misclassified points.
(after Guenther, 1977)

TABLE 1. COMPARISON OF AVERAGE SAMPLE NUMBER (ASN) VALUES

Actual Accuracy	*ASN*
0.80	10
0.85	13
0.90	19
0.95	32

Average sample number tends to be reduced as the actual classification accuracy of the maps being tested is reduced and as the required accuracy increases. Table 1 compares average sample sizes for a test of 90 percent minimum accuracy using a sample size N of 46 and allowable errors X of 1. Consumer's risk for this test is 5 percent.

The selection of a sampling design should consider the consumer's and producer's risks, the accuracy required, and cost of performing the test. Where verification of each sample point is costly, the average sample number will be an important factor in the selection of a sample design. However, to estimate the average sample size, some prior knowledge of the accuracy of the maps to be tested is required.

The normal approximation to the binomial. Under certain conditions the normal distribution can be used to approximate the binomial distribution. Rosenfield and Melley (1980) have discussed their use of this approximation in map accuracy assessment. The approximation provides a convenient way of determining a cumulative binomial probability. The confidence interval for this approximation is given by Snedecor and Cochran (1967) as follows:

$$Z = \frac{(|Q_T - Q_L|) - 1/2N)}{\sqrt{Q_L(1 - Q_L)/N}} \tag{7}$$

where z is the normal deviate of the standard normal distribution and other terms are as previously defined. For a one tailed test this equation can be written as

$$\Pr\left[\frac{(|Q_T - Q_L|) - 1/2N)}{\sqrt{Q_L(1 - Q_L)/N}} > Z\right] < A \tag{8}$$

(where Pr means 'probability of'). The producer's risk can be calculated as follows:

$$\Pr\left[\frac{(|Q_T - Q_H|) - 1/2N)}{\sqrt{Q_H(1 - Q_H)/N}} < Z\right] = B \qquad (9)$$

The value 1/2*N* is a continuity correction which some authors (e.g., Hord and Brooner, 1976) choose not to use in order to simplify calculations. The equation better approximates the binomial as the proportion approaches 0.5 and as sample size increases. Proportions at the extremes (e.g., 0.9) have highly skewed binomial distributions, whereas the normal distribution is symmetrical. Cochran (1977, p. 58) has suggested the following minimum sample sizes for using the normal approximation with continuity correction: for a proportion of errors of 0.1 (e.g., to test for a map accuracy of 90 percent) the minimum sample size is 600; and for $P = 0.05$ the minimum sample size is 1400. Sample sizes this large are generally considered prohibitively expensive for map accuracy testing.

Comparison of models. To illustrate the differences between the various models, Table 2 compares the consumer risks of *X* or fewer misclassifications for a sample size *N* as calculated by the cumulative binomial (Equation 3), the normal approximation with continuity correction (Equation 7), and the normal approximation without continuity correction.

The normal approximation without continuity correction underestimates the consumer's risk as calculated by the binomial, thereby overestimating the classification accuracy. The normal approximation with continuity correction overestimates the binomial, underestimating the accuracy level. The same biases apply to calculations of producer's risks. The errors introduced by the normal approximations vary from about 25 to 50 percent of the values for the binomial. As a result, sampling designs based on the normal approximation with continuity correction will require larger sample sizes for a given number of allowable errors, thereby increasing the producer's risk unnecessarily.

With the availability of computers and Ginevan's convenient tabulations, it would seem more convenient to use the exact values of the binomial than the approximations. Another advantage of Ginevan's tables is that the designs have been optimized to give the lowest producer's risk for a given number of allowable errors and a given consumer's risk. (For a given number of allowable errors, the smallest sample size that gives an acceptable consumer's risk will minimize the producer's risk.)

TABLE 2. COMPARISON OF CONSUMER'S RISK CALCULATIONS FOR MINIMUM ACCURACY OF 85 PERCENT

N	*X*	Binomial*	Normal with cc**	Normal without cc**
30	1	0.0480	0.0618	0.0367
35	1	0.0243	0.0375	0.0222
40	2	0.0486	0.0606	0.0384
46	2	0.0234	0.0344	0.0217
50	3	0.0460	0.0571	0.0375

* taken from Ginevan (1979)
** cc, continuity correction

ASSESSING ACCURACY FOR INDIVIDUAL CLASSES

The sampling designs discussed above give an overall estimate of accuracy but do not differentiate between errors of omission and commission, nor do they take into account the distribution of these errors. This distinction may be important, as for example when it is important to accurately identify only one or two classes. This type of information can be obtained from an error matrix.

