The Application of ANOVA and Bonferroni *post-hoc* t-tests in the Comparison of Analytical Data from Four Laboratories on a Chromite Concentrate

Allan Fraser

Allan Fraser & Associates, P.O. Box 369, Fourways, 2055, South Africa
allan@allanfraserandassociates.co.za

Abstract

In establishing the final purchase price, mining companies supplying chromite concentrate to buyers will typically have a sample of the concentrate analysed by a referee laboratory to confirm the Cr_2O_3 content. The sample taken is also analysed by the mine laboratory for comparison with the analysis result obtained by the referee laboratory and often two or three additional laboratories may also participate. As there is uncertainty in all analytical measurement it would be expected that there would be differences in the means on a sample analysed by the different laboratories. Typically, the result of the referee laboratory is taken as the value of the consignment and the mining company pays penalties on any differences. This approach is very subjective and could result in the mine paying unnecessary penalties and loosing valuable revenue. This is because a valid decision cannot be made based on a single result alone and all analysis is conducted to make an informed decision.

Therefore an objective statistical test is needed to determine if all the means are statically different, or some are different or none are different. In this work it is shown that should independent replicates of a sample be analysed by several laboratories then an analysis of variance (ANOVA) can be performed, primarily to establish if there are any significant statistical differences between the means of the participating laboratories. Should the ANOVA be significant, *i.e.,* at least one of the means is different then the use of multiple comparison procedures can be applied to establish which mean(s) is different or statistically similar at a specified confidence level. The Bonferroni *post-hoc* t-tests procedure is used to make the comparisons between the participating laboratories.

1. Introduction

In order to determine the final purchase price, mining companies supplying chromite concentrate to buyers are contractually obliged to have a sample of the concentrate analysed by a referee laboratory to confirm the Cr_2O_3 content. Sampling of the concentrate typically would take place during the mechanical transfer of the material into a bulk carrier. The sample taken is also analysed by the mine laboratory for comparison with the analysis result obtained by the referee laboratory. In the case of a dispute between the miner's result and that of the referee, several other laboratories may also be used in order to broaden the scope of the analyses for the consignment. As there is uncertainty in all analytical measurement it would be expected that there would be differences in the means on the identical sample analysed by different laboratories.

Typically each laboratory reports a single value for the shipment sample, which does not allow any meaningful comparative statistical analysis to be done and ultimately may result in subjective conclusions being made about the Cr_2O_3 content of a consignment. Penalties are then paid by the mining company based on the absolute differences between reported values. For example, should the mine laboratory report 44.85% and the referee laboratory 44.01% then the latter is accepted as the consensus value of the consignment. Similarly, should a consignment result be reported as 44.20% by the mine, and the referee laboratory 43.95% then the Cr_2O_3 shipment content is in dispute as the result is below the specification of 44%.

An informal approach to observation of differences in the means can be demonstrated by a graphical investigation, such as using, side-by-side box plots or multiple bars with ±2 standard deviation.

However, to make an objective decision about the differences in means generated on a single sample a more formal statistical approach is needed. Firstly, at least six independent replicates of the sample need to be analysed by each of the participating laboratories so that there is sufficient statistical discernibility between any differences in the data sets. With multiple replicates each data set is subjected to an analysis of variance (ANOVA) with the following hypotheses:

H_0: The means of all the laboratories are equal.

H_1: Not all the means of the laboratories are equal.

In the ANOVA a 95% confidence level is typically used with a level of significance (α) of 5%, or 0.05. A p-value is calculated in the ANOVA from an ANOVA F-statistic and compared to the significance, where the following rules apply:

- A p-value >0.05 is evidence in support of the H_0, since it indicates that there is no significant statistical difference between the means of the different laboratories.
- A p-value <0.05 is evidence for the support of H_1, since it indicates that there is a significant difference between laboratories and within laboratory groups.

By convention, p-values of less than 0.05 are considered 'small'. That is, if p is less than 0.05 there is a less than 1 in 20 chance that a difference as big as that seen in the analysis of the data could have arisen by chance if there was really no true difference. Therefore, p-values with these small (or smaller) results are considered to be statistically significant (unlikely to have arisen by chance). Smaller p-values (*e.g., p*<0.01) are called 'highly significant' because they indicate that the observed difference would happen less than one in 100 times if there was really no true difference (Thompson & Lowthian, 2011).

