Statistics for Product Development

Jerome Wynne

UNIVERSITY OF BRISTOL

Abstract

Analyzing a product's performance during development is essential to making informed design decisions, yet many engineers are uncomfortable using statistics. This should not be the case: statistical tools offer a means of improving the quality and consistency of design decisions, and of developing exceptionally robust products. Here, DCA's current use of statistics is compared to modern statistical practice. Experimental, analytical, and graphical methods are suggested that would allow DCA to realize the benefits of statistical methods. The report concludes with an assessment of the practical feasibility of these tools.

JUNE 19, 2017

Table of Contents

*

List of Tables

List of Figures

Glossary

Confounding factor	A difference between treatment groups that could be affecting the repsonse, thereby making it unclear whether this difference or the treatment is affecting the response.
Event	A set of outcomes.
Outcome	The result of an trial.
Experiment	Varying factors in a controlled manner, and recording their effects on a response.
Factor	A controlled variable that may affect the response in an experiment.
Level	The degree or class to which a factor is set to, such as the volume of lubricant applied, or the mechanism subassembly chosen.
Nuisance variable	A source of variation in the response that is not the subject of an experiment.
Probability	The proportion of evidence in favour of a particular event, relative to the overall evidence for any event.
Probability density	The ratio of an interval's probability to its width.

Population	A collection of units subject to the same sources of variation.
Randomness	Variation caused by unmonitored or unknown factors.
Ranges	The region of a factor's domain over which to choose levels.
Random variable	A function mapping physical events onto real numbers.
Response	The outcome of testing a unit.
Treatment	The design variation being studied in an experiment.
Unit	An individual example of the population being studied.

Acknowledgements

I would like to thank Matt Edwards and Paul Harper: Matt, for providing feedback that improved every section of this report, and Paul, for the support that he has given and continues to give to all students of Engineering Design.

Declaration

I confirm that the work presented here is wholly my own and has been generated as a result of my own thought and study. Where I have consulted the work of others it is mentioned, and where my work was part of a group effort my contribution is made clear. Where the work of another is quoted, the source is given.

1 Introduction

DCA Design International is a 150-person product design consultancy based in Warwick. It orients itself towards the mechanical design of medical and consumer products for clients such as Unilever, Sanofi, and GSK. Figure 1 shows two of the company's most prolific designs. DCA's competitors are other global medium-to-large technical product design consultancies focusing on the medical, consumer, and transport sectors such as Seymourpowell and Cambridge Consultants.

Figure 1: Products designed by DCA.

DCA employs approximately sixty mechanical engineers, each of whom are general-purpose technical consultants capable of fulfilling the engineering demands of a project. These demands will consist of design, analysis, and prototyping work, in proportions varying according to project. DCA's substantial investment in engineering distinguishes it from other product design consultancies, many of which have less expertise available to handle a product's technical development [3]. This investment is manifest in both its engineering workforce and its ownership of four test labs. The experimental work that makes use of these labs is one of DCA's value propositions to prospective clients: DCA's handling of experimental data

contributes to its success in securing new clients and satisfying current ones. The way in which DCA's engineers collect, analyze, and present experimental data is the focus of this report. These data-oriented activities constitute the scientific discipline of statistics.

Statistics makes it possible to profile, understand, and predict variation in a product's performance. Engineers can use statistics to design products that are more robust, less costly to manufacture, and of a higher quality than they would be otherwise. Being able to use statistical tools fluently would elevate DCA's capacity as a technical consultancy.

The first half of this report evaluates how DCA currently applies statistics; the second half then suggests statistical methods that the company might find useful. Section 2 explains DCA's current investigatory framework and how statistics is applied to it. Section 3 compares the company's approach to modern statistical practice. Software packages for implementing the recommended methods are also evaluated. The report concludes with an evaluation of how actionable the suggested methods are, and responds to reasonable criticisms of the relevance of statistics in a product design consultancy. An appendix summarizing the most useful statistical results, along with definitions of notation and technical terms, is available for reference.

2 Overview of DCA's Use of Statistics

This section contextualizes DCA's uses of statistics and explains what methods are currently being applied by its engineers. The strengths and shortcomings of each method are listed, and the section closes with a summary and appraisal of DCA's approach.

2.1 The Structure of a Lab Investigation in DCA

A lab investigation in DCA consists of a series of experiments to understand the behaviour of a product or process. It begins with identifying the knowledge needed. Experiments are then designed, executed, and analyzed until the knowledge is acquired or is deemed no longer relevant. This process is depicted in Figure 2. All aspects of an investigation - from experiment design through to presenting the results to a client - are handled by engineers assigned to the associated project.

Engineers assigned to medical projects perform the majority of the lab work that takes place at DCA. Occasionally engineers working on fast-moving consumer goods (such as toothbrushes or lotion bottles) will run tests to compare design variations or verify performance relative to some baseline. In general, the timeframes and functional requirements of such products limit the relevance of extensive experimental investigations. Medical products, on the other hand, are subject to strict regulation relating to product failure modes and must be extensively tested.

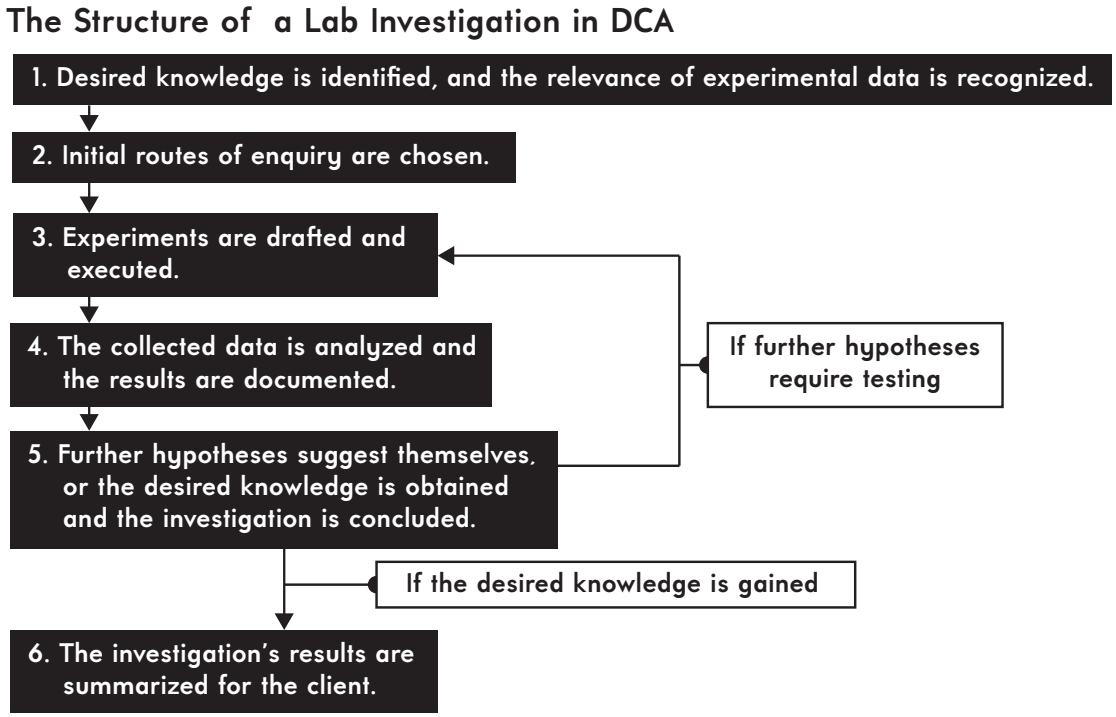

Figure 2: Diagram of DCA's experimental process.

Tests can be conducted at any point in a product's development, however as can be seen in Figure 3, the rate of testing increases as a product develops and varies dramatically according to project stage. Towards a product's release date is when resolving minor performance issues becomes a worthwhile pursuit, exploration for future product variants becomes a possibility, and rehearsal for fast-approaching regulatory tests becomes essential.

The equipment supporting this work includes axial and torsional testing machines, environment chambers, coordinate measuring machines, mass balances, and high-speed cameras, among other engineering instruments. Investigations commonly revolve around a particular experimental set-up, but other experiments are sometimes conceived to provide supplementary information. With that in mind, this report attempts to be data-agnostic in its recommendations of analytical techniques.

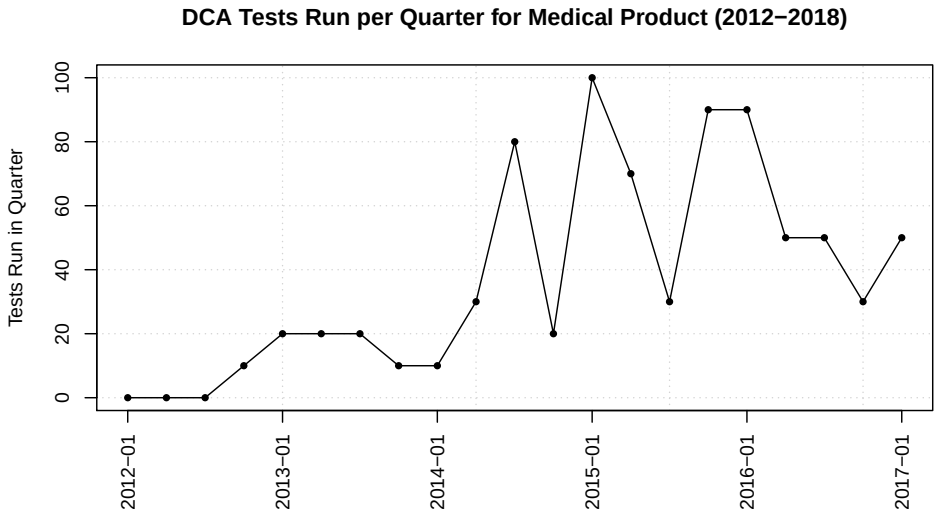

Figure 3: Testing frequency over the development of a single medical product.

Aside from lab work, DCA's engineers also use statistics to analyze tolerances and, increasingly, create predictive user interfaces. Unfortunately, the latter cannot be discussed for confidentiality reasons; the former can be and is. Statistics is used to varying degrees within the Human Factors and upper management of DCA, but these applications are not evaluated here.

To make sense of how statistics is applied in DCA's lab investigations, each step of an experimental procedure is reviewed in turn. These steps are:

1. Planning and execution.
2. Analysis.
3. Presentation and visualization.

Investigatory strategy guides the choice of all of the above and is examined in Subsection 2.2.

2.2 Experiment Design

Experiment design can be used to make products that perform better, are more reliable, less risky to develop, and have a uniquely justifiable development process. Design of Experiments refers to both experiment designs and a broader philosophy of systematic experimentation. An experiment design is a particular structure of experiment; good experimental design produces data that is unambiguous and relevant to an experimental objective.

Three principles form the backbone of robust experimentation [10]:

Replication Testing a particular treatment on more than one unit. Replication allows experimental error to be estimated and, since unbiased errors cancel on being averaged, provides a more precise estimate of a treatment's effect.

Randomization Randomly allocating treatments to units and the sequence in which units are tested averages out the effects of nuisance variables, and validates the analytical assumption that observations are randomly drawn from a population.

Blocking Blocking accounts for the effects of nuisance variables when assigning treatments - units are grouped on the basis of a shared attribute, then treatments are distributed evenly among the groups. See Figure 4 for further explanation.

These principles constitute the makings of any well-designed experiment, and they are evident in DCA's labwork: units are blocked according to factors such as component batches and time of assembly, testing and assembly sequences are randomized, and engineers are keen to provide replicates in their tests. The experimental designs used in the company are enumerated, explained, and critiqued in Table 2. Despite the company's awareness of these aspects of experiment design, two considerations seemed to be overlooked: how the data collected will be analysed, and how an investigation should be structured.

