Effectively teaching statistics to chemical engineers as a core professional attribute

Damien Batstone[a], Bronwyn Laycock[a], Greg Birkett[a], Richard Wilson[b], and Maria Jose Farre[a]
Department of Chemical Engineering, The University of Queensland, Australia[a], Department of Mathematics, The University of Queensland, Australia[b]
Corresponding Author Email: damienb@awmc.uq.edu.au

Structured abstract

BACKGROUND
Engineering statistics has been historically a challenge to teach due to a focus in engineering on uncertainty observation and process optimisation rather than population analysis and active experimentation. This is particularly the case in chemical engineering statistics, with a strong process focus, continuous rather than discrete data sets, and underlying system (and model) non-linearity. However, to properly understand and apply advanced techniques, a basic understanding of inferential statistics is required. This offers an opportunity to approach statistics from a process orientated approach at an early stage to allow chemical engineers to develop statistics capabilities as a core attribute.

PURPOSE
In this paper, we approach statistics from a process engineering point of view, to implement process experimentation and statistics at an early stage (2nd year undergraduate level).

DESIGN/METHOD
A course was developed linking inferential statistics to chemical engineering from the ground up, separating major topics into observation, estimation, and propagation of uncertainty. This was a mixed-mode course with lecture, practical, computing tutorial, and project management elements addressing all components. Due to classic inferential and engineering statistics not covering major elements at a basic level (e.g., non-linear regression, uncertainty propagation), a substantial amount of new material needed to be developed.

RESULTS
The course has now run consecutively for two years (126 students year 1, 145 students year 2) as CHEE2010 at UQ. 2013 results are not yet available, but 2012 indicated very strong student satisfaction with clear understanding of the links between professional attributes and the material being delivered. Pass rate was high (92%), and engagement was very high.

CONCLUSIONS
There are strong benefits to teaching statistics in a process engineering context and this can provide vertical integration and ancillary skills development that provide a better subsequent programme experience, and substantially assist with key graduate attributes, particularly related to addressing risk and uncertainty.

KEYWORDS
Statistics, parameter estimation, uncertainty, risk

Introduction

Engineering statistics has been historically a challenge to teach due to a focus in engineering on uncertainty observation and process optimisation rather than population analysis and active experimentation. This is particularly the case in chemical engineering statistics, with a strong process focus, continuous rather than discrete data sets, and underlying system (and model) non-linearity. As such, there is a strong need for advanced techniques such as non-linear parameter estimation and I/O uncertainty propagation, and a weaker need for elements such as binomial statistics that are commonly a core part of engineering and general statistics. However, to properly understand and apply advanced techniques, the inference framework needs to be fully developed. As an example, the basis of uncertainty analysis in non-linear parameter estimation is an ANOVA (and related hypothesis test) to identify the limits of model validity. Without understanding the principles of hypothesis testing and model fit analysis through ANOVA, it is not possible to understand how non-linear parameter uncertainty can be derived, without which parameter estimation in general is not meaningful.

However, this also offers an opportunity, as core process engineering concepts such as I/O systems, non-linear models, and I/O propagation response can be developed at an early stage and fully integrated with an understanding of the significance of statistics to chemical engineering experimentation and system analysis.

Key elements that are very important to process engineering that are not covered in more general and widely used engineering statistical texts (Devore & Berk, 2012; Ryan, 2007) are:-

- Non-linear regression (non-linear parameter estimation) (Bevington & Robinson, 2003) and extension of inferential statistics to estimating uncertainty in parameters obtained from non-linear regression (Dochain & Vanrolleghem, 2001). The majority of chemical engineering problems are non-linear and particularly extension of inferential statistics to the non-linear problem, including estimation of model and parameter uncertainty is critical.
- Propagation of uncertainty. This assesses the impact that uncertainty in parameters or variability in inputs has on outputs from a process. The basic approaches are analytical propagation (through either addition of variance or Taylor series approximations (Bevington & Robinson, 2003; Wikipedia_contributors, 2012)) , or numerical propagation through Monte-Carlo simulations (Fishman, 1996; Metropolis & Ulam, 1949).