Error matrices. An error matrix is a tabulation of accuracy test results which shows the number of points correctly and incorrectly identified. Commission errors (erroneously including a point from a class) and omission errors (erroneously excluding a point from a class) are clearly presented. For example, in the error matrix shown in Table 3, for class A, 26 points were correctly classified, and there were two commission errors and seven omission errors. A test to estimate the classification accuracy of a single class can be done using the same methods described above for all classes of a map by randomly selecting points from a single class.

Error matrices clearly show accuracy test results; however, when a single measure of quality is needed, as for example when two remote sensing system products are to be compared, it is convenient to have a value to represent the information of the entire matrix. One approach to this problem is that of Congalton *et al.* (1981) who have used multivariate analysis techniques to analyse and compare error matrices. A second approach is the use of analysis of variance as discussed in Rosenfield and Melley (1980) and Rosenfield (1981).

Comparison of classification results. Congalton *et al.* (1981) used multivariate analysis techniques to generate a normalized error matrix and to measure the agreement between two error matrices. They are continuing the development of these methods, and the limited space only allows for a brief description here.

A normalized matrix has each row and each column sum to one. This allows different elements in an error matrix or in different matrices to be directly compared despite differences in the way an accuracy test sample has been distributed among the classification categories. The theory of normalizing a matrix is presented in Biship *et al.* (1975, p. 85). By normalizing the error matrices

TABLE 3. A MAP ACCURACY ERROR MATRIX

				CLASSES					
				VERIFIED					
		A	B	C	D	E	Total	% Correct	% Commission
OBSERVED	A	26	1	0	0	1	28	93	7
	B	1	5	0	0	3	9	56	44
	C	2	0	43	1	2	48	90	10
	D	4	1	2	76	13	96	79	21
	E	0	0	2	1	29	32	91	9
Total		33	7	47	78	48	213		
% Omission		21	29	9	3	40			

generated by testing each of two remote sensing systems, corresponding values for each cell in the matrix can be directly compared. However, this also tends to hide the data on sample size, which is important because a sample accuracy of four out of five points correct has less confidence than 40 out of 50 correct.

Congalton *et al.* (1981) also defined a measure of agreement termed KHAT, a statistic calculated for each error matrix. Confidence limits can be calculated, allowing the significance of the difference between KHAT values for two matrices to be evaluated.

One advantage of this method over an analysis of variance (ANOVA) is that it does not assume that the accuracy levels (factor levels) observed in each category are independent (as is necessary in order to use an ANOVA test (Neter and Wasserman, 1974)). If one class is consistently confused with one other class (as commonly occurs), the proportion correct for those classes will not be independent. Rosenfield and Melley (1980) and Rosenfield (1981) have addressed these problems and illustrate the use of ANOVA for comparison of error matrices.

It is important to note that, in using error matrices to compare remote sensing systems, it is the entire system from image acquisition through image interpretation to map compilation which is being compared. A valid comparison requires that all these variables be controlled; otherwise, differences in map compilation, for example, could completely mask the differences between the interpretability of the image products.

The comparative analysis of error matrices may provide better methods of comparing land-use classification mapping methods than a simple comparison of overall estimated map accuracy. However, a user faced with a specific application may find that any single measure of quality does not provide the information he needs in order to understand the relative advantages of the two systems. In these cases an error matrix can be more valuable by allowing class by class comparisons. It clearly shows the actual number of samples drawn from each class, allowing the user to select a suitable test to judge the confidence of conclusions drawn from the data.

CONCLUSION

This paper has reviewed the theory of map accuracy estimation. It has been shown that the null hypothesis can be stated in two ways. Though both statements will give the same estimates for consumer and producer risks, the calculations must be done differently. The design of an accuracy test should consider both types of risk because there are significant costs associated with delays in receiving information caused by unnecessarily rechecking maps, as well as costs in accepting substandard products. It appears that the method described by Ginevan (1979) is statistically valid, easy to use, and considers both the consumer's and producer's risks.

More detailed analysis of error matrices using measures of agreement or analysis of variance may be valuable for comparing remotely sensed data acquisition systems. As use of the tests are better documented, they may be more commonly used.

