ACTA STEREOL 1983; 2: 239-249
STEREOLOGY AND MORPHOMETRY IN PATHOLOGY
Proc 2nd Symp Morphometry in Morphol Diagnosis
Collan Y. et al., eds.
Kuopio University Press, Kuopio 1983

STATISTICS IN STEREOLOGY AND MORPHOMETRY

Kalevi Selkäinaho

Department of Applied Mathematics,
University of Kuopio, P.O.B. 6,
SF 70211 Kuopio 21, Finland

ABSTRACT

This paper is a short review of some basic principles of statistics that is useful in stereology and morphometry. The chapter I deals with sampling theory concerning the stereological analysis of microstructures. The problem is how to construct optimal estimators and how to estimate optimal sample sizes in several stages of a sampling design. The chapter II deals with the numerical study of the reliability of measurements obtained by methods above. It is concerned with two kinds of coefficients that indicate the degree of that reliability: the intraclass correlation coefficient (ICC) for continuous measurements and the kappa coefficient for discrete measurements.

I. ON SAMPLING THEORY CONCERNING THE STEREOLOGICAL ANALYSIS OF MICROSTRUCTURES

0. Introduction

The quality of the estimation of the stereological parameters of microstructures is usually measured in terms of a) bias and b) variance of the final estimator:

a) The bias depends upon the sampling design adopted and upon the statistical model underlying the sampling design. No manipulation of the sample data will reveal the bias, as it is known from mathematical statistics.

b) The variance depends also upon this statistical model and, in addition, upon the sample sizes in several sampling levels.

Now, the problem is (1) how to construct optimal stereological estimators, and their variances, from replicted observations; and (2) how to estimate optimal sample sizes which minimize the variance of the estimator to be considered,for a given cost.

1. The principles of hierarchical sampling designs

In the stereological analysis of microstructures there can be two kinds of hierarchical sampling designs: a nested design and a cascade design.

(i) A nested design (by Sokal & Rohlf, 1969) is carried out in several stages. For example, we have n microscopic sections, which constitute the first stage. But in several cases these sections cannot be observed as a whole of the required final magnification. Consequently, each section must be subsampled by a number of microscope fields or micrographs (quadrats), which can be analysed as a whole. This subsampling is called the second stage. In general, we can have n_a animals as the first stage, n_b blocks from some organ of each animal as the second stage, n_s sections from each block as the third stage and so on.

(ii) A cascade design (by Cruz-Orive & Weibel, 1981) is based on two preliminary factors, which must be taken into account:

- firstly, very often, observing and measuring the object phase of ultimate interest (denoted by Ω, say) in a section requires a high final magnification;

- secondly, a global stereological parameter $\gamma = \gamma(\Omega)$ is best estimated via an intermediate ratio to the volume of a reference phase, which contains Ω.

Then, γ can be estimated if the volume of the reference phase is known.

Taken together, these two factors pose the initial question of how to make an optimum choice of the reference phase, or a "cascade" (serie) of several reference phases at different magnifications. The final parameter is then estimated as the product of the intermediate ratios with the volume of the specimen, which is estimated independently. Each level in this design (also called multi-level design) can be regarded as an independent sampling design.

Example: (Cruz-Orive & Weibel, 1981) Consider the estimation of the total capillary surface area in a given lung. The phase of interest, capillaries (denote by Ω_3), is contained in the thin walls (Ω_2) between the air spaces, which together constitute a foam-like domain called lung parenchyma (Ω_1). Coarser structures ("non-parenchyma") bind the subdomains of parenchyma to make the whole lung (Ω_0). Now, the phase Ω_3 is rather inhomogeneous within Ω_0, representing only a small volume fraction of it (0.04-0.09). So a section for electron microscopy (which must be used) is necessarily small, and reducing the variance of the estimator of $\gamma(\Omega_3)$ would require a large number of sections, this rendering the sampling design too expensive, Also, there is a danger that $\gamma(\Omega_3)$ will become overestimated. So, it is necessary to know more specific properties of the different phases Ω_0, Ω_1, Ω_2 and Ω_3 in order to construct a suitable sampling design:

(i) The non-parenchymal phase is observable at a low magnification M_1 in a section through Ω_0. The parenchymal volume fraction $V(\Omega_1)/V(\Omega_0)$ is usually high (about 0.8 or more).