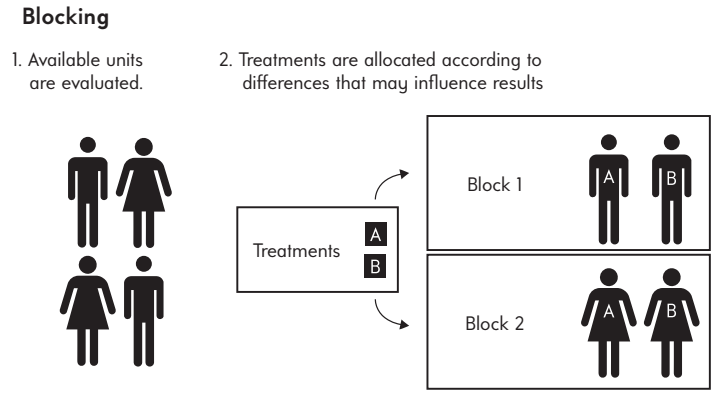

Figure 4: Blocking allocates treatments evenly amongst units with some difference that affects the response, to prevent this difference confounding the results.

Giving forethought to how the collected data should be analyzed allows confounding factors to be identified and subsequently controlled. Moreover, thinking about what analytical results are desirable focuses an experiment on its objective.

DCA lacks a framework for planning experimental investigations. As a consequence of this, basic activities such as verifying that experimental set-ups produce repeatable results, scoping an investigation, and running screening experiments to close off unpromising avenues of enquiry are routinely forgotten. The dominant experimental strategy within the company is a best-guess approach: one treatment is tested in each experiment, chosen based on the expert insight of the engineering team. This method's greatest shortcoming is that if the treatment does not elicit the desired effect, then the next factor to vary must be guessed at, a process that can continue almost indefinitely. Furthermore, if a treatment is successful, then its investigation may be concluded when in fact a better solution exists [10].

Experimental Design	Description	Evaluation
Randomized complete block design	The units are split into blocks according to differences that may influence the response, then each treatment is randomly assigned to at least one unit from every block.	Allows the effects of nuisance variables to be eliminated during analysis, provided the block factor and treatment do not interact.
		Lends itself to established analytical techniques (e.g. ANOVA).
		Can be extended to block on more than one factor (such a design is called a Latin square)
		Not possible if the number of units in a block is fewer than the number of treatments to be tested.
Factorial design	The factors and their respective levels are chosen, then all possible combinations of the factors are tested.	More time-efficient than testing one factor per experiment.
		May be limited by resources if there are many factors
		Allows interaction effects to be estimated.

Table 2: Experimental designs applied in DCA.

After an experiment is designed, it needs to be executed. As has been mentioned, DCA is well endowed to run experiments, and has a system in place for documenting the date, purpose, and conditions of them. These documents are scanned and stored on a local network, where they can be referred to at a later date. These documents are useful, but could be improved by:

- Listing what factors were controlled and which were uncontrolled. This would help future users of the test data avoid mistakes when interpreting results.
- Providing a reference to a test in which the experimental apparatus and procedure used was verified.
- Including an image of the experiment so that it can be understood by other engineers.

- A specific statement of how the collected data will be analyzed. This is to ensure that the results of an experiment are used in the way intended by the overall investigation.

2.3 Analysis

Analyses in DCA rely heavily on expert knowledge of the systems being tested and rarely on statistical results. This is probably because the relevance of statistics may not be clear, and how it might be applied even less so, which is understandable. It is widely agreed that most people's experience with statistics is one of discomfort and bemusement. Having said this, relying on intuition alone risks falling prey to cognitive biases, missing valuable information that is not superficially obvious, and being unable to relate physical behaviours to experimental observations. Foregoing statistics when analyzing product behaviour handicaps the ability of an engineer to design a robust product.

Several hundred of DCA's test reports were surveyed to understand how DCA currently uses statistics. Many of these reports contained summary statistics, such as arithmetic means, variances, maximums, minimums, and so on. A few made use of interval estimates as informed by a regulatory standard. In general, the focus was on observed extremes and point estimates of population means. These tools will now be explained, and their usefulness and possible weaknesses detailed.

Summary Statistics

Randomness refers to variation in a response as a result of uncontrolled factors. If a coin is always flipped according to the same process, then it will always land in the same orientation. By contrast, if a coin is flipped by hand, then the person doing the flipping has a limited amount of control over the factors affecting the flip. As a result, there is uncertainty as to what the

outcome of that flip will be. There are two ways to reduce this uncertainty: Either make assumptions about how those flip-influencing factors vary, or measure them. Assumptions allow real-world knowledge to dictate what outcomes are possible what their relative propensities are. Conversely, measurements allow factor effects to be estimated, accounting for the sources of variation in the response [8].

It may be suggested that no particular outcome is favoured by a system, in which case uniformity could be assumed. Alternatively, if many independent effects, no one of which is especially large, are responsible for a response's variation then assuming normality may be appropriate. In both instances, a distribution of outcomes is suggested based on physical knowledge, making it possible to analyze the system for useful information.

Mathematical analysis can be used to understand systems that vary randomly. As with any theory, some tools need to be defined to make the maths possible: a random variable is one of these tools. Despite its name, a random variable (r.v.) is a function that maps events onto real numbers [1]. For example, an r.v. X could be defined that maps the outcomes of a coin toss onto the numbers 1 and 0:

$$X(\texttt{Coin lands Heads}) = 1 \tag{1}$$

$$X(\texttt{Coin lands Tails}) = 0 \tag{2}$$

Usually, the choice of mapping is quite natural - an r.v. might, for instance, be used that counts the number of successes in many trials, or that takes on the value of a measurement.

Variation in the events that an r.v. maps from is described using a probability distribution. Each value is weighted according to its probability or, in the case of continuous-valued r.v.s, the ratio of the width of an interval to that interval's probability. Figure 5 highlights this difference.

The essential problem of experimental statistics is trying to understand the distribution of a random variable from just a sample. In product design, this

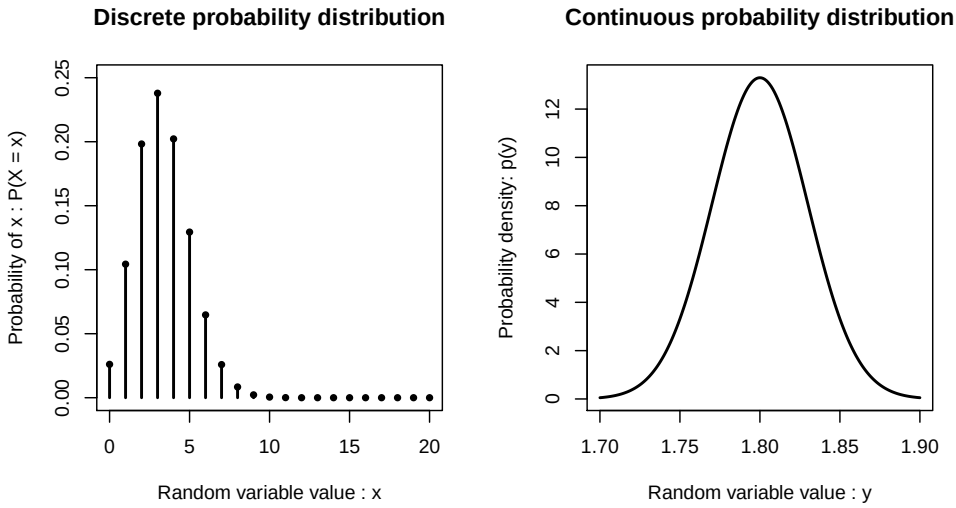

Figure 5: Left: Probability mass function. Right: Probability density function.

means using measurements from just a limited number of prototypes to estimate the behaviour of a much larger number of units. The attributes of a random variable's distribution - such as its spread and average - can be estimated using summary statistics.

One particularly useful summary statistic is a sample mean, which approximates the mean response of a larger population of units. DCA use sample means to discriminate between the performance of two or more populations, each representing possible design variants. A sample may not be representative of its population. Bigger samples tend to represent their populations better, and tightly clustered responses imply that any one response is likely to be close to the population mean.The standard error of the sample mean - the average difference between a sample mean estimate and the population mean - reflects these observations, as it corresponds to $\frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation[1] and n is the number of units in the sample. Change in standard error with sample size is shown in Figure 6.

Certain summary statistics can be thought of as estimates of a distribution's

[1]Standard deviation is the average difference between the members of a population and that population's mean.

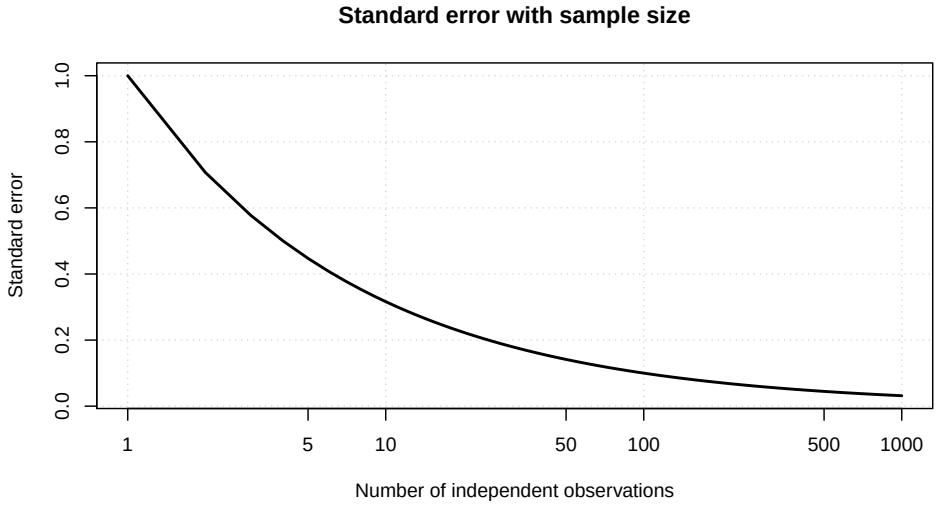

Figure 6: The sample mean converges on the population mean as sample size increases.

parameters. These are values that constrain a particular distribution's shape. The normal distribution's shape, for example, can be specified by supplying just two values: the variance (spread) and mean (location). An estimate of a parameter's distribution - the relative probability of possible values - represents the certainty warranted by a test better than small-sample point estimates.Section 3 contains techniques for making distributional estimates, because DCA's experiments are constrained by the number of units that can be tested.Interval estimation is used occasionally in DCA to quantify estimate uncertainty: Its results need to be interpreted carefully, however, and its assumptions need to be verified.

Tolerance Intervals

DCA base their most sophisticated statistical analysis on ISO 16269-6, *Determination of statistical tolerance intervals* [11]. This standard outlines how to construct tolerance intervals under the assumption that the random variable in question is normally distributed.

A tolerance interval is a range of values that's likely to contain a particular fraction of the population. Because this interval is an estimate, it can only contain the advertised fraction of the population part of the time. The proportion of intervals that would, on average, contain the specified fraction of the population is called the confidence level. If many 99% tolerance intervals were constructed for different samples of the same size to a 95% confidence level, then 95% of the intervals estimated would contain at least 99% of the population. Figure 7 plots 99% tolerance intervals for 20 synthetic samples, all drawn from the same population.

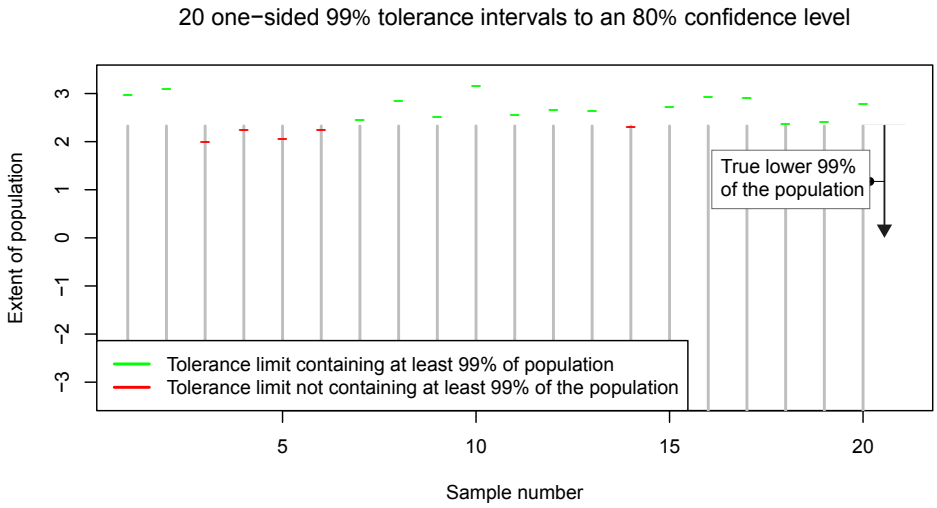

Figure 7: Tolerance intervals estimate where the bulk of a population lies.

