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 shouldn't 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 tools are suggested that would allow DCA to realize the benefits of statistical methods.

JUNE 8, 2017

Table of Contents

*

List of Tables

List of Figures

Notation & Glossary

Attribute A measurable property of a *unit*.

Acknowledgements

Beyoncé, J.D.Sallinger, and Santa Claus.

1 Introduction

DCA Design International is a 150-person product design consultancy based in Warwick. Their work is oriented towards the mechanical design of medical and consumer products: their past clients include Unilever, Sanofi-Aventis, and GSK. The products they engineer are usually hand-held items such as insulin injector pens or deoderant cans: Figure 1 shows two of their 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.

DCA employs about sixty mechanical engineers. Each of them are general-purpose technical consultants that are capabl of fulfilling the engineering demands of a project, be that design, physical prototyping, or analysis. DCA's substantial investment in engineering distinguishes it from other product design consultancies, many of which do not have the expertise to handle a product's technical development. This investment is manifest in both the its engineering workforce and its ownership of four test

Figure 1: Products designed by DCA.

labs. The experimental work that makes use of these labs is one of DCA's value propositions as a consultancy, and it is the collection, analysis, and presentation of this experimental data that is the focus of this report. Data-oriented activities constitute the scientific discipline of statistics.

Statistics makes it possible to profile - and therefore understand and predict - the variation of real-world systems. This is critical to robust product design: knowing what factors affect performance, and by how much, is essential to making a product reliable.

The first half of this report evaluates how DCA currently applies statistics: the second half then suggests methods from modern statistics that would address the weaknesses identified. In particular, Section 2 explains the company's current investigatory framework and how statistics is currently applied within it. In Section 3 the company's approach is compared to modern statistical practice, in the process of which these alternative methods are detailed and evaluated. Tools for implementing statistical methods are also discussed in the context of DCA's needs. The report concludes with an evaluation of how actionable the suggested methods are, and responds to possible 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 the needed knowledge being identified. Experiments are then be 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 project associated with that investigation. Typically investigations focus on a particular product parameter, such as the volume of fluid dispensed by an injector, or the propensity of a inhaler to fail upon being dropped.

Most lab work is by engineers on medical projects. 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 however, the timeframes and functional requirements of such products limit the relevance of extensive experimental

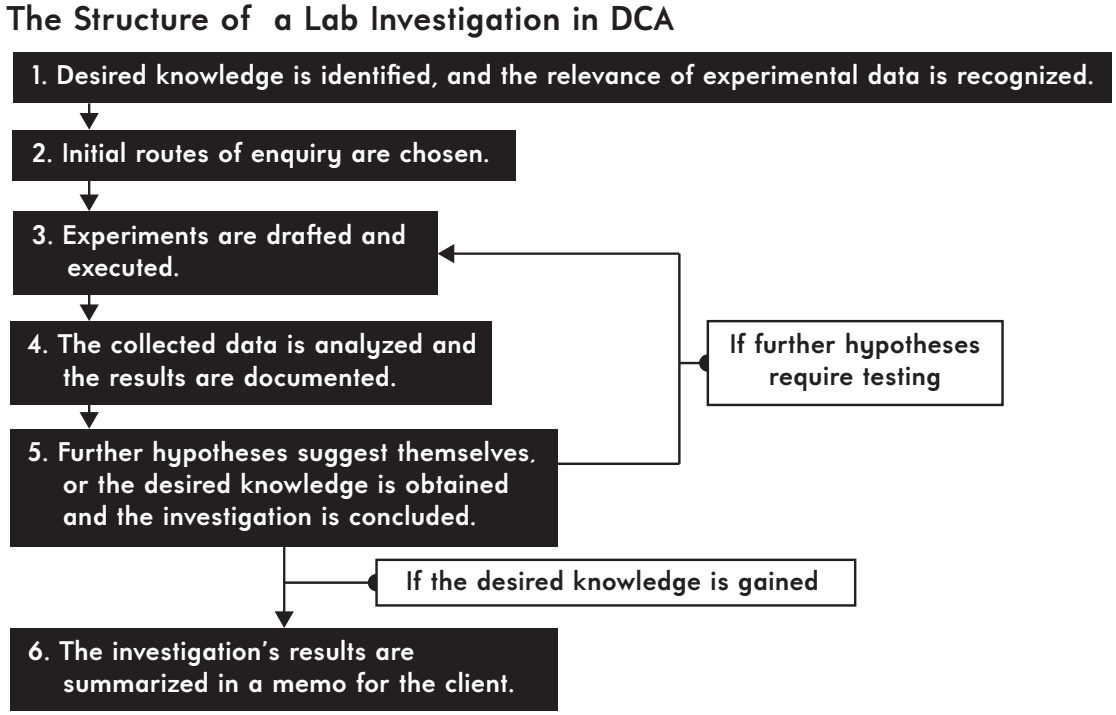

Figure 2: Investigation diagram.

investigations to them: medical products on the other hand, see a good deal of the test lab.

Product testing can be done at any point over a development cycle, however as can be seen in Figure 3, the rate of testing increases as a product develops. Towards a product's relase date is when resolving minor performance issues becomes a worthwhile pursuit, exploration for future product variants become 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, however other experiments are sometimes conceived to provide supplementary information. With that in mind, this report attempts to be

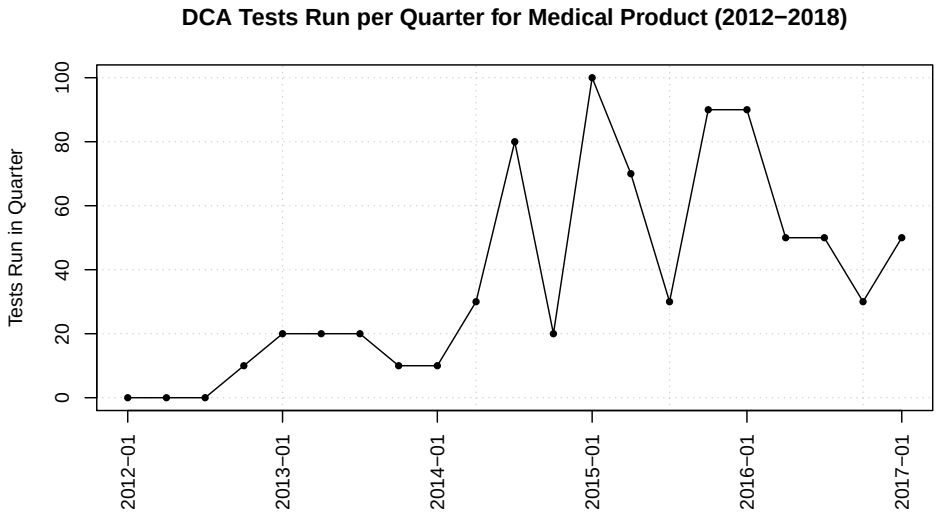

Figure 3: Frequency of tests in terms of time until project launch.

data-agnostic in its recommendations of analytical techniques.

Other engineering activities that DCA's engineers apply statistics to are tolerance analysis and, increasingly, predictive user interfaces. Unfortunately, the latter cannot be discussed for confidentiality reasons; the former can be, and is - see the section on Monte Carlo estimation. Outside of engineering, statistics is used to varying degrees within the Human Factors and upper management od DCA, however these applications aren't talked about here.

To help make sense of how statistics is applied in DCA's lab investigations, its use at each step of an experimental procedure - planning and execution, analysis, and presentation - is discussed in order. For any experimental investigation to be successful, the experiments themselves need to be given careful thought, and should be conducted in a similar spirit. Experiment design aims to guide such thought.

2.2 Experiment Design

Experimental design and analysis can be used to make products that perform better, are more reliable, less risky to develop, and have a uniquely

justifiable development process. It is expertise that would elevate DCA's capacity as a technical consultancy.

Design of Experiments refers to both experiment designs and a broader philosophy of systematic experimentation. An experiment design is a particular structure of experiment, such as comparing the effects of two factors each at two levels. Good experimental design produces data that is unambiguous and relevant to an experimental objective.

Three principles are crucial to robust experimentation:

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 validifies the important analytical assumption that observations are randomly drawn from a distribution.

Blocking Blocking allows the effects of a nuisance factor to be averaged out during analysis, by accounting for unit differences when assigning treatments - see Figure 4. A block is a set of similar units.

These 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.

