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 can be invaluable for recognizing patterns in experimental data, and therefore offer a means of improving the quality and consistency of design decisions. Here, DCA's current use of statistics is evaluated relative to modern statistical practice. Experiment design, analysis, and presentation tools are suggested that would enhance DCA's testing process. These tools are evaluated against the realities of DCA's work by considering how they might be implemented in DCA's experimental procedures.

MAY 23, 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 facilities 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 experiments run in these labs and the data they produce is the focus of this report.

This equipment is regularly used by DCA's engineers to generate experimental data. The way in which this data is collected, analyzed, and

Figure 1: Products designed by DCA.

presented is the focus of this report: data-oriented activities constitute the scientific discipline of statistics. Statistical methods allow resources and information 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 hypothetical 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. Each method is critiqued according to its use case.

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 predictive modelling for product interfaces.

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.

As can be seen in Figure **??**, most lab work is conducted by engineers on

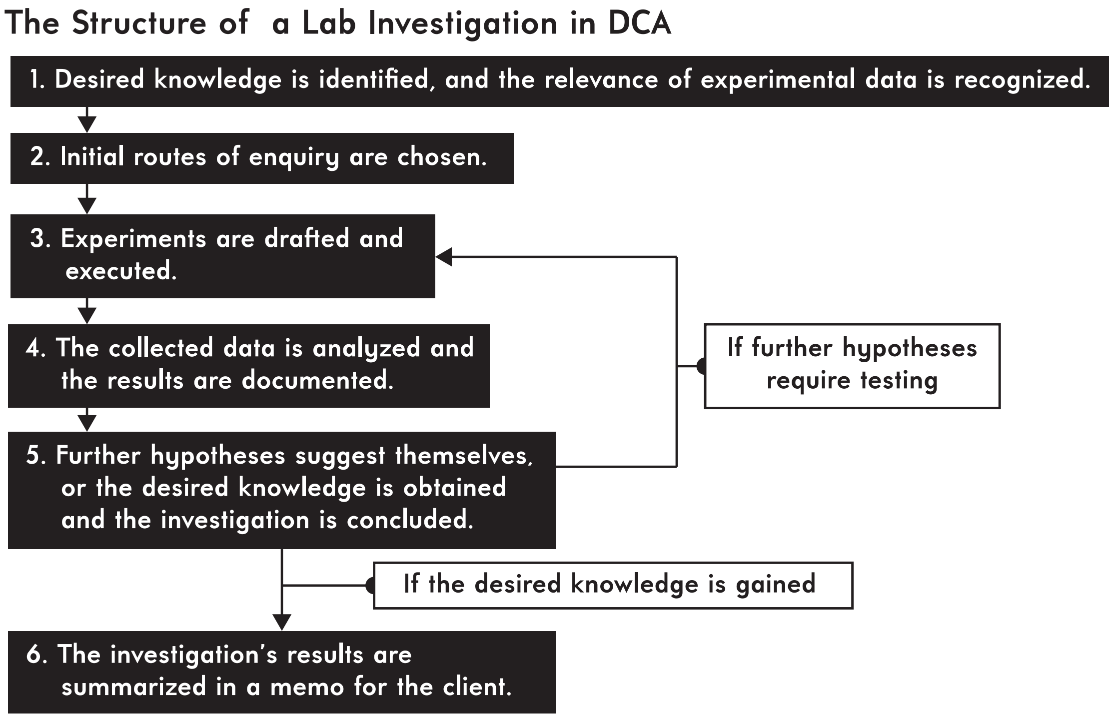

Figure 2: Investigation diagram.

medical projects. Occasionally 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 **??**, most experiments are run in the late stages of the design process. This 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 obsersvations are randomly drawn from a distribution.

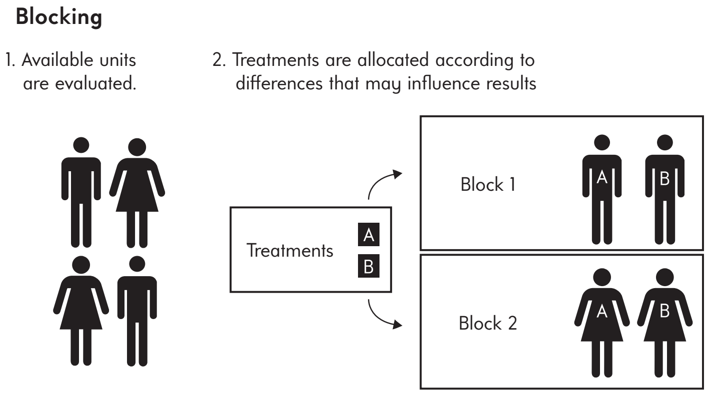

Figure 3: Diagram of what blocking involves.

