Received: 17 January 2024 | Revised: 5 March 2024 | Accepted: 21 March 2024

DOI: 10.1002/cem.3549

TUTORIAL

Testing differences in predictive ability: A tutorial

Tom Fearn

Department of Statistical Science, University College London, London, UK

Correspondence
Tom Fearn, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK.
Email: t.fearn@ucl.ac.uk

Abstract

This paper describes some statistical tests for comparing the predictive performance of two or more prediction rules. It covers the cases of both quantitative and qualitative predictions, that is, both regression and classification problems. Worked examples are included for both cases.

KEYWORDS

model comparison, prediction, qualitative, quantitative, statistical tests

1 | INTRODUCTION

Many papers in the area of chemometrics compare two or more methods for generating prediction rules, usually with the result that the new method developed by the authors is claimed to predict better than the ones it is compared with. Sometimes the difference in predictive performance is so large that statistical methods are not really needed to back up this claim. More often, the difference is small to modest, and it would be useful to have some reassurance that small differences are not being over interpreted as meaningful. The aim of this paper is to describe some statistical approaches to this problem in the cases of both quantitative and qualitative predictors, giving enough detail to enable the tests to be implemented without the need to consult the cited references.

The distinction made here between a method, examples of which would be partial least squares regression (PLSR) or linear discriminant analysis, and a prediction rule, which is used here to denote a fixed recipe for converting raw input data into a quantitative or qualitative prediction, perhaps needs some elaboration. Using the method of PLSR on near infrared (NIR) spectroscopic data, for example, a rule would typically comprise taking the data in a specified spectral range, applying a specified spectral pretreatment and using a completely specified linear equation to convert this treated data into a prediction. The step from method to rule is that of training. The point of making the distinction is that the statistical tests described in what follows do not compare methods; they compare rules, and this should be remembered when discussing results.

The context is the comparison of the performance of two or more prediction rules on a set of samples that have not been used to train them. This might be achieved by predicting a completely separate test set of samples to those used for training, or it might be achieved by cross-validation. In either case, the comparison is only valid if the predicted samples have not been used at all in the tuning of the rules. It would be valid to use the tests described below as part of tuning a rule, for example, to establish whether there really is any difference between the cross-validatory performance with various candidate spectral pretreatments. However, once the cross-validation or separate test set has been used in this way, it has become part of the training and cannot be used to compare the trained rule with another rule. It is particularly important not to make this mistake when methods of different complexity are compared. The method with more adjustable parameters has more chance to adapt to data it has seen. Its rule will almost always win unless the comparison with the rule produced by the simpler method is made on genuinely unseen samples.

The statistical tests described below are all of traditional form, t-tests, F-tests and analysis of variance (ANOVA), together with some nonparametric alternatives. What is not included are computer intensive tests involving permutations or Monte Carlo. The purpose of this paper is to encourage the use of tests in cases where performance improvements are not clearly and obviously real ones by presenting some methods that can be used without a lot of extra work.

2 | QUANTITATIVE PREDICTIONS

When two rules make quantitative predictions for the same set of n samples, the raw data on which to compare their performance consist of a set of n paired prediction errors, the two errors on each sample forming a pair. Let e_{ij} be the error of rule i on sample j. At first sight, a tempting approach might be to compute the mean squared errors (MSEs) for each of the rules separately, using

$$MSE_i = \frac{\sum_{j=1}^{n} e_{ij}^2}{n}, \tag{1}$$

and compare them by applying an F-test to their ratio, this being the standard way of testing equality of variances in the case of normally distributed data.[1] There are two problems with this approach. One is that if either or both of the rules has a bias that contributes more than a very small part of the errors, then the ratio of their MSEs will not have an F-distribution even if the errors are normally distributed. They need to be normally distributed about zero, not about some nonzero bias, for the distribution to hold. The other is that the F-test requires two independent estimates of variance, and the pairing violates this assumption. These objections are not statistical nit-picking: The tests described below will have more power for proving differences than an F-test of the MSEs, in some cases considerably more power.

2.1 | Test bias and variance separately

Fearn[2] describes an approach in which bias and variance are tested separately, with the test for comparing variances allowing for correlation between the two sets of errors. The first step is to separate the MSE into its two components: bias and variance. Separately for each rule, we calculate its bias

$$b_i = \frac{\sum_{j=1}^{n} e_{ij}}{n}, \tag{2}$$

and variance

$$v_i = \frac{\sum_{j=1}^{n} \left(e_{ij} - b_i\right)^2}{n-1}. \tag{3}$$

These would combine to give the MSE in Equation (1) via

$$MSE_i = b_i^2 + \frac{n-1}{n} v_i; \tag{4}$$

thus, MSE = bias squared + variance, give or take an $(n-1)/n$.