This being said, DCA lacks a framework for planning experimental investigations. As a consequence of this, fundamental activities such as verifying that experimental set-ups produce repeatable results, scoping an investigation, and running screening experiments to systematically close off avenues of enquiry are sometimes forgotten. The dominant experimental strategy within the company is a best-guess approach: one treatment is

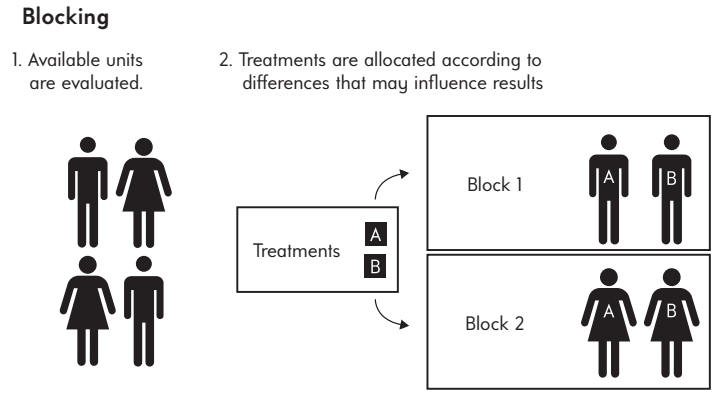

Figure 4: Diagram of what blocking involves.

tested in each experiment, chosen based on the expert insight of the engineering team. This method's most serious 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 it may be tempting to stop the investigation when a better solution is actually available, allowing luck to play too large a role in a product's design.

DCA's experimental procedure is compared against conventional experimental steps in Table 1.

Table 1: Comparison of DCA's experimental procedure with conventional practice.

Experimental step	DCA's implementation	Strengths	Suggestions
Recognition and statement of the problem.	A problem is usually identified in either other experiments or design-side activities. It is not formally stated, but is agreed in loose terms among the engineering team. There is no mechanism for assessing whether a problem is well-suited to being addressed by a lab investigation, as opposed to other analytical methods.	The benefits that experimental investigations provide, such as empirical validity and flexibility, are recognized.	A precise problem statement focuses an investigation towards a particular end, and allows progress towards this end to be gauged. It also makes it clear to the team what an investigation aims to achieve.
			Specific problem statements make it easier to review previous work - without them, it is difficult to determine where one investigation begins and another ends.
Choice of factors, levels, and ranges.	This choice is made in engineering team meetings. Factors can be identified haphazardly: there is no labelling of those that are identified as design or nuisance factors, or whether they are controlled or uncontrolled. Ranges and levels are usually chosen according to expert knowledge. The number of levels is usually kept small (2 or 3) because differences rather than overall responses are of interest.	Level choices have a rational motivation which is justified via physical reasoning or previous experimental results.	A list of factors guides the systematic elimination of sources of variation from an experimental set-up, and can be used to survey for possible confounding factors.
		Deciding which factors are relevant in a meeting uses the entire team's engineering knowledge and critical thinking skills.	A meeting should be a place to review a choice, not to generate it - conversation is not methodical and can easily miss relevant information.
Selection of the response variable.	The response variable is usually evident from the problem statement (e.g. torque output of mechanism). More than one way to measure the response variable will almost always be considered.	Engineers in the company can generate many possible response variables, and analyze their respective merits.	Precursor experiments comparing response variables may reduce the length of investigations.
		The response variable chosen is chosen carefully to closely represent the system under study.	Documenting the alternative response variables would make it easier to clearly outline why one method is superior, and to transition between testing methods as circumstances change.
Choice of experimental design.	The experimental design is also chosen in a meeting. They are usually one from a small selection (detailed in the text body). The choice made is incidental, as reflected by the absence of planning documents.	Simple experiment designs are easily communicated, executed, and documented.	A well-chosen experimental design can reduce the resources (time, materials, and effort) expended in satisfying the investigation's objective.

			Considering analysis beforehand makes it possible to ensure analytical assumptions are met.
Performing the experiment.	Engineers run their experiments in a laboratory. Frequently run experiments have protocols; hand-written observations are mandated for all experiments.	Labs afford flexibility in the level of experimental control.	Experimental fixtures should be shown to generate repeatable, reliable results before being used.
		Blank observation sheets encourage critical thinking about the experiment Experiments are run by engineers solving the problem - this makes engineers personally responsible for their results, and exposes them to undocumented experimental information.	Images of test set-up would make it much easier to retrospectively understand an experiment. Observation sheets could provide guidance on what factors to monitor.
Analysis of the data collected.	Analyses are run as soon as data is available, and will be handled by the engineer that ran the experiment. Excel - and occasionally Matlab - is used. The conclusions tend to be judgemental as opposed to statistical. Compared to the time spent running the experiment, analysis is brief. Analysis is discussed in more detail in the next section.	Engineering expertise is applied to explain experimental results in a physically meaningful way.	Statistics should be used to ensure that sound conclusions are made - subjective assessment alone is susceptible to various biases that can lead to lost time and confusion.
			Time invested in analyses should be seen as what makes an experiment practically useful, rather than a formality between experiments.
Conclusions and recommendations	Conclusions are incorporated into client memos and presentations. Interim results are presented at internal meetings - graphics play an important role in communicating results.	The importance of graphics is realized and put to good effect in client presentations.	Charts exist beyond those being used that may make it easier to demonstrate experimental results.
		Conclusions are presented in a way that is accessible and avoids needless technicalities.	Experimental results are not supplemented by estimates of uncertainty.

The experimental designs used in the company are enumerated, explained, and critiqued in Table 2. As mentioned, the core problem with DCA's experimental strategy is its narrow focus on testing specifics - running a few large experiments considering many different treatments at once can be much more efficient than many very specific experiments. Moreover, there is no understanding of what constitutes a statistically efficient design -...

Experimental Design	**Description**	**Evaluation**
Randomized complete block design	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	Applied to experiments in which more than one factor is varied - all combinations of factor levels 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 must be run. DCA is very well endowed to run experiments, and has a system in place for documenting the date, purpose, and conditions of an individual experiment. Systematically identifying and stating whether factors are controlled would allow the company to make more effective use of its experimental equipment - things are easily forgotten, and mishaps can easily be avoided through a little forethought.

Once an experiment has been planned and run, it remains to analyze the

raw data to extract useful information. A well-designed experiment's analysis should write itself - the purpose should be evident from the beginning. Methods for constructing an analysis are now discussed.

2.3 Analysis of Experimental Data

The content of several hundred test reports was tabulated to inform this discussion, which focuses on summary statistics and interval estimates.

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's 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 isn't superficially obvious, and being unable to properly relate physical behaviours to experimental observations. Foregoing statistics when analyzing product behaviour severely handicaps the ability of an engineer to design a robust product.

The reports surveyed 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, and one report applied a t-test. These tools will now be explained, and their usefulness and possible weaknesses detailed.

Summary Statistics

Randomness simply means variation in an event as a result of unseen factors. A random variable (r.v.) is a function that maps events onto real numbers. For example, we could define an r.v. X 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 - for example, we might use an r.v. that counts the number of sucesses 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 of values to its probability. Figure 6 highlights this difference.

The essential problem of experimental statistics is trying to understand the behaviour of a broader population from just a sample. In product design, this means estimating the distribution of a population using measurements from just a limited number of prototypes. The attributes of this distribution - such as its spread and average - can be estimated using summary statistics.

A sample's mean response, for example, approximates the mean response of a population. This result is regularly used in DCA to discriminate between the performance of two or more populations, each representing possible design variants. The accuracy of this estimate improves as more samples are tested, with diminishing returns, a relationship that is shown in Figure 5. Because a sample may not be representative of its population, it's usually useful to understand how far off the sample mean could be from the true mean.The sample mean's average difference to the true mean is called the standard error, and it corresponds to $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the population and n is number of units in the sample. DCA implicitly appeal to this relationship when they choose to run more units in a test, recognising that bigger samples tend to better represent their populations.

Certain summary statistics can be thought of as estimates of a distribution's 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). Viewing statistics as an exercise in estimating a distribution makes handling the uncertainty inherent in test results more natural, and avoids misleading people with small-sample point estimates. DCA's engineers at particularly at risk of the latter because their experiments are usually constrained by the number of units that can be tested.