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	Images of test set-up would make it much easier to retrospectively understand an experiment.
		Experiments are run by engineers solving the problem - this makes engineers personally responsible for their results, and exposes them to undocumented experimental information.	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 4 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 product's performance. The attributes of this variation - such as its spread and average - can be estimated using summary statistics.

For instance, an average of independent observations approximatates the mean of the underlying population's distribution. The accuracy of this estimate improves as we test more samples, with diminishing returns, a relationship that is shown in Figure 5. The standard deviation of the sample mean, $\frac{\sigma}{\sqrt{n}}$, is the average difference between the sample mean and population mean for a given sample size, and is called standard error. It's useful as a tool for assessing how far the sample mean is from the actual mean of the population's distribution, a use that will be discussed more when we talk about interval estimates.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \approx E[X] = \mu = \int_{-\infty}^{\infty} x \cdot p(x) \cdot dx \tag{3}$$

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'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

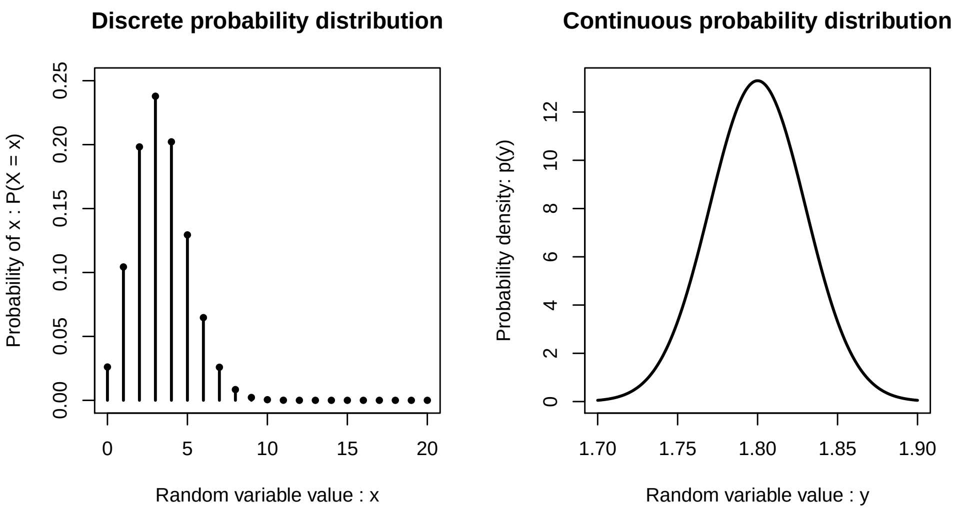

Figure 4: Left: Probability mass function. Right: Probability density function.

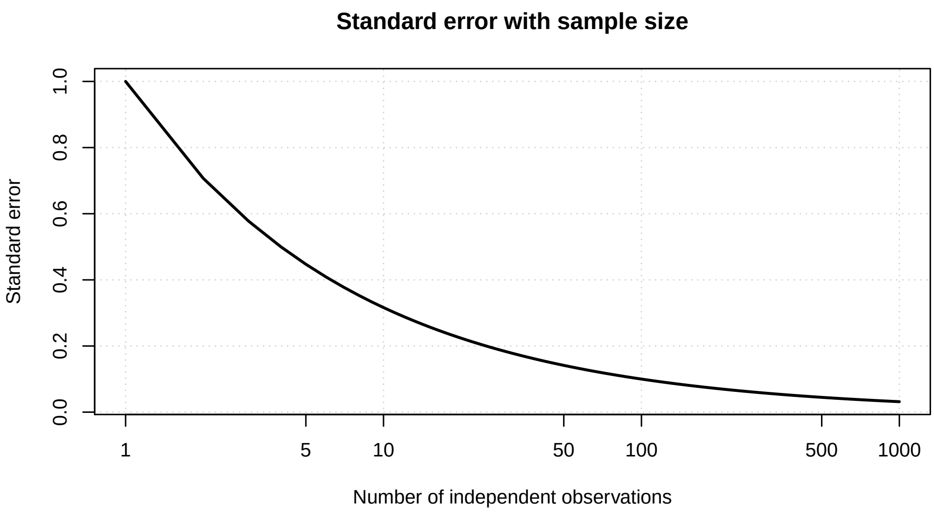

Figure 5: Convergence of sample mean to population mean.