The disadvantage of ANOVA is that if the means of the laboratories are shown not to be the same (*i.e.* H_1 is supported), it does not indicate which ones differ. In order to indicate which means differ, multiple pair-wise comparisons need to be done in the form of a two-sample, two-tailed t-test assuming unequal variances between each data set from which a p-value for each of the comparisons is found. Such after-the-fact t-tests are called *post-hoc* comparisons (Clark, 2013). The disadvantage of multiple comparisons is that the more comparisons made between data sets, the larger the risk of making a Type I Error, which is the potential error in the rejection of a null hypothesis that is actually true. For a given number of data sets *k,* the number of pair-wise comparisons (*j*) is determined from:

$$j = k\left(\frac{k-1}{2}\right) \quad [1]$$

For *k* number of data sets in the experiment and j pair-wise comparisons, there is a 0.95^j chance, denoted as $\propto_{pc}$ of correctly accepting all true null hypotheses (Jaccard, 1984). The chance of making at least one Type I error is:

$$(1 - 0.95^j) \qquad [2]$$

The risk of a making at least one Type I error increases quadratically with an increasing number of comparisons (Figure 1). To fix this problem requires reducing the probability of making a Type I error without reducing the ability to detect effects or differences in the data sets and ensuring that it never exceeds an exact threshold (*e.g.,* 0.05). This is achieved by using *post-hoc* procedures, of which there are several available including the Bonferroni, Tukey and Scheffe methods. In this work, the Bonferroni method is used as it is the most conservative of the aforementioned procedures and is simple in its application (Olejnik *et al.,* 1997).

In the Bonferroni method a two-sample, two-tailed t-test is done between each data set and a p-value for each of the comparisons is found. A Bonferroni corrected α value is then calculated for each p-value from the t-test as

$$\alpha' = \frac{\alpha}{j} \qquad [3]$$

where; α = level of significance *i.e.* 0.05.

α' = Bonferroni corrected α value.

j = number of data sets compared in the *post-hoc* t-tests.

Therefore, should the Bonferroni corrected α' value be larger than the strict threshold of α=0.05 then there is a 95% chance that the two groups in the t-test are different from one another (Seaman, *et.al.,* 1991).

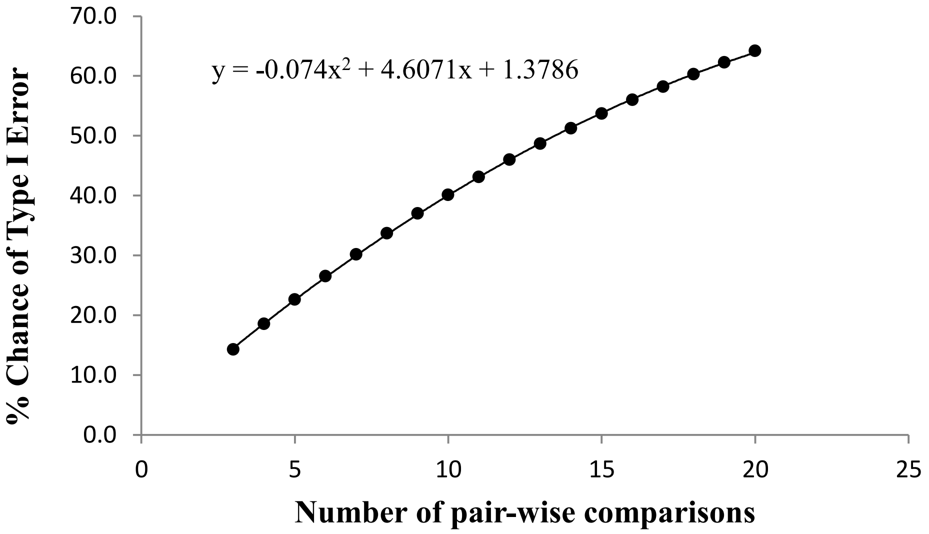

Figure 1. Function of the number of pair-wise comparisons and the %chance of making at least one Type I error.

2. Results and Discussion

2.1 Sample Preparation and Analysis

In this study, a chromite concentrate sample taken from a consignment at the port of loading was crushed and pulverised to 66% passing minus 75µm. The sample was then mechanically split using a rotary splitter with one portion being retained by the referee laboratory (Lab B) and the remaining portion further split and sent to three additional participating laboratories, *i.e.,* Labs A, C and D. Labs A, B and C are accredited with the standard of laboratory competency, *viz,.* ISO17025:2005. The %Cr_2O_3 content was determined on six independent replicates per laboratory using Inductively Coupled Plasma Optical Emission Spectrometry (ICP-OES) and the data obtained are presented in Table 1 and shown graphically in Figure 2.