Both of these topics would be regarded as very advanced in inferential statistics but are a core requirement of process engineering, and are highly engaging as analytical problems for process engineering students (Crosthwaite, Cameron, Lant, & Litster, 2006). This paper proposes an approach that addresses the requirements of inferential statistics education with process engineering statistics requirements to engage process engineering students at an early stage in their education.

Methodology

The course was approached from the point of view of uncertainty management in process engineering with the three core modules of:-

(a) Uncertainty observation, covering core inferential statistics, including source and application of distribution, point estimation, and hypothesis testing.
(b) Uncertainty and parameter estimation, covering ANOVA, and linear and non-linear regression and parameter estimation.
(c) Uncertainty propagation, covering analytical propagation of variance in linear and non-linear I/O equations, as well as numerical Monte-Carlo propagation.

Assessment was split evenly between individual assessment and group assessment, with major practicals addressing each of the components above, as well as a capstone practical covering all components, and mid-semester (10%), computing practical (10%) and final (30%) examinations.

Wastewater treatment problems were used heavily as case studies, as an example of an uncertainty dominated system (2010 flooding being an extreme example – Figure 1), and a field trip to the Oxley Creek Wastewater plant was used to provide practical relevance and context (Figure 1).

Figure 1: Oxley WWTP used as an example of uncertainty management issues (2010 flood event, photo by Aleks Atrens, copyright preserved, and photo used with permission).

The subject has now been run for two consecutive years 2012-2013 (126 and 146 students respectively) as the 2nd year Chemical Engineering course CHEE2010 at The University of Queensland. Course evaluation results are available for 2012 and will be presented for 2013 in the presentation. The course was previously taught in 1st semester 3rd year as CHEE3010, and was a more conventional engineering statistics course, but was completely rewritten as presented in this paper as CHEE2010.

Uncertainty observation

The components of uncertainty observation were broken into the following major components:-

(a) Concept of inferential statistics
(b) Data visualisation and basic analysis (scatter plots, box plots, histograms etc).
(c) Basic parameters of location and dispersion.
(d) Binomial theory.
(e) Source of normal distributions and central limit theory.
(f) Using normal distribution for probability prediction.
(g) Point estimation and confidence in mean.
(h) Hypothesis testing and single- and two- tailed t, Z, and F-testing.

These concepts are covered exhaustively in the educational literature as they form the basis of most engineering statistics courses, and Ryan in particular (Ryan, 2007) provides excellent teaching material.

For computing, we used common tools such as Microsoft Excel and Matlab to analyse data rather than specialised tools such as Minitab, which was largely to provide vertical integration with subsequent 3rd year courses. A number of computing tutorials were used that addressed the learning objectives of this module including visualisation, calculation of point estimators and confidence intervals, and hypothesis testing (non-parametric, one, and two-tailed t-testing).

The practical orientated towards uncertainty observation was a temperature measurement practical, where students were required to test three different resistive temperature detectors (RTDs) with data logged every second, and to assess the number of samples required to achieve a specific error (95% confidence interval) at different temperatures.

Uncertainty estimation – ANOVA, linear and non-linear regression

Analysis of variance (ANOVA) was used as the keystone technique for factorial analysis, design of experiments, and regression (linear and non-linear). ANOVA is a topic that is taught superficially, but is not generally focused on in engineering statistics (Ryan, 2007; Walpole, Myers, & Myers, 1998). This is an issue since ANOVA is the basis for all future analysis and proportioning of variance, including formal tests for model validity and model comparison.

In its basic form, ANOVA allows for proportioning variance to either variance due to variation in factors, or due to residual variance. If variance due to factors is large enough compared to residual variance (tested by comparison of F value vs the critical F value in a hypothesis test), one can conclude that the factor has a significant impact. Interaction between factors can also be included if there are sufficient residual degrees of freedom.

