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 4, 2017

Table of Contents

List of Tables

List of Figures

Notation & Glossary

Attribute	A measurable property of a *unit*.
Block	A set of *units* thought to share some common *attribute* that influences their *response*.
Event	A set of *outcomes*.
Experiment[1]	The controlled collection of data.
Experiment[2]	Physically realizing an outcome of the system under study.
Factors	*Treatments* that are discrete. For example, lubricated/unlubricated.
Outcome	A possible result of a *trial*.
Probability	A method for quantifying uncertainty, or a value representing the uncertainty of an event.
Response	The measured performance of a *unit*.
Treatment	A modification applied to a *unit*.
Unit	A single test specimen - in the context of product testing, this is likely to be a prototype build of the product.

Acknowledgements

~~Beyoncé, J.D.Sallinger, and Santa Claus.~~
~~This report would not have been possible without the encouragement of Paul Harper and technical supervision of Sophie Sladen. More broadly, I am grateful to DCA Design International for providing me a place to work on these ideas and develop professionally. Thank you to DCA's engineers - especially Will Marsh, Matthew Jones, and Matthew Edwards - for setting the bar so high and for helping me to improve as an engineer.~~

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. Their work is oriented towards the mechanical design of medical and consumer products. Much of what they develop are hand-held items such as insulin injector pens or deoderant cans: Figure 1 shows two of their most prolific designs. DCA's competitors are ~~[DCA's COMPETITORS AND THEIR CAPABILITIES].~~

DCA employs about sixty mechanical engineers. Each of them are general-purpose technical consultants ~~and experts in a particular engineering subdiscipline.~~ 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 ~~[REFERENCE].~~ This investment is manifest in both the its engineering workforce and its ownership of four test labs.

The way in which DCA's data is collected, analyzed, and presented is the focus of this report: data-oriented activities constitute the scientific discipline of statistics. Statistical methods allow resources and information

Figure 1: Products designed by DCA.

to be used efficiently, ~~in both a mathematical sense and a practical one.~~

Section X1 explains the company's current investigatory framework and how statistics is currently applied within it. In Section X2 the company's approach is compared to modern statistical ~~methods,~~ in the process of which ~~these~~ alternative methods are detailed and evaluated. Tools ~~(i.e. software and tangibles)~~ 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.

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

It's convenient to split the statistical methods that DCA apply into two factions: those brought to bear in lab investigations, and those used in other engineering activities~~, in particular tolerance analysis and~~ .

A lab investigation in DCA consists of a series of experiments to understand the behaviour of a product or process. It begins with the required knowledge being identified. Experiments ~~will~~ 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 relevant project.~~ Typically an investigation focuses 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 ~~conducted~~ by engineers on medical projects. Occasionally

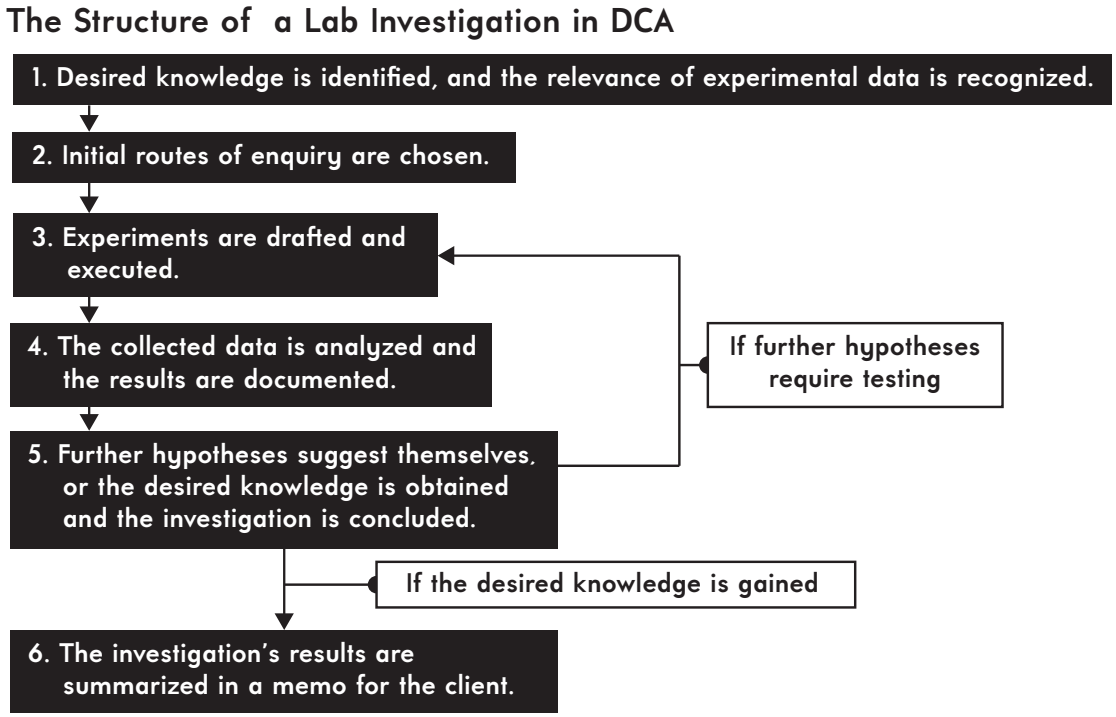

Figure 2: Investigation diagram.

engineers working on fast-moving consumer goods (such as toothbrushes or lotion bottles) will run ~~one off~~ 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 investigations to them: medical products on the other hand, see a good deal of the test lab.

As can be seen in Figure **??**, 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.

~~DCA's engineers have access to~~ 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 ~~ancillary~~ experiments are often designed to provide supplementary information. With this in mind, ~~it is worth noting that~~ this report attempts to be data-agnostic in its recommendations of analytical techniques.