2.2 Grubbs Outliers Test

The data obtained from the four different laboratories (Table 1) were subjected to a Grubbs test for outliers and datum 44.89 from Lab C was found to be an outlier (Z_{calc}=1.99 and Z_{crit} for N=6 is 1.89). This value was therefore not included in the mean value for Lab C.

2.3 Pair-Wise Comparisons and Probability of Type I Error

Since there are four different laboratories viz., A, B, C, D, there will be six pair-wise comparisons differences *i.e.,* Labs: A-B, A-C, B-C, A-D, B-D, C-D, or using Equation [1], the number of comparisons will be:

$$j = 4\left(\frac{4-1}{2}\right) = 4\left(\frac{3}{2}\right) = 6$$

For four data sets with six pair-wise comparisons, there is $0.95^6 = 0.735$ or a 73.5% chance of correctly accepting all true null hypotheses. From the six pair-wise comparisons the probability of a making at least one Type I error is found using Equation [2]:

$$(1 - 0.95^6) = (1 - 0.735) = 0.265$$

Therefore, there is a 26.5% probability of making at least one Type I error should a Bonferroni correction not be applied.

2.4 Bonferroni Corrected *p*-value

For the six pair-wise comparisons the Bonferroni corrected *p*-value is found using Equation [3]:

$$\alpha' = \frac{0.05}{6} = 0.0083$$

Therefore the Bonferroni corrected *p*-value is 0.0083.

2.5 ANOVA and Bonferroni

Following the Grubbs test, the four data sets were subjected to ANOVA which gave an ANOVA F-statistic of 37 and a *p*-value of <0.001 (Table 2). Since the ANOVA *p*-value is less than 0.05, the null hypothesis that there is no difference in the data sets is rejected and at least one of the data sets shown in Table 1 is statistically significant. Applying the correction of Bonferroni, the mean %Cr_2O_3 of Lab A and Lab D were found not to be statistically different (*p*=0.48 and greater than the Bonferroni corrected *p*-value of 0.0083) in contrast to the remaining comparisons that all show statistical significance, *i.e., p*<0.001 and less than the Bonferroni corrected *p*-value of 0.0083 (Table 3). Therefore, since the Lab A and D's data sets are not statistically different they belong to the same population with a pooled mean of 44.48% Cr_2O_3 for the test sample.

If the mean obtained of the referee laboratory were accepted simply because of its status as a 'referee laboratory', based on the outcome of the *post-hoc* tests, it could be argued, that the shipment value would be underestimated in terms Cr_2O_3.

Table 1. %Cr_2O_3 on sample from four different laboratories. These data are shown graphically in Figure 2. Datum 44.89 is an outlier (Z_{calc}=1.99 and Z_{crit} for N=6 is 1.89) and is not included in mean of Lab C. *s* is the sample standard deviation. Lab B is the referee laboratory.

	Lab A	**Lab B**	**Lab C**	**Lab D**
	%Cr_2O_3	**%Cr_2O_3**	**%Cr_2O_3**	**%Cr_2O_3**
	44.45	44.17	43.84	44.29
	44.42	44.28	43.73	44.66
	44.26	44.28	43.93	44.44
	44.40	44.19	43.84	44.55
	44.66	44.22	44.00	44.49
	44.52	44.24	***44.89***	44.62
Mean=	44.45	44.23	43.87	44.51
s=	0.133	0.0456	0.102	0.134

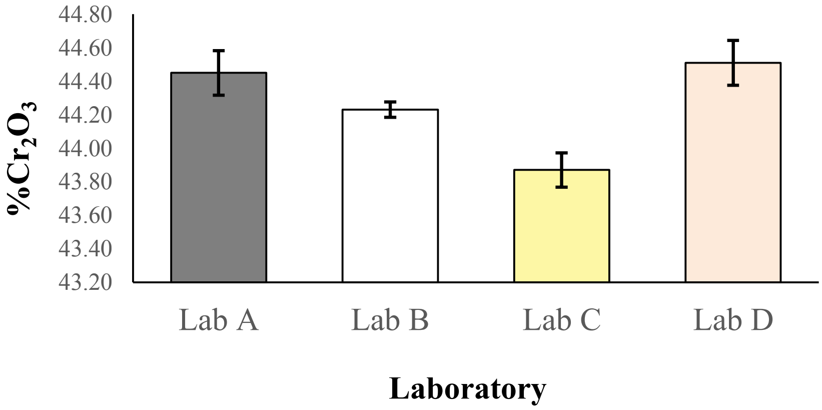

Figure 2. Mean %Cr_2O_3 for six replicates on a chromite sample from four different laboratories. The error bars represent ±2 standard deviations.