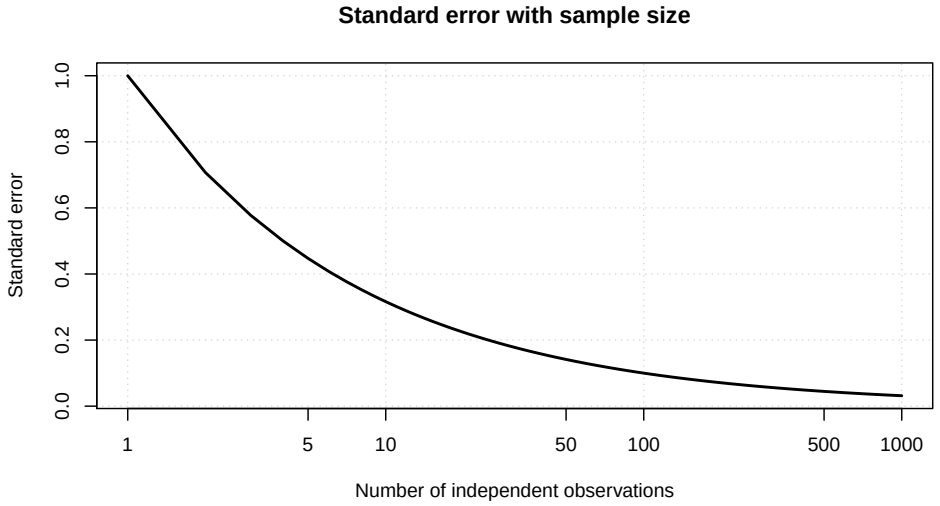

Figure 5: Convergence of sample mean to population mean.

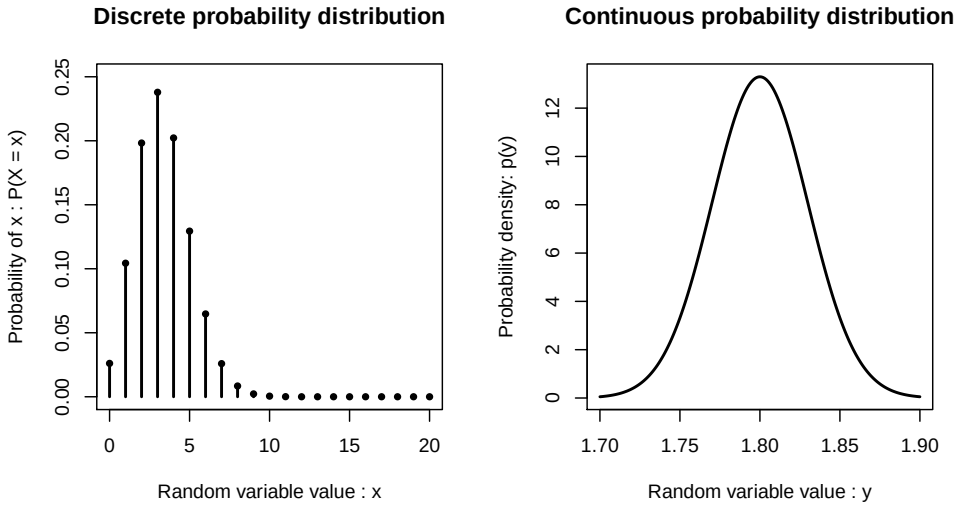

Figure 6: Left: Probability mass function. Right: Probability density function.

Estimating a distribution will be seen again later, in the section on Bayesian inference. Interval estimates are a tool for quantifying the uncertainty on an estimate, and can be made with assumptions about the distribution's shape, or without. They need to be interpreted carefully, and an appropriately careful explanation of two of the most often used interval estimates - confidence and tolerance intervals - will now be presented.

Tolerance Intervals

DCA's most sophisticated statistical analysis is based on ISO 16269-6, *Determination of statistical tolerance intervals*. This standard outlines how to construct tolerance intervals under either no assumptions about the random variable's distribution, or the assumption that the random variable has a normal distribution. A tolerance interval is a range of values that contain a particular fraction of the population. Because this interval is only an estimate, they can only contain the advertised fraction of the population most of the time. The proportion of time that an interval would, on average, contain the specified fraction of the population is called the confidence level. To elaborate a little further, 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 shows what this means.

So tolerance intervals allow us to make statements about the performance of a population, with clear limits on that statement's uncertainty. Tolerance intervals are derived by thinking about 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 $(mu + usigma)$ is greater than k sample standard deviations from the sample mean is eqaul to $1 - a$.

$$P(\bar{x} + ks \geq \mu + u_p\sigma) = 1 - \alpha \tag{3}$$

   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.
   In words, k is such that $\bar{x} + ks$ will be greater than the p% of the population $(1 - \alpha)$% of the time.

2. Assuming x has a normal distribution then u_p can be read off a table of normal values, and by definition $\frac{(n-1)s^2}{\sigma^2}$ will have a chi-square distribution (see Appendix A). What this means is that k has the same distribution as a t-distributed r.v. centered at $\sqrt{n}u_p$ and scaled by $\frac{1}{\sqrt{n}}$:

$$k = \frac{1}{\sqrt{n}} t_{n-1}(\sqrt{n}u_p) \tag{4}$$

This is a normal distribution that's more spread out - the bigger spread represents the uncertainty on account of the fact that the true variance isn't known.

3. Bonanza! The lower interval containing at least 95% of the population is

$$\Big(-\infty, \bar{x} + \frac{t_{1-\alpha}(\sqrt{n}u_p, n-1) \cdot s}{\sqrt{n}} \Big] \tag{5}$$

A fraction $1 - \alpha$ of these intervals will contain less than p% of the population, as shown in Figure 7.

DCA use tolerance intervals in a particular test to check that a certain proportion of the population satisfies a performance threshold. Their use isn't widespread, and previously it's been the case that k values haven't been properly updated according to the conditions of new tests.

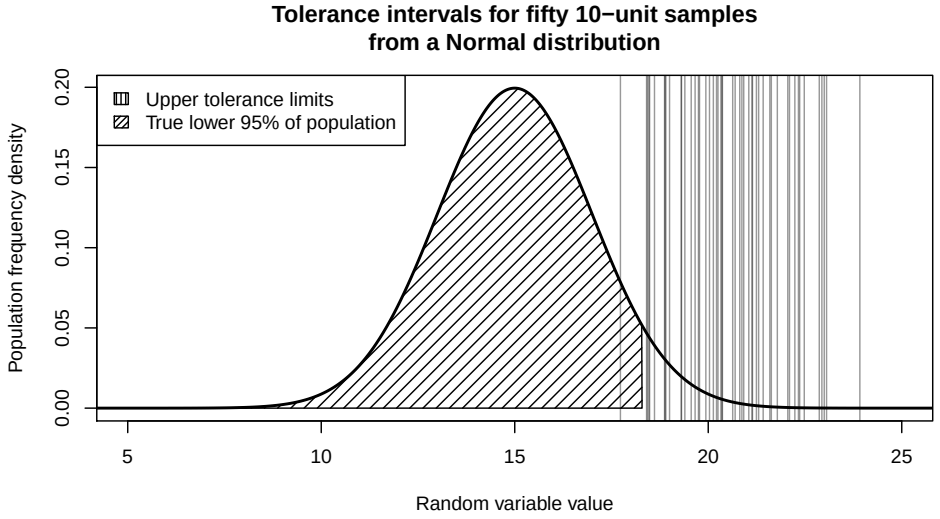

Figure 7: Twenty tolerance limits versus population limit.

The strengths and suggestions for DCA's tolerance intervals are listed in Table 3.

Table 3: Evaluation of tolerance intervals.

Strengths	**Shortcomings**
Provides a threshold indicating roughly where a certain fraction of the population is.	Confidence levels can be misinterpreted as specifying the probability that a constructed interval contains at least 95% of the population.
In line with the expectations of regulatory standards	Normality assumption needs checking
Can be setup to be applied without theoretical understanding	Repeated use at a low confidence level increases the probability that the limit will be under-estimated.
	k-values need to be looked up - could easily be done incorrectly if the theory isn't understood.