DCA used tolerance intervals in to check that a given proportion of a population satisfied a performance threshold. Their use was not widespread, and in some of the reports studied the k-values[2] were

[2] A k-value is a term used in the ISO standard to represent the number of sample standard

incorrectly calculated. No effort was made to test the assumption of normality.

Confidence Intervals

Another tool that DCA's engineers occasionally used was the confidence interval, which indicates a range of values that an estimated parameter is likely to fall within [7]. As with tolerance intervals, this range only contains the population parameter a certain fraction of the time, a problem that's unavoidable since there will always be a chance that an unrepresentative sample is drawn. If a confidence interval were to be constructed for the population mean of 100 samples, each of 5 units, the confidence level would indicate how many of these intervals would - on average - contain the mean's actual value (as opposed to its estimated value). Figure 8 demonstrates this idea. Confidence intervals can be placed on any parameter estimate, and represent the number of standard errors from the estimated parameter that's likely to contain the population parameter. Their derivation is supplied in the Appendix.

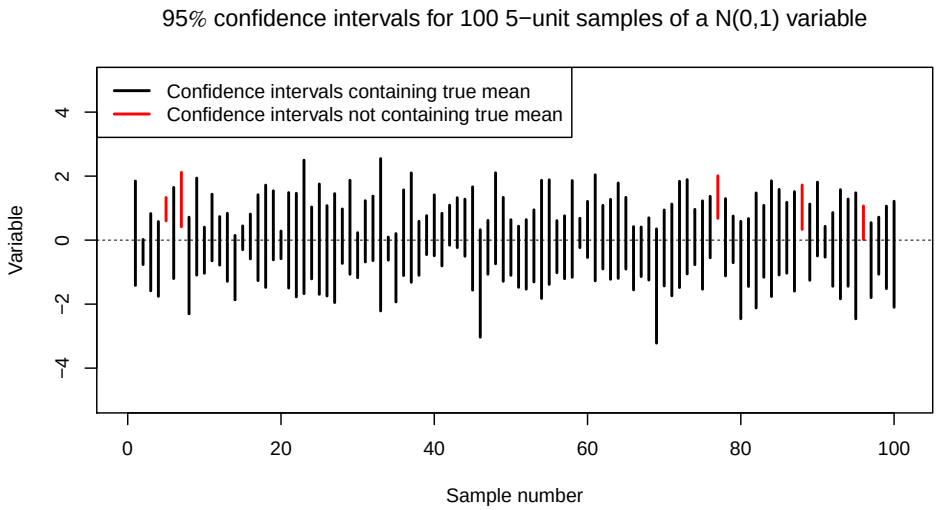

Figure 8: Confidence intervals indicate the range of values likely to contain an estimated parameter.

deviations from the mean that will contain at least p% of the population $(1-\alpha)$% of the time, where $(1-\alpha)$ is the confidence level.

Confidence intervals suffer from similar problems to tolerance intervals. DCA's engineers used them rarely, meaning they made decisions with a limited appreciation of the uncertainty in their estimates. When large differences in effects were of interest, this was not a problem. However, when subtle differences were important, or sample sizes limited, then quantifying an estimate's uncertainty would have helped avoid erroneous or inappropriately confident conclusions.

Monte Carlo Estimation

Monte Carlo estimation approximates a quantity by simulating the random process generating it. In DCA it was used to analyze tolerance chains in products. One use case was somewhat similar to the following: a gap between two components affected a product's function, so the distribution of gap values needed to be known. A computer was used to simulate values for each dimension affecting the gap, according to their distributions, then calculate a value for the gap produced. The frequencies of the simulated values were then used to approximate the distribution of the gap. Figure 9 is a diagram of this process.

Monte Carlo estimation is a powerful statistical tool that is easily implemented and interpreted. Enlarging its use beyond tolerance analysis to other areas of work at DCA would allow the effects of variation in manufacturing processes on product performance to be predicted. Mathematical models were used in the company to analyze a product's performance: randomly sampling the inputs to these models would give DCA's engineers the ability to profile the overall variation in performance, rather than estimate the worst or best case alone.

2.4 Visualization

Visualization is essential to clearly and convincingly summarizing an experiment's results. Graphics allow engineers and clients to see for themselves what's been discovered. A plot should be relevant, easily

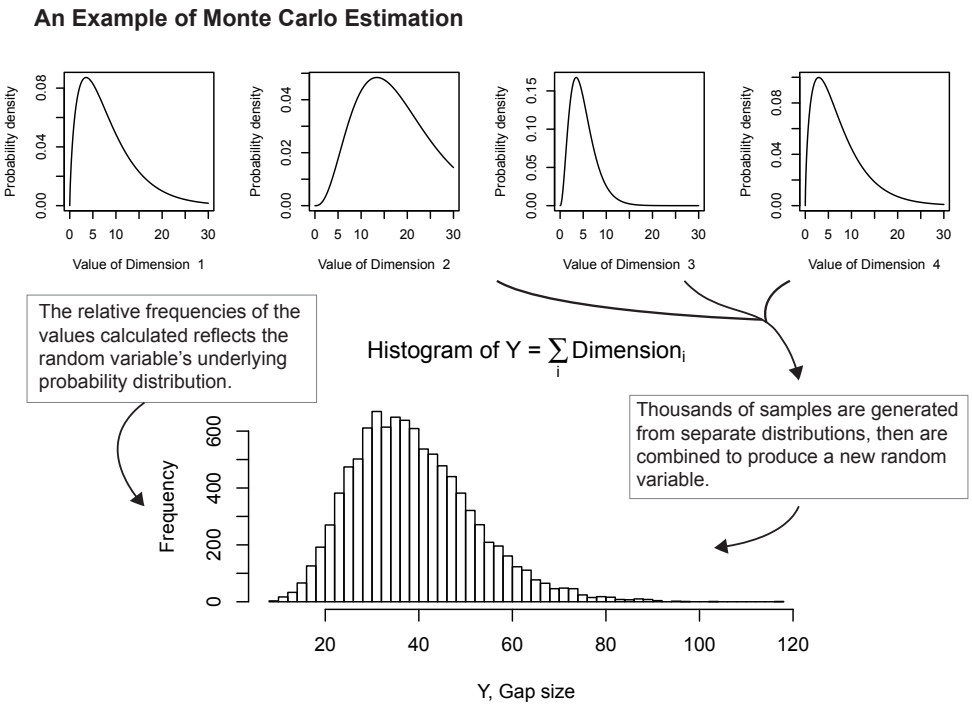

Figure 9: Monte Carlo estimations solve statistical problems by simulating their underlying random variables.

interpretable, and accurately convey its underlying data.

DCA's reports and client presentations contained plots of experimental data generated using either Microsoft Excel or Matlab. These plots were line or bar charts, along with the occasional scatterplot. Line charts were frequently used in DCA because they are directly plottable from the raw data output by DCA's axial and torsional testing machines. Bar charts, on the other hand, tended to be used to highlight differences between groups, such as their means and maximums.

Directly plotting raw data requires very little time, and produces plots that are easily related to patterns observed over the course of an experimental run. These are major benefits in favour of using line charts to understand the results of an experiment quickly. However, there are several reasons why these types of charts are not suitable for presenting at team meetings or to clients. Figure 10 was taken from a DCA test report and is of a type that was regularly used to facilitate technical discussions. In this test, the

response of the units at 10.8 seconds was of interest, yet all raw data were plotted. Plotting irrelevant data has various negative corollaries: the redundant information away from 10.8 seconds is distracting and of limited use, since it is not possible to see how individual units are behaving. Even if the other displacements were relevant, the plot would still be somewhat misleading as attention is directed to the group extremes, and cannot be used to see anything besides gross differences between units. Moreover, the data's resolution is not apparent: this problem can be exacerbated if a spline fit or graphical smoothing is applied[3]. Subsection 3.3 suggests solutions to these problems.

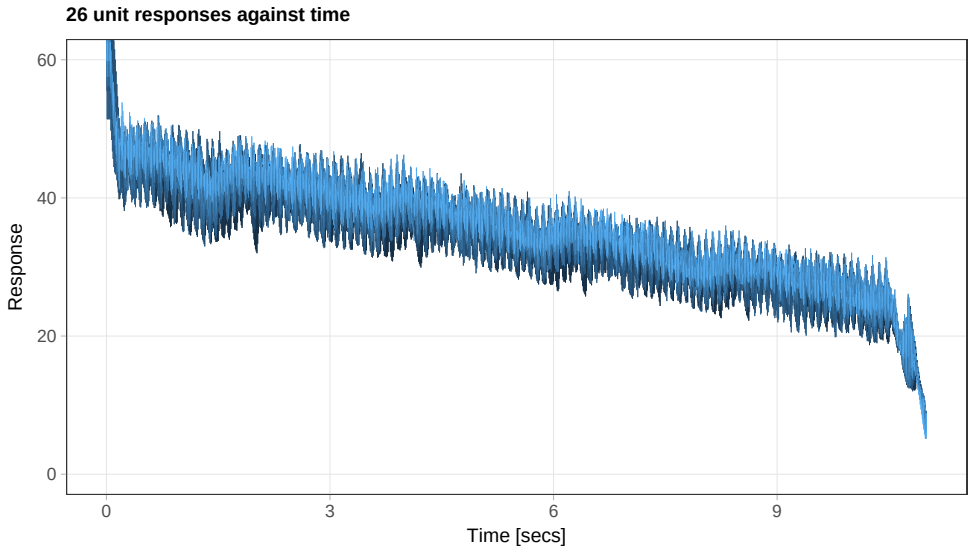

Figure 10: A representative use of line charts from a DCA test report. Each shade of blue corresponds to a distinct unit.

As has been mentioned, bar charts and scatter plots were also used within DCA, albeit less frequently than line charts. There are several ways in which the use of these tools could have been improved. Plotting errors bars on bar charts would allow the uncertainty in the estimates to be represented, making group comparisons less prone to unwarranted confidence. Alternatively, the individual unit responses could be included as a scatter plot. Scatter plots are capable of displaying location, spread, density, and sample size information, making them a valuable tool for summarizing

[3]Excel automatically smooths line charts.

experimental data: the response at 10.8 seconds in Figure 10 could easily be captured for each unit then presented in a scatter plot, making the results easier to interpret and discuss.

2.5 Review

The report so far has presented and critiqued the statistical methods used by DCA's engineers at each step in an experimental procedure. The next section suggests methods that will help DCA use statistics more effectively by:

- Using regression to relate factors to a response.
- Estimating distributions of plausible parameter values rather than their "most-likely" or "worst-case" values.
- Planning investigations that systematically identify the main factors influencing a response, and the factor levels needed to optimize that response.
- Visualizing results in a way that clearly reveals experimental patterns.

These recommendations were chosen to allow a product's behaviour to be methodically analyzed and optimized, and to transparently represent the uncertainty associated with experimental work.

3 Suggested Methods

Having assessed DCA's current use of statistics, it is possible to suggest methods from modern statistical practice that would give them new capabilities and improve their existing ones. Like Section 2, this section is categorized by the steps in an experimental process.

3.1 Analysis

Regression Analysis

Experimental work often tries to an answer questions such as:

- Which factors are having a big effect on the response?
- How does a factor affect the response? How do several factors interact?
- Which design variation is better?

Regression analysis would allow DCA's engineers to answer questions such as these. Knowing what parameters affect the product, and by how much, focuses development on the things that matter, resulting in a better quality product.

Regression estimates how a continuous response changes with respect to certain inputs. Linear regression uses a linear function to predict the response: This may sound limiting, since in the physical world many relationships are nonlinear, but nonlinear relationships can be made linear by transformation.