A difference in biases between the two rules may be tested using a paired t-test.[1] If we let $d_j = e_{1j} - e_{2j}$ be the difference between the errors for sample j, then

$$\overline{d} = \frac{\sum_{j=1}^{n} d_j}{n} \tag{5}$$

is the difference in biases, $b_1 - b_2$, and a t-statistic for testing whether the true difference is zero is

$$t_{bias} = \frac{\sqrt{n}\bar{d}}{s_d}, \tag{6}$$

where

$$s_d = \sqrt{\frac{\sum_{j=1}^{n}(d_j - \bar{d})^2}{n-1}} \tag{7}$$

is the standard deviation of the d_j. The statistic t_{bias} should be compared with the percentage points of a t-distribution on $n-1$ degrees of freedom to establish the significance level.

To test for a difference between variances, we need to first calculate the squared correlation between the two sets of errors

$$r^2 = \frac{\left[\sum_{j=1}^{n}(e_{1j} - b_1)(e_{2j} - b_2)\right]^2}{\sum_{j=1}^{n}(e_{1j} - b_1)^2 \sum_{j=1}^{n}(e_{2j} - b_2)^2}. \tag{8}$$

Then if F is the larger of the variance ratios v_1/v_2 and v_2/v_1, a t-statistic for testing equality of variances is given by

$$t_{var} = \frac{F-1}{2}\sqrt{\frac{n-2}{(1-r^2)F}}. \tag{9}$$

The statistic t_{var} should be compared with the percentage points of a t-distribution on $n-2$ degrees of freedom. The originator of this test of equality of correlated variances was Pitman.[3] It is described by Snedecor and Cochran.[4] The treatment above differs from that in Fearn,[2] where a confidence interval for the true ratio is presented, rather than a test statistic, but the two versions are equivalent.

Both of these tests rely on assumptions of normal distributions. For the t-test, the differences should be normally distributed; for the variance test, the errors should follow a bivariate normal distribution. The t-test is generally considered not to be particularly sensitive to departures from normality. Tests on variances are more easily affected by outliers or other departures from normality. The biases could be tested avoiding the normality assumption by applying a nonparametric alternative to the t-test, the Wilcoxon signed rank test,[1] to the differences. There is no obvious nonparametric alternative to the F-test. The approach described in Section 2.2 does have nonparametric versions and would be preferable if there is serious doubt about the normality assumption. The obvious 'safe' option of always using the nonparametric test comes at the price of a loss of power, with a nonparametric test usually having a larger p-value than the corresponding parametric test on the same data. Many statistical texts suggest that one should always carry out appropriate tests for normality before using any test that assumes normality. This approach is not without its problems. With large amounts of data, tests of normality become very powerful and are quite likely to flag up departures from normality that have no practical importance. The author's personal preference is not to carry out routine normality tests but to plot the data in some way; see, for example, the plots in Section 2.5 and check visually for any large outliers or unexpected patterns, on the grounds that if you cannot see it, it probably does not matter very much. Very large prediction errors on one or two samples are the main problem to look out for. Not only would this invalidate the tests; the presence of these errors may be the main reason one rule appears better than another. If you do prefer to carry out tests of normality, they are easy enough to find in statistical packages.

2.2 | Work with absolute or squared errors

An alternative approach, which sidesteps the bias issue and copes in a different way with the pairing, has been described by Indahl and Næs[5] and explored in more detail by Cederkvist et al.[6] We begin by describing the case of two rules, where this approach reduces to a paired t-test using either the absolute values of the errors or their squares.

If we let $d_j = |e_{1j}| - |e_{2j}|$ or $d_j = e_{1j}^2 - e_{2j}^2$, then a t-test carried out using these differences and the formulas in Equations 5–7 will test not for a difference in bias but for a difference in the average size of errors between the two rules. This test has two advantages over the approach described in Section 2.1. It avoids having to separate bias and variance, and it is simpler to calculate. The one possible drawback is that even if the errors themselves are normally distributed, the differences of absolute values or squares used to calculate t will not be. Cederkvist et al[6] present some evidence based on real data sets suggesting that using the absolute values of the errors is the preferred option for getting closer to the correct distribution for t. If you were really nervous about violating distributional assumptions, you could apply the Wilcoxon signed rank test[1] to the differences.