Confidence Intervals

Another tool that DCA's engineers occasionally use is confidence intervals, which indicate a range of values that a parameter is likely to fall within. As with tolerance intervals, this range will only contain the population parameter a certain fraction of the time however, a problem that's unavoidable since there will always be a chance that an unrepresentative sample is drawn. For example, if we were to construct a confidence interval for the population mean of 100 samples, each of 5 units, the confidence level would tell us how many of these intervals would - on average - contain the actual value of the population mean. Figure 8 demonstrates this idea. Confidence intervals can be placed on any parameter estimate, although they're usually used to quantify the uncertainty on an estimate of the population mean. Their derivation is similar to a tolerance interval's:

$$\begin{aligned} \mathrm{P}(\bar{x} - ks \leq \mu \leq \bar{x} + ks) &= 1 - 2 \cdot \mathrm{P}(\bar{x} - ks \leq \mu) = 1 - \alpha \\ \implies \frac{\alpha}{2} &= \mathrm{P}\left(\frac{\bar{x} - \mu}{s} \leq k\right) \end{aligned} \tag{6}$$

The last line implies that k has a t-distribution with $n - 1$ degrees of freedom, so that the confidence limit that contains the true mean $(1 - \alpha)\%$ of the time is

$$[\bar{x} - t_{n-1}(\alpha) \cdot s,\ \bar{x} + t_{n-1}(\alpha) \cdot s] \tag{7}$$

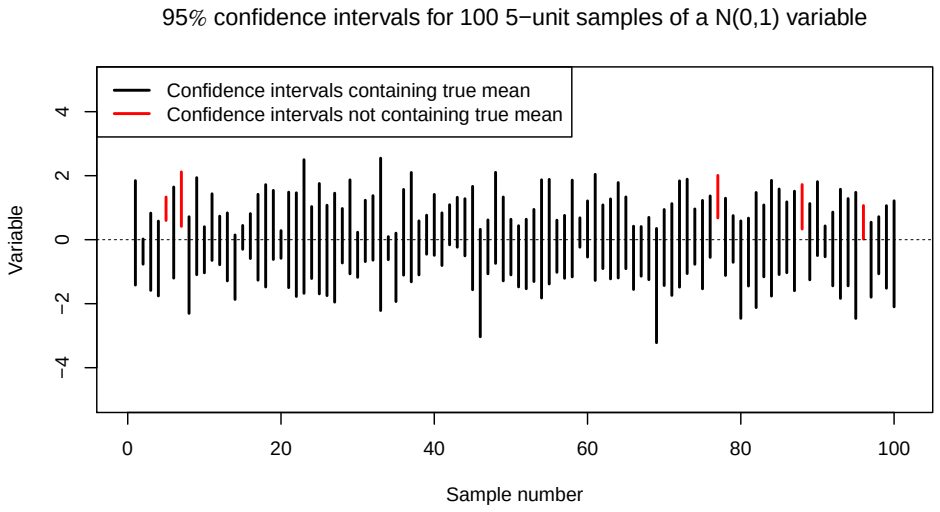

Figure 8: 99% confidence intervals, to a 5% confidence level, for 100 10-unit samples.

Confidence intervals suffer from similar problems to tolerance intervals. DCA's engineers use them very rarely.

Monte Carlo Estimation

Monte Carlo estimation approximates a quantity by simulating the random process generating it . In DCA it's been used to analze tolerance chains in products. The use case was somewhat similar to the following: the tolerance limit of a combination of distributions, each corresponding to a part dimension, was needed. Take $X \sim \text{Binom}(n = 10, p = Y)$ as an example, where $Y \sim \text{Beta}(a = 7, b = 3)$[1]. Rather than attempt to derive the distribution of this dimension's value directly, tens of thousands of values of y were first generated according to Y's distribution using a computer. Each of these values were then used to generate a value of x from $\text{Binom}(n = 10, p = y)$. The resulting frequencies of the x values then represented the dimension's distribution. It was then possible to calculate the mean by averaging over all the x values obtained. A diagram of this process is shown in Figure 9.

[1]The beta distribution is a continuous and generates a number between 0 and 1, which makes it useful in modelling the distribution of a probability.

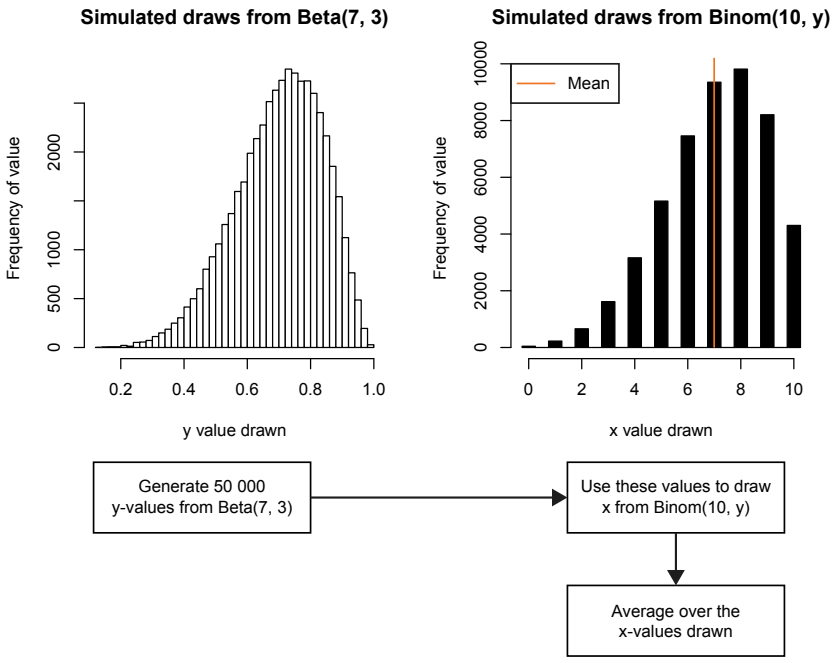

Figure 9: Process diagram for a Monte Carlo simulation.

2.4 Visualizing Experimental Data

Visualization is essential to clearly and convincingly summarizing an experiment's results. Graphical tools allow engineers and clients to see for themselves what's been discovered. A plot should relevant, easily interpretable, and accurately convey its underlying data.

DCA's reports and client presentations frequently contain plots of the data collected from an experiment, typically generated using either Microsoft Excel or Matlab. The plots used are line and bar charts, along with the occasional scatterplot. Line charts are particularly ubiquitous in DCA because they're directly plottable from the raw data provided by axial and torsional testing machines. As a consequence of this many graphical summaries are usually overlaid line plots, similar to that shown in Figure 10. The effectiveness of this use-case is evaluated in Table 4: in short, this type of plot can include a lot of redundant information and can make it difficult to see how individual units are behaving. On the other hand, it takes very little time to prepare, and can be related to patterns observed over the course of an experimental run. Specific problems that DCA suffer from are: not providing sufficient resolution, smoothing, and using an inappropriate line thickness.

Bar charts are also used fairly frequently to display performance relative to a nominal value, and a particular format of scatterplot is used to present the results of a tolerance limit analysis. The latter is shown in Figure 11.

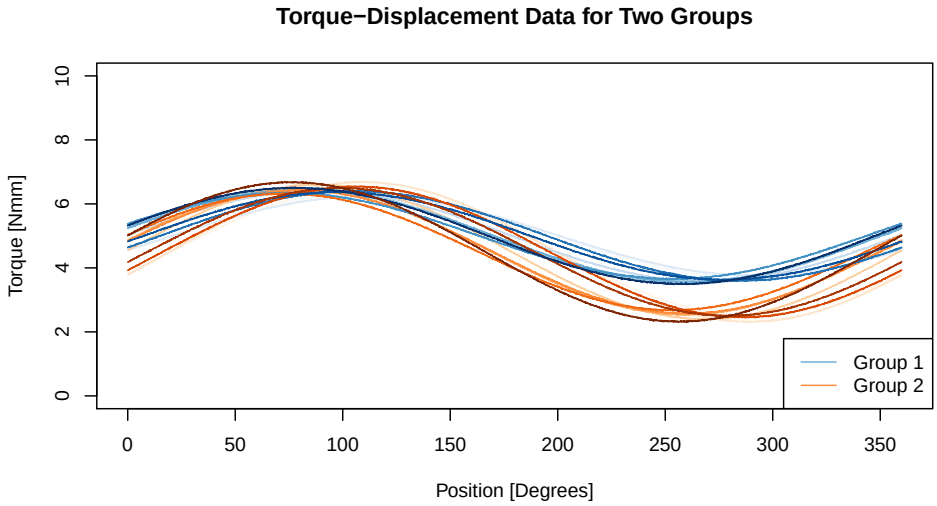

Figure 10: Sample summary line chart from a DCA test report.