Linear models describe a response y as a linear function of some parameters $\hat{\beta}_i$, each weighted by a corresponding input x_i [4]. The rate of change of the response with respect to a particular input is embedded in that input's $\hat{\beta}$ value. An example of such a model would be:

$$\hat{y} = f(x_1, x_2) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot e^{x_1} + \hat{\beta}_3 \cdot x_1 \cdot x_2 \tag{3}$$

Where $\hat{y}$ is the estimated response, x_i is an input, and $\hat{\beta}_j$ is the coefficient of the jth input. Note that while the coefficients $\hat{\beta}_j$ are linear, the inputs $g(x_i)$ can be nonlinear functions of the measurements.

This model can be fit by adjusting the $\hat{\beta}_i$ values so that $\hat{y}$ is a good estimate of the true response - that is, the mean response for a given set of inputs. To do this, it is necessary to measure how inaccurate $\hat{y}_i$ is relative to $E[y|x_i]$. One intuitive measure of fit is the overall distance between the estimated and the observed responses. This metric is known as the residual sum of squares[4]:

$$\text{RSS}(\hat{\beta}) = \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{p} \hat{\beta}_i x_{ij} \right)^2 \tag{4}$$

Where y_i is the response of the ith observation and x_{ij} is the jth predictor value of the ith example. Figure 11 shows the results of regressing a response onto one and two inputs. The sum of the distances between the points and the regression line/plane corresponds to fits' RSS. There are two alternative justifications for least squares, which are both presented in the Appendix.

From a practical standpoint, regression models should be fit using software. Matlab, R, and Octave can all be used. By setting up the problem such that the N observed responses are in a column vector $\mathbf{y}$, their associated p inputs form the rows of a matrix $\mathbf{X}$, which is $N \times p$, and the p coefficients are in a column vector $\hat{\beta}$, it is possible to succinctly write then minimize the RSS

[4]The residual refers to the difference between an estimated response and an observed response at a particular set of inputs. Error is the to deviation of a response around its expected value for a given input.

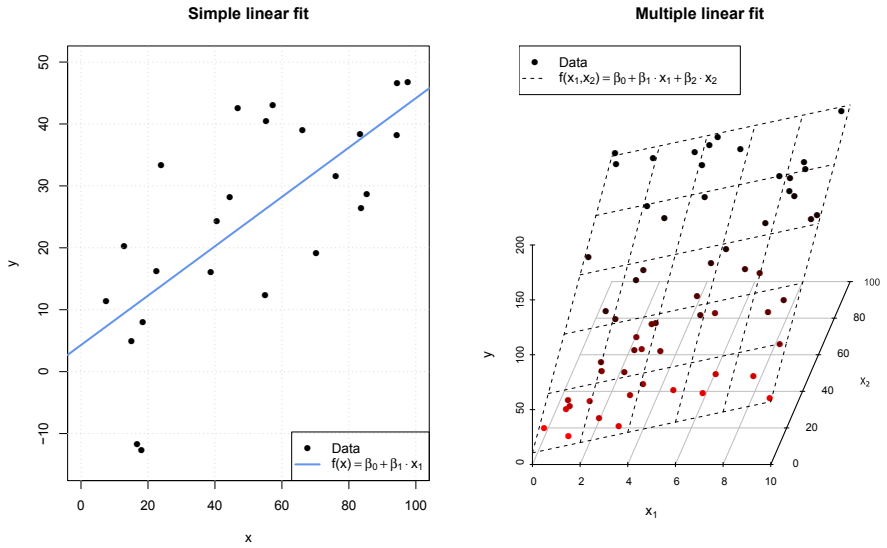

Figure 11: A simple linear and a multiple linear fit.

criterion [7]:

$$\mathrm{RSS}(\hat{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) \tag{5}$$

$$\frac{\partial \mathrm{RSS}}{\partial \hat{\beta}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0 \tag{6}$$

$$\implies \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{7}$$

Being able to represent linear models using matrix notation means that any software capable of matrix manipulation can perform linear regression. This terse notation will also be useful in the example that follows.

Consider a test in which the load delivered by three groups of ten units is measured. The differences between the groups are categorical: the groups correspond to three design variations A, B, and C. Since there is no natural order to the groups, it is necessary to encode them in a sensible way. There are several ways to do this, but a simple indicator stating whether a unit has a particular modification is sufficient. Table 5 shows this encoding. Choosing how to encode categorical variables matters affects how the regression coefficients should be interpreted.

The model to be fit is then defined according to the available inputs and

Table 3: Dummy coding of groups.

Unit ID	x_A	x_B	x_C	Load, y [N]
1	1	0	0	5.43
2	0	1	0	7.48
		⋮		
30	0	1	0	6.47

theories about relationships between those inputs and the response:

$$\hat{y} = \hat{\beta}_0 x_A + \hat{\beta}_1 x_B + \hat{\beta}_2 x_C = \mathbf{X}\hat{\beta} \tag{8}$$

$$\mathbf{X} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(30)} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 5.43 \\ 7.48 \\ \vdots \\ 6.47 \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \tag{9}$$

Note that $x^{(i)}$ corresponds to the ith observations set of inputs. The normal equation (Equation 7) can be used to determine the model parameters according to the least-squares criterion[5]. As it happens, in this case $\hat{\beta}_A$, $\hat{\beta}_B$, and $\hat{\beta}_C$ are each group's average response, as is shown in Figure 12.

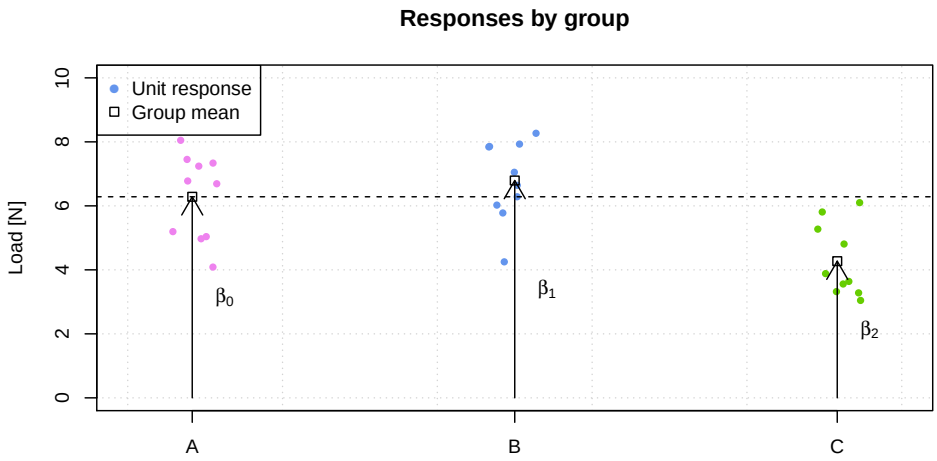

Figure 12: The coefficients of Equation (8) estimate the group sample means upon being fit.

[5]Least-squares is synonymous with minimizing the RSS

A reasonable criticism of the above example is that it is effectively just a drawn-out calculation of the group sample means. This is true, until the experiment is extended to involve another variable, this time a continuous one, say the volume of a lubricant applied to each design variant. Then the model can be adjusted to estimate the effects of the lubricant on the load output of each mechanism:

$$\hat{y} = \hat{\beta}_0 x_A + \hat{\beta}_1 x_B + \hat{\beta}_2 x_C + \hat{\beta}_3 x_A x_l + \hat{\beta}_4 x_B x_l + \hat{\beta}_5 x_C x_l \tag{10}$$

Where x_l is the volume of lubricant applied, and $\hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$ are the change in output force for each design variant per unit volume of lubricant. This model would allow an engineer to see not only how good design variants are relative to one another, but also how much lubricant would need to be applied to bring the performance of one in line with another. Linear models are powerful because they can be adapted to many situations, and because they provide a consistent way to disentangle treatment effects.

Something to be quite careful of is redundantly encoding the groups. If the term $\hat{\beta}_3 x_A$ were to be included in the model above, then there would be many equivalent ways to express the group effects: $\hat{\beta}_0$ could be any constant value, and $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ would be set to deviate from this constant to the group averages. This means that attempts to evaluate Equation 7 will be unsuccessful or unstable[6].

The fit of a linear model attempts to estimate the true relationship between the inputs and response. In other words, $\hat{\beta}$ are estimates of the parameters β:

$$y = X\beta + \varepsilon \tag{11}$$

Where ε is the error in the response - variation caused by unmonitored variables. In the example, $\beta_1, \beta_2, \beta_3$ are the true group means, and $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ are estimates of them. These estimates are not going to be perfect, and their standard errors can be calculated to understand how accurate they really are. To reemphasize, a standard error is the average difference between an estimate of a parameter over many samples, and the true parameter value.

[6]The instability is caused by numerical errors in calculating the inverse directly.

Standard error can be made smaller by reducing the group variance, or by making the sample sizes bigger. DCA's engineers should seek to minimize variation that is not relevant to the investigation because this will make it easier to see the effects of the inputs on the response for a given sample size. General calculation of standard errors is described and explained in the Appendix.

As an aside, assessing the significance of differences between groups is frequently called Analysis of Variance (ANOVA). It is called this because the analysis focuses on decomposing variation into its sources. Two test statistics, the t and F statistics, can be used to run hypothesis tests on possible sources of variation. Under certain assumptions, these tests provide guidance as to whether a treatment affects the response. It is relevant here because ANOVA can be viewed as linear regression with an emphasis on hypothesis testing. As discussed in the next section, hypothesis tests can be misleading. Furthermore, in complicated models, the canned formulae provided by some statistics resources can become unwieldy and confusing, which could make their results suspect and difficult to explain. By contrast, the structure and basic principles of a linear model are consistent across a variety of model sizes. For this reason it is suggested that DCA focus on learning to use linear models rather than ANOVA.

Linear models would be useful to DCA because they can be used to establish the sources of a response's variation. Quantifying how factors affect performance is the first step in understanding how to change a design to make it both better-performing and more robust. Used in combination with response-surface experiment designs (described in the next section), linear models could reduce the time and resources taken to conclude an experimental investigation, and would offer a tool for aligning theoretical understanding with empirical evidence.

Bayesian Inference

Summary statistics such as the coefficients of a linear model or a sample variance state what the most likely value for a parameter[7] is, based on the data alone. They do not indicate how much more likely this value is is than its peers, or let knowledge beyond the data be included. In reality, a sample will suggest a distribution of plausible values, and it will be known roughly what values are realistic. Bayesian inference combines readily available knowledge with experimental data to estimate a distribution of possible parameter values. Figure 13 compares distributional estimates to point estimates.

Bayesian methods would be useful to DCA because they are easier to understand, visualize, and explain than classical methods[8], and are relevant to a broader range of situations. They also allow expert knowledge to be used, making it possible to reach a sanitary compromise between gut-feel and experimental observation. Finally, their emphasis on distribution rather than point estimates better reflects the underlying uncertainty in analytical results. These claims will be explained in the context of an example.

[7] Reminder: a parameter is a number that controls the shape of a distribution.
[8] "Classical methods" here refers to hypothesis tests and interval estimates.

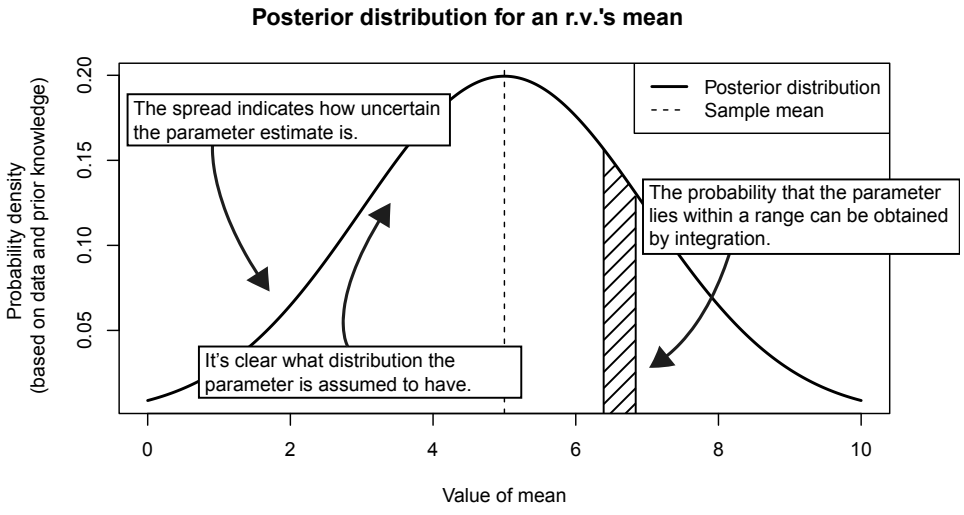

Figure 13: Distributional estimates offer many benefits over point estimates.