(ii) The phase Ω_2 can be regarded as a system of septa extending all over the containing phase Ω_1 with a varying degree of homogeneity. In a section, Ω_2 has to be observed by light microscope at least (the magnification M_2=100x to 200x). The volume fraction $V(\Omega_2)/V(\Omega_1)$ may vary between 0.10 to 0.15 in different specimens.

(iii) Identifying the phase of interest Ω_3 in a section requires a final magnification M_3=7000x or more. Now, the volume fraction $V(\Omega_3)/V(\Omega_2)$ is of the order of 0.4-0.7, which means that Ω_3 is fairly abundant within Ω_2. These properties and circumstances suggest a three-level, "cascade" sampling design: At the first level, the ratio $R_1=V(\Omega_1)/V(\Omega_0)$ is estimated at a low magnification; at the second level, $R_2=V(\Omega_2)/V(\Omega_1)$ is estimated by light microscope and at the third level, $R_3=\gamma(\Omega_3)/V(\Omega_2)$ is estimated by electron microscope. Finally,

$$\gamma(\Omega_3)=V(\Omega_0)\cdot R_1\cdot R_2\cdot R_3.$$

The ordinary ratio-of-sums estimator of the ratios R_1, R_2 and R_3 is based on point countings from uniformly positioned (integral) test systems of independent uniform random sections (IUR-sections). This estimation method is very optimal in the case of replicated ratio sampling, as Jensen & Gundersen, 1982, shows. How to generate IUR-sections and uniformly positioned test systems, see Cruz-Orive & Weibel, 1981, and Weibel, 1979.

2. Optimum sampling sizes at the different stages of a nested design

Each of the levels of a cascade design can be studied separately and, if necessary, using a nested design (how to allocate the sampling sizes for a given cost when all these levels will be taken together, see Cruz-Orive & Weibel, 1981).

The aim of a sampling design is to obtain maximal amount of quantitative structural information at a given total cost or effort. Gundersen & Østerby, 1981, discuss principles of such optimal designs and illustrate methods for generating them.

In general, the variation between different sampling units at the highest stage of a nested design is the major determinant of overall efficiency, whereas the variation between single microscopic features is less important. The expenditure of time and/or money in order to increase the precision of the individual measurements (at the lowest level) is irrational in almost all studies where the emphasis is, for example, on the biological results.

If we denote by O_s2 the observed variance between n patients (or blocks if we have only one patient) and by $\bar{x}$ their average value, the aim is to reduce the relative standard error RSE = $\sqrt{O_s2}/(\bar{x}\sqrt{n})$ to the level 0.1, say. For example, a little increase in the number of blocks and/or sections may reduce O_s2 significantly. But even a marked increase in the number of fields in sections or in precision in measuring them may not cause sufficient reduction in O_s2.

3. Sampling by point counting methods

As Jensen & Gundersen, 1982, shows, the fact that the estimation is based on counts (as opposed to complete 2-d observations) does not necessarily mean a reduction in information. For certain types of stereological ratios, the ordinary ratio-of-sums estimator based on complete observation has shown even to be less accurate than that based on simple and fast counting.

If it is not possible to have a sufficiently great number of patients and/or blocks, then one must pay attention, especially, to the precision of the measurements in lower stages.

a) The computation of the number of test points in estimation of V_V:

Here we are sampling, from a microscopic section, for proportions between two spaces: the object space (also called object phase or structure) a and the containing space (phase, structure) c. A certain number P_c of test points is applied to the containing space, and for each point it is determined whether it is in a or not. The number of test points which are in a is denoted by P_a. Now V_V equals to P_a/P_c the more accurately the greater is P_c. Here P_a is a random variable having binomial distribution with parameters P_c and V_V. Hence, the expectation of $P_p=P_a/P_c$ is V_V and its standard deviation is

$SD=\sqrt{V_V(1-V_V)/P_c}$, which can be estimated by $\sqrt{P_p(1-P_p)/P_c}$.