2.3 | More than two rules

The approach of Section 2.1 does not generalise to the simultaneous comparison of multiple rules unless one is prepared to assume that all the pairwise correlations between predictions are equal.[7] The approach of Section 2.2 does and indeed is presented in its general form in the two references cited. With $r > 2$ rules, the statistical analysis becomes an $r \times n$ two-way ANOVA on the either the absolute values $|e_{ij}|$ or the squares e_{ij}^2 of the errors, with the absolute values being the preferred option.[6] This analysis, available in almost any statistics package, will provide an overall F-test for no difference between any of the r rules. When $r = 2$, this F-test is equivalent to the t-test described in Section 2.2. There are no replicates, that is, there is just one prediction for each combination of sample and rule, so the analysis is a two-way ANOVA with no interaction. What would have been the interaction sum of squares is used to estimate the variance of what are assumed to be normally distributed random errors in the observations. Whether the sample effects are regarded as fixed or random makes no difference to the test for rule effects in this situation.[5,6] Most packages will also offer a set of pairwise comparisons, typically adjusting the significance levels for the fact that multiple tests are being carried out. Some packages will also offer to carry out appropriate tests of normality, in this case on the residuals from an additive fit (mean+sample effect+rule effect) to the observed data. A simple alternative is to look at plots like those in Section 2.5 for each of the rules and check for outliers. There is a nonparametric alternative, the Friedman test,[8] which is also supported by many statistical packages.

2.4 | Ignore the problems and use the *F*-test anyway

Both of the above approaches require access to all of prediction errors, not just the usual summary statistics. These data are not always available, the obvious example being when one wishes to assess the significance of results in a published paper. Fearn[9] discusses the implications of using a simple F-test on the ratio of either variances or MSEs in this situation.

The correlation between the two sets of errors is induced by the presence of common sources of variability, the most obvious one being that there is usually a contribution from the error in the reference measurement with which both predictions are compared. As a result, the correlation will typically be positive, and in this situation, the F-test on the variances that ignores the correlation will be conservative, in the sense that it has less power than the correct test. That is to say that if the simple F-test is significant, the test on variances in Section 2.1 would also have been significant, but if the simple F-test is not significant, one does not know whether the other test would or would not have been. The loss of power is analogous to that which usually occurs when a two-sample t-test is used instead of a paired t-test when the latter is appropriate. If the simple F-test is all you can do, it is worth trying, but it may leave you with no firm conclusion.

The statement in Fearn[9] that the F-test applied to the ratio of MSEs will tend to be even more conservative is an over simplification of what is quite a complicated situation. Applying the simple F-test to the MSEs is even more problematic than applying it to the variances and is best avoided unless one is confident that any biases present are very small.

2.5 | Example

The corn data, a favourite example for algorithms seeking an application, may be downloaded from the Eigenvector website.[10] The data set comprises near infrared spectra measured on 80 corn samples on each of three instruments, together with four compositional variables for each sample. For this illustration, the first instrument (m5) and the first compositional variable (moisture) were selected. The first 60 of the 80 samples were taken as a training set and the remaining 20 as a prediction set. This is not a recommended way to split data; a random split would be preferable, but this one is easy to reproduce. Two calibrations were made using the PLS Toolbox (Eigenvector Research, Manson WA, USA), with the only pretreatment being mean centering of the X (spectral) data and with 10-fold cross-validation using venetian blinds. One used PLSR,[11] accepting the software's recommendation of six factors; the other used principal component regression (PCR),[11] this time accepting the software's recommendation of seven factors. These two calibrations were then used to predict the 20 split off samples. The prediction errors for the two calibrations, rounded to four decimal places, are shown in Table 1. The errors were calculated as predicted minus observed. Taking the difference the other way round would change the signs of the errors and biases but nothing else.

Using these rounded errors and the formulas in Equations (1)–(3) gives the statistics in Table 2.

The PLSR predictions look better than the PCR predictions on all of these measures. This superiority may be seen in the two scatterplots in Figure 1, where both the tendency for the PCR errors to be larger than those for PLSR and for the PCR to predict a little low are visible. There are no obvious outliers that might invalidate the tests, so we proceed to carry them out.