The proportion of units passing a test is a useful measure of a design's quality. Using the results from a test sample, the expertise of an engineering team, and Bayes' theorem, it is possible to estimate what pass rates would be likely if the design were to be produced in larger volumes [5].

Say that a sample of n units are tested, and y pass. The engineering team collude to sketch out a distribution for the passing rate θ that's tall near values they think probable and low near ones that seem unlikely. The team is to calculate the probability of a unit passing, given their experimental data and preliminary distribution: Bayes' theorem can be used to do this: a passing proportion is more probable if it makes the the number of units that really did pass in the sample more likely (high $p(y|\theta)$) and seems sensible to the engineering team (high $p(\theta)$):

$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{\int_\theta p(y|\theta) \cdot p(\theta) \cdot d\theta} \tag{12}$$

The denominator of this expression is constant w.r.t. θ, making it possible to express the above as:

$$p(\theta|y) \propto p(y|\theta) \cdot p(\theta) \tag{13}$$

$$\texttt{Posterior} \propto \texttt{Likelihood} \cdot \texttt{Prior}$$

(13) makes it clear that to estimate $p(\theta|y)$, two things are used:

- The probability of the data - y in n units passing - given a particular population passing proportion, $p(y|\theta)$ (the *likelihood*).
- The probability of a passing proportion according to the engineering team, $p(\theta)$ (the *prior*).

In this case, the likelihood is the probability of y units passing and $(n - y)$ units failing. Assuming that passes and failures and independent and that the units come from the same population, then the probability of y passes assuming that θ% of that population would pass is:

$$p(y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \tag{14}$$

In a more general sense, the likelihood is the probability of observing the data given that it was being generated according to the model parameterized by θ. The prior distribution, $p(\theta)$, encodes knowledge of what passing proportions are probable. If the engineering team is unsure what the passing proportion would be, then they may assume that all values are equally likely:

$$p(\theta) = 1 \qquad \theta \in [0, 1] \tag{15}$$

Figure 14 displays these prior and likelihood distributions. At this point the engineering team can do one of two things: they can evaluate the posterior analytically, or approximate it using a computer. Irrespective of the method chosen, the expression being evaluated is:

$$p(\theta|y) = \texttt{constant} \cdot p(y|\theta) \cdot p(\theta) \tag{16}$$

In practice, (16) is calculated using a computer. A grid of θ values is defined, and their prior probabilities and likelihoods are calculated in line with the functions in (15) and (14). This process can be described by the pseudocode:

$$\begin{aligned}
n &:= \texttt{No. of units tested} \\
y &:= \texttt{No. of units that passed} \\
\boldsymbol{\theta} &:= (0,\ 0.01,\ \ldots,\ 1) \\
\texttt{Prior} &:= \texttt{Uniform}(\boldsymbol{\theta}, [0, 1]) \\
\texttt{Likelihood} &:= \texttt{Binomial}(y, n, \boldsymbol{\theta}) \\
\texttt{Posterior} &:= \texttt{Prior} \odot \texttt{Likelihood}
\end{aligned}$$

Where `Uniform` returns the probability density of the uniform distribution for each value in $\boldsymbol{\theta}$ (i.e. a list of ones), `Binomial` returns the probability of y in n units passing given each of the passing probabilities in $\boldsymbol{\theta}$, and $\odot$ is the element-wise product. The posterior list would contain the probability density for each passing proportion θ, and is again shown in Figure 14. The peak indicates more probable values - a taller, pointier peak represents a more certain estimate because a few values have a much higher probability than their peers. In the same spirit, the flat prior that was used can be understood as highly uncertain.

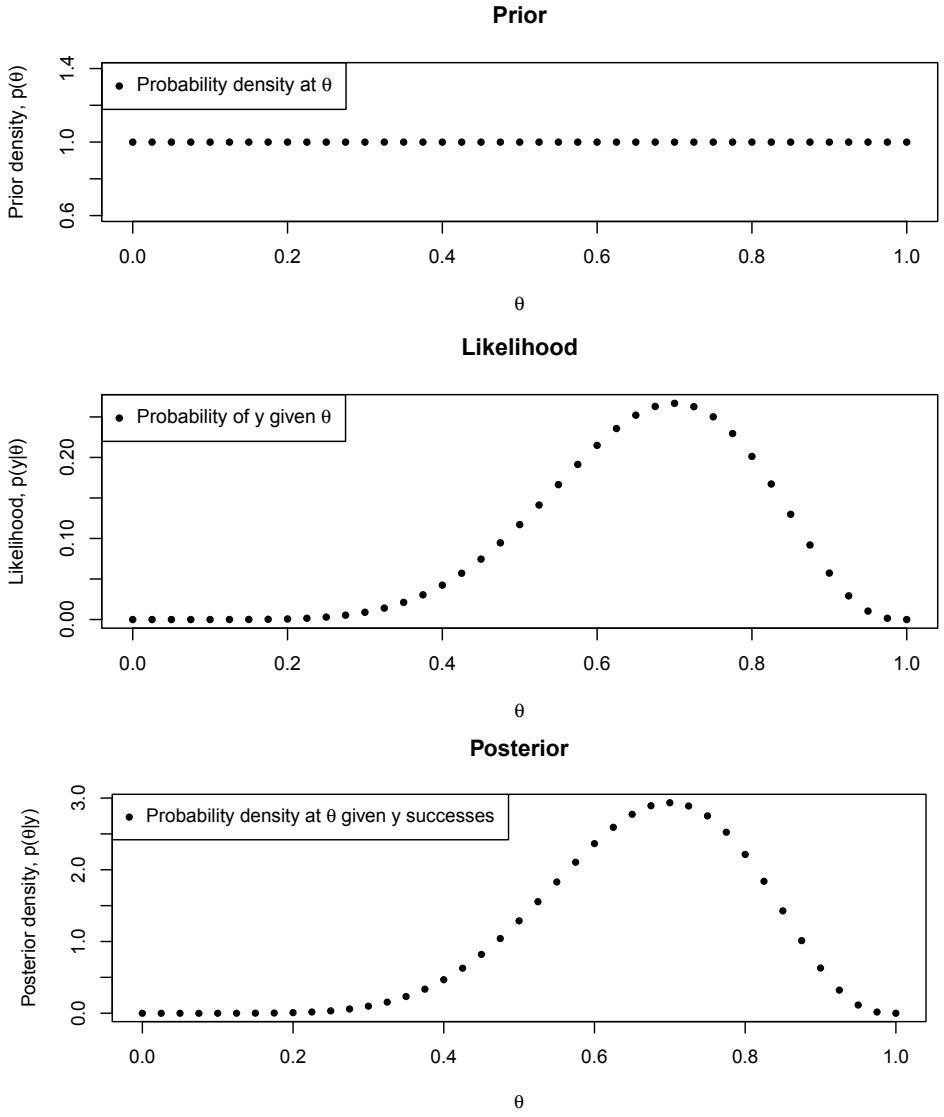

Figure 14: Combining prior knowledge with a likelihood distribution produces a posterior distribution of plausible parameter values.

Once the posterior has been calculated, it can be used to predict the behaviour of future units. $\tilde{y}$ denotes the number of future units that pass, $\tilde{n}$ is the number of units tested. The probability of $\tilde{y}$ successes, based on the observed data and additional information, is the weighted average of $\tilde{y}$ successes over all possible values of θ:

$$p(\tilde{y}|y) = \int_{\theta} p(\tilde{y}|\theta, y) \cdot p(\theta|y) \cdot d\theta \tag{17}$$

Once again, some potentially mischievous mathematics can be avoided by approximating this integral using a computer: draw samples of θ based on $p(\theta|y)$, then sample a value of $\tilde{y}$ from $p(\tilde{y}|\theta, y)$. Do this many times and the

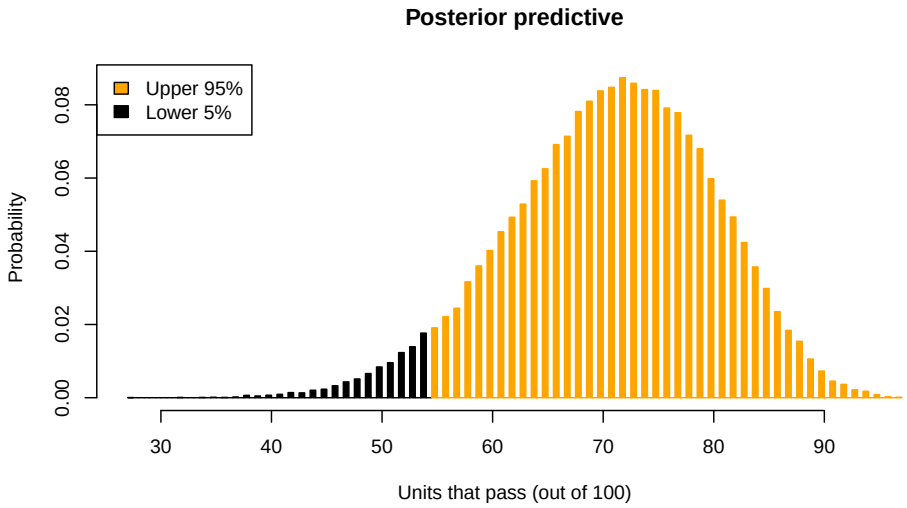

Figure 15: Posterior predictive distribution for the number of passing units in a 100-unit run.

relative frequency of $\tilde{y}$ values will tend towards $p(\tilde{y}|y)$.

```
for (i in [1, 10 000]) {
                    ñ := No.  future units to be tested
                    ỹ := (0, 1, ..., ñ)
                    θ := Sample(θ, Posterior)
Posterior predictive[i] := Sample(ỹ, Binomial(ỹ, ñ, θ))
                    }
```

Figure 15 shows a plot of the posterior predictive, along with the 5% lower limit on the number of units that will pass. This limit can be interpreted as a bound on the plausible number of units to pass, according to the evidence,

The process of inference just described would be valuable to DCA because it would provide a direct representation of how likely particular values are for key performance parameters. Unlike with standard errors, the uncertainty in the estimate is immediately apparent, and the effect of increasing sample size on accuracy is also clear, since the posterior of one analysis can be used as the prior of the next: this means that the information from tests is able to accumulate. This principle is shown in Figure 16, where the flat prior

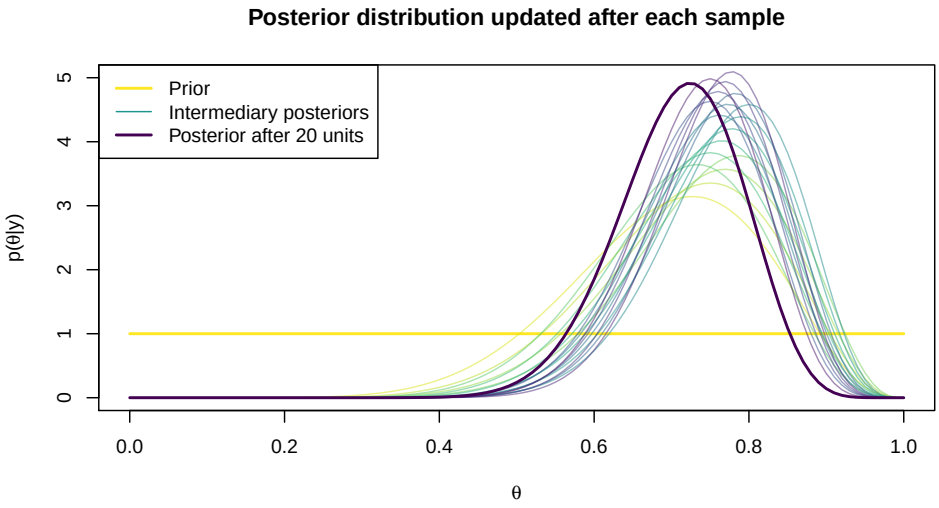

Figure 16: The uncertainty in the posterior decreases as more samples are conditioned upon.

represents initial ignorance about whether a unit will pass: as more units are run, an increasingly narrow peak forms around the most probable passing probability.