The current interest in map accuracy estimation will probably lead to the development of more or less standard test methods. Such a test is needed as remote sensing systems are compared for selecting the "best" one for a specific operational application. As the need for better management of diminishing natural resources becomes increasingly more important, so will the need for information which meets a certain minimum accuracy level. Accurate estimation of map accuracy then becomes essential.

ACKNOWLEDGMENTS

The author gratefully acknowledges the critical reviews by Dr. R. N. Colwell of the Department of Forestry, and Mr. Andrew Benson of the Remote Sensing Research Program at the University of California, Berkeley. The author would like to thank the Department of Forestry of the University of California at Berkeley for supporting this research.

REFERENCES

Bishop, Y., S. Fienberg, and P. Holland, 1975. *Discrete Multivariate Analysis: Theory and Practise,* MIT Press, Cambridge, Mass., 575p.

Cochran, W. G., 1977. *Sampling Techniques,* John Wiley and Sons, New York, N.Y.

Congalton, R. G., *et al.,* 1981. *Analysis of Forest Classification Accuracy,* Remote Sensing Research Report 81-1, Lyndon B. Johnson Space Center, Houston, Texas.

Ginevan, M. E., 1979. Testing Land-Use Map Accuracy: Another Look, *Photogram. Engin. and Remote Sensing* 45(10):1371-1377.

Guenther, W. C., 1977. *Sampling Inspection in Statistical Quality Control,* Charles Griffin and Company Ltd., London.

Hord, R. M., and W. Brooner, 1976. Land Use Map Accuracy Criteria. *Photogram. Engin. and Remote Sensing* 42(5):671-677.

Neter, J., and W. Wasserman, 1974. *Applied Linear Statistical Models,* Richard D. Irwin Inc., Homewood, Illinois.

Rosenfield, G. H., 1981. Analysis of Variance of Thematic Mapping Experiment Data, *Photogram. Engin. and Remote Sensing* 47(12):1685-1692.

Rosenfield, G. H., and M. L. Melley, 1980. Applications of Statistics to Thematic Mapping, *Photogram. Engin. and Remote Sensing* 46(10):1287-1294.

Snedecor, G. W., and W. G. Cochran, 1967. *Statistical Methods,* 6th edition, Iowa State University Press, Ames, Iowa.

Received 2 July 1981; revised and accepted 8 February 1982)

APPENDIX

FORMAL STATEMENT OF THE NULL HYPOTHESIS

Statement 1. The classical statement of the null hypothesis for a one tailed test would take the form

H_0: $Q \geqslant Q_m$ (i.e., the map is of acceptable accuracy)
and
H_1: $Q < Q_m$ (i.e., the map is not of acceptable accuracy)

where Q is the inferred accuracy of the map being tested and Q_m is the minimum required accuracy. α is defined as the probability of incorrectly rejecting H_0, which in this case means the probability of incorrectly considering the map to be inaccurate. β is defined as the probability of incorrectly rejecting H_1, which in this case means the probability of incorrectly considering the map to be sufficiently accurate. Using this statement of the null hypothesis, the consumer's risk (the probability of accepting an unsuitable product) is β and the producer's risk is α.

Statement 2. An alternative statement of the null hypothesis could take the form

H_0: $Q \leqslant Q_r$ (ie the map is not of acceptable accuracy)
and
H_1: $Q > Q_r$ (ie the map is of acceptable accuracy)

where Q is the inferred accuracy of the map being tested and Q_r is the highest accuracy level that would be *rejected*. (Ginevan (1979) uses a maximum rejected accuracy for his calculations.)

In this case α, the probability of incorrectly rejecting H_0, becomes the probability of incorrectly considering the map to be accurate. Similarly β, in this case, becomes the probability of incorrectly considering the map to be inaccurate. Here the consumer's risk is α and the producer's risk is β, the reverse of the situation in statement 1.

Choice of null hypothesis. It does not matter which statement of the null hypothesis is used as long as consumer and producer risks are appropriately defined as above. The reason for this is that Q_r and Q_m are points on a continuum and thus, for purposes of calculation, "greater than or equal to Q_m" is the same as "greater than Q_r." For example, consider the calculation of the maximum number of misclassifications in a sample of size N to give a consumer's risk of 5 percent that a map is 85 percent accurate. Using the first statement of the null hypothesis, the maximum number of allowable misclassifications X would be determined by finding the maximum value of X such that the cumulative probability of X or fewer misclassifications in a sample of size N is less than or equal to 5 percent (in this case β is being set at 5 percent). The value for Q_m used in this calculation would be 85 percent.