One way of judging that P_c is sufficiently large is to compute the relative standard error of P_p,

$$RSE(P_p) = \sqrt{\frac{1-P_p}{P_c \cdot P_p}} ,$$

for several values of P_c, adding the number of quadrats and/or sections until RSE remains under 0.1. Another way is to apply the normal approximation of binomial distribution and compute a confidence interval for V_V, which is in the form

$$P_p-Z_\alpha \cdot SD \leq V_V \leq P_p+Z_\alpha \cdot SD.$$

Here V_V in the formula of SD must be estimated from a pilot survey, for example, and $\pm Z_\alpha$ are the abscissas of the normal curve, which cut a total area fraction α at the tails. For $\alpha=0.05$ (95 % probability of being within confidence interval) $Z_{0.05}=1.96$, $Z_{0.01}=2.57$ etc.

If we want that the deviation $Z_\alpha \cdot SD$ is at most d % of the true V_V, we must have

$$P_c \geq \frac{Z_\alpha^2}{d^2} \cdot \frac{1-\hat{V}_V}{\hat{V}_V} ,$$

where $\hat{V}_V$ is the estimated (in a pilot survey) V_V.

Remark. For some problems concerning e.g. the optimal density of test points, the inhomogeneity of the object space and the section thickness, see Weibel, 1979.

b) The computation of the test line length in estimation of S_V:

The lines of a square lattice (grid) may be used to estimate the surface density S_V if components (of the object space) by counting the intersections I_a, with profile boundaries. In a coherent test system (a test system formed by a lattice of fundamental figures) S_V is connected to the total test line length L_t. S_V equals to $2I_a/L_t$ the more accurately the greater is L_t.

Now I_a is a random variable with a Poisson distribution, depending on the true S_V. The standard deviation of I_a can be estimated by $\sqrt{I_a}$; hence the relative standard error of the estimator $S_V=2I_a/L_t$ is $SD(S_V)/S_V=1/\sqrt{I_a}$. In cases of non-contiguous convex particles of low volume and surface density the formula $RSE(S_V)=\sqrt{2/I_a}$ is recommended (see Weibel, 1979). Now we get that

$$L_t \gtreqless \frac{4}{\hat{S}_V \; RSE_0 2}$$

if RSE is wanted to be lower than RSE_0 and $\hat{S}_V$ is an estimated S_V (by a pilot survey, for example).

II. DERIVING COEFFICIENTS OF INTERNAL CONSISTENCY OF MEASUREMENTS

0. Introduction

The quality of data critically depends on the reliability with which primary observations are assigned to categories, scaled, or measured. This chapter is concerned with the numerical study of that reliability (also called reproducibility, repeatability, internal agreement etc.) which in this paper is called internal consistency.

This is a difficult field, but a field of growing importance. For example, the recent rapid increase in data-cathering in the social and medical sciences is containing several variables which are difficult to measure. In order for such data to be empirically meaningful, a "high"-degree of internal consistency must be demonstrated.

The problem is to asses the discrepancies between repeated measurements of the same experimental unit and to express the results in a concise way. We are concerned with two kinds of coefficients to indicate the degree of the internal consistency of those measurements: the intraclass correlation coefficient (ICC) for continuous measurements and the kappa-type coefficient for discrete measurements. These coefficients seem to be the most useful in practice.

The confidence intervals of these coefficients are of major importance. The normal theory and the jackknife procedure will be used. The author also suggest some lables to be assigned to the corresponding ranges of the ICC, similar to that of the kappa coefficient suggested by Landis and Koch (1977).

In what follows, the experimental units are referred as "persons" and the repeated measurements are referred as "instruments". (For example, the instruments may be the repeated scalings by one observer). The data format in this paper is always as in Table 1.

Table 1. Notation for analysis of measurements

	Instruments				
Persons	1	2	...	m	means
1	x_{11}	x_{12}	...	x_{1m}	$\bar{A}_1$
2	x_{21}	x_{22}	...	x_{2m}	$\bar{A}_2$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
i	x_{i1}	x_{i2}	...	x_{im}	$\bar{A}_i$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n	x_{n1}	x_{n2}	...	x_{nm}	$\bar{A}_n$
Means	$\bar{B}_1$	$\bar{B}_2$	...	$\bar{B}_m$	$\bar{T}$

For an introduction to reproducibility problems in medical diagnostics, see Collan, 1982.

