THE APPLICATION OF SIMPLE STATISTICS IN GRAINS RESEARCH

Phil Williams

PDK Projects, Inc., Nanaimo, Canada
philwilliams@pdkgrain.com

INTRODUCTION

It is helpful to remember two descriptions. The first is the "observation". An observation is a single piece of information, such as the result of a moisture test. The second is the "population", which is the total number of observations. A population can simply mean all of the observations that have been recorded, or a set of observations that have been selected from a bigger population in order to prove the validity of a hypothesis.

The word "Statistics" often creates a feeling of distrust. It has two meanings. The first is that statistics represent information recorded about something that has been done, such as a series of chemical analyses, or a survey. The other meaning is that Statistics is a form of science, and a valuable tool that can be used to determine and report the validity of the results of such exercises. The saying that "statisticians always lie" is sometimes aired. But statisticians only report the mathematical evaluation of the results of the operation. They do not create the data, and in most cases are not aware of the operation until someone asks them to evaluate what has happened, as a result of the work. They may be consulted in the design of experiments, but often are not involved until the experiments have been done. As a result, sometimes unsatisfactory reports from statisticians are the result of poor experimental design, which is the result of experimental work that has been carried out by scientists without consulting the statisticians before carrying out the work.

STATISTICS FOR APPLICATION TO GRAIN RESEARCH AND ANALYSIS

There are five statistical options that can be used to evaluate most of the work that has to be carried out on grains. These are the mean, the standard deviation, the coefficients of correlation and regression, and the analysis of variance (ANOVA). If application of these fails to arrive at a conclusion, either the experiment may not have been well designed, or the hypothesis needs re-visiting!

The mean is the average of all of the observations. There are three main forms of the mean. The first is the Arithmetic mean, which is calculated by summing all of the observations and dividing the result by N (the number of observations). Table 1 gives examples of the running and weighted means. In the table the weighted mean represents deliveries of wheat to a country elevator.

Table 1. Examples of Running and Weighted Means

Running mean				Weighted mean		
N	Protein	N	Protein	Protein	Weight (tonnes)	Protein x Weight
1	12.8	2	13.6	12.8	14.6	186.88
2	13.6	3	13.1	13.6	21.2	288.32
3	13.1	4	13.5	13.1	20.3	265.93
4	13.5	5	13.2	13.5	44.6	602.10
5	13.2	6	13.7	13.2	28.7	378.84
6	13.7	7	12.9	13.7	16.9	231.53
7	12.9	8	14.4	12.9	33.3	429.57
8	14.4	9	13.3	14.4	24.9	358.56
9	13.3	10	13.4	13.3	18.3	243.39
10	13.4	11	13.8	13.4	26.6	356.44
11	13.8	12	13.0	13.8	38.7	534.06
12	14.0	13	14.0	13.0	27.8	361.40
Sum	160.7	Sum	161.9	160.7	315.9	4237.02
Mean	13.39	Mean	13.49	(Σ(protein x weight)/315.9)		13.41 (Weighted mean)

The second is the Running mean. The running mean is based on a sequence of e.g. 12 observations. When the next (13^{th}) observation is made the first observation is left off, and the running mean calculated from the 12 observations that have been retained, and so on for as long as the running mean is to be maintained. The third is the Weighted mean. This is used in calculating the average of a series of observations based on e.g. the different quantities of grain that makes up a silo-full, or a cargo.

The standard deviation (SD) expresses the variance in a set of data. Table 2 gives an example of the standard deviation of the results of testing a check sample for protein content:

Table 2. Example of standard deviation

Protein	Statistic	Value
13.26	N	12
13.37	Mean	13.47
13.57	ΣX	161.61
13.74	ΣX^2	2176.6791
13.49	SD	0.134
13.48		
13.34		
13.60		
13.35		
13.40		
13.55		
13.46		

The SD merits further explanation. The significance of the SD is that the value of e.g. ± 0.134 in Table 2 represents only about 67% of the total population. The value of $\pm$1.96 times SD represents 95% of the population and is referred to as the 95% Confidence Limit. This is a rather better way to describe the variance in a population. The value of

± 2.56 times SD represents 98% of the population. This means that 2% of the population will lie outside even these limits. Table 3 shows the range in results for testing a check sample for protein content that will happen (for a constant SD):

Remember that 2% of the population will fall **outside** the maximum ranges shown in the previous table and the range between which all samples fall is surprisingly big. The significance of this is that in a population of 500 (a good number for a stable calibration model), 10 samples will have values of greater than 2.56 times the SD. These will NOT be outliers. Removal of them will improve the appearance of the statistics, but will not change the picture.