Similarly for the second statement of the null hypothesis, suppose that Q_r, the highest accuracy level to be *rejected*, is set some small amount ϵ less than 85 percent, i.e., (85-ϵ) percent. Then the maximum number of allowable misclassifications for a consumer risk of 5 percent would be determined by finding the maximum value of X such that the cumulative probability of X or few misclassifications for a sample size N is less than or equal to 5 percent (in this case α is set equal to 5 percent.) Since ϵ is taken to be very small, the limit as ϵ approaches zero is for Q_r to approach 85 percent. Thus, because the values of Q_r and Q_m are essentially two points on a continuum of accuracy values, the cumulative probability up to and including that point is the same as the cumulative probability up to but not including the point. Thus, it does not matter which statement of the null hypothesis is used. The confusion arises in deciding whether α or β should be set to the consumer's risk.

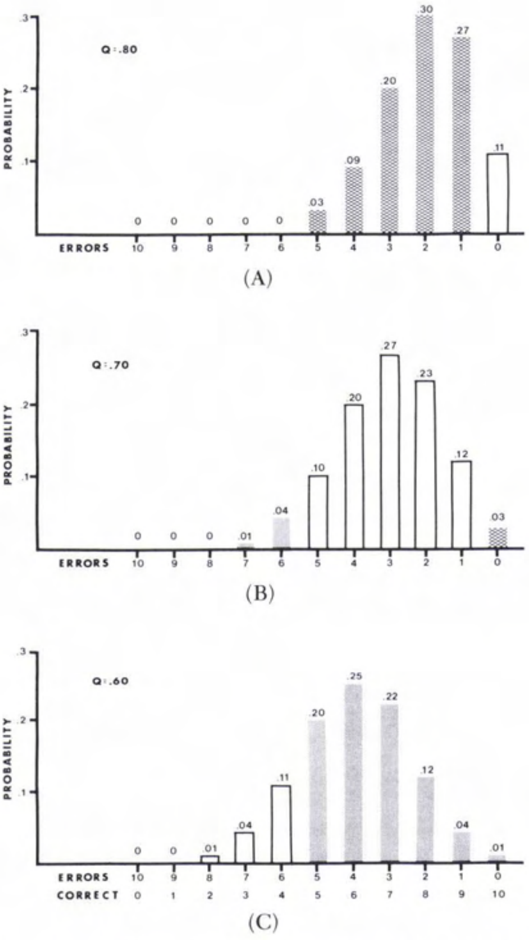
If an accuracy test is to be designed so that 95 percent of the time a map which is accepted in fact meets the accuracy criterion, then the consumer risk is 5 percent. Using the first statement of the null hypothesis, this means that $\beta = 5$ percent and $1-\beta = 95$ percent. Technically, this means setting the risk of a type II error β at 5 percent and the power of the test $(1-\beta)$ at 95 percent. However, in popular parlance the test might also be regarded as one at the "95 percent confidence level," when technically speaking it is at the 95 percent power level. Confidence level is defined as being equal to $1-\alpha$ and α in this case is the *producer's risk.* If the second statement is used, then the reverse is true. That is, to achieve the same 5 percent consumer's risk α must be set to 5 percent and $1-\alpha =$

95 percent. Technically this test would be considered at the 95 percent confidence level but the 5 percent consumer risk is now the probability of a Type I error.

In using the terms consumer and producer risk throughout this paper, the author has attempted to avoid this confusion. Also, by choosing to use the second formulation of the null hypothesis, the confusion of saying 95 percent confidence and meaning 95 percent power is avoided.

A NUMERICAL EXAMPLE

The following numerical example will be used to illustrate the effect on calculations in using the two formulations discussed above. The example considers designing an accuracy test involving the selection of ten points from a map (a sample of ten is used for ease of illustration; normally, a larger sample would be used). The allowable number of misclassifications X is to be found such that, if the sample has X or fewer misclassifications, it can be inferred with 90 percent confidence that the map has an accuracy of at least 70 percent.



(A)

(B)

(C)

FIG. A-1. Binomial sampling distribution in bar graph form for a sample size of ten and true map accuracies of 80 percent, 70 percent, and 60 percent. The probability represented by each bar is shown above the bar. See text for explanation of shading.