The *t*-statistic for comparing the biases is $t_{bias} = 5.13$, with 19 degrees of freedom and a *p*-value of 0.00006. To compare the variances, we need their ratio, $F = 2.3919$, and the squared correlation between the two sets of errors, $r^2 = 0.8884$. Putting these into Equation (9) gives $t_{var} = 5.715$ with 18 degrees of freedom and a *p*-value of 0.00002. The observed differences in bias and variance are very unlikely to be due to chance. The paired *t*-test of Section 2.2 using absolute values of the errors gives $t = 2.46$ with 19 degrees of freedom and a *p*-value of 0.024, larger than those from the other approach but still strong evidence for a real difference in performance between the two calibrations. For completeness, comparing the variance ratio, $F = 2.3919$, with an *F*-distribution on 19 and 19 degrees of freedom give a *p*-value of 0.065. As discussed in Section 2.4, ignoring the correlation between the two sets of errors leads to a loss of power.

Using the Matlab Statistics and Machine Learning Toolbox, and with the two sets of prediction errors in 20×1 vectors e1 and e2, the command [h,p,ci,stats]=ttest(abs(e1),abs(e2)) will return the value of the *t*-statistic (+2.46 or −2.46 depending on which set of errors is in which of the two vectors) as one of the contents of the structure stats and the *p*-value, 0.024, in p. As an alternative, the command [p,tb]=anova2([abs(e1),abs(e2)],1) returns an ANOVA table in the

TABLE 1 Prediction errors for PLSR and PCR.

Sample	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
PLSR	−0.0830	−0.1034	−0.0856	−0.1235	−0.1315	−0.1056	0.0955	0.0545
PCR	−0.1809	−0.2683	−0.2631	−0.2820	−0.2709	−0.1872	0.1204	−0.0214
Sample	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
PLSR	−0.0506	0.0181	0.0795	0.0588	0.0192	0.0241	0.1718	0.0643
PCR	−0.1739	−0.0505	0.1287	0.0893	−0.0165	0.0126	0.1179	−0.0459
Sample	**17**	**18**	**19**	**20**				
PLSR	0.1656	0.0751	0.0987	0.0145				
PCR	0.1113	0.0229	0.0366	−0.1090				

Abbreviations: PCR, principal component regression; PLSR, partial least squares regression.

TABLE 2 Prediction statistics for PLSR and PCR.

	PLSR	**PCR**
Bias	0.012825	−0.061495
Variance	0.008826	0.021111
MSE	0.008548	0.023837

Abbreviations: MSE, mean squared error; PCR, principal component regression; PLSR, partial least squares regression.

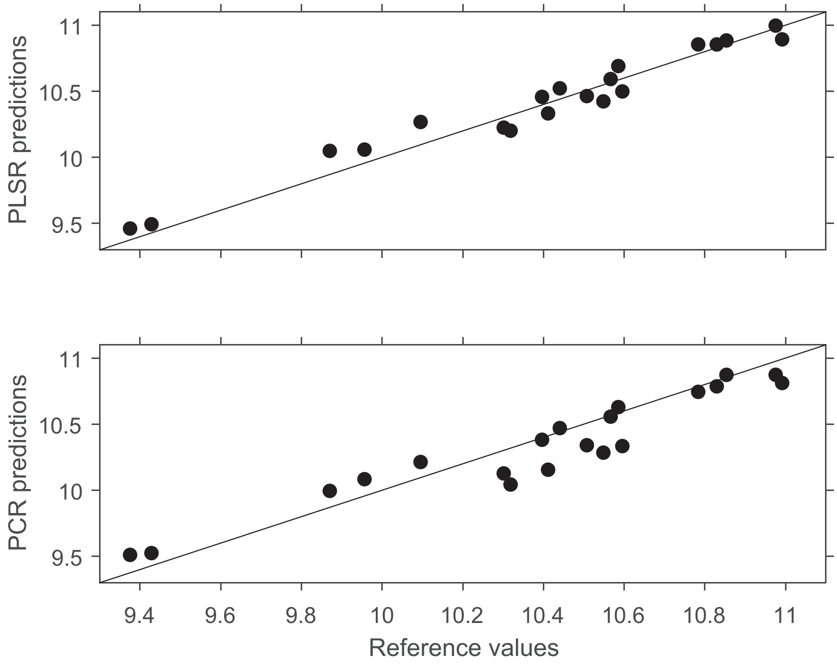

FIGURE 1 Predicted versus reference values for moisture using PLS and PCR. PCR, principal component regression.

cell array tb. The first row of this ANOVA table ends with the value, 6.05, of the F-statistic for testing for a difference in performance between the two calibrations and the corresponding p-value, once again 0.024 because the t-test and F-test are equivalent in the case of two rules. The F-statistic is the square of the t-statistic. With more than two rules, just add more columns of errors, for example, [abs(e1),abs(e2),abs(e3)].