The other engineering activities that DCA's engineers apply statistics to are tolerance analysis and, increasingly, predictive user interfaces. ~~Before talking about this work however, DCA's use of statistics in its lab investigations will first be summarized and critiqued.~~

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 ~~the experiment's~~ objective.

Robust experiments are designed with three principles in mind:

Replication Testing a particular ~~combination of factor levels with~~ more than one unit. ~~It~~ allows experimental error to be estimated and, since unbiased errors cancel on being averaged, ~~gives us~~ a more precise estimate of a ~~particular factor's influence.~~

Randomization Randomly ~~determining the allocation~~ of treatments to units and the sequence in which units ~~are tested averages out the effects of~~ nuisance variables, and validifies ~~the~~ assumption that obser~~sv~~ations are randomly drawn from a distribution.

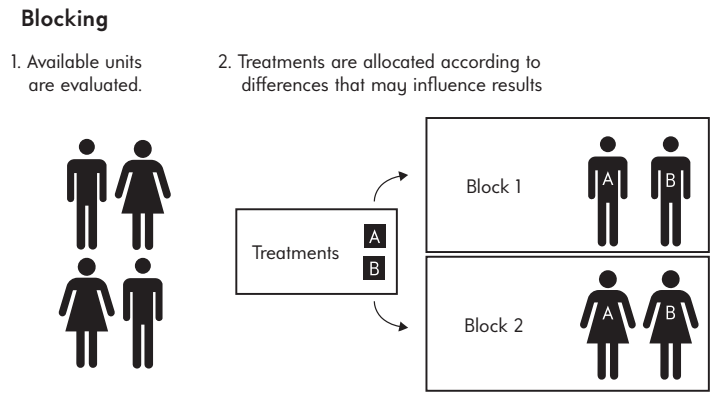

Figure 3: Diagram of what blocking involves.

Blocking Blocking accounts for unit differences when assigning treatments ~~- see Figure 3.~~ It allows the effects of a nuisance factor to be averaged out during analysis. A block is a set of similar units.

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 fret about their sample sizes.

This being said, DCA ~~neglect pre-experimental planning~~ and do not verify that their experimental set-ups produce repeatable results. Further, it has an inconsistent approach to screening for important factors, ~~resulting in mired investigations.~~ The dominant experimental strategy within the company is a best-guess approach: one factor is tested in each experiment, chosen based on the expert insight of the engineering team. Shortcomings of this method are that if the factor does not elicit the desired effect, then the next factor to vary must be guessed at, and if it is successful, then it may be tempting to stop the investigation when a better solution may be available.

To systematically detail DCA's experimental procedure, it 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.

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.

2.3 Analysis of Experimental Data

~~This section provides a short overview of the statistical tools applied to experimental data in DCA:~~ the content of several hundred test reports was tabulated to inform this discussion, which focuses on summary statistics and interval estimates.

Analyses in DCA ~~were found to~~ 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.

~~Many of~~ 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.

Summary Statistics

A summary statistic is a value describes an aspect of a random variable's distribution. 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, ~~its contribution per unit length to the cumulative probability.~~ Figure 5 highlights this difference.

The essential problem of experimental statistics is ~~understanding~~ the behaviour of a broader population from just a sample. In product design, this means using measurements from a limited number of prototypes to estimate the variation in ~~a much larger population of~~ units. The attributes of this variation - such as its spread and average - can be estimated using summary statistics.

~~A mean of several independent measurements, for example, approximates the mean of the underlying population's distribution.~~ The accuracy of this estimate improves as more samples are tested, with diminishing returns, a relationship that is shown in Figure 4. This estimate'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, ~~although it did not seem to be used actively to gauge appropriate sample sizes.~~ Certain summary statistics can be thought of as

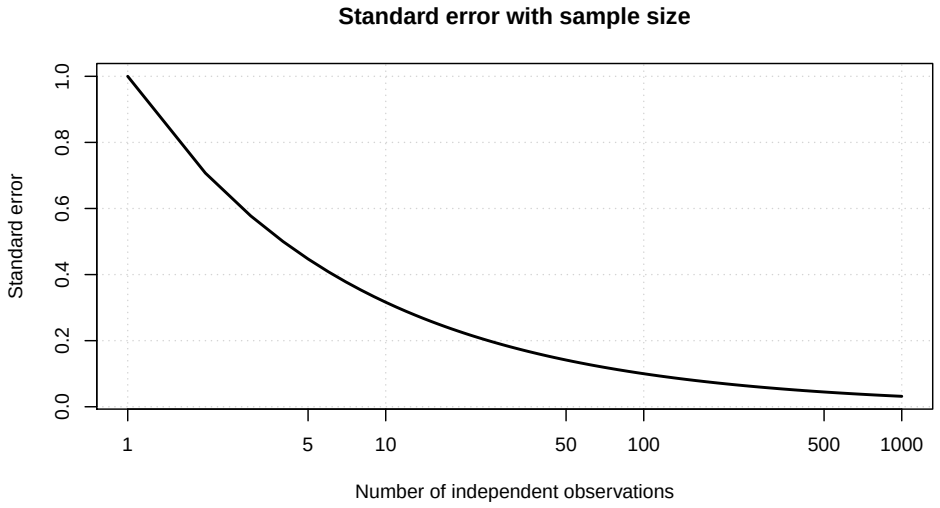

Figure 4: Convergence of sample mean to population mean.