Another advantage of Bayesian methods over hypothesis testing is that it is relatively easy to build models that estimate many parameters simultaneously. An instance of this might be when estimating the mean and the variance of data that's assumed to have a normal distribution. The only change relative to the single-parameter scenario is that the prior and likelihood in Equation 13 need to be defined over two parameters instead of one:

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2) \cdot p(\theta_1, \theta_2) \tag{18}$$

Where θ_1 is the data's mean, θ_2 is its standard deviation, and y is the dataset. Using a joint prior that weakly favours a range of mean and variance values (based on sensible physical estimates) and a normal likelihood $p(y|\theta_1, \theta_2) = \mathrm{N}(\theta_1,\ {\theta_2}^2)$ results in a distribution of parameter values like that shown in Figure 17. This distribution shows how the data reduced uncertainty about what values are reasonable for the mean and standard deviation: it would probably be easier to explain to clients and colleagues than confidence intervals or hypothesis tests directed at a similar purpose.

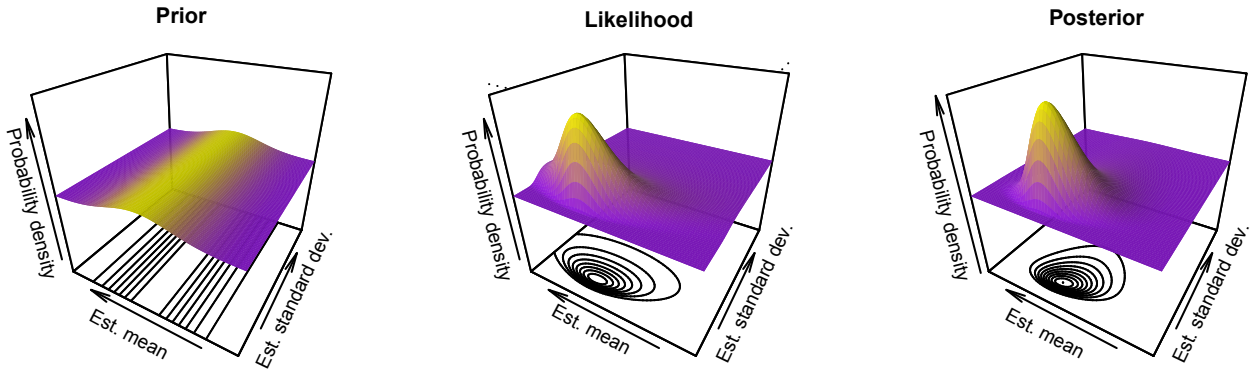

Figure 17: Bayesian inference can be extended to estimate multiple parameters simultaneously.

Shifting from classical methods to Bayesian ones would make statistics within DCA more transparent to both its engineers and clients. Hypothesis tests and interval estimates are easily misinterpreted and are opaque representations of the certainty in parameter estimates, since they are presented simply as numerical values. It is generally recognized that the scientific community as a whole needs to reconsider the practical relevance of hypothesis testing [9, 6]. The jargon of classical statistics can make it unclear what's relevant to the problem, and makes an honest explanation of its methods to non-technical team members difficult. Bayesian methods make it clear how the data and prior knowledge are being combined, and provide results that are easier to interpret. On the other hand, sophisticated Bayesian models require careful thought to be applied effectively, and may take longer to set up as a result. However, a few models could probably answer most questions asked in experimental work, such that once a model is set up it can be used to analyze data from many different experiments.

3.2 Experiment Design

Response Surface Methodologies

In any given one of DCA's products, it is likely that most parameters will either need to be:

- Kept within a range - for example, criticial dimensions, or
- Maximized/minimized - such as mechanism friction or split line prominence.

Response surface methodologies (RSMs) try to achieve these objectives by sequentially identifying the factors that affect the response parameter. The factors that have the biggest effect on the response are measured and approximately optimized first. Once this has been done, subtler effects - such as factor interactions - are explored, allowing the system to then be optimized with respect to all factors. Figure 18 is included to support this explanation[9][10].

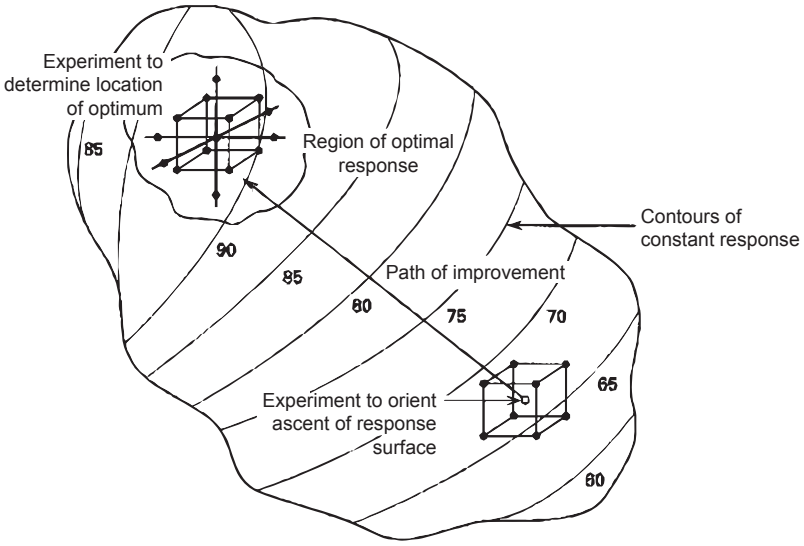

Figure 18: RSMs progressively optimize a system's response.

Response surface methodologies are built around linear models. The link is clearest when an RSM is broken down into steps:

1. Start by determining what the first-order effects of factors are on the response. In other words, estimate the coefficients β_i in the model:

$$f(x_1, ..., x_n) = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot x_i \tag{19}$$

[9]This Figure was adapted from Montgomery, 2000 [10].

2. Incrementally optimize the response based on the model. Were the response being maximized this would mean adjusting the factors x_i until the response $f(x)$ stopped increasing.
3. Determine what the second-order effects are in the vicinity of the first-order optimum. That is, fit a model of the form:

$$f(x_1, ..., x_n) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j + \sum_{i=1}^{n} \beta_{ii} x_i^2 + + \sum_{i=2}^{n} \sum_{j<i} \beta_{ij} x_i x_j \qquad (20)$$

   The first two terms correspond to the mean response and main effects determined in step 1, whose estimates are, in this step, refined in the vicinity of the optimum. The last step accounts for interactions.
4. Optimize the response according to these second-order effects.

Factors that have a negligible effect on the response need to be discarded to avoid having to test a staggering number of second-order treatments to test.

A 2-level factorial design can be used in step (1). This design provides enough information to estimate the intercept, main effects, and interaction effects - the left panel of Figure 19 explains this in detail. By using center points it is possible to estimate the response variance and to check whether there is any curvature in the response surface. The test for curvature is quite simple: if the response at the center point is substantially above or below the response plane described by the boundary points, then it suggests that the response surface in the region is curved. Curvature implies that quadratic (or possibly even higher order) effects are important to the response's variation.

At the second step, where the response is optimized in the vicinity of the optimum, three points along each factor's axis are needed to estimate quadratic effects. The design shown in the right panel of Figure 19 does exactly this: the outer elements can be used to estimate curvature (averaged across two paths), while the center point can be used to estimate error in the region of interest.

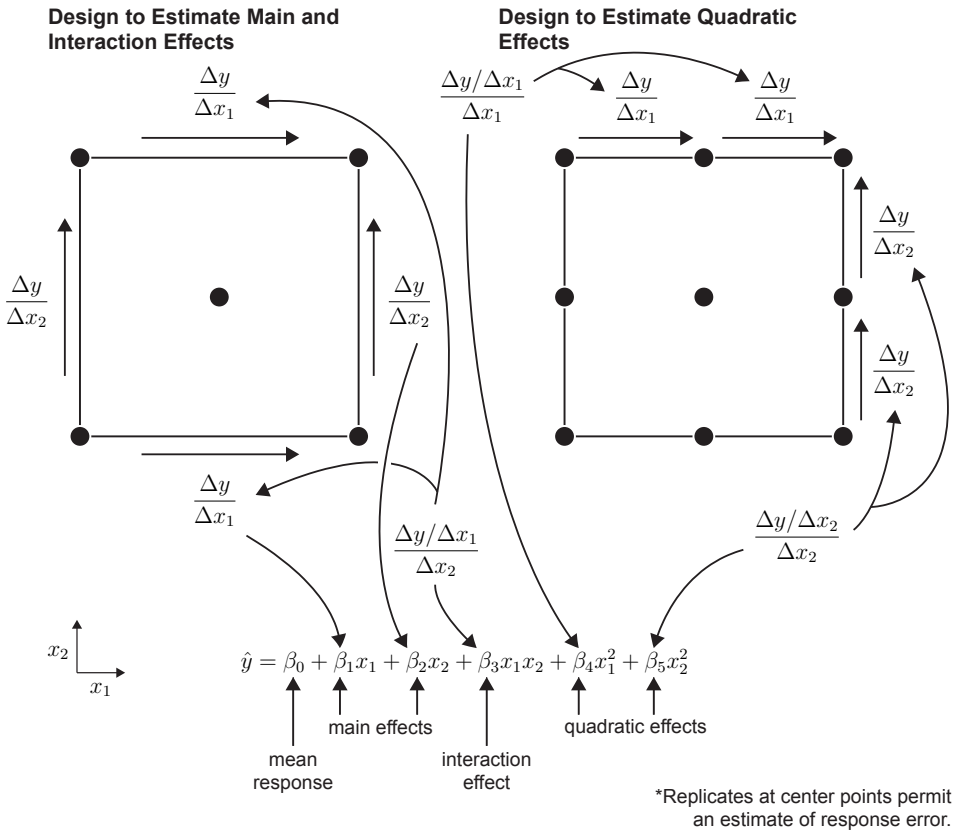

Figure 19: Experiment diagrams for the first (left) and third (right) steps in an RSM procedure.

RSMs would converge on solutions more quickly than DCA's current best-guess approach. It would also enable product performance to optimized. It can be difficult obtain the resources to run a large experiment, but perseverance will repay itself in the risk avoided and quality gained as a result. Investigations relying on expert guessing will tend to consume more resources than planned investigations. This design provides an idea of the minimum cost to understand the effects of a set of factors on a response. If this cost is too high and factors cannot be judiciously discounted, then the objectives of the investigation should be refined.

3.3 Visualization

Good visualization exposes patterns in data in a way that's immediately interpretable and precise. A visualization is a mapping from the numerical domain of a dataset to the visual domain of a plot. Data is not stored in a way that can be interpreted easily, so it needs to be transformed for it to be useful - mathematics is one means of making this transformation and plots are another.

Designing a visualization means giving thought to [12]:

- Data - are the variables continuous, categorical, or ordered categorical?
- Datapoint geometry - points, lines, or bars.
- Datapoint aesthetics - color, size, shape, and position of the geometries.
- Scales - mappings from the data's units of measurement to aesthetics.
- Summary statistics - plottable summaries of the data.
- Facets - plotting subsets of the data separately.
- Themes - typeface, non-data colours, axes and so on.

This many elements may seem overwhelming, but the most important are choice of geometry, aesthetics, and scales. Aesthetics can represent dimensions that are not displayed via a plot's axes: group membership, for example, is often depicted by a datapoint's colour or shape. Datapoint geometry, meanwhile, carries information such as resolution, chronology, and connectivity. Lines alone may obscure the sampling rate in a time series, and suggest some form of relationship or order between points. Bars are often used to represent an accumulated quantity (such as frequency), whereas points have the capacity to convey information about density and the number of measurements made. Scales describe how to map from numerical values in a dataset to an aesthetic's property. Consequently, scales affect a reader's perception of gradients and ordering. Several scales are shown in Figure 20.