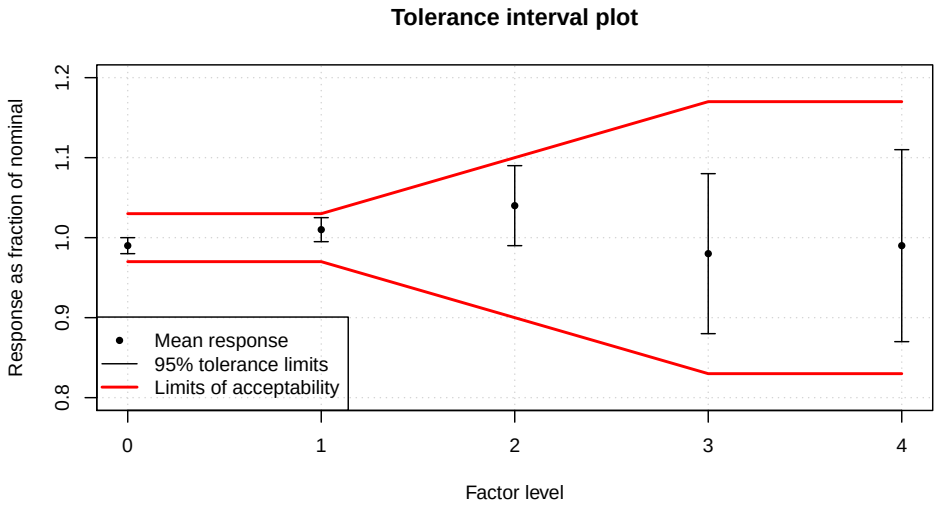

Figure 11: Tolerance interval plot.

Table 4: Evaluation of the line chart.

Strengths	Shortcomings
Allows an entire test to be viewed simultaneously, providing a high-level summary of results.	May provide irrelevant information - it's often the case that only the peak or average values are of interest
Is easily relatable to physical observations during a test.	Directs focus to extremes of group ranges, rather than the distribution of each group's performance.
Its meaning can be understood without explanation - it is a universally familiar chart.	Obfuscates data artifacts that aren't related to location or dispersion (such as harmonic content).
	Can obscure the behaviour of individual units.

2.5 Review

The report so far has presented and critiqued the statistical methods currently being used by DCA's engineers. ...CONTINUE

3 Suggested Methods

Having assessed DCA currently uses statistics, it's possible to suggest methods from modern statistical practice that would give them new capabilties and improve their existing attempts. This section again is structured according to the steps in an experimental process.

3.1 Experiment Design

Experiments design is motivated by trying to find efficient ways to explain variation in performance. To understand what factors determine a product's performance, this variation must be split into components attributable to product factors. The factors can be tuned to produce a desirable response.

Experimental investigations should be planned to methodically work out what the effects of factors are on the response, with the available resources. Response surface methodologies would allow DCA's engineers to do this in a way that is resource-efficient, and is in some measure plannable.

3.1.1 Response Surface Methodologies

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

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

Response surface methodologies try to achieve these objectives by sequentially identifying the factors that affect the response parameter. The general idea is to:

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

$$f(x_1, ..., x_n) = x_0 + \sum_{i=1}^{n} \beta_i \cdot x_i \tag{8}$$

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

$$f(x_1, ..., x_n) = x_0 + \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} \cdot x_i \cdot x_j \tag{9}$$

4. Optimize the response according to these second-order effects (described in the Appendix).

Something missing from the above steps is that a shot of common sense is needed at each stage - factors that have a negligible effect on the response

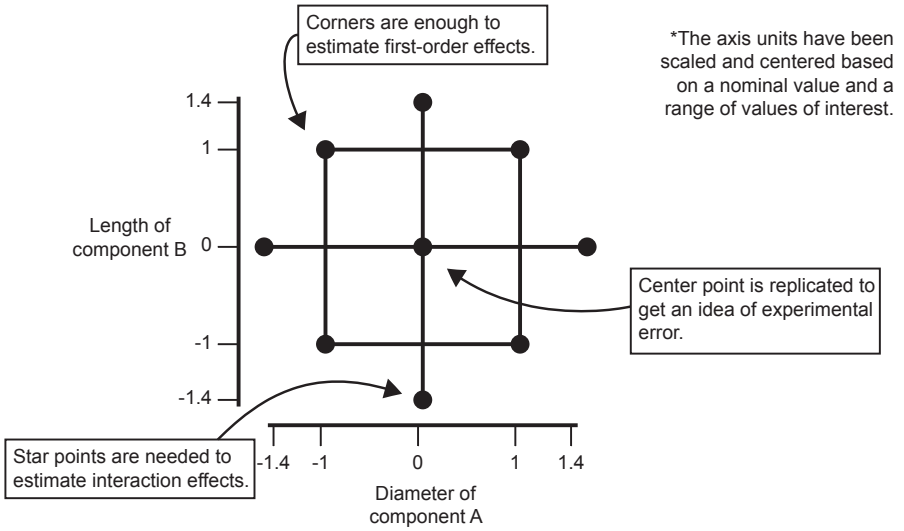

Figure 12: A central composite design - the points indicate factor combinations to be tested, and the lines suggest the experiment's intentions.

need to be discarded to avoid having to test a staggering number of second-order treatments to test. The structure of an RSM assumes that first-order effects are likely to be larger than interactions.

To estimate first-order effects, two factor levels need to be defined, then every combination of factor levels - 2^k in total - need to be tested. To estimate error of course, replicates of each combination of factors need to be tested. To make these estimates more efficient, replicates at only a few points can be run. If these points are centered relative to the other factors, then they can suggest whether the changes seen moving towards the surrounding points are probably attributable to error or an effect. Figure **??** shows what a center point is.

Central composite designs can be used to fit the second-order model: these consists of a testing all combinations of factors, replicating at the center points described above, then displacing one factor at a time from each of these center points. This type of experiment provides just the right amount of information to characterise the reponse surface around the center points (see the Appendix's discussion of orthogonality).

DCA could use response surface methodologies to converge more quickly on a solution than their current best-guess approach. Additionally, the solutions found using this method would be approximately optimal, rather than sufficient - this means better products, and more satisfied clients. It can be difficult to get hold of the resources to run a large experiment, and to convince a client that it will be worthwhile, but perserverance will repay itself in the risk, possible embarassment, and possible loss of business avoided as a result. Furthermore, investigations relying on - admittedly expert - guessing will consume more resources than planned investigations. The experiment designs presented here provide an idea of the minimum up-front cost needed to understand the effects of a set of factors on a response. If this cost is too high, and the factors can't be reasonably discounted, then the objectives of the investigation should be refined.

3.2 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 reponse? How do several factors interact?
- Which design variation is better?

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

Regression estimates how a continuous response changes with some inputs. Linear regression uses a linear function to predict the reponse: This may sound limiting, since in real life lots of 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. An example of such a model would be:

$$\hat{y} = f(x_1, x_2) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot e^{x_1} + \hat{\beta}_3 \cdot x_1 \cdot x_2 \tag{10}$$

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 predictors x_j 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. To do this, it's necessary to measure how inaccurate $\hat{y}$ is relative to y_i. One way of measuring fit inaccuracy is the residual sum of squares:

$$\text{RSS}(\hat{\beta}) = \sum_{i=1}^{m} \Big(y_i - \sum_{j=1}^{p} \hat{\beta}_i x_{ij} \Big)^2 \tag{11}$$

Where y_i is the response of the ith example, and x_{ij} is the jth predictor value of the ith example. This measure of fit is a good one for several reasons, the

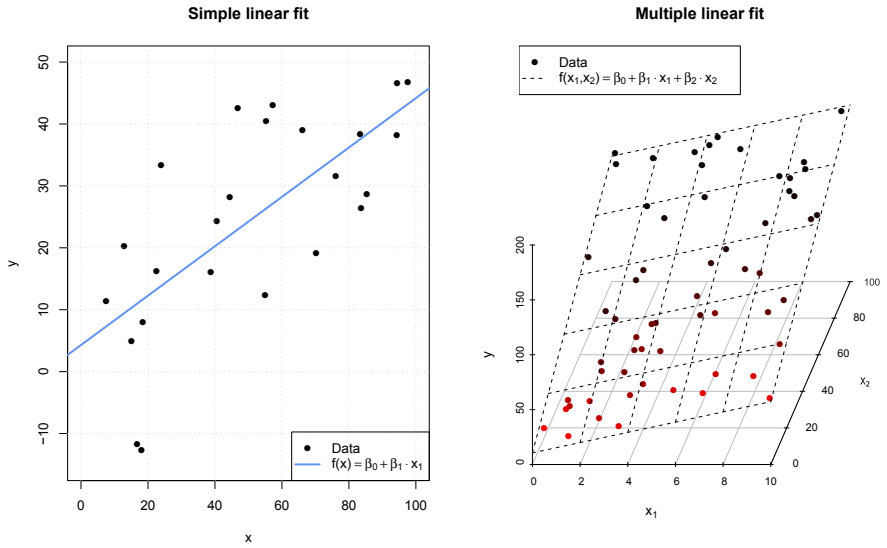

Figure 13: A simple and a multiple linear fit.

most intuitive of which is that it minimizes the overall distance between the estimates $\sum_{j=1}^{p} \hat{\beta}_j x_{ij}$ and the responses y_i [2] . There are two alternative justifications for least squares, which are both presented in the Appendix.