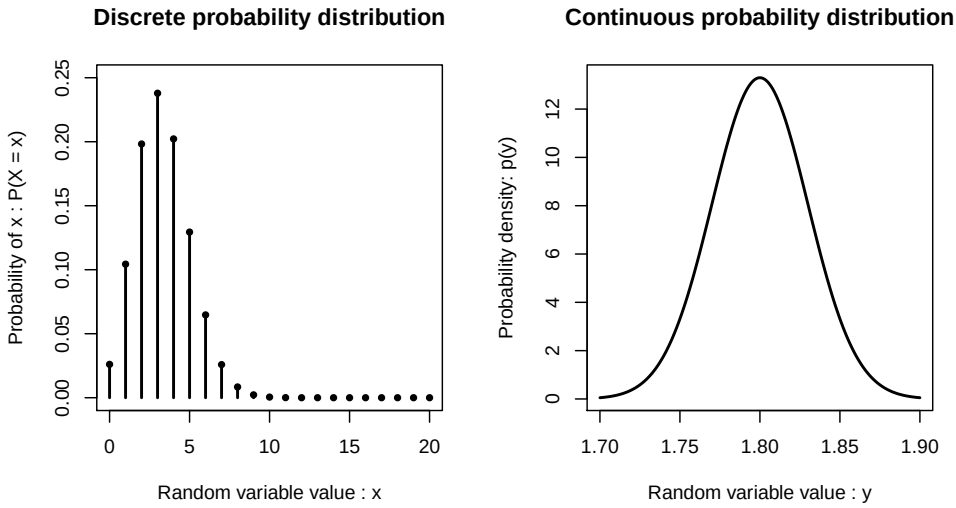

Figure 5: Left: Probability mass function. Right: Probability density function.

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's parameters will be seen again later, in the section on Bayesian inference.

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 to a given confidence level; a confidence level is the proportion of intervals constructed that will contain that fraction of the population. So tolerance intervals allow us to make statements about the performance of a population, with clear limits on that statement's uncertainty. A derivation of a one-sided interval is given in Appendix A, and can be summarized as:

1. Defining k such that

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

   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. If x is assumed to have 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 $\frac{1}{\sqrt{n}}$:

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

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 6.

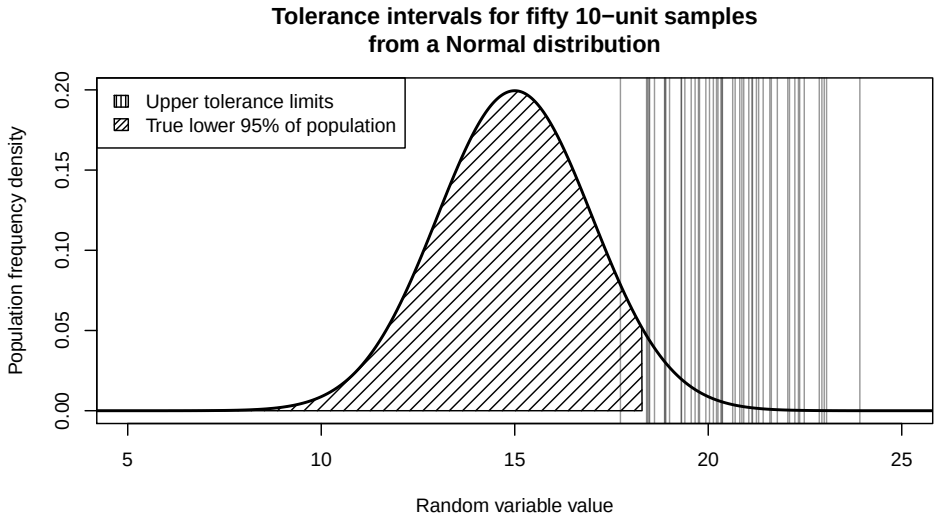

Figure 6: Twenty tolerance limits versus population limit.

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.	Can be misinterpreted as specifying the probability that a constructed interval contains at least 95% of the population.
Plug-and-play	Normality assumption needs checking
	Repeated use at a low confidence level increases the probability that the limit will be under-estimated.

Confidence Intervals

Another tool that DCA's engineers occasionally use is confidence intervals, which indicate a range of values that a population parameter is likely to fall within. 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 7 demonstrates this idea. Confidence intervals can be placed on any parameter estimate, although they're usually used to quantify the uncertainty on a mean. Their derivation is similar to a tolerance interval's: the two-sided one is as follows:

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

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

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

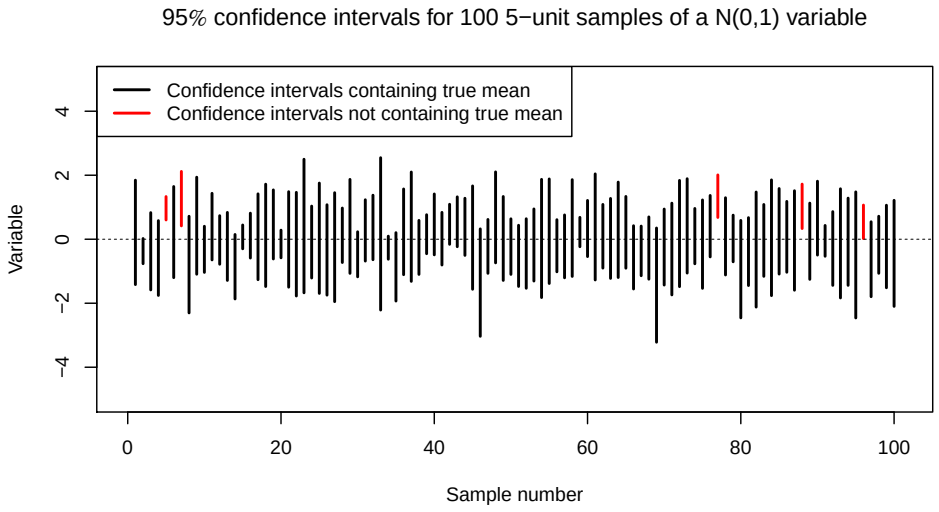

Figure 7

Monte Carlo Simulation

Monte Carlo simulation 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 95th percentile of some analytically inconvenient combination of distributions, each corresponding to a part dimension, was needed. Take $Y \sim \text{Binom}(n = 10, p = X)$ as an example, where $X \sim \text{Beta}(a = 7, b = 3)$[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 8.