1. The ICC in the one-way model

Techniques for the numerical study of internal consistency of continuous measurements x_{ij} are mainly based upon the analysis of variance and the estimation of variance components. For the general theory of variance components see e.g. Searle, 1971. It is also referred to Cochran, 1968 and Bartko, 1966.

The simplest case is the one-way random effects model. Here we have m repeated independent measurements by one instrument, for each person. The usual assumption is the model

(1) $x_{ij}=u+a_i+e_{ij}$ $(i=1,\ldots, n;\ j=1,\ldots,m)$,

where u is the overall effect common to all observations; a is a random variable, with zero mean and variance σ_a^2 common to the i:th person and e_{ij} is the random error, with zero mean and variance σ_e^2, associated with observation (i,j) and independent of a_i. The usual analysis of variance table is given in Table 2.

Table 2. Analysis of variance: One-way random model

Source of variation	Degrees of freedom	Sums of squares	Mean squares	Expected mean squares
Persons	n-1	SSA	MSA	$\sigma_e^2 + m\cdot\sigma_a^2$
Error	n(m-1)	SSE	MSE	σ_e^2
Total	nm-1	SST		

Here $SST = \Sigma_{i=1}^{n}\Sigma_{j=1}^{m}(x_{ij}-\overline{T})^2$

$SSA = m\Sigma_{i=1}^{n}(\overline{A}_i-\overline{T})^2$

$SSE = SST-SSA$,

and the mean squares are obtained by dividing the sums of squares by the corresponding degrees of freedom. From the expected mean squares we get the unbiased estimators for σ_e^2 and σ_a^2 (for any distributions of a_i and e_{ij}):

$$\hat{\sigma}_e^2 = MSE, \quad \sigma_a^2 = (MSA-MSE)/m.$$

The ICC ρ_x for the measurements x_{ij} is defined by $\rho_x = \sigma_a^2/var(x_{ij})$, which becomes now $= \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$ and its analysis of variance estimator

$$\hat{\rho}_x = \frac{MSA-MSE}{MSA + (m-1)MSE}$$

is obtained by replacing σ_a^2 and σ_e^2 by their estimators above.

If a_i and e_{ij} are normally distributed the confidence interval (ρ_1, ρ_2) for ρ_x can be computed from $\emptyset$=MSA/MSE, which is distributed as a multiple of an F-distributed variable. The limits ρ_1 and ρ_2 work out as follows (see e.g. Searle, 1971):

$$\rho_k = (\emptyset-F_k)/(\emptyset+(m-1)F_k),\ k=1,2,$$

where (F_2,F_1) is the usual interval of the F(n-1, n(m-1))-distribution for a given confidence probability.

For ICC in other models and the computation of jackknife confidence interval, see Selkäinaho, 1983. The relative strength of internal consistency associated with ICC is shown in Table 3.

2. The kappa coefficient for discrete measurements

If the measurements in Table 1 have nominal or ordinar scale, we introduce the kappa-coefficient K_0 as suggested by Kraemer (1980). In this case each observation x_{ij} is a choice of one category among K possible categories. To each x_{ij} there corresponds a K-dimensional vector of ranks. For example, the usual single choice of one category C_k imposes a rank 1.0 on category C_k and a rank (K+2)/2 on the other K-1 categories, hence we get the vector (3.5, 1.0, 3.5, 3.5, 3.5) if we have K=5 categories C_1, C_2,..., C_5 of which C_2 has been chosen. An equivocal response A/B (equally A or B) imposes a rank of 1.5 on categories A and B, and (K+3)/2 on the other K-2 response categories. A ranked response AB (A primary) imposes a rank 1.0 on A, 2.0 on B and (K+3)/3 on the other K-2 categories. And so on.