Figure 13.

From a practical standpoint, regression models should be fit using software. Matlab, R, or Octave can all be used to fit regression models. 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's possible to succinctly write the RSS criterion, then minimize it:

$$\text{RSS}(\hat{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) \tag{12}$$

$$\frac{\partial \text{RSS}}{\partial \hat{\beta}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0 \tag{13}$$

$$\implies \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{14}$$

As a case study to show why linear models are useful, consider a test in which the load delivered by three groups of ten units is measured. The

[2] The residual refers to the difference between an estimated response and a true response at a particular set of inputs. Error refers specficially to deviation of a response around its expected value.

Table 5: 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

differences between the groups are categorical: the groups correspond to three design variations A, B, and C. Since there isn't a natural order to the variations, it's necessary to encode this difference in a sensible way. There are several ways to do this, and a simple indicator stating whether a unit has a particular modification is sufficient. Table 5 shows this encoding. As will hopefully become apparent, choosing how to encode categorical variables matters a good deal with regards to how the regression coefficients beta should be interpreted.

Next the model is set up:

$$\hat{y} = \hat{\beta}_0 x_A + \hat{\beta}_1 x_B + \hat{\beta}_2 x_C = x\hat{\beta} \tag{15}$$

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{30} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 5.43 \\ 7.48 \\ \vdots \\ 6.47 \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \tag{16}$$

We can then use the normal equation (Equation 14) to make the least-squares fit. As it happens, in this case $\hat{\beta}_A$, $\hat{\beta}_B$, and $\hat{\beta}_C$ are the average responses of each group, as is shown in Figure 14.

Something to be quite careful of is redundantly encoding the groups. If a constant term $\hat{\beta}_4$ were to be included in the model above, then there would be many equivalent ways to express the group effects: $\hat{\beta}_4$ 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 14 will

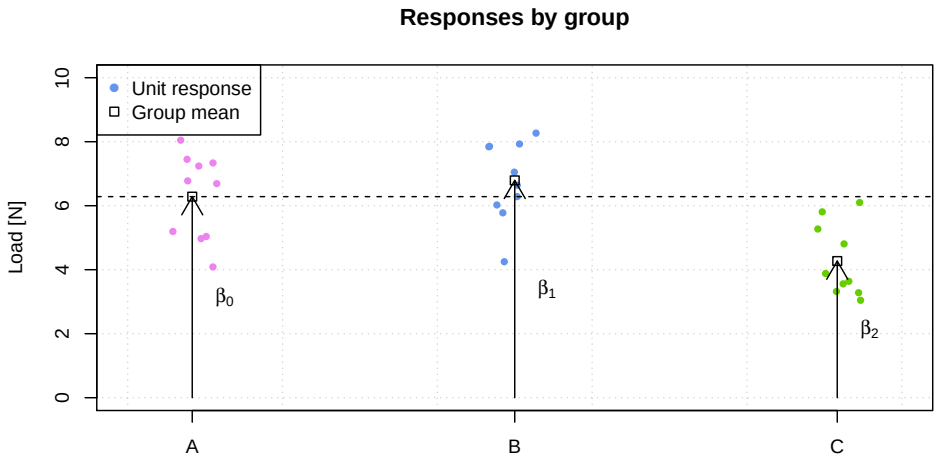

Figure 14: Unit responses by group.

be unsuccessful or unstable[3].

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:

$$y = X\beta + \varepsilon \tag{17}$$

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 aren't going to be perfect, and their standard errors can be calculated to get a feel for 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. 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 isn't relevant to the investigation because this will make it easier to see the effects of the inputs on the reponse for a given sample size. Calculation of standard errors is described and explained in the Appendix.

A reasonable criticism of the above example is that it effectively just a drawn-out calculation of the group sample means. This is true, until the

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

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 adjustedl to estimate the effects of the lubricant on the load output of each mechanism:

$$\hat{y} = \hat{\beta}_0 x_A + \hat{\beta}_1 x_B + \hat{\beta}_2 x_C + \hat{\beta}_3 x_A x_l + \hat{\beta}_4 x_B x_l + \hat{\beta}_5 x_C x_l \tag{18}$$

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. Point being, linear models are powerful because they can be adapted to many situations, and because they provide a consistent way of disentangling the effects of experimental treatments.

As an aside, assessing the significance of differences between groups is frequently called Analysis of Variance (ANOVA). It's 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 induces an effect or not. It's 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, certain websites in particular, can become unwieldy and confusing, which could make its 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's suggested that DCA focus on learning to use linear models rather than attempt to use ANOVA.

Linear models would be useful to DCA because they can be used to figure out where variation in a response is coming from, which is the fundamental objective of most experiments. Identifying the factors that have the largest

effect on performance, and quantifying those effects, is the first step in understanding how to change a design to make it both better-performing and more robust. Used in combination with multiple-factor experiment designs, 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[4] is, based on the data alone. They don't say how much more likely this value is is than other values, or let knowledge besides the data be included. In reality, a sample will suggest a distribution of plausible values, and there will be expert knowledge that can be used since it will be known roughly what values are realistic.. Bayesian inference combines readily available knowledge with the experimental data to estimate a distribution of possible parameter values. Figure 15 points out the benefits of making distributional rather than point estimates.

Bayesian methods would be useful to DCA because they're more easily understood, visualized, and explained than classical methods[5], and are relevant to a broader range of situations. They also allow expert knowledge to be used, making it possible to reach a sanitary comprimise between gut-feel and experimental observation. Finally, their emphasis on distribution rather than point estimates better reflects the underlying

[4]Reminder: a parameter is a number that controls the shape of a distribution.
[5]"Classical methods" here refers to tools such as hypothesis tests and interval estimates

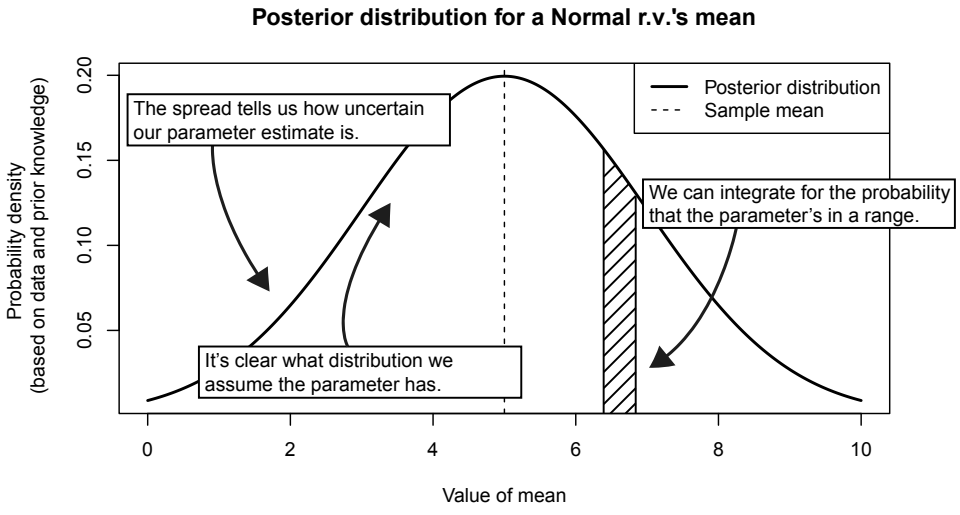

Figure 15: An annotated posterior distribution.

uncertainty in analytical results. These claims will be explained in the context of an example.

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

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.