[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.

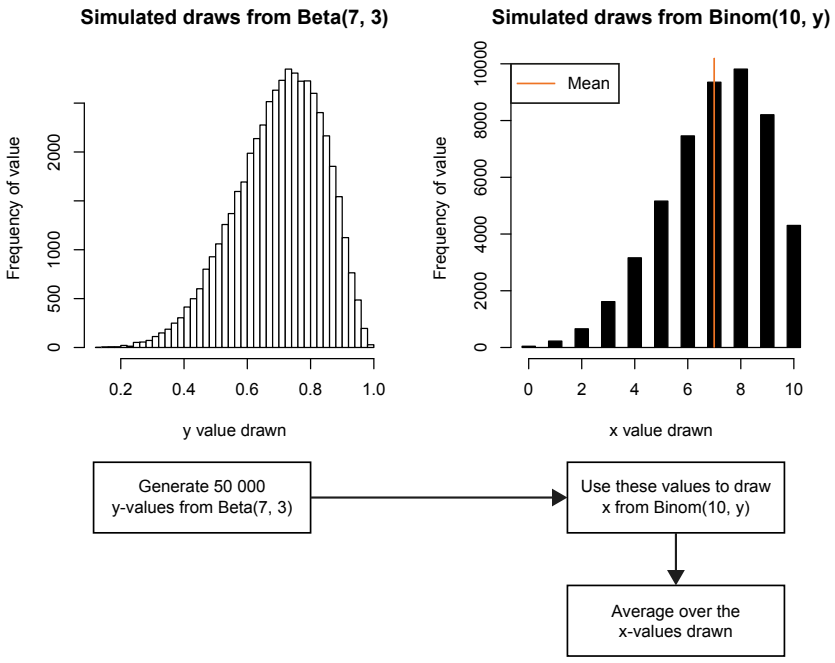

Figure 8: Process diagram for a Monte Carlo simulation.

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 9. 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.

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 10.

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.

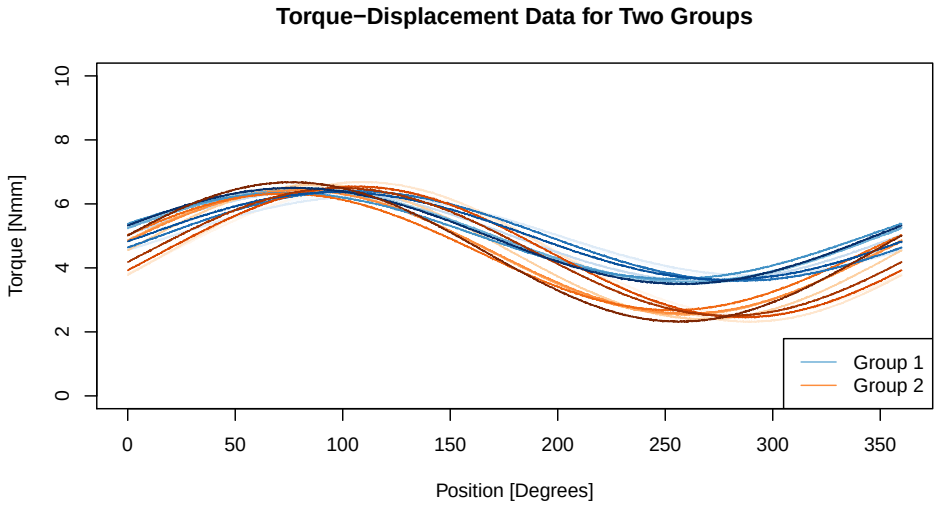

Figure 9: Sample summary line chart from a DCA test report.

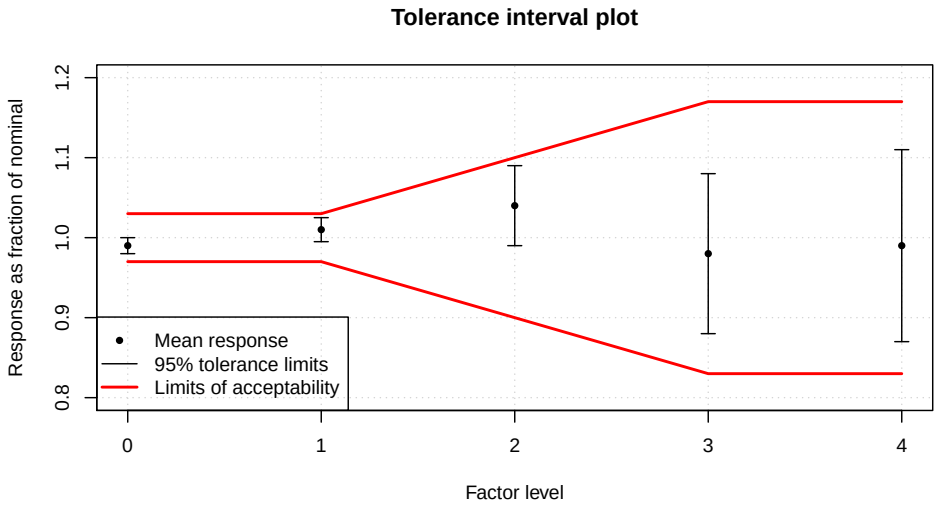

Figure 10: Tolerance interval plot.