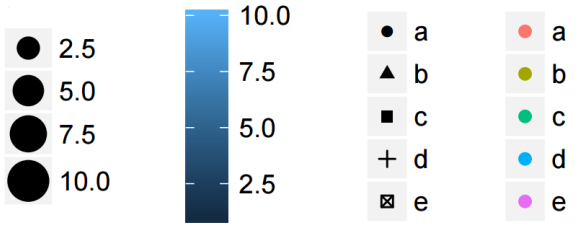

Figure 20: Scales describe how to map from numeric values to aesthetic attributes such as size, colour, and shape.

To show how the elements listed above provide a framework for presenting data, consider the following example. In DCA, experimental results are regularly discussed in team meetings. These discussions are guided by charts of the results; the conclusions of these meetings will guide an experimental investigation and can only be as well-informed as those charts permit. If charts contain irrelevant or misleading information, an investigation may stall.

Figure 21 is an example of how DCA currently summarizes test data. In this case, the performance of each unit at 10.8 seconds needed to be known. The behaviour of the units over the rest of the time span was thought to provide information as to why there were differences at that point. Each group had a distinct waveform and within groups the waveforms were similar. The shortcomings of presenting raw data directly were discussed in Subsection 2.4 - crude differences can be made apparent, but relevant information is otherwise subsumed by noise.

Figure 22 shows how splitting out specific observations into separate charts would have allowed this experimental data to be understood more deeply[10]. Each of its charts facilitates a specific technical discussion and provides information that supports the questions posited by its neighbors. The second chart, a strip plot, focuses on a key measure of each group's performance. The number of units tested is shown by this plot, as is the spread in the response, two pieces of information that are valuable to a reader because they measure the test's precision. Including the group

[10]The response units have been witheld to preserve confidentiality.

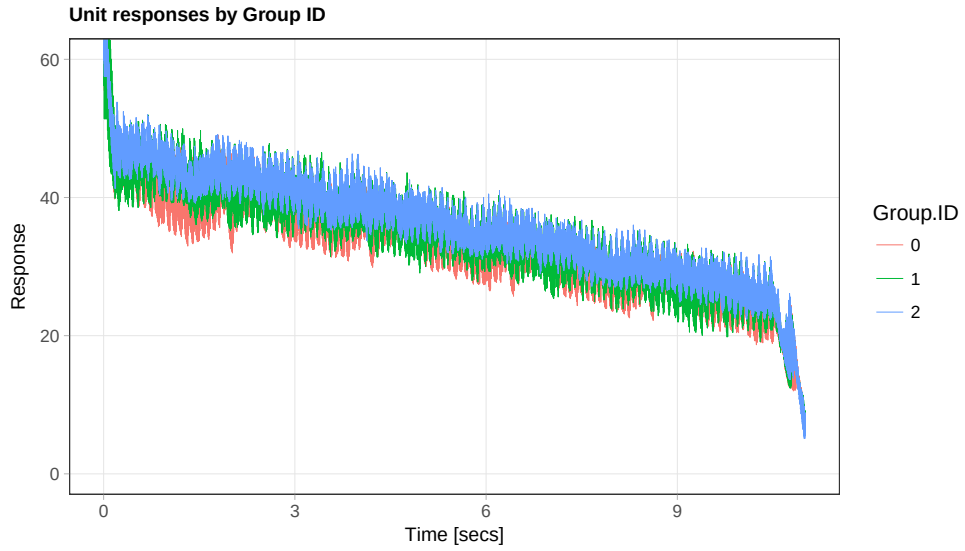

Figure 21: An example of how DCA currently summarizes experimental results by plotting raw data.

sample means also prevents outliers skewing how the groups relative performance is perceived. The colours mapped to each group in these first two plots are of a similar luminance and are colorblind-safe. Using colors of a different luminance[11] can unintentionally skew attention towards a particular datapoint or group. Additionally, the colors do not have a perceptual ordering, as say a sequence of red shades would. This is important since there was no natural ordering to the experiment's group treatments. The final plot uses color to emphasize an observation, and provides a diagram so that the physical system under study can be related to the experimental data. As with the first plot, an annotation states the message being conveyed.

Figure 22 stresses that if graphs are used to direct technical discussions then they should form a narrative. If a particular result is of interest then that result should be abstracted into as simple a plot as possible. Properties affecting the precision of an experiment - such as its size and controlled variables - can be embedded in a plot through aesthetics, geometry, facets, and scales. Where needed, diagrams and annotations should be used to provide the information needed to understand a plot. Applying these

[11] Luminance is a measure of a surface's brightness.

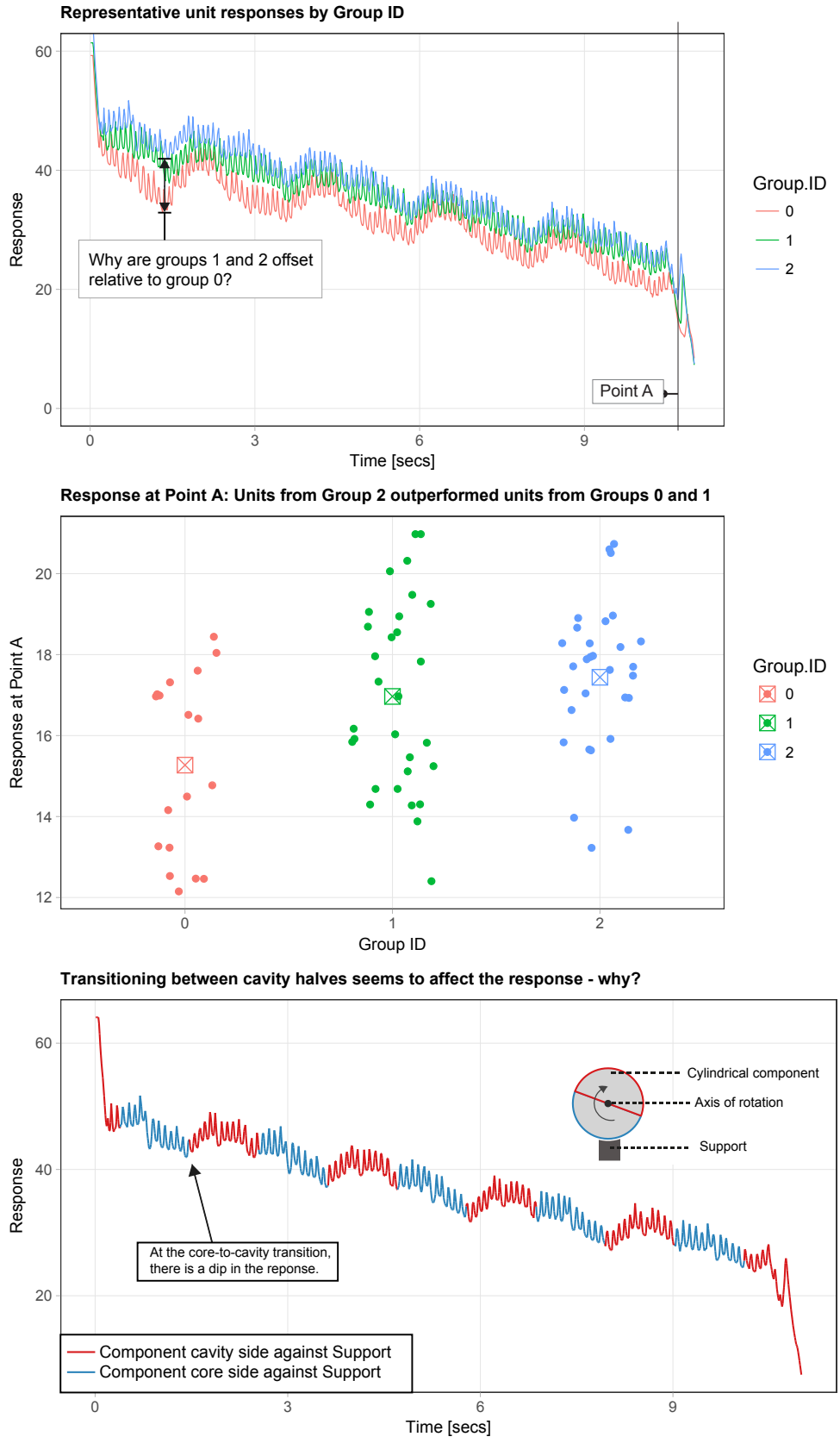

Figure 22: Splitting information across multiple charts, and considering how plot attributes will affect its perception, makes the results of an experiment clearer.

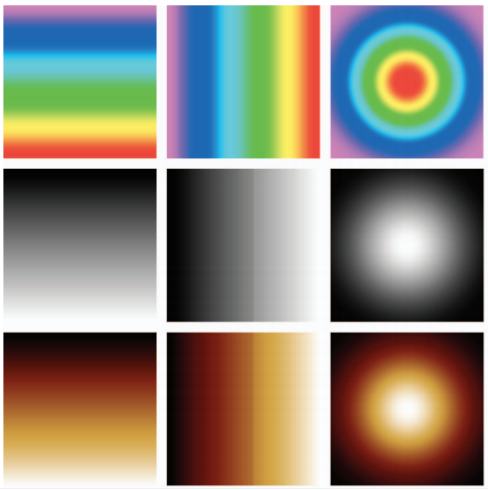

Figure 23: Each column corresponds to a set of data. Column 1 is a vertical linear scale; 2 is linear with a step at the center; 3 maps radius. Grayscale and black-body colormaps accurately represent the underlying data: jet does not.

principles would make it easier for DCA's engineers to discern nuanced patterns and to present their findings to their colleagues and clients.

As an aside, it is worthwhile to highlight some flaws in the jet colormap, the default used by many of DCA's simulation packages. Jet has been criticized as being actively misleading and confusing [2]. Luminance allows the visual system to perceive boundaries; across the jet colormap, luminance does not change at a constant rate. In some sections it changes very rapidly, whereas in others it is almost constant. This has the dual effect of suggesting sharp boundaries where they may not exist in the data, and obscuring actual borders in the map's isoluminant regions. Figure 23 demonstrates these qualities[12]. It is advisable that DCA use alternative colormaps where available, such as the black-body or grayscales maps also shown in Figure 23. Doing so would make it easier to identify genuine patterns in simulation results.

3.4 Software

Software is required to implement the tools presented so far. With regards to DCA's business purposes, this software would ideally be:

[12]This image was taken from Borland, 2007 [2]

- Easily learnt - can be picked up without requiring specialist training.
- Powerful - is able to implement the methods presented.
- Succinct - generates results quickly.
- Readily understood by other engineers and clients.
- Inexpensive.

Excel and Matlab are already used within DCA, and the company also owns a Minitab license. Excel, Matlab, R, and Minitab will now be compared according to the above criteria.

Excel is an excellent tool for quickly plotting test results, but it can only be used for basic analyses and plotting. It is clumsy for filtering data in comparison to Matlab and R. It is also prone to slowdown when large datasets are being handled. That said, it requires no previous programming experience to use, which is a signficant advantage when results need to be passed on to clients. Excel does not have features for experiment design, linear modelling, or Bayesian analysis.

Matlab's tools for statistical analyses are contained within Mathworks' Statistics and Machine Learning Toolbox. For the analytical methods presented here to be used, this Toolbox would be necessary. It costs upward of £1800 per year[13]. DCA's graduate engineers are already familiar with Matlab and do much of their mathematical modelling in the software. Consequently, Matlab is well-positioned for easing statistics into DCA's existing simulation work. Matlab's documentation is particularly clear and its debugging aids are very good, as would be expected from a piece of commercial software.

R is an open-source programming language designed for statistics. It is free to use, as are its many extensions that provide advanced modelling, visualization, and data-manipulation capabilities. Academic researchers

[13] An individual annual Matlab license costs £1800, with the Toolbox costing an additional £900. DCA already own five networke88d licenses (these are more expensive than individual licenses).

and professional analysts rely on R for their work. All of the plots in this report were produced using R, some of which would have been difficult to produce using Matlab; R offers data-reshaping tools that Matlab does not, and has an exceptionally flexible plotting extension. RStudio, a free development environment, is able to automatically format analyses into PDF reports. A major drawback to using R is that it has quite a steep learning curve. This is due to its unusual syntax, which can be difficult to read, and the cryptic error messages that result from its ability to store different data types in one variable. As an estimate, it would probably take one of DCA's engineers at least 20 hours to become familiar enough with base R to use it to run the analyses presented here. This would cost at least £1200, but would be a one-time investment (per engineer).