ANOVA is also very important for regression (linear and non-linear). During regression, ANOVA is used to proportion variance predicted by the model vs residual variance. Model validity is defined by an F-test on model vs residual variance, and parameter and model uncertainty are defined by another F-test that identifies all models “as good as” the optimal model. Non-linear regression is taught as an extension of linear regression that successively optimises the objective function (J=residual sum of squares). This allows for demonstration of the Jacobian principle, and estimation of model and parameter uncertainty from the J-p Jacobian (Dochain & Vanrolleghem, 2001). A key component running through the whole estimation module is that estimation of parameters is not useful without estimation of uncertainty in model and parameter values.

Log-transformation for parameter estimation is discouraged on the basis that it transforms residuals and is only applicable to a limited range of problems (i.e., exponential decay systems).

Teaching ANOVA and regression simultaneously allows for demonstration of powerful techniques such as mixed categorical/continuous ANOVA (ANCOVA) where categorical factors are analysed by proportioning variance and continuous factors are analysed by regressing (using only one DOF per factor). This is done effectively using the `anovan` command in Matlab.

Tutorials were provided on linear and non-linear regression, and the practical based around this was measurement of oxygen concentration during batch oxygenation, as well as estimation of mass transfer coefficient k_La and saturation oxygen concentration C^* in the equation:-

$$C_{O_2} = C^*(1 - e^{-k_L a.t}) \qquad (1)$$

where C^* is the saturation concentration, k_La is the mass-transfer coefficient, and t is the time.

A variety of k_La values could be obtained which need to be assessed against factors (aerator size, stirring speed, and air flow) through an ANOVA.

Uncertainty propagation

The principles of uncertainty propagation as applied in this course have been extensively presented in (Batstone, 2013) (email corresponding author for a copy), but as a summary, the uncertainty in an output from a model or process can be estimated from uncertainty in input in two different ways:-

(a) Analytical propagation, in which weighted arithmetic propagation of variance is applied to linear combinations of variables, and Taylor series approximations to evaluate non-linear combinations of variables.
(b) Numerical propagation (Monte-Carlo simulations (Fishman 1996)), in which the mathematical formula is repeatedly evaluated a large number of times while applying pseudo-random values to the input to estimate output population properties.

In both forms, correlation in inputs can be readily accounted for, by either inclusion of the correlation term in (a), or generating correlated input vectors in (b).

Both methods were taught in the course, and are shown to result in the same outcomes for systems linear in input.

The practical applied to propagation was blending of salty (0.1M NaCl) solution with clean water and measurement in a conductivity probe. The salty solution was subject to normally distributed random noise.

Results

Technical Outcomes

A number of major skills were demonstrated through the course, including observing and estimating uncertainty, estimating parameters, and estimating impact of model uncertainty in outputs. Basic competency (needed to pass the exam) included determining confidence intervals, fitting a model (including parameter confidence), and plotting the model vs the data. Advanced competency included determining model confidence intervals, and propagating uncertainty through the model.

Figure 2 demonstrates an example output from the practical (computing examination), which could be done in either Microsoft Excel (using the solver add-in for optimisation and parameter confidence calculation) or Matlab. It demonstrates how both model and data uncertainty can be presented, as well as parameter confidence (inset). Along with these integrated skills, project management skills, uncertainty reporting (including to non-technical experts), and data processing skills were developed.