An additional comment on the data is that the cluster of samples that predict below the diagonal line for PLSR and even further below for PCR are actually samples 1–6 of the test set of 20, so 61-66 of the set of 80. This suggests that there is some structure in the data set and that the systematic split was indeed less than ideal.

The comparison above is one between one particular PLSR calibration and one particular PCR calibration, neither of which has been optimised with great care and has little or nothing to contribute to the probably unresolvable question as to whether one of these methods is to be preferred to the other in general.

3 | QUALITATIVE PREDICTIONS

The situation in which the prediction rules to be compared simply assign unknowns to one of two or more classes is much easier to deal with. We begin with the case of two rules and two classes.

3.1 | Two rules and two classes

For the purpose of comparing the rules, the results may be summarised as in Table 3 below. The entries in the table are the numbers of test samples falling in each category; for example, b is the number of samples correctly classified by rule 1 but incorrectly classified by rule 2.

It is the samples on which the rules disagree that provide the information for comparing them. If $b = c$, then the error rates for the two rules will be identical. Usually b and c will not be exactly equal, but it is easy to test whether their difference is larger than could reasonably be ascribed to chance. As the labelling of the two rules is arbitrary, we may arrange that b is the smaller of b and c. In the case of no real difference, b will be a draw from random variable B with a binomial distribution with index $b + c$ and probability 0.5. Then the probability that $B \leq b$ is given by the cumulative distribution function (cdf) of the binomial distribution

TABLE 3 Classification success for two rules.

	Rule 2	
Rule 1	**Correct**	**Error**
Correct	a	b
Error	c	d

TABLE 4 Critical values for binomial test.

$b+c$	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
b_{crit}	0	0	0	1	1	1	2	2	2	3	3	4	4	4	5

TABLE 5 Classification success for LDA and PLSDA on the iris data.

	LDA	
PLSDA	Correct	Error
Correct	95	3
Error	0	2

Abbreviations: LDA, Linear discriminant analysis; PLSDA, partial least squares discriminant analysis.

$$p(B \leq b) = \sum_{r=0}^{b} \binom{b+c}{r} 0.5^{b+c}, \tag{10}$$

which needs to be multiplied by two to give the two-tailed p-value. The formula for the binomial cdf is included here for completeness, but it is widely available as a function in statistical packages, as well as in Excel. For example, the Matlab command p=2*cdf('Binomial',b,b+c,0.5) computes the two-tailed p-value using the cdf function in the Statistics and Machine Learning Toolbox.

The days of publishing extensive statistical tables are over, but one short table does seem worth including here.

For values of $b+c$ up to 20, Table 4 shows b_{crit}, the largest value of b for which a significance level of $p<0.05$ is achieved. These numbers may come as a surprise to readers accustomed to interpreting small differences in classification accuracy as meaningful. It is not possible to achieve statistical significance at $p<0.05$ with fewer than six samples on which the rules differ and even with 20 such samples, only a split as extreme as or more extreme than 5:15 achieves such significance.

3.2 | Example

The Fisher iris data[12,13] are perhaps the most famous classification example data set. Taking just two of the three iris species, Versicolour and Virginica, gives a data set with $n=100$ samples, 50 of each species, and 4 predictor variables. Linear discriminant analysis (LDA)[12] gets 95/100 predictions correct using 10-fold venetian blinds cross-validation, while partial least squares discriminant analysis (PLSDA)[14] with my choice of two factors beats this with 98/100 correct. Is 98% really better than 95% here?

Examining the predictions, 95 of the 100 samples are classified correctly by both rules, and two Versicolour samples (21 and 34) are incorrectly classified as Virginica by both rules. PLSDA make no further errors, but LDA misclassifies a further two Versicolor samples (19 and 23) and one Virginica sample (84). These results are shown in Table 5. Note that the table shows numbers of samples. These happen to be equal to the percentages for this particular data set, because $n=100$, but this is not true in general, and it is the numbers of samples that are needed, not the percentages.

Comparing this with Table 3, we have $b+c=3$ and $b=0$, so

$$p(B \leq 0) = \sum_{r=0}^{0} \binom{3}{r} 0.5^3 = \binom{3}{0} 0.5^3 = 0.125, \tag{11}$$

and doubling this gives a p-value of 0.25. Thus, a 0:3 or 3:0 split of the samples on which the rules disagree would not be at all surprising if the rules were equally accurate, and to claim that PLSDA is truly better than LDA here would be to overinterpret a difference that may well be due to chance.