3 Suggested Methods

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 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:

$$\hat{y} = f(x_1, x_2) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot e^{x_1} + \hat{\beta}_3 \cdot x_1 \cdot x_2 \tag{8}$$

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 can be a nonlinear function 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 is to use the residual sum of squares:

$$\text{RSS}(\hat{\beta}) = \sum_{i=1}^{m} \Big(y_i - \sum_{j=1}^{p} \hat{\beta}_i x_{ij})^2 \tag{9}$$

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

A common linear model is a simple linear fit, where a single reponse variable is regressed onto a measurement and a constant:

$$f(x_1) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 \tag{10}$$

This model describes a line: it can be extended to more than one variable, where it becomes known as multiple linear regression:

$$f(x_1, x_2) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 \tag{11}$$

The coefficients of a three-parameter linear model describe a plane; Linear regression of a dummy dataset onto one and two predictor variables is shown in Figure 11.

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

[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 random variation of the response around its expected value.

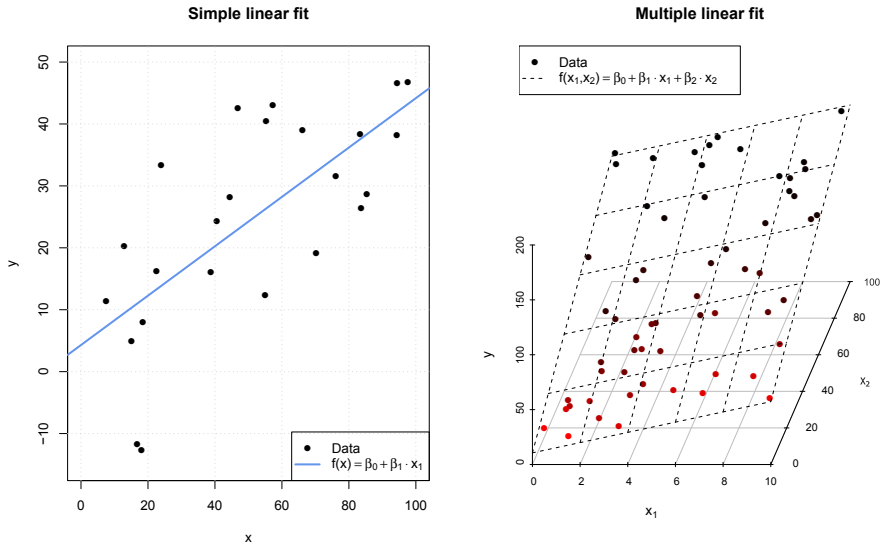

Figure 11: A simple and a multiple linear fit.

the RSS criterion, then minimize it analytically:

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

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

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

The uncertainty in the fit can be quantified, although this would probably be of limited relevance to DCA's work since its engineers are understandably interested in practical rather than statistical significance.

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 difference between the groups is 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 code the difference between the units in a sensible way. There are several ways to do this, although a simple indicator stating whether a unit has a particular modification is sufficient.

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

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 $\hat{\beta}_C$ would be the average responses of each group, as is shown in Figure 12. Something to be

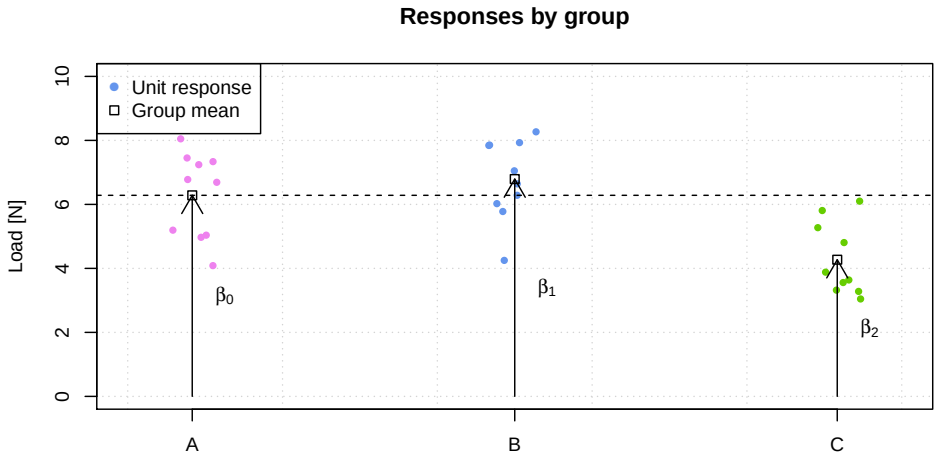

Figure 12: Unit responses by group.

quite careful of is redundant encoding of groups. For example, if we were to include a constant term $\hat{\beta}_4$ in the model above, then there would be many 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 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 this example, $\beta_1, \beta_2, \beta_3$ are the true group means, and $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ are going to be 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 coefficient over many samples, and the true coefficient. They 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 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 experiment is extended to involve another variable, this time a continuous one, say the volume of a lubricant applied to each design variant. Then we can adjust model 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. $\hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$ are the change in output force for each design variant per unit volume of lubricant. The results of this model would allow an engineer to see not only how good design variants are relative to one another, but also how much lubricant

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

would need to be applied to bring the performance of one in line with another.

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. 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.

Bayesian Inference

Summary statistics 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. Bayesian inference combines this knowledge with the collected data to estimate a distribution of possible parameter values. Figure 13 points out the benefits of determining how probable particular parameter values are, compared to the point estimates provided by summary statistics.

Bayesian methods would be useful to DCA because they're more easily understood, visualized, and explained than classical methods, 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. These claims will can be explained and justified in the context of an example.