Minitab costs £1300 per license and provides a graphical interface for designing and analysing experiments. It guides the user towards appropriate experiment designs and analyses based on their needs, but is less flexible and extensible than Matlab and R. Its tools for filtering and reformatting data are also, by comparison, limited. It may be the most approachable choice here, besides Excel, for engineers without prior programming experience.

As is probably clear, no one software package stands out as being exceptionally well-suited to the company's needs. Matlab is more expensive than R, however the time taken teaching engineers to use the latter may offset this cost. R is the most capable piece of software considered and could probably support the most effective business processes. It would probably be more efficient to upskill a small number of engineers in statistical methods rather than attempt to teach all engineers in the company. Specialist courses are available, most of which will use one of the tools mentioned her, although these courses are geared towards traditional statistical methods.

4 Conclusions & Recommendations

Statistics allows engineers to:

- Design robust products.
- Optimize product performance.
- Minimize risk in product development.
- Conduct efficient experimental investigations.
- Develop a coherent understanding of a product's behaviour.
- Present the uncertainty inherent in experimental work honestly.

It is for these reasons that DCA should continue to develop its expertise in statistics. Professional experimental work is necessary to deliver high-quality products. Statistical expertise is pre-eminently marketable, particularly given the recent surge in popularity of machine learning algorithms. Being able to meet the technical demands of clients in this area would allow DCA to surpass the offerings of other technical product design consultancies: statistics provides a lucid approach to product development that is uniquely able to handle risk and meet strict performance criteria. Asking DCA's engineers to become statisticians is unrealistic: engineers are paid to deliver products, not to study. Consequently, it may be advisable for the company to hire a statistician or a testing development engineer. DCA's engineers are almost always involved with project work, giving them very little time to prepare the kinds of tools described in this report. A contractor dedicated to enhancing DCA's statistical offering could resolve the

company's existing experimental problems, and support the development work of DCA's core engineering staff.

A reasonable criticism of the methods proposed is that products do not need to be optimal, and that experimental investigations do not need to be absolutely efficient. This is true - they do not. Products can be "good enough". However, neglecting to profile a product's performance in the face of manufacturing and environmental variation exposes projects - and therefore the company - to severe risk. DCA claims to strive to "deliver market-leading products through world-class engineering". As clients demand increasingly sophisticated products, DCA's engineers will need to manage mounting complexity and uncertainty, particularly if they wish to match up to this bold claim. Statistics may allow them to do this.

Bibliography

[1] J.K. Blitzstein and J. Hwang. *Introduction to Probability*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2014.

[2] D. Borland and R.M Taylor. Rainbow color map (still) considered harmful. *IEEE Computer Graphics and Applications*, 27(2):14–17, March 2007.

[3] DesignWeek. Design week top 100, 2017 (accessed 13-06-2017). https://www.designweek.co.uk/top-100/.

[4] J.J. Faraway. *Linear Models with R*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2004.

[5] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[6] G. Gelman and J. Carlin. *Some natural solutions to the p-value communication problem— and why they won't work*, 2017.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York, 2013.

[8] E.T. Jaynes and G.L. Bretthorst. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[9] John K Kruschke and Torrin M Liddell. The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, pages 1–29, 2017.

[10] D.C. Montgomery. *Design and Analysis of Experiments*. Wiley, 2000.

[11] ISO Technical Committee 69: Applications of Statistical Methods. Iso 16269-6:2014 statistical interpretation of data - part 6: Determination of statistical tolerance intervals. 2014.

[12] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

A

A.0.1 What is a Probability?

Probability allows us to analyze a system without requiring complete mechanical knowledge of it. ''Randomness" refers to sources of variation that are not measured. You may have heard of probabilities as representing ''Degrees of belief". To understand what a belief is, consider this example. We machine a coin that we check is a symmetric disk of homogeneous density. I flip the coin ten times, and it comes up heads every single time. You might be surprised by this, and accuse me of flipping it in a controlled way. I then ask you how I can flip it in a way that is fair. What is your response?

If you say that it should come heads as many times as tails, then the experiment is no longer random, as we know what the outcome will be. You may gesticulate and say "You need to flip it *randomly*". I would press you to tell me what this means - I require a mechanism to decide how to flip the coin, and physical mechanisms are deterministic.

Point being, the probability of an outcome can only be evaluated relative to a set of assumptions you make about the mechanism generating those outcomes. You had a preconceived notion that the way I flipped the coin would favor neither heads nor tails, and therefore saw ten heads as supremely improbable. It is exactly these kinds of assumptions that form the basis of statistical analyses. Being able to express physical assumptions mathematically gives an analytical voice to our physical understanding of the world, and should ideally feel as natural as that understanding.

A.0.2 Closed-Form Solution to the Bayesian Inference Example

As was mentioned in the Section's main body, it is possible to solve some Bayesian inference problems directly, provided the prior and likelihood functions are convenient. As it happens, choosing a prior density function that has a similar form to the likelihood function makes solving for the posterior much easier. Priors chosen for this reason are referred to as conjugate priors. Equation (16) does not have a conjugate prior. Instead, its prior is uniform:

$$p(\theta|y) \propto \binom{n}{y}\theta^y(1-\theta)^{n-y} \cdot 1$$

This can still be solved directly. To find the constant of proportionality, the r.h.s. needs to be scaled by its integral over all values of θ, such that $\int_0^1 p(\theta|y) \cdot d\theta = 1$. In this the r.h.s. has the form of what's called a beta distribution

$$p(x; a, b) = \frac{1}{\beta(a,b)}x^{a-1}(1-x)^{b-1}$$

Where $\beta(a, b)$ denotes the beta function, which varies according to the exponents to ensure the function integrates to 1. The practical implication of this link is that the posterior corresponds to a closed-form function - the beta distribution - that is a valid probability density function (non-negative, integrates to 1). This means that rather than using a grid approximation, $p(\theta|y)$ can be obtained directly via:

$$p(\theta|y) = \frac{1}{\beta(y+1,\ n-y+1)}\theta^y(1-\theta)^{n-y}$$

A.0.3 Standard Error of Linear Regression Coefficients

Using the matrix notation introduced in the main report, it's straightforward to obtain a general formula for the standard error of a regression coefficient. Referring back to the normal equation

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

Where $\mathbf{X}$ is a matrix of inputs - p columns, one for each input, n rows, one for each observation - and y is a vector of n responses. Assuming that the reponses are uncorrelated and have a constant variance with respect to the

inputs:

$$E(y^T y) = \sigma^2$$

Then it's possible to deduce the covariance matrix of the vector of regression coefficients:

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= E[(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)] \\
&= \Big((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(y - E[y|\mathbf{X}])\Big)^T \Big((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(y - E[y|\mathbf{X}])\Big) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}(y - E[y|\mathbf{X}])^T(y - E[y|\mathbf{X}]) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2
\end{aligned}$$

The diagonal entries of this matrix correspond to the standard errors of the regression coefficients. Note that the estimate of common variance σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n - p - 1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

A.0.4 Why Normality is a Reasonable Assumption

The normality assumption can reduce an estimator's uncertainty, provided that the underlying variation satisifies several conditions. [Jaynes, 2003] provides two particularly intuitive explanations, known as the Herchel-Maxwell and Landon derivations. A more general (but longer and more technical) justification is given by the central limit theorem, where the random variable need only be a sum of independent random variables with finite variances that satisfy some condition. A simpler, but less general, version of the central limit theorem can be proven using moment generating functions (see [1]).

A.0.5 Derivation of the Sample Mean's Confidence Interval

Let $\bar{x}$ denote the sample mean, s the sample standard deviation, μ the population mean, α the confidence level. k is sought such that the interval

$[\bar{x} - k\frac{s}{\sqrt{n}}, \bar{x} + k\frac{s}{\sqrt{n}}]$ contains the population mean with a probability $1 - \alpha$:

$$\begin{aligned} \mathrm{P}(\bar{x} - k\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + k\frac{s}{\sqrt{n}}) &= 1 - \alpha \\ &= 1 - 2 \cdot \mathrm{P}(\bar{x} - k\frac{s}{\sqrt{n}} \leq \mu) \\ \implies \frac{\alpha}{2} &= \mathrm{P}\Big(\frac{\bar{x} - \mu}{s/\sqrt{n}} \leq k\Big) \end{aligned}$$

The last line implies that k has a t-distribution with $n - 1$ degrees of freedom. A t-distributed random variable is defined as the ratio between a standard normal r.v. and the square-root of a scaled chi-square random variable[14]:

$$t = \frac{\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right)}{s}$$

so that the interval that contains the true mean $(1 - \alpha)\%$ of the time is:

$$[\bar{x} - t_{n-1}(\tfrac{1}{2}\alpha) \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{n-1}(\tfrac{1}{2}\alpha) \cdot \frac{s}{\sqrt{n}}]$$

A.0.6 Derivation of Tolerance Intervals assuming a Normal Population

Tolerance intervals permit statements about the performance of a population to be made, with clear limits on that statement's uncertainty. Tolerance intervals are derived by considering the probability that a member of the population will be within a particular range. For a one-sided tolerance limit - a value for which at least $p\%$ of the population is greater than or less than - this means:

1. Defining k such that the probability a member of the population $(\mu + u_p\sigma)$ is greater than k sample standard deviations from the sample mean is equal to $1 - \alpha$.

   $$P(\bar{x} + ks \geq \mu + u_p\sigma) = 1 - \alpha$$

   Where $\bar{x}$ is the sample mean, s is the sample standard deviation, μ is the true mean of the population, σ is its true standard deviation, and u_p is such that $\mu + u_p\sigma$ is greater than $p\%$ of the population. α is the confidence level.

[14] Nb. $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_n$

2. Assuming x has a normal distribution, then by definition $\frac{(n-1)s^2}{\sigma^2}$ will have a chi-square distribution. This implies that k has the same distribution as a t-distributed r.v. centered at $\sqrt{n}u_p$ and scaled by $\frac{1}{\sqrt{n}}$:

$$k = \frac{1}{\sqrt{n}} t_{n-1}(\sqrt{n}u_p)$$

Consequently, the distribution of the r.v. k is effectively a more spread-out version of a normal distribution - the larger variance is a consequence of having to model the variance as another r.v.

3. The lower interval containing at least 95% of the population over α% of constructed intervals is therefore:

$$\Big(-\infty, \bar{x} + \frac{t_{1-\alpha}(\sqrt{n}u_p, n-1) \cdot s}{\sqrt{n}} \Big]$$

A.0.7 Other Justifications for Least-Squares Regression

As mentioned, there are two alternative reasons for minimizing the residual sum of squares besides that doing so brings the predicted and observed response vectors as close to one another as possible. The first of these is that the RSS criterion follows from maximizing the likelihood of the observed responses, given that the input-response relationship is linear and that the errors in the response are normal with constant variance and mean zero:

$$\text{Assume} \quad y = \mathbf{X}\beta + \epsilon \quad \text{such that} \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$
$$\text{i.e.} \quad p(\epsilon|\hat{\beta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{\epsilon^2}{2\sigma^2})$$

This implies that $E(y|\mathbf{X}) = \mathbf{X}\hat{\beta}$. The likelihood function is, therefore:

$$p(y|\hat{\beta}, \mathbf{X}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(y - \mathbf{X}\hat{\beta})^2}{2\sigma^2} \Big)$$

Maximizing the probability of the observations w.r.t. the estimated regression coefficients can be seen as a reasonable thing to do. By taking logarithms and dropping the constants, the likelihood function becomes:

$$l(\hat{\beta}) \propto -(y - \mathbf{X}\hat{\beta})^T (y - \mathbf{X}\hat{\beta}) \tag{21}$$

Such that maximizing the likelihood is equivalent to minimizing the RSS criterion.

Alternatively, it is possible to jusitfy the least-squares fit without a normality assumption. This is because least-squares provides the lowest variance estimate of EY for a given X. In other words, a least squares fit will, on average, be closer to the true response than a linear fit made in some other way. This result is known as the Gauss-Markov theorem.