3.3 | More than two classes

When there are more than two classes, the overall performance of two rules can be compared exactly as above, because Table 3 can still be constructed. Alternatively, it would be possible to focus on any one class by pooling the other classes after the classification results are computed, thus constructing Table 3 by counting a prediction of a sample not from the class of interest as correct so long as it assigns the sample to any one of the other classes.

3.4 | More than two rules

It is not obvious, to this author at least, how to extend the above approach to the comparison of several rules, except by carrying out multiple pairwise comparisons. If this is done, consideration should be given to adjusting significance levels to account for the multiple testing. The simplest way to do this is to multiply the achieved p-values by the number of tests carried out. This approach, the so-called Bonferroni correction,[1] protects against the worst case of independent tests and so is generally rather conservative. If you do not make any formal adjustment to the p-values, you should at least be aware that if you carry out 20 significance tests, you could reasonably expect one of them to be significant at $p < 0.05$ just by chance.

3.5 | Other comments

When $b+c$ is large enough, greater than 30 say, an alternative test may be applied to the data in Table 3. McNemar's test,[1] essentially a chi-squared goodness of fit test of the binomial distribution with probability 0.5, compares

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{12}$$

with the percentage points of a chi-squared distribution on one degree of freedom. This test is approximate, and there seems little value in preferring it to the exact binomial calculation unless access to the binomial cdf is not available.

The fact that both rules predict the same samples means that the seemingly obvious option of directly comparing the overall accuracies of the rules is not valid. The usual tests for this require independent, not paired, predictions. This is the discrete analogy of the issue of correlated predictions that complicates the quantitative case. For this reason, assessing the significance of differences in performance in a published paper is not usually possible if the author has not done this. Unless one rule has 100% success, it is not possible to construct Table 3 from the two overall accuracies alone.

There are some implications here for experimental design. It is the borderline samples that distinguish between rules, not the easy or the impossible ones. Including quite a few borderline samples in a validation set may reduce the reported overall accuracy of your novel classification method, but it will improve your chance of demonstrating its superiority over the standard approach it is compared with, assuming of course that it really is superior. The other point to note is that classification rules need to be compared on large numbers of samples to have any chance of demonstrating the superiority of one over another.

4 | A FINAL COMMENT

The aim in presenting this tutorial was not to propose that chemometric papers should be peppered with p-values but to encourage researchers to think twice before making claims that are not supported by the evidence and to provide some appropriate tools for checking the validity of those claims.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/cem.3549.

ORCID

Tom Fearn https://orcid.org/0000-0003-2222-6601

REFERENCES

1. Mohr DL, Wilson WJ, Freund RJ. *Statistical Methods*. 4th ed.: Academic Press; 2021.
2. Fearn T. Comparing standard deviations. *NIR News*. 1996;7(5):5-6.
3. Pitman EGG. A note on normal correlation. *Biometrika*. 1939;19:9-12.
4. Snedecor GW, Cochran WG. *Statistical Methods*. 7th ed.: Iowa State University Press; 1980.
5. Indahl UG, Næs T. Evaluation of alternative spectral feature extraction methods of textural images for multivariate modelling. *J Chemometrics*. 1998;12:261-278.
6. Cederkvist HR, Aastveit AH, Næs T. A comparison of methods for testing differences in predictive ability. *J Chemometrics*. 2005;19:500-509.
7. Han C-P. Testing the homogeneity of a set of correlated variances. *Biometrika*. 1968;2:317-328.
8. Glover T, Mitchell K. *An Introduction to Biostatistics*. McGraw-Hill; 2002.
9. Fearn T. Comparing standard deviations (continued). *NIR News*. 2009;20(7):24-25.
10. https://www.eigenvector.com/data/Corn/
11. Martens H, Næs T. *Multivariate Calibration*. Wiley; 1989.
12. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen*. 1936;7(2):179-188.
13. Fisher RA. Iris. *UCI Machine Learn Reposit*. 1988. doi:10.24432/C56C76
14. Barker M, Rayens W. Partial least squares for discrimination. *J Chemometrics*. 2003;17(3):166-173.

How to cite this article: Fearn T. Testing differences in predictive ability: A tutorial. *Journal of Chemometrics*. 2024;e3549. doi:10.1002/cem.3549