The proportion of units passing a specification test is a useful measure of a

[4]Reminder: a parameter is a number that controls the shape of a distribution.

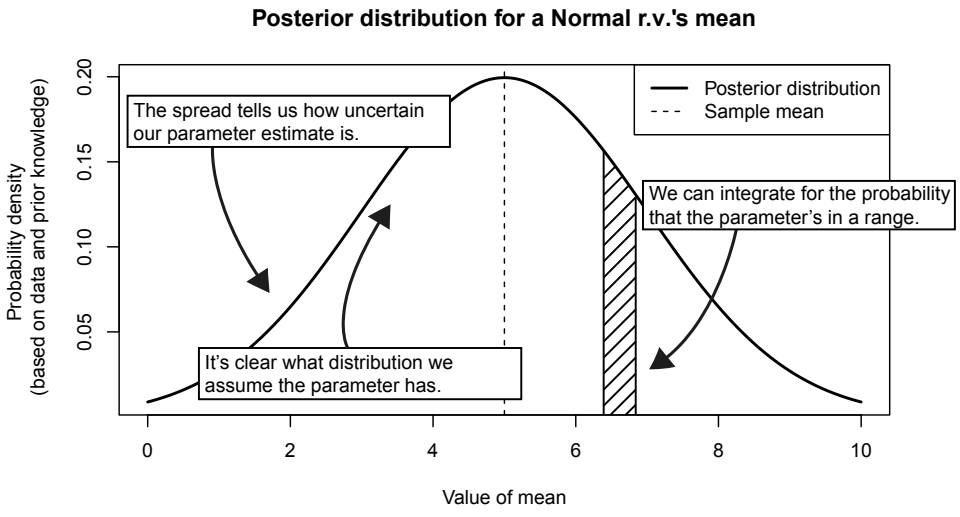

Figure 13: An annotated posterior distribution.

design's suitability for the problem at hand. Using the results from a test sample and the expertise of an engineering team, it's possible to estimate what pass rates would be likely if the design were to be produced in larger volumes by a similar method.

Say that a sample of n units are tested, and n pass. The engineering team also collude to produce a distribution for the passing rate θ that's tall near values they think probable and shallow near ones that seem unlikely. The team's aim is then to calculate the probability of a unit passing, given their data and preliminary estimates: Bayes' theorem can be used to do this.

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

Expression (19) means that a passing proportion is more probable if it makes the number of units seen to pass more likely and seems sensible to the engineering team. The denominator of this expression is constant w.r.t. θ, so that

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

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

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

- The probability of the data assuming a particular passing proportion, $p(y|\theta)$ (the *likelihood*).
- The probability of a passing proportion according to the engineerig team, $p(\theta)$ (the *prior*).

The likelihood is the probability of y units passing and $(n - y)$ units failing. Assuming that passes and failures and independent and the units come from populations with the same underlying passing probability, then the probability of y passes given that θ of that population would pass is:

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

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:

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

Figure 14 displays the prior and likelihood distributions. At this point it's possible to do one of two things: the posterior can either be evaluated analytically, or it can be computed. 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:

$$n := \texttt{No. of units tested} \tag{24}$$

$$y := \texttt{No. of units that passed} \tag{25}$$

$$\boldsymbol{\theta} := (0,\ 0.01,\ \ldots,\ 1) \tag{26}$$

$$\texttt{Prior} := \texttt{Uniform}(\boldsymbol{\theta}, [0, 1]) \tag{27}$$

$$\texttt{Likelihood} := \texttt{Binomial}(y, n, \boldsymbol{\theta}) \tag{28}$$

$$\texttt{Posterior} := \texttt{Prior} \odot \texttt{Likelihood} \tag{29}$$

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 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.

One of the neat things about Bayesian inference is that 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 idea is shown in Figure **??** - the flat