$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{\int_\theta p(y|\theta) \cdot p(\theta) \cdot d\theta} \tag{19}$$

In words, this means that a passing proportion is more probable if it makes the the number of units that really did pass in the sample more likely and seems sensible to the engineering team. The denominator of this expression is constant w.r.t. θ, making it possible to express the above as:

$$p(\theta|y) \propto p(y|\theta) \cdot p(\theta) \tag{20}$$

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

(20) 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 engineerig 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

given that θ of that population would pass is:

$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \tag{21}$$

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:

$$p(\theta) = 1 \qquad \theta \in [0, 1] \tag{22}$$

Figure 16 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:

$$p(\theta|y) = \texttt{constant} \cdot p(y|\theta) \cdot p(\theta) \tag{23}$$

In practice, (23) 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 (22) and (21). 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 16. The

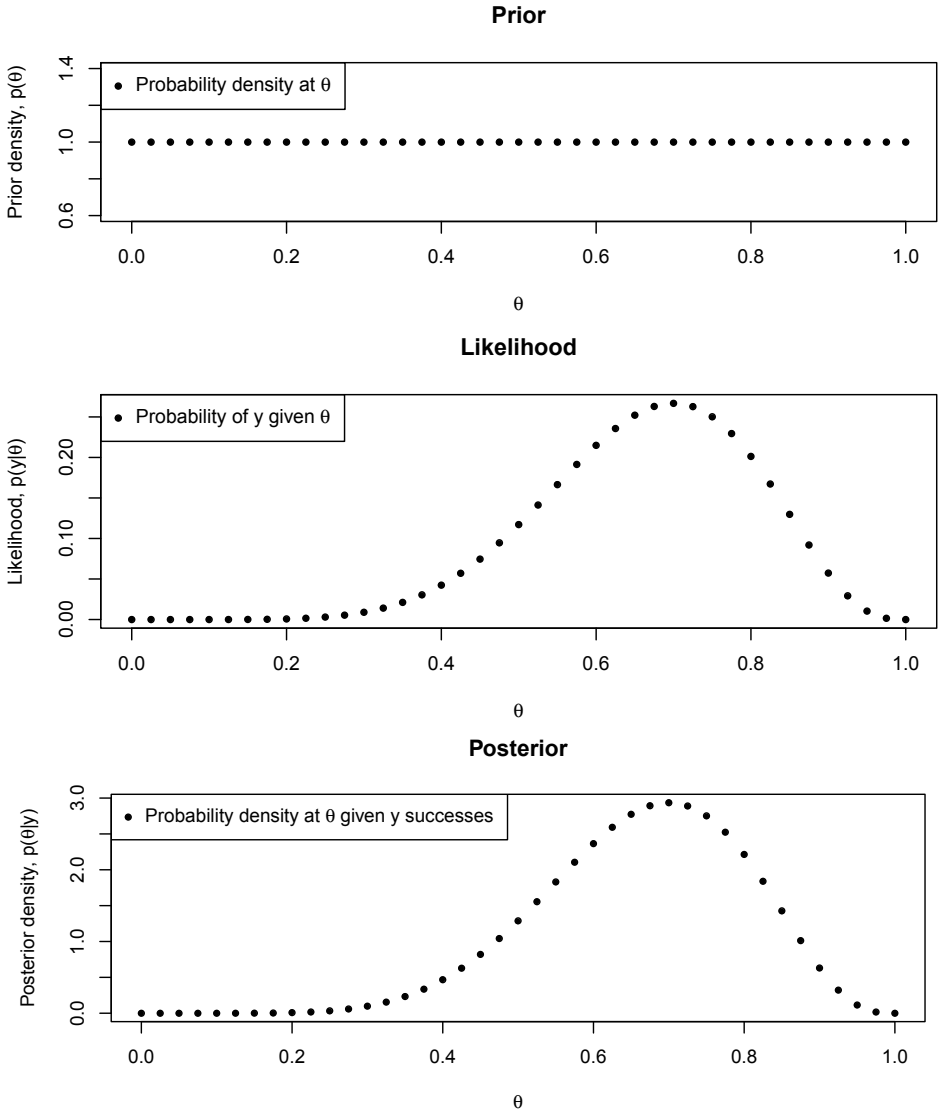

Figure 16: Combining an uninformative prior with a likelihood distribution to form a posterior distribution of plausible parameter values.

pointy part indicates more probable values - a taller, pointier peak represents a more certain estimate because a few values have a much higher probability than lots of others. In the same spirit, the flat prior that was used can be understood as highly uncertain.

To recap what's been do so far, the prior knowledge of the engineering team was combined with experimental observations to come up with a likely range of values for the true passing proportion for a population. Taking the product of the prior probability of parameter values with the likelihood of

the data produced a posterior distribution indicating the probability of values based on both osurces of information.

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:

$$p(\tilde{y}|y) = \int_{\theta} p(\tilde{y}|\theta, y) \cdot p(\theta|y) \cdot d\theta \tag{24}$$

Once again, we can avoid some potentially mischievous mathematics 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 relative frequency of $\tilde{y}$ values will tend towards $p(\tilde{y}|y)$.

```
for (i in [1, 10 000]) {                                          (25)
```

$$\tilde{n} := \texttt{No. future units to be tested} \tag{26}$$

$$\tilde{\boldsymbol{y}} := (0, 1, \ldots, \tilde{n}) \tag{27}$$

$$\theta := \texttt{Sample}(\boldsymbol{\theta}, \texttt{ Posterior}) \tag{28}$$

$$\texttt{Posterior predictive[i]} := \texttt{Sample}(\tilde{\boldsymbol{y}}, \texttt{ Binomial}(\tilde{y}, \tilde{n}, \theta)) \tag{29}$$

```
                                  }                               (30)
```

Figure 17 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 **??**, where the flat prior represents initial ignorance about whether a unit will pass: as more units are

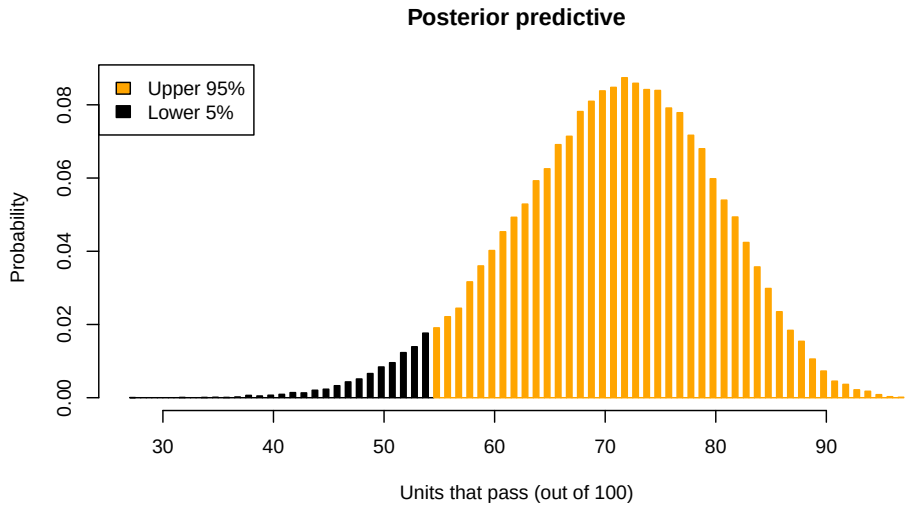

Figure 17: Posterior predictive distribution for the number of passing units in a 100-unit run.

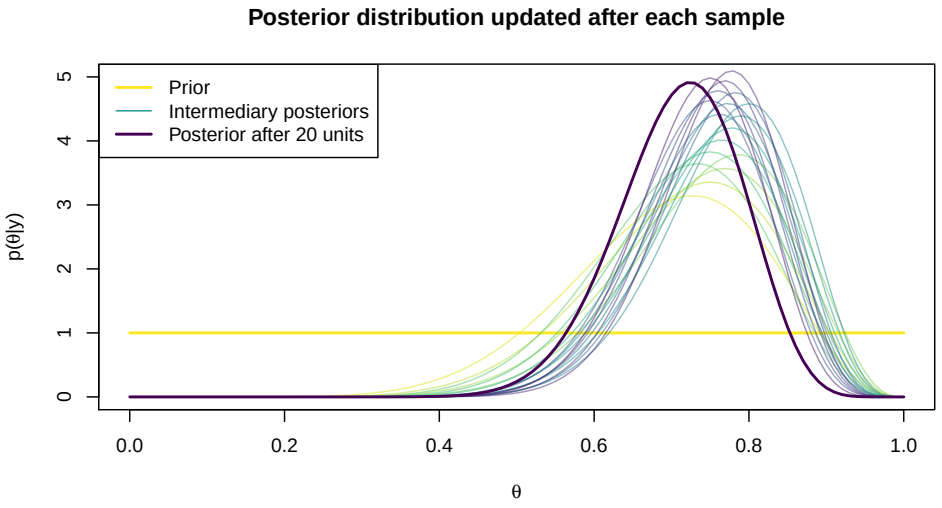

Figure 18: The uncertainty in the posterior decreases as more samples are conditioned upon.

run, an increasingly narrow peak forms around the most probable passing probability.