Table 3. Illustrating the true meaning of the standard deviation

Mean protein = 12.9 %				
SD = 0.134	Low %	High %	Range	% Outside range
± 1 x SD	12.78	13.03	0.25	67 %
± 1.96 x SD	12.64	13.16	0.52	5 %
± 2.56 x SD	12.56	13.24	0.68	2 %
SD = 0.234				
± 1 x SD	12.67	13.13	0.46	67 %
± 1.96 x SD	12.44	13.36	0.92	5 %
± 2.56 x SD	12.30	13.50	1.20	2 %

The standard error of a single test (SET) is the precision or reproducibility of testing by any method. It is the SD of the results of repeated testing of a check sample, including all of the steps, from sampling, sample preparation to testing. It should be determined for any chemical or other type of test likely to be applied to a check sample or samples of every commodity and constituent or parameter to be tested during the project. The SET values are the foundation stones of all applications of analytical work.

The coefficient of correlation, r and coefficient of determination, r^2 show the degree to which any two sets of results are related to each other. The value of r may be positive or negative. In most cases operators do not need to know the formulæ for calculation of r, or the regression coefficient (see below), because any statistical package will include the correlation and regression options, and r^2 is easily calculated. Coefficients of correlation can become statistically significant at very low values. This depends on the number of observations. For example for a population of 40 a correlation coefficient of only as high as 0.33 is significant at P = 0.05. The r^2 value indicates that only about 10 % of the variance has been accounted for, so the relationship is of little practical value.
The coefficient of determination shows the proportion of variance in y data that is attributable to variance in the x data. For example an r value of 0.97 will give an r^2 value of 0.941. This means that 94.1 % of the variance in Y can be attributed to variance in X (and 5.9 % of the variance is attributable to all of the other factors). The value of r^2 is always positive. Table 4 gives some information on interpretation of the values of r and r^2:

Table 4. Interpretation of Coefficients of Correlation and Determination

Value of r	Value of r^2	Interpretation
Up to ± 0.50	Up to 0.25	Not recommended for use
± 0.51 - 0.70	0.26 - 0.49	Poor correlation: research the reasons
± 0.71 - 0.80	0.50 - 0.64	Suitable for rough screening
± 0.81 - 0.90	0.66 - 0.81	OK for screening and approximate work
± 0.91 - 0.95	0.83 - 0.90	Usable with caution for most applications
± 0.96 - 0.98	0.92 - 0.96	Usable in most applications
± 0.99+	0.98+	Excellent, usable in any application

The regression coefficient b_{yx} (sometimes called the slope) and the intercept (a) show the degree to which values of y can be predicted from those of x. The regression coefficient and intercept can be positive or negative. If the slope is within 0.05 of 1.00, a slope adjustment will not improve the data very much. Slope changes are not usually recommended. If the slope differs from 1.00 significantly (e.g. 0.85 - 1.15) or greater, this means that the calibration will not be stable. The data will be improved by slope adjustment, but the model may not be reliable for prediction of different sample sets. The intercept should not be confused with the bias (see below).

The Analysis of Variance (ANOVA) means just that. The standard deviation is an indication of the variance (basically the degree of homogeneity) that exists in a system. The ANOVA is a system for determination of the main sources of the variance. Again, most statistical packages include an ANOVA option, but the operator needs to identify the believed sources of variance, and set up the data in a form that can be analyzed by the ANOVA system. The experimental design should include replication for these sources of variance.

The simplest arrangement is then to arrange the data corresponding to a source of variance, such as variety, in blocks that correspond to the replication. The ANOVA will then result in three sums of squares, corresponding to varieties, replicates, and a residual sum of squares. The mean sums of squares are obtained by dividing the sums of squares of each source of variance by the associated degrees of freedom (N-1): e.g. if there are 5 replicates, there will be 4 degrees of freedom.

The residual mean sum of squares represents unexplained variance in the data. If the residual mean sum of squares is bigger than the mean sums of squares for varieties or replicates it is likely that the experiment has not been successful in verifying the original hypothesis. This is where the importance of determining the SET for all analytical methods becomes apparent. If operators know that their SETs are reliable they can relate the results of ANOVA testing to the SET for a given constituent or parameter with confidence. Without this assurance it is impossible to assess the value of an ANOVA because a high mean residual sum of squares may be the result of poor analytical data.

From the foregoing, it is apparent that interpretation of the results of any analytical work on grains (or any other commodity) is fundamentally dependent on the precision (reproducibility) of the analytical work carried out. Without this vital information it is impossible to evaluate the validity of differences between groups of samples that represent different sources of variance. To communicate the results of an application of any type of research on grain effectively, in a report or a scientific paper the results of all of these five statistics should be reported.