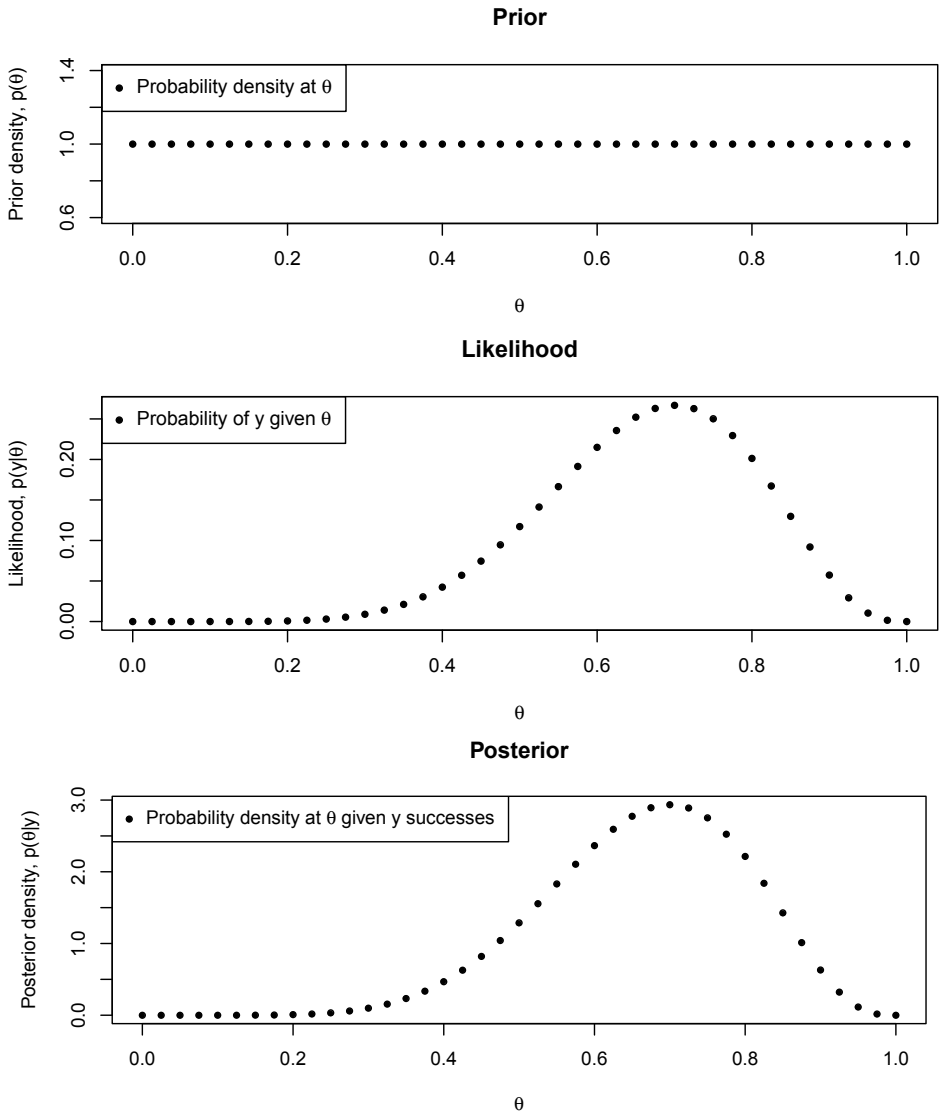

Figure 14

prior 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.

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:

$$p(\tilde{y}|y) = \int_{\theta} p(\tilde{y}|\theta, y) \cdot p(\theta|y) \cdot d\theta \tag{30}$$

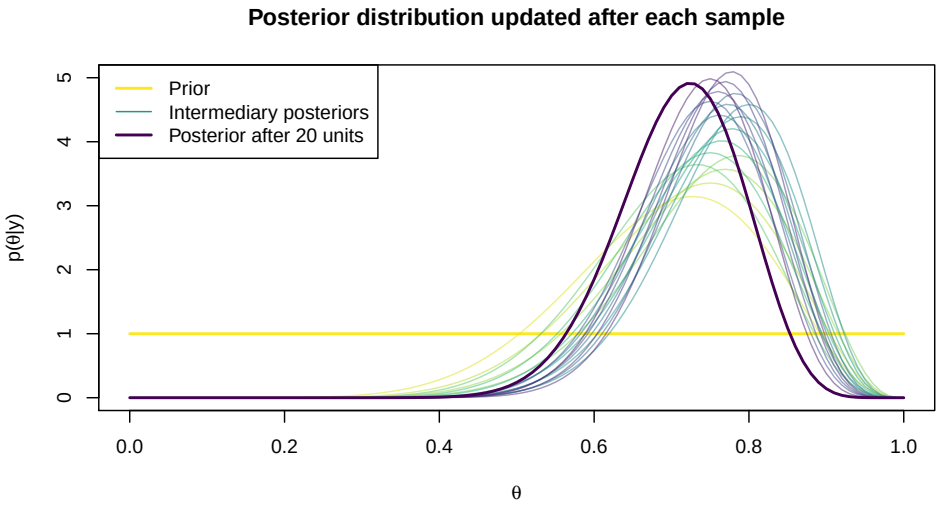

Figure 15: The uncertainty in the posterior decreases as more samples are conditioned upon.

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 $p(\tilde{y}|y)$.

$$\texttt{for (i in [1, 10 000]) \{} \tag{31}$$

$$\tilde{n} := \texttt{No. future units to be tested} \tag{32}$$

$$\tilde{\boldsymbol{y}} := (0, 1, \ldots, \tilde{n}) \tag{33}$$

$$\theta := \texttt{Sample(}\boldsymbol{\theta}\texttt{, Posterior)} \tag{34}$$

$$\texttt{Posterior predictive[i]} := \texttt{Sample(}\tilde{\boldsymbol{y}}\texttt{, Binomial(}\tilde{y}, \tilde{n}, \theta\texttt{))} \tag{35}$$

$$\texttt{\}} \tag{36}$$

Figure 16 shows a plot of this posterior predictive distribution, 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 weighted evidence, or it can be given a frequency interpretation: if we were to run 20 samples of 100 units, we would expect one of these samples to have a pass rate of less than 54 units. Another advantage of Bayesian methods over hypothesis testing is that it's relatively easy to build models

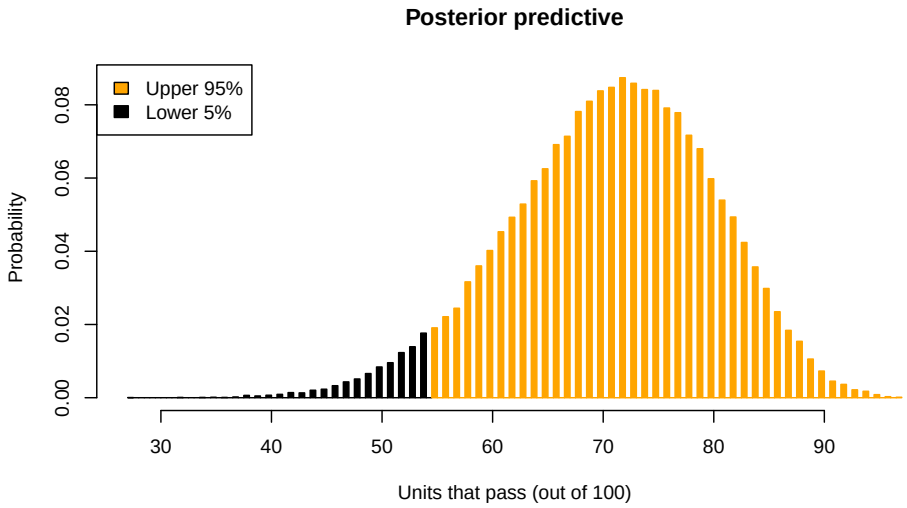

Figure 16: Posterior predictive distribution for the number of passing units in a 100-unit run.

that estimate many parameters simultaneously. An instance of this might be when estimating the mean and 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:

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2) \cdot p(\theta_1, \theta_2) \tag{37}$$

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 makes it immediately clear how much the data reduced uncertainty about what values are reasonable for the data's mean and standard deviation.