Now, the average Spearman rank correlation coeffient r_i among the m(m-1)/2 pairs of observation of subject i (i=1, ..., n) is calculated from the rank vectors above. Also the average r_I of $r_{1,2}$, ..., r_n and the average Spearman rank correlation coefficient among all possible pairs are calculated. The K_0 is defined as $K_0 = (r_I-r_T)/(1-r_T)$. If there is no agreement among the instruments, r_I=r_T and hence K_0=0. At the other extreme, K_0=1 if and only if there is absolute agreement among all observations of any single person, i.e. r_I=1 (and also some heterogeneity between persons, i.e. $r_T \neq 1$).

How to calculate the correlation coefficients $r_1,r_2,\ldots r_n$ and r_T in a handy way, see Kraemer, 1980. In practice, we can usually assume that r_T is fixed, and hence the standard error of K_0 is readily estimated as:

$$SE(K_0) = S_r/(\sqrt{n}(1-r_T)),$$

where $S_r^2 = \Sigma_{i=1}^n (r_i - r_I)^2/(n-1)$. For moderate sample sizes n the t(n-1) -distribution is sufficiently robust to justify computation of a confidence interval for the "true" value of K_0, say κ, as

$$K_0 - t_\alpha(n-1)\cdot SE(K_0) \leq \kappa \leq \kappa_0 + t_\alpha(n-1)\cdot SE(K_0),$$

where $\pm t_\alpha(n-1)$ are the abscissas of the t-distribution curve (with n-1 degrees of freedom), which cut a total area fraction α at the tails.

Remark. In the case of single choice of a category it is very simple to make a program that computes K_0 and its confidence interval, using Table 1 directly. It needs about 70 lines by Fortran. In other cases, the generation of the rank vectors is more complicated.

The relative strength of internal agreement associated with kappa is shown in Table 3.

Table 3. Labels of internal consistency associated with ICC and kappa

ICC	kappa	strength of internal consistency
≤ 0.50	< 0.00	poor
0.51-0.60	0.00-0.20	slight
0.61-0.70	0.21-0.40	fair
0.71-0.80	0.41-0.60	moderate
0.81-0.90	0.61-0.80	substantial
0.91-1.00	0.81-1.00	almost perfect

As a practical example of the use of ICC and kappa in morphometry we refer to Kosma et al., 1983.

REFERENCES

Bartko JJ: The intraclass correlation coefficient as a measure of reliability. Psychological Reports 19: 3-11, 1966

Cochran WG: Errors of measurements in statistics. Technometrics 10: 637-666, 1968

Collan Y: Reproducibility, the neglected corner-stone of medical diagnostics. In: Collan Y and Romppanen T (editors): Morphometry in Morphological Diagnosis. Kuopio University Press, Kuopio 1982

Cruz-Orive L-M and Weibel ER: Sampling designs for stereology. J Microscopy 122: 235-257, 1981

Gundersen HJG and Østerby R: Optimizing sampling efficiency of stereological studies in biology: or "Do more less well!" J. Microscopy 121: 65-73, 1981

Jensen EB and Gundersen HJG: Stereological ratio estimation based on counts from integral test systems. J. Microscopy 125: 51-66, 1982

Kosma V-M, Selkäinaho K, Collan Y, Syrjänen K, Aalto M-L and Seppä A: Observer variation and reproducibility of grading: Analysis of the postcapillary venules in human axillary lymph nodes using subjective and morphometric methods. In: Collan Y et al (editors): Morphometry and stereology in pathology. Kuopio University Press, Kuopio 1983. Also published as a special issue of Acta Stereol 2, 1983

Kraemer H: Estension of the kappa coefficient. Biometrics 36: 207-216, 1980

Landis JR and Koch GG: The measurement of observer agreement for categorial data. Biometrics 33: 159-174, 1977

Selkäinaho K: Deriving coefficients of internal consistency of measurements: ICC and kappa. Reports on Statistics, University of Jyväskylä, 1983 (in press)

Searle SR: Linear Models. John Wiley & Sons, New York, San Francisco, 1969

Sokal RR and Rohlf FJ: Biometry. Freeman and Company, San Francisco, 1969

Weibel ER: Stereological methods. Vol 1. Academic Press, London, 1979