Calculation using statement 1. Using statement 1, the test strategy is to accept the null hypothesis, in this case "the map is at least 70 percent accurate," unless the test results indicate there to be less than a 10 percent chance of this being true. Because the binomial distribution is discrete (number of misclassifications must be an integer), the largest number of misclassifications giving a cumulative probability in the left hand tail less than or equal to 10 percent is 6 (shaded area in Figure A-1B). Thus if five or fewer misclassifications occur in the sample of ten points, the map will be accepted.

Suppose that the map actually has an unacceptably low accuracy of 60 percent. The probability of this substandard map passing the test is the cumulative probability of a sample of ten points from such a map having five or fewer misclassifications, shown as the shaded area in Figure A-1C. The cumulative probability is the sum of the individual probabilities shown as shaded bars, the value in this case is 0.84. Thus, the probability of accepting this substandard map (the consumer risk) is 84 percent. The 10 percent chance of incorrectly accepting the null hypothesis is in this case the *producer's risk*! Using statement 1 of the null hypothesis is misleading in that the "90 percent confidence" does not indicate a 10 percent consumer's risk but a 10 percent producer's risk.

Calculation using statement 2. Using statement 2, the test strategy is to accept the null hypothesis, in this case "the map is less than 70 percent accurate," unless the test results indicate there to be less than a 10 percent chance of this being true. To reject H_0 at the 90 percent level of confidence, the highest number of misclassifications X is found such that the cumulative probability of X or fewer misclassifications is 10 percent or less, in this case the cross hatched area of Figure A-1B. Thus, zero misclassifications are permitted in the sample of ten points and consumer's risk is at most 10 percent. The exact consumer's risk in this case is 0.03 or 3 percent. Had a larger sample size been used, this value could have been closer to the required 10 percent.

Suppose that the map is actually 80 percent correct. The probability that it would be rejected by such a test (i.e., the producer's risk) is the cumulative probability of having one or more misclassifications in a sample of ten points for Q = 80 percent. From Figure A-1A, this is seen to be 0.89 or 89 percent (the cross hatched area). In this case, the producer's risk is very high, but the assertion that the map is at least 70 percent accurate at the 90 percent confidence level is true in the sense that consumer's risk has indeed been reduced to 10 percent or less. To reduce the producer's risk while maintaining the same consumer's risk, a larger sample size could be used.

CONCLUSION

Either of the two statements of the null hypothesis, as presented here, will give the same result as long as the appropriate calculation of consumer and producer risks is used. Confusion occurs because the consumer's risk could be either a Type I or Type II error, depending on the way the null hypothesis is formulated. By using the terms consumer's risk and producer's risk, this confusion is eliminated.

Thermosense V

An International Conference on Thermal Infrared Sensing Diagnostics

Cadillac Hotel, Detroit, Michigan
25-27 October 1982

Sponsored by the International Society for Optical Engineering (SPIE) in cooperation with The Department of Energy/Oak Ridge National Laboratory and the American Society of Photogrammetry, Thermosense V, the fifth in an annual series of conferences on applications of thermal infrared sensing diagnostics to buildings and industrial uses, will provide the opportunity to present and exchange technical information on all aspects of infrared thermography and thermal sensing. As applied to buildings, thermography has come of age as a tool for preventive maintenance and for energy use analysis, and the time is right for more standardized approaches.

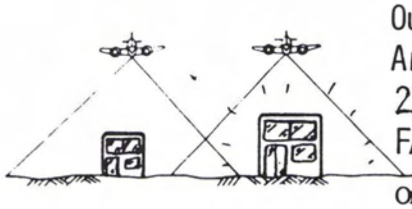
At the same time, infrared thermal sensing is finding increased application in industry as a solution to problems involving preventive maintenance and process control in which noncontact temperature measurement is required. In recognition of this latter interest, Thermosense V will highlight analysis and application of thermography in industrial processing.

For further information please contact

SPIE/Thermosense V
P.O. Box 10
Bellingham, WA 98227-0010

JUST FLYING BY TO TELL YOU WE'VE MOVED!

We've moved about 2 blocks away to larger offices to accommodate our growing society!



OUR NEW ADDRESS:
AMERICAN SOCIETY OF PHOTOGRAMMETRY
210 LITTLE FALLS STREET
FALLS CHURCH, VA 22046

Effective May 24, 1982

Only the street address has changed;

The City, State, Zip code, and
Phone number (703) 534-6617
are still the same as before!