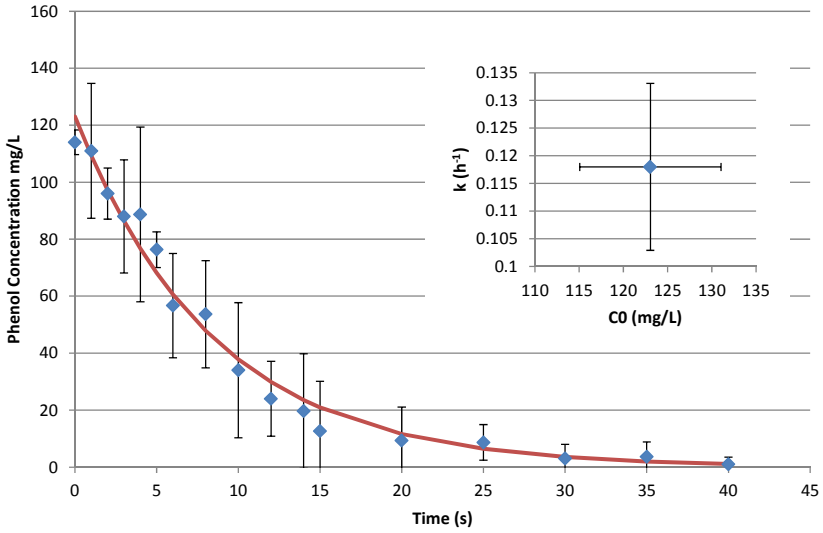

Figure 2: Actual (symbol) and predicted (line) phenol concentration over time during a dilution event (error bars indicate 95% confidence); inset graph shows 95% confidence intervals for the model parameters C_0 and *K*.

Student proficiency

In-line with the overall high level of engagement, attained student proficiency was high, with a fairly top-heavy grading outcome (Figure 3). This was across both computing and theoretical examination, which explains why the scoring in non-linear estimation was so high (it was a core skill to the practical exam).

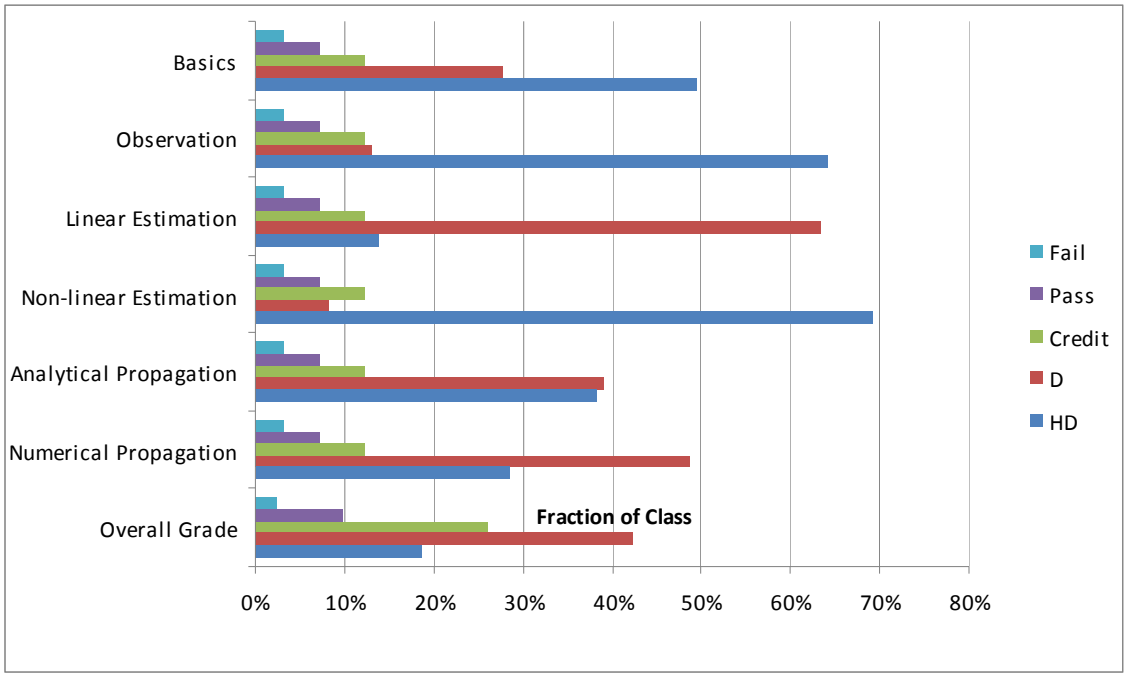

Figure 3: Performance in individual assessment (HD=high distinction, D=Distinction), including practical (computing) examination and end of semester exam.