STATISTICS FOR APPLICATION TO NEAR-INFRARED SPECTROSCOPY ANALYSIS

To record the results of a Near-infrared study up to 16 pieces of information should be included. These are:

1. Source of samples
2. Number of samples
3. Sample preparation and storage method(s)
4. Reference method(s)
5. Standard error of reference methods
6. Standard error of NIRS testing
5. Mean reference data
6. Standard deviation (SD) of reference data
7. Mean NIRS data
8. Bias
9. Standard deviation of NIRS data
10. Calibration evaluation method
11. Standard error of cross-validation (SECV if cross-validation is used) 12. Standard error of prediction (SEP if test-set is used)
13. r and r^2
14. b-value (slope)
15. a-value (intercept) (this is optional)
16/16. RPD or RER (to be explained later)

Details of the samples should include the sources of all samples including e.g. different materials involved in compiling an animal feed mix, growing locations and seasons, classes of grains or seeds, types and sources of ingredients, and others, specific to the application. How many samples were used? Useful guidelines for sample assembly for development of a NIRS calibration model is to use at least 20 samples per wavelength in a Multiple Linear Regression (MLR) calibration and 15-20 samples for a Partial Least Squares (PLS) regression. At least 200 samples should be used in the development of a reliable calibration model. Sample number information should include how many samples were used in the calibration and validation sample sets. Was the application carried out on whole or processed commodities? Details should include whether and how the samples were freed of foreign material. If processed, how were they processed and how were the samples stored before use?

The two main methods for evaluation of the performance of a calibration model are:

a). cross-validation
b). test-set validation

The real value of cross-validation lies in the evaluation of small (up to 100) sample sets, or during optimization of wavelength range and mathematical pre-processing of spectra data. Cross validation involves eliminating samples from model development either singly or in blocks, developing the model without them, then predicting them. The samples that have been removed are then replaced, another sample or samples removed, and so on until all samples have been predicted, while not having been used to develop the model used to predict them. Theoretically, cross-validation is acceptable (particularly "one-out" cross-validation, where samples are eliminated one at a time). In practice, because all of the samples for validation derive from the same population the validation statistics provide no data on bias or slope. When cross-validation is used the reporting should include the method of sample elimination, i.e. the number of groups of samples and the number of samples within a group.

Test-set evaluation means setting up a set of samples separate from those to be used in model development, but carrying the same sources of variance. These may or may not have been identified from the original population. A calibration model is developed and

the test set used to evaluate it. The test set statistics will indicate any slope or bias inherent in development of a calibration model from that population

The error level in NIRS testing is reported in terms of the standard error of cross-validation (SECV) or the familiar standard error of prediction (SEP). These are respectively the SD of differences between reference and NIRS results for cross-validation or test-set validation of the application

When calculated from the NIRS predictions of data in a test set (validation set) the bias (the mean difference between reference and NIRS data) is a measure of the overall accuracy of the calibration (relative to the reference values). Bias can be positive or negative. They can occur even when the SEP and Coefficient of Correlation (see below) indicate that an excellent calibration has been developed. The bias is a very important statistic. In the world of commerce, where premiums and discounts are applied, and feeds are mixed on the basis of analytical results biases mean money (usually big money).

The ratio of the SEP (or SECV) to the SD of the reference data used in validation is called the RPD (Williams et al 1993). It is calculated by dividing the SD of the reference values used in validation or prediction (SD_X) by the SEP. If the SEP is equal to the SD_X (RPD = 1.0) the calibration model is not predicting the reference values at all. The RPD statistics relates the error of prediction to the variance in the sample set, and is a useful statistic to report.

The RER statistic is calculated by dividing the range in reference data by the SEP.

It was invented by Carol Starr and her co-workers in Cambridge, England (Starr et al 1981), and differs from the RPD in that it can be inflated by a single extreme result, whereas the method of calculation of the SEP reduces this source of possible mis-interpretation. Both the RPD and RER statistics should be high if the calibration model is effective. Table 5 illustrates the significance of both of these simple, but very useful statistics

RPD Value	RER value	Classification	Application
0.0 - 2.3	Up to 6	Very poor	Not recommended
2.4 - 3.0	7 - 12	Poor	Rough screening
3.1 - 4.9	13 - 20	Fair	Screening
5.0 - 6.4	21 - 30	Good	Quality control
6.5 - 8.0	31 - 40	Very good	Process control
8.1+	40+	Excellent	Any application

The coefficient of determination (r^2), the bias, and the RPD are the most useful statistics for “instant” appraisal of the efficiency of a NIRS calibration model. The “Background” information gives full details of all aspects of the application.

REFERENCES

1. Williams, P.C. and Sobering, D.C. 1993. J. Near-infrared Spectroscopy 1: 25-32
2. Starr, C., Morgan, A.G. and Smith, D. B. 1981, J. Agric. Sci. 97: 107-118