In summary, 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 needlessly obscure the amount of certainty in parameter estimates. The jargon of classical statistics can make it unclear what's relevant to the

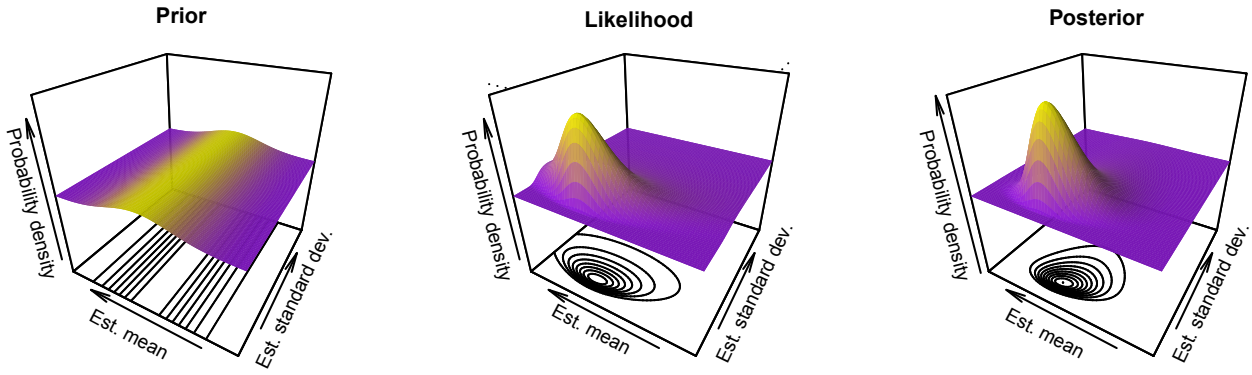

Figure 17: Multiparameter Bayesian inference.

problem at hand, and makes an honest explanation of its methods to non-technical team members difficult and comical. Bayesian methods make it clear how the data and prior knowledge are being combined, and provide results that are more easily interpreted.

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.

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 an outrageous number of grid points. Instead, the posterior can be approximated using a Markov chain Monte Carlo method. This entails randomly drawing a sequence of values according to their relative probabilities, such that their relative frequencies reflect the posterior distribution. [TBC]

3.2 Presentation & Visualization

3.2.1 Why Visualization is Important

Good visualization exposes patterns in data in a way that's immediately interpretable and precise. Bad visualization can be an effective tool for lying: Figure **??** is an example of this. A visualization is a mapping from the numerical domain of a dataset to the visual domain of a plot.

DCA's use of visualization could be improved by:

- Seeing plots as a deliverable for communicating quality and professionalism.
- 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.
- Presenting summaries rather than raw data where appropriate.
- 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.

The most important of these are the choice of geometry, aesthetics, and scale. 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.

3.2.2 Scatter plots

3.2.3 Box plots

3.2.4 Separation Plots

http://mdwardlab.com/sites/default/files/GreenhillWardSacks.pdf

3.3 Software

- Excel
- Matlab
- R
- Python
- Minitab

4 Conclusions & Recommendations

A

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.

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.

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:

$$\text{Var} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \text{Var} X_i \tag{38}$$

$$\text{Var} n\bar{X} = n\text{Var}X \tag{39}$$

$$\implies \sqrt{\text{Var}\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{40}$$

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

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

Confidence intervals

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

Justifying the least-squares fit.

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

(41)

$$\epsilon \sim \text{N}(0, \sigma^2) \tag{42}$$

$$p(\epsilon|\hat{\beta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{\epsilon^2}{2\sigma^2}) \tag{43}$$

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:

$$l(\hat{\beta}) \propto -(Y - X\hat{\beta})^2 \tag{44}$$

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:

$$E[y|X] = X\hat{\beta} \tag{45}$$

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:

$$E[X\hat{\hat{\beta}}] = E[\boldsymbol{X(X^TX)^{-1}X^Ty}] \tag{46}$$

$$= \boldsymbol{X(X^TX)^{-1}X^TX}\hat{\beta} \tag{47}$$

$$\implies E[X\hat{\hat{\beta}}] = \boldsymbol{X}\hat{\beta} \tag{48}$$

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:

$$\text{Var}(\boldsymbol{X}\hat{\hat{\beta}}) \leq \text{Var}(\boldsymbol{c^Ty}) \tag{49}$$

The proof for this is:

$$E[(a^T\hat{\hat{\beta}} - a^T\hat{\beta})^2] \leq E[((c^Ty - a^T\hat{\hat{\beta}}) - (a^T\hat{\beta} - a^T\hat{\hat{\beta}}))^2] \tag{50}$$

$$\leq E[((c^Ty - a^T\hat{\hat{\beta}}) + (a^T\hat{\hat{\beta}} - a^T\hat{\beta}))^2] \tag{51}$$

$$\leq E[(c^Ty - a^T\hat{\hat{\beta}})^2] + E[(a^T\hat{\beta} - a^T\hat{\hat{\beta}})^2] \tag{52}$$

$$\iff 0 \leq E[(c^Ty - a^T\hat{\hat{\beta}})^2] \tag{53}$$

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