The fail rate was approximately 4%, with those individuals failing being those who demonstrated consistently low engagement. The subject final grading is slightly top-heavy, with excessive distinctions, which can be corrected by increasing difficulty in both end of

semester and practical exams (probably by introducing additional challenge questions). Thresholding was applied in grading, with student grades being derived from the maximum of their work on overall assessment, or individual assessment (quiz, exam, practical exam). Strong individual achievement could improve poor performance on group work, but students needed to score highly individual to achieve a distinction or high distinction.

Student satisfaction

Student satisfaction outcomes were extremely good (particularly for a 2nd year subject). A summary of the overall course ratings are given in Figure 4, indicating 98% found the overall course to be good or very good. The course was one of the highest rated 2nd year engineering courses in the Faculty, with an overall course ranking of 4.4. The previous equivalent course (CHEE3010) is shown for reference. This was generally at or below 3.0, and had been one of the lower ranked courses in the Faculty, consistently identified as requiring correction.

In individual comments relating to CHEE2010, students focused on a number of issues that generated high levels of satisfaction. These include:-

(a) Focus on ancillary skills such as computing and project management.
(b) Constant demonstration of relevancy of statistics and uncertainty management to graduate attributes.
(c) Analytical focus, particularly in application to statistics.
(d) Multi-mode teaching, and focus on group based project delivery.

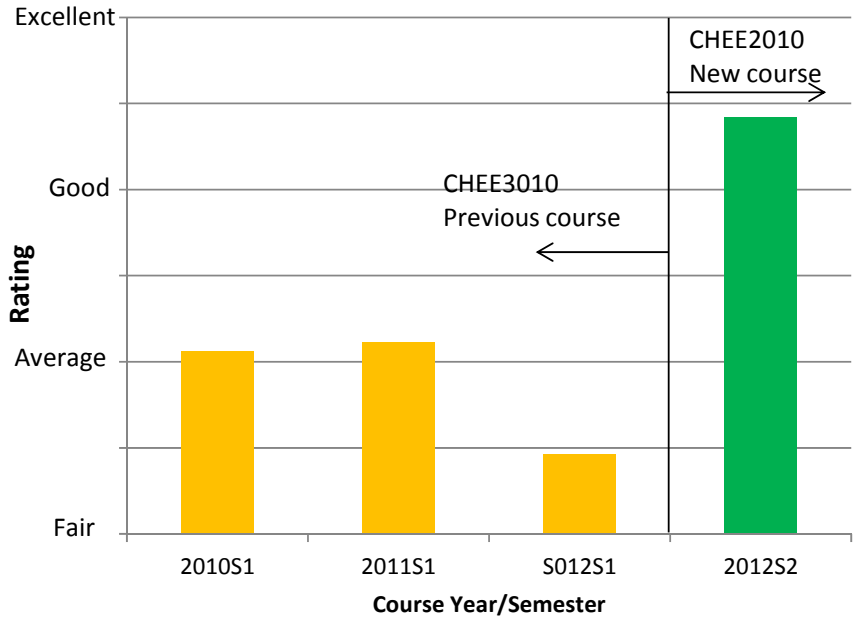

Figure 4: Student course rating through the standardised student evaluation of course and teacher (SECAT), with an overall rating of 4.4/5. Previous equivalent course (CHEE3010) shown for reference.

Discussion

While references to formal surveys have not been found, based on the authors experience, application of statistics in process engineering practice is not widespread, or is the province of experts rather than integrated with general engineering practice. This is likely related to the limited extent to which more general engineering statistics courses are applicable to process engineering statistics. Increasing the general knowledge and capacity to apply contextualised statistics is becoming more important given its fundamental link to quality

management systems and accreditation. Indeed, addressing these challenges for engineers is the basis for approaches such as Six Sigma (Dale, Wiele, & Iwaarden, 2007), but this focuses on general manufacturing and its link to QC rather than the basic understanding of theory and linking it to discipline and graduate attributes.