Table 2. Results of an ANOVA on the four data sets per laboratory. Since p<0.001 and less than 0.05, the null hypothesis that there is no difference in the data sets is rejected and at least one of the data sets shown in Figure 1 is statistically significant. SS is the sum of squares, MS the mean sum of squares, *df* the degrees of freedom, F is the ANOVA *F*-statistic, and F_{crit} is the F critical value.

Source of Variation	SS	*df*	MS	F	*p*	F_{crit}
Between Laboratories	1.35	3	0.451	37	<0.001	3.12
Within Laboratories	0.23	20	0.0121			
Total	1.58	22				

Table 3. Results of Bonferroni corrected *post-hoc* t-tests for each group of the four data sets. The Bonferroni corrected *p*-value is 0.0083, therefore the reported *p*-value from the t-test if less than the corrected *p*-value indicates that there is a 95% chance that there is significant difference between the groups compared.

Pair-wise Comparison	*p*-value	Outcome
Lab A and Lab B	0.0032	Different
Lab A and Lab C	<0.001	Different
Lab A and Lab D	0.48	Not different
Lab B and Lab C	<0.001	Different
Lab B and Lab D	<0.001	Different
Lab C and Lab D	<0.001	Different

3. Conclusion

The use of ANOVA indicates that there is at least one of the means for Labs A-D to be different. The use of *post-hoc* comparisons showed that the mean values reported for Cr_2O_3 from Lab A and D are statistically the same at a 95% confidence level. In contrast, the remaining mean comparisons are all significantly different. A pooled mean is calculated from the mean values obtained on six replicates each of Lab A and D giving a consignment value of 44.48% $Cr_2O_{3.}$

This work shows that in accepting the mean value obtained by a referee laboratory simply because the laboratory has the status of a 'referee laboratory', that the value of a consignment may be underestimated (or overestimated) in terms of Cr_2O_3. It is preferable to determine which data sets are statistically similar and use a pooled mean as the Cr_2O_3 value for the consignment of chromite concentrate.

The combination of ANOVA significance testing and the use of Bonferroni *post hoc* t-tests isolate significant differences between data sets and could be adapted to other commodities in which there is a likelihood of subjective decisions being made in the quality.

4. References

Clark, J. (2013). Performing a one-way ANOVA in Excel with post-hoc t-tests. Retrieved May 31, 2014, from YouTube:https://www.youtube.com/watch?feature=player_detailpage&v=tPGPV_XPw-o

Jaccard, J., Becker, M. A., Wood, G. (1984). Pairwise multiple comparison procedures: A review. Psychological Bulletin **96 (3):** 589

Olejnik,S., Li, J., Supattathum, S., and Huberty, C.J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. Journal of educational and behavioral statistics, ***22***, 389-406.

Seaman, M.A., Levin, J.R., & Serlin, R.C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. Psychological Bulletin, **110**, 577-586.

Thompson, M., & Lowthian, P. (2011). Notes on statistics and data quality for analytical chemists. Imperial College Press. 15-115.

5. Acknowledgements

I wish to thank Mr. Silas Hlapolosa and Mrs. Patience Sedidi as well as the laboratory personnel at Dwarsrivier mine for their kind assistance in the generation of the analytical data and for Assmang's permission to reproduce the data for this paper.

Allan Fraser

Allan Fraser is a consulting analytical chemist. He holds a national diploma in analytical chemistry and a master’s degree in geology from the University of Johannesburg. Allan is a former director and board member of Panalytical (Pty) Ltd., a position he held for six years. In the early part of his career he worked for Matthey Rustenburg Refiners as a research chemist and later at Sandvik Hard Materials as laboratory manager. Allan gained further experience as operations manager of three laboratory sites at SGS Société Générale de Surveillance and commercial experience with LECO Africa as sales engineer and spectroscopy manager.

In 2008, Allan established Allan Fraser & Associates, an analytical and geochemical consulting company. Allan’s area of expertise is in analytical method development and in the use of statistical methods in analytical chemistry and he has a good working knowledge of a variety of analytical methods including, XRF, XRD, LECO combustion, AAS, GF-AAS, GC, GC-MS, ICP-OES, ICP-MS and GD-OES.

He facilitates regular public courses on statistical method validation for test laboratories and measurement uncertainty. In his spare time Allan enjoys collecting aesthetic mineral specimens and has been doing so for the good part of four decades and today boasts a collection of some 4500 specimens from mainly southern Africa and other international localities.

ALLAN FRASER
& associates