

Reviewer's comments

3 April 1995

Math Geology ms 95-31

The Properties of Variances

J.W. Merks

Regrettably this manuscript is filled with errors, the same is true of the enclosed CIM Forum articles "Abuse of Statistics" and "Simulation models for mineral processing plants". Clearly neither of the latter were adequately reviewed, they should never have appeared in print.

The following is a partial list of errors and mis-conceptions that appear in the manuscript, it is not complete.

It is NOT true that in "applied statistics" the population variance is replaced by the sample variance. It is true that in statistics the population variances are in general not known and one uses the sample variance to estimate the population variance. The same may be true of the population mean vs the sample mean. In such cases one should distinguish between the variability that is inherent (population variability) and that related to the sampling. More generally however the objectives in probability theory are simply different than in statistics. In statistics one will generally want to infer some property of the population, in particular properties of the distribution such as the type or parameters such as the mean or variance. Statistical inference also includes testing of hypotheses. One should not confuse statistical inference with descriptive statistics. Descriptive statistics is concerned with describing a data set and in general does not require any statistical assumptions nor is it based on probability theory. Statistical inference and the validity of the conclusions is very much dependent on the validity of the underlying statistical assumptions. In some cases this validity can be assured by the sampling plan, i.e., the way the data is generated. While the author has attempted to make statements that are related to statistical inference he has totally neglected the question of the validity of the underlying assumptions and in fact has even neglected to mention or identify them.

1. The first sentence of the INTRODUCTION is at best incomplete and at worst nonsensical. "its" requires an antecedent but there is none. One could model the wet mass of a quantity of mill feed per hour, per day, per month, etc as a random variable but some form of limitation is crucial. The same problem occurs with the other purported examples.

2. Probability Theory doesn't "dictate" anything.

3. If wet mass, moisture and grade are to be random variables (presumably the author wants to treat them as such since there is a reference to variances) then it is nonsensical to say that the variance of "metal contained" is a function of these three variables and their variances. The (population) variance is a constant not a random variable yet the author is claiming that this constant is a function of the three random variables.

4. Presumably in the second paragraph of the INTRODUCTION when the author refers to independent variables he is referring to stochastic independence. The formula given on page 2 (eq [1]) is not correct or at the very least is incomplete. Consider an elementary example

Let X, Y be uniformly distributed on the interval $[0,1]$ and independent. Now let $Z = XY = f(X,Y)$. By direct computation $E\{X\} = 1/2 = E\{Y\}$, $E\{X^2\} = 1/3 = E\{Y^2\}$, $E\{XY\} = 1/4$, $E\{(XY)^2\} = 1/9$. Then $\text{Var}\{X\} = \text{Var}\{Y\} = 1/12$ and $\text{Var}\{XY\} = 7/[(9)(16)]$. Note that in computing derivatives one has to consider a deterministic function $z = f(x,y) = xy$. Then $\partial f/\partial x = y$, $\partial f/\partial y = x$. Now the partial derivatives have to be evaluated at some point, presumably in this case at the "point" X, Y . Applying eq [1] one would obtain

$$7/[(9)(16)] = Y^2 (1/12) + X^2 (1/12)$$

which is nonsensical since the right hand side is still a random variable. Clearly the stated formula is wrong.

The results quoted later as being special case applications of eq [1] are true but not because of eq [1]. The author has confused a Taylor series approximation to the variance with an equality.

5. The last several sentences at the bottom of page 2 are gibberish. Eq [1] is NOT referred to as the additive property of variances. The following is true and is sometimes referred to as the additive property of variances.

Let X_1, \dots, X_n be independent random variables with finite variances and a_1, \dots, a_n arbitrary constants. Then

$\text{Var}\{a_1 X_1 + \dots + a_n X_n\} = a_1^2 \text{Var}\{X_1\} + \dots + a_n^2 \text{Var}\{X_n\}$. The proof is direct using only the definitions of expected value, variance and independence. It does not depend on eq [1].

Note that it is crucial to assume that each of the random variables separately has finite variance, the sum might have finite variance even when the individual terms do not. Note also this is a theorem pertaining to population variances and not to sample variances.

Commutativity is something quite different and is completely irrelevant in this context. In the last sentence the author has mixed up population variances and sample variances. He neglects to say what the difference of two (sample?) variances is to be an estimator FOR, it is obviously irrelevant to ask whether the ratio of two population variances is statistically significant. Statistical significance is related to a particular Type I error probability level (the author does not mention one) and hence the claim about statistical significance is at best incomplete. The constant reference to "statistical significance" is a clear indication that the author does not understand about testing of hypotheses. While many authors use .05 for the Type I error probability there is nothing sacred about this value.