There is certainly a copious amount of material at the senior undergraduate and postgraduate level, but this is generally focused on more advanced topics such as dynamic process modelling and model identifiability and uncertainty (Dochain & Vanrolleghem, 2001; Hangos & Cameron, 2001). This certainly presents statistics as a contextualised issue, but often in a way that is inconsistent with previous early stage courses (i.e., process rather than observation focused). Based on this course, we propose a bottom up approach. That is, at an early stage, focusing on analytical statistics, and presenting simplified chemical engineering problems (including dynamic non-linear problems) that demonstrate applicability of inferential statistics. This also offers the opportunity to show how different approaches (e.g., confidence intervals vs t-testing vs paired t-testing), are derived from the same principles, relate to each other and can be used in a hierarchical way. Finally, it uses a consistent approach across a very broad range of practical engineering problems that can then scale and be applied in later chemical engineering focused courses to provide better vertical integration.

Conclusions

The outcomes indicate that there are strong benefits to teaching statistics in a process engineering context and that this can provide vertical integration and ancillary skills development that provide a better downstream experience, and substantially assist with key graduate attributes, particularly related to addressing risk and uncertainty.

References

Batstone, D. J. (2013). Teaching uncertainty propagation as a core component in process engineering statistics. *Education for Chemical Engineers, in-press*.

Bevington, P. R., & Robinson, D. K. (2003). *Data reduction and error analysis for the physical sciences*: McGraw-Hill.

Crosthwaite, C., Cameron, I., Lant, P., & Litster, J. (2006). Balancing Curriculum Processes and Content in a Project Centred Curriculum. In Pursuit of Graduate Attributes. *Education for Chemical Engineers, 1*(1), 39-48.

Dale, B. G., Wiele, A. v. d., & Iwaarden, J. v. (2007). *Managing quality* (5th ed. ed.). Malden, MA :: Blackwell Pub.

Devore, J. L., & Berk, K. N. (2012). *Modern mathematical statistics with applications* (2nd ed. ed.). New York, NY :: Springer.

Dochain, D., & Vanrolleghem, P. (2001). *Dynamical Modelling and Estimation in Wastewater Treatment Processes*. London: IWA Publishing.

Fishman, G. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*: Springer.

Hangos, K. M., & Cameron, I. T. (2001). *Process modelling and model analysis*. San Diego :: Academic Press.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association, 44*(247), 335-341. doi: 10.1080/01621459.1949.10483310

Ryan, T. P. (2007). *Modern engineering statistics*. Hoboken, N.J. :: Wiley-Interscience.

Walpole, R. E., Myers, R. H., & Myers, S. L. (1998). *Probability and statistics for engineers and scientists* (6th ed. / by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers. ed.). London ; Upper Saddle River, N.J. :: Prentice Hall International.

Wikipedia_contributors. (2012). Propagation of uncertainty. . . *Wikipedia, The Free Encyclopedia* Retrieved October 30, 2012. , from http://en.wikipedia.org/w/index.php?title=Propagation_of_uncertainty&oldid=518173749

Acknowledgements

Dr. Damien Batstone is the recipient of an ARC Research Fellowship. Teaching activities of Dr. Batstone, Dr. Laycock, and Dr. Farre were supported by the UQ ResTeach Scheme for teaching into the Department of Chemical Engineering.

Copyright statement

Copyright © 2013 Batstone *et al.*: The authors assign to AAEE and educational non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is used in full and this copyright statement is reproduced. The authors also grant a non-exclusive licence to AAEE to publish this document in full on the World Wide Web (prime sites and mirrors), on Memory Sticks, and in printed form within the AAEE 2013 conference proceedings. Any other usage is prohibited without the express permission of the authors.