Abuse of Statistics

J.W. Merks, President, Matrix Consultants Limited,
Vancouver, British Columbia

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." H.G. Wells, 1866-1946

Introduction

Scientists and engineers have used statistical tools and techniques for hundreds of years. Without statistics the world of averages, indices and trends would collapse. Statistical parameters are used to explain why a certain course of events will occur, and the same parameters are later used to explain why the events took a different course. H.G. Wells (as quoted by Huff, 1954) was impressed with statistical thinking but Huff (1954) and Reichman (1961) presented sobering perspectives on abuse of statistics.

Sampling theory and its practical applications are based on probability and statistics. Because sampling theory plays an essential role in geostatistics, one would expect geostatistical theory to be based on sound principles of probability and statistics. This technical brief presents geostatistical applications that violate fundamental requirements of probability theory.

Discussion

The variances for sets of randomly distributed and ordered data in one-, two- and three-dimensional sample spaces have found application in exploration, mining and metallurgy. On-stream measurements in mineral processing plants are a typical example of ordered data in a one-dimensional sample space. Variances for ordered sets of on-stream data at intervals of 15 minutes or less are invariably lower than variances for randomly distributed sets. Lower variances result in more precise metallurgical balances, and in more realistic simulation models for mineral processing plants.

Covariances and kriged variances dominate geostatistics. Yet, the degrees of freedom for a data set, a concept even more fundamental than the variance, seems irrelevant in geostatistical theory. The concept of degrees of freedom is mentioned in many elementary textbooks on statistics (Huntsberger and Billingsley, 1973; Moroney, 1951; Reichman, 1961; Spiegel, 1961) while advanced textbooks address it in greater detail (Davies and Goldsmith, 1947; Mandel, 1964). Koch and Link (1970) refer to the phase rule in physical chemistry when discussing degrees of freedom which is not surprising. After all, degrees of freedom are fundamental in thermodynamics.

Neither Clark (1979) nor David (1977), or Journel and Huijbregts (1978) for that matter, even mention the concept of degrees of freedom let alone caution against the perils of violations. Just how fundamental a concept it is becomes obvious upon realizing that the denominator in the formula for the variance is the degrees of freedom for a set of statistically independent data points. The following formula shows that the variance for a randomly distributed set of n data has n-1 degrees of freedom:

$$\text{var}(x) = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

in which:

var(x)	=	variance for a random set
$\bar{x}$	=	mean for the set of data
x_i	=	ith datum
n	=	number of data in the set
n − 1	=	degrees of freedom for the set

Even though Clark (1979) presented a different formula for the terms of a sampling variogram, the denominator in the formula for the variance terms of a set of ordered data reflects 2n − 2 degrees of freedom:

$$\text{var}_j(x) = \frac{\Sigma(x_{i+j} - x_i)^2}{2n - 2}$$

in which:

$\text{var}_j(x)$	=	jth variance term for the ordered set
x_{i+j}	=	(i+j)th datum
x_i	=	ith datum
j	=	spacing between data
n	=	number of data for the jth variance term
2n − 2	=	degrees of freedom for the jth variance term

The concepts of dependencies and degrees of freedom are closely related. For example, the kriged datum is the dependent variable of the function that its relationship with the measured data set defines. In probabilistic terms a kriged datum does not have a single degree of freedom. Hence, the kriged variance for a set of measured (independent) data and some arbitrary set of kriged (dependent) data is invalid because it violates the requirement for statistical independence. Similarly, covariances for all equations that the set of dependent and independent data defines also violate the requirement for statistical independence.

Analysis of variance is one of the most powerful tools in applied statistics. It is based on applying Fisher's F-test to a pair of variances to check whether they differ significantly or are statistically identical. For example, the existence of a spatial correlation at spacing j can be verified by comparing the calculated F-ratio between the variance for a set of randomly distributed data and the jth variance term for the ordered set with a tabulated F-value, either at 95% or 99% probability and with the appropriate degrees of freedom. If the calculated F-ratio does not exceed the tabulated F-value, the variances are statistically identical, and their difference is just a random number. Applying mathematical analysis to the difference between a pair of statistically identical variances is an abuse of statistics. In geostatistical applications, however, such differences are routinely entered into smoothing relationships to predict tonnages and grades.

Another example of abuse of statistics in geostatistical applications is to calculate a correction factor to predict grades of bulk samples from grades of coincident drill core sections. However, Student's t-test showed

Jan Merks is an author, consultant, lecturer and publisher with twenty-five years of experience in metrology, the science of measurement as applied to mining and metallurgy. He is the president of Matrix Consultants Limited, a company providing consulting services in measurement technology. He is the author of "Sampling and Weighing of Bulk Solids", wrote papers and reports on topics ranging from precision estimates for ore reserves to trace elements in refined metals, and participates in the activities of various Technical Committees of the International Organization for Standardization.

that the mean difference between paired grades of bulk samples and coincident drill core sections was statistically identical to zero. Hence, either set can be used to predict the other without any risk that their mean difference becomes statistically significant.

Almost a hundred years ago H.G. Wells was already impressed with statistics (Huff, 1954). Forty years ago Huff (1954) took a whimsical look at the abuse of statistics. Thirty years ago Reichman (1961) observed that "very few people nowadays can progress very far without at some point coming into contact with statistics", and he concluded that "the Age of Statistics is upon us". One cannot help but feel that Huff and Reichman would be amused with our creative abuse of statistics but that H.G. Wells would be disappointed with our progress.

REFERENCES

CLARK, I., Practical geostatistics; Applied Science Publishers, London, 1979.

DAVID, M., Geostatistical ore reserve estimation; Elsevier, Amsterdam, 1977.

DAVIES, O.L., and GOLDSMITH, P. L., Statistical methods in research and production; Longman Group, London, 1947.

HUFF, D., How to lie with statistics; Penguin Books, Harmondsworth, 1954.

HUNTSBERGER, D.V., and BILLINGSLEY, P., Elements of statistical inference; Allyn and Bacon, Boston, 1973.

JOURNEL, A.G., and HUIJBREGTS, CH. J., Mining Geostatistics; Academic Press, London, 1978.

KOCH, G.S., and LINK, R.F., Statistical analysis of geological data; John Wiley & Sons, New York, 1970.

MANDEL, J., The statistical analysis of experimental data; Dover Publications, New York, 1964.

MORONEY, M.J., Facts from figures; Penguin Books, Harmondsworth, 1951.

REICHMAN, W.J., Use and abuse of statistics; Penguin Books, Harmondsworth, 1961.

SPIEGEL, M.R., Theory and problems of statistics; McGraw-Hill, New York, 1961.