Another advantage of Bayesian methods over hypothesis testing is that it's relatively easy to build models that estimate many parameters simultaneously. An instance of this might be when estimating the mean and

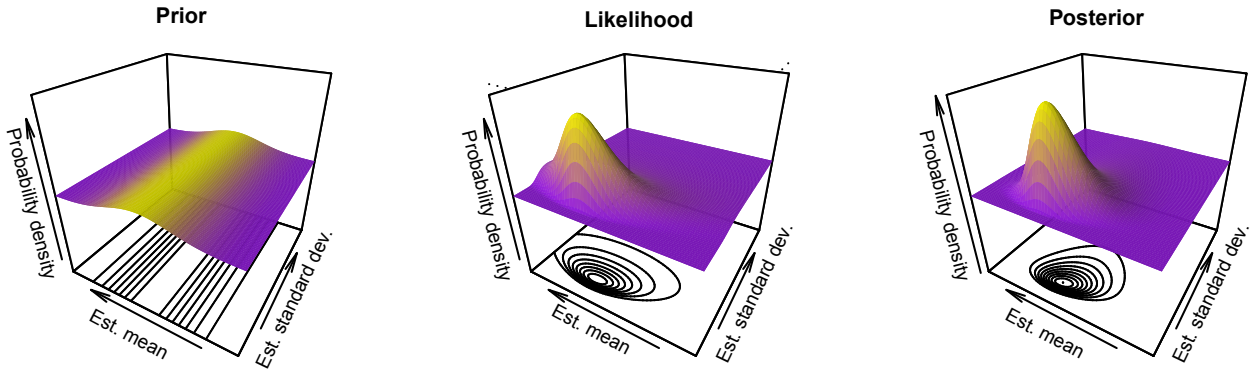

Figure 19: Multiparameter Bayesian inference.

the variance of data that's assumed to have a normal distribution. The only change relative to the single-parameter scenario is that we need to define the prior and likelihood in Equation 20 over two parameters instead of one:

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2) \cdot p(\theta_1, \theta_2) \tag{31}$$

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 19. This distribution shows how the data reduced uncertainty about what values are reasonable for the mean and standard deviation. This distribution would probably be easier to explain to clients and colleagues than confidence intervals or hypothesis tests.

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 usually presented simply as numerical values. The jargon of classical statistics can make it unclear what's relevant to the problem at hand, 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 more easily interpreted.

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 couple of 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.

It is generally recognized that the scientific community as a whole needs to reconsider the practical relevance of hypothesis testing: DCA can hardly be faulted for neglecting to apply ineffective methods. Tolerance and confidence intervals are useful, but can be prone to misinterpretation.

Markov Chain Monte Carlo

As mentioned, DCA have previously used Monte Carlo simulation to approximate the distribution of a tolerance chain's dimension. For problems with many parameters - such as a subassembly of many components - Monte Carlo simulation isn't feasible because it would require too many values to be stored in a computer's memory. Instead, the posterior can be approximated using a Markov chain Monte Carlo method, which provides a method for estimating the posterior even if the domain of parameter combinations is extremely large. The details won't be presented here, but in future if DCA should choose to construct models containing many parameters they will probably need to use it. The basic idea is to generate a sequence of values according to their relative probabilities, such that the relative frequency of values in the chain converges towards the posterior distribution.

3.3 Presentation & Visualization

3.3.1 Why Visualization is Important

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 isn't stored in a way that can be interpreted easily, so it needs to be transformed for it to be useful

- maths is one means of transforming it to reveal patterns, and plots are another.

DCA's use of visualization could be improved by:

- Presenting summary plots rather than raw data where appropriate.
- Providing only as much information as is needed to understand the results - is the entire group's results necessary, or just one unit's?
- Using plots to understand data as well as present it - in the same way that fiddling with prototypes inspires ideas, visual representations of data can suggest things that aren't clear in formal analyses.
- Faceting line plots - overlaid line charts are crude representations of data that will generally only show obvious differences. This is mentioned because of DCA's reliance on line plots.

Designing a visualization means giving thought to:

- Data - are the variables continuous, categorical, or ordered categorical?
- Geometry of datapoints - points, lines, or bars
- Datapoint aesthetics - color, size, shape, and position of the geometries
- Scales - mappings from data to aesthetics
- Summary statistics - plottable summaries of the data
- Facets - using a grid to separate plots
- Themes - typeface, non-data colours, axes and so on.

This many elements may seem overwhelming, and has a justification, although the most important ones are choice of geometry, aesthetics, and scales. Aesthetics can represent dimensions that aren't displayed via a plot's axes. Geometry carries information such as resolution, chronology, and connectivity. Scales can skew perception of effects, and should be chosen carefully.

[1], [2], [3], [**?**], [4], [5], [6], [7], [8].

3.4 Software

- Excel
- Matlab
- R
- Python
- Minitab

4 Conclusions & Recommendations

Bibliography

[1] 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.

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

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

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

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

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

[7] J. Kruschke, *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, Academic Press, 2015.

[8] I. T. C. . A. of Statistical Methods, "Iso 16269-6:2014 statistical interpretation of data - part 6: Determination of statistical tolerance intervals," 2014.

A

A.0.1 What is probability theory, and 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 aren't 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 Standard Error of a Sample Mean

We can see this by recognizing that the sum of independent observation's variances is equal to the variance of the variance of the sum of the observations:

$$\mathrm{Var}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \mathrm{Var}X_i \tag{32}$$

$$\mathrm{Var}n\bar{X} = n\mathrm{Var}X \tag{33}$$

$$\implies \sqrt{\mathrm{Var}\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{34}$$

Analytically, we can solve for $p(\theta|y)$ directly

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

To find the constant of proportionality, we need to divide the r.h.s. by its integral over all values of θ, such that the $\int_0^1 p(\theta|y) \cdot d\theta = 1$. As it happens, the r.h.s. has the form of what's called a beta distribution

$$p(x; a, b) \propto x^{a-1} \cdot x^{b-1}$$

A.0.3 Confidence intervals

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

A.0.4 Justifications for Least-Squares Regression

Assume that the errors in y given x are normal with constant variance and mean zero

(35)

$$\epsilon \sim \mathrm{N}(0, \sigma^2) \tag{36}$$

$$p(\epsilon|\hat{\beta}) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{\epsilon^2}{2\sigma^2}) \tag{37}$$

Assume that the true response is a linear function in $\hat{\beta}$, such that $E(Y|X) = X\hat{\beta}$, then take the log of the likelihood:

$$l(\hat{\beta}) \propto -(Y - X\hat{\beta})^2 \tag{38}$$

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

The final way to justify a least squares fit does not require a normality assumption: instead, it is rationalized as providing 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.

Start by assuming that y truly is a linear function of X, which is a vector of inputs, so that:

$$E[y|X] = X\hat{\beta} \tag{39}$$

Next note that it follows from the above that the least squares fit will be an unbiased estimate, provided that $\boldsymbol{y}$ is drawn i.i.d. at a given X:

$$E[X\hat{\hat{\beta}}] = E[\boldsymbol{X(X^TX)^{-1}X^Ty}] \tag{40}$$

$$= \boldsymbol{X(X^TX)^{-1}X^TX}\hat{\beta} \tag{41}$$

$$\implies E[X\hat{\hat{\beta}}] = \boldsymbol{X}\hat{\beta} \tag{42}$$

Almost there. Now we need to think about other possible estimates of $E[y|X]$. We can see from (16) that our estimate is a linear function of y (Estimable functions?). Imagine another function $\boldsymbol{c^Ty}$ that's also linear in $\boldsymbol{y}$. The Gauss-Markov theorem states that the variance of the least-squared estimate is guaranteed to be less than this other linear estimate:

$$\text{Var}(\boldsymbol{X}\hat{\hat{\beta}}) \leq \text{Var}(\boldsymbol{c^Ty}) \tag{43}$$

The proof for this is:

$$E[(a^T\hat{\hat{\beta}} - a^T\hat{\beta})^2] \leq E[((c^Ty - a^T\hat{\hat{\beta}}) - (a^T\hat{\beta} - a^T\hat{\hat{\beta}}))^2] \tag{44}$$

$$\leq E[((c^Ty - a^T\hat{\hat{\beta}}) + (a^T\hat{\hat{\beta}} - a^T\hat{\beta}))^2] \tag{45}$$

$$\leq E[(c^Ty - a^T\hat{\hat{\beta}})^2] + E[(a^T\hat{\beta} - a^T\hat{\hat{\beta}})^2] \tag{46}$$

$$\iff 0 \leq E[(c^Ty - a^T\hat{\hat{\beta}})^2] \tag{47}$$

Is this is a proof? How can you be sure?