6. On page 3 the author has a section on functional dependence yet seems to be talking about

stochastic dependence which is quite different. Functional dependence and stochastic independence are very different concepts. This entire subsection is wrong.

7. What is the "degree of association" of two variables? At one point the author refers to the "degree of association" as being statistically significant and then in the next sentence he says that the degree of association is "high" when the covariance is statistically significant. Obviously the covariance has to be a sample covariance. Statistical significance is terminology related to the testing of hypotheses but the author has neglected to identify the null and alternative hypotheses.

If testing a hypothesis on a population mean then the important quantity is the difference between the value of the mean given in the null hypothesis and the value of the sample mean, if the value is large statistically (i.e., the probability of a value that large is less than the prescribed Type I error probability) then one would say that the difference is statistically significant (in this case one would reject the null hypothesis but now it is crucial to identify the alternative hypothesis). However in an ANOVA test it is not a difference that is important but rather a ratio between the hypothesized variances and the sample variances. The author is being unbelievably sloppy.

8. The entire subsection "Spatial Dependence" is wrong.

9. Page 3, last paragraph. The author refers to set of "ordered measurements", note that he is referring to the data being collected in a manner that is ordered with respect to time. In statistics ordered data or ordered measurements would refer to ordering the data according the magnitudes of the data values which is quite different.

10. It is definitely NOT true that eq [5] is referred to as the Central Limit Theorem. There are actually several versions of the Central Limit Theorem but the best known one is as follows:

Let X_1, \dots, X_n, \dots be a sequence of independent, identically distributed random variables with finite moments of order $2 + \delta$, $\delta > 0$. Let $S_n = (1/n)[X_1 + \dots + X_n]$, then

$$[S_n - E\{X\}]/(\sigma/n^{1/2})$$

converges in distribution to a Normally distributed random variable with mean 0 and variance 1. The Normal approximation to the Binomial is a special case of this but can be proved directly somewhat easier (than the general theorem).

11. The author has mis-placed the emphasis on degrees of freedom and based on the manuscript one would have to say that he does not understand. "Degrees of freedom" is an older terminology that is not relevant to the modern development of statistics. The term was carried over from physics and is useful in motivating or intuitively explaining certain results but it is not used in the rigorous development of the statistics. For example, degrees of freedom are often referred to in connection with the Student-t distribution and the F distribution but in both cases these are simply parameters.

It is true that in many reference works, the statistical tables are labeled in terms of "degrees of freedom" but this is simply a continuation of long standing practice. It is simply a way of distinguishing the various parameters that appear in the table.

As a consequence of this mis-understanding of the terminology "degrees of freedom" the entire manuscript is based on a fallacious premise.

12. Page 6, second paragraph. It is NOT true that "Probability theory dictates that the distribution of diamonds of uniform mass, when measured in homogeneous samples of kimberlite, can be described with the Poisson distribution". Probability theory does not "dictate" anything. It is not true that the observed number of events and the variance are numerically identical. Perhaps the author is referring the property of the Poisson wherein the expected value and the variance are the same but that is quite different.

Aside from the other errors in this paragraph the author has used the Normal distribution to compute the width of the 95% confidence interval without saying so. He has confused confidence intervals (which pertain to estimating parameters such as the mean and variance) with prediction intervals. A correctly determined confidence interval (for the mean for example) using the Poisson distribution would not be symmetric since the Poisson is not a symmetric distribution.

Note that if the mean and variance are known then Chebychef's Inequality can be used can be used to generate a symmetric prediction interval but it will be much wider than the true interval using the Poisson distribution.

The Poisson distribution has a very important property that is relevant here but not mentioned. If X_1, \dots, X_n, \dots are independent Poisson distributed random variables with means $\lambda_1, \dots, \lambda_n, \dots$ then $X_1 + \dots + X_n, \dots$ is Poisson distributed with mean $\lambda_1 + \dots + \lambda_n, \dots$. For example, if the X_i 's are identically distributed and each represents the count for a fixed volume or area then the arithmetic average is a good estimator of the common mean and one can "adjust" the mean for other volumes or areas, moreover since the Poisson has only one parameter the distribution can then be fit to different volumes or areas. This is rather different from what the author says in the manuscript.