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 contain at least the stated fraction of the population. Consequently, tolerance intervals allow us to make statements about the performance of a population, with clear limits on the uncertainty of the statement's uncertainty.

We seek k such that $\bar{x} + ks$ is greater than at least a proportion p of the population with probability $1 - \alpha$. Let $\mu + u_p\sigma$ be greater than exactly a fraction p of the population. Therefore:

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

After some algebraic wrangling, it's possible to find:

$$P\Big(\frac{\sqrt{n}(\sigma u_p - \bar{x} + \mu)}{s} \leq \sqrt{n}k\Big) = 1 - \alpha$$

The term on the l.h.s. of the inequality has a t-distribution with $n - 1$ degrees of freedom and location $\sqrt{n}u_p$. This means that

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

where $t_{1-\alpha}(...)$ is the value corresponding to the $1 - \alpha$ percentile of the t-distribution with $n - 1$ degrees of freedom centered at $\sqrt{n}u_p$. The interval containing at least p of the population with probability $1 - \alpha$ is therefore:

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

This interval contains at least a fraction p of the population with probability $1 - \alpha$. Example tolerance intervals are shown in Figure 6.

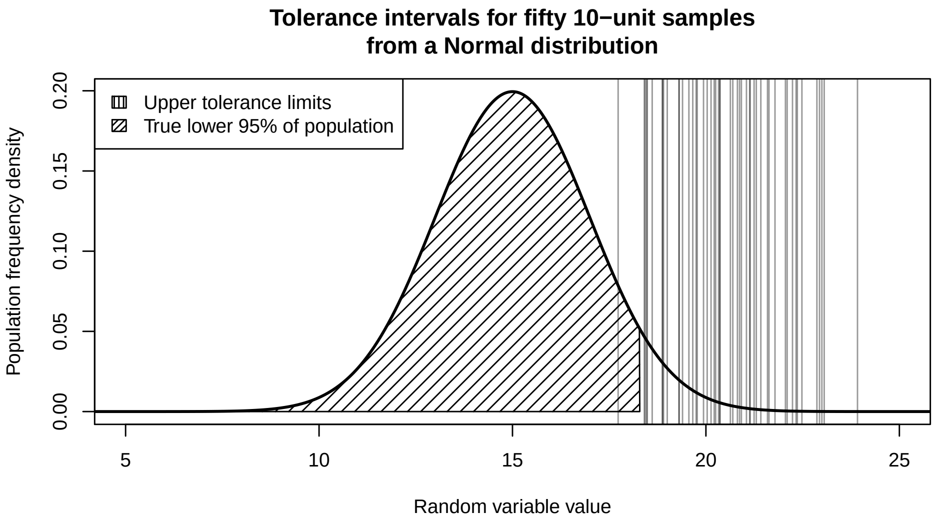

Figure 6: Estimated tolerance limits versus true population limit.

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. The fraction of intervals constructed that will contain the population parameter is called the confidence level of an interval. For example, if we were to construct a confidence interval for the population mean for 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 **??** demonstrates this. Confidence intervals can be placed on any parameter estimate, although they're usually used to quantify the uncertainty on a mean.

One-sided confidence limit. We want to find k so that

$$P(\bar{x} + k\frac{\sigma}{\sqrt{n}} \leq \mu) = 1 - \alpha$$

Subtacting the population mean from both sides and dividing through by the standard error gives:

$$P(k \leq \frac{\mu - \bar{x}}{\sqrt{\sigma^2/n}}) = 1 - \alpha$$

This implies that k has a Normal distribution, since the r.h.s. of the inequality is just a shifting and scaling of a Normal r.v. (the sample mean). Usually the variance isn't known, in which case the constant σ is replaced by the square root of the sample variance. Qualtitatively, this means that the r.h.s. has a t-distribution with $n - 1$ degrees of freedom.

Monte Carlo estimation

Monte Carlo simulation approximates a quantity by simulating the random process generating it . It has previously been applied within DCA to understand worst-case tolerances 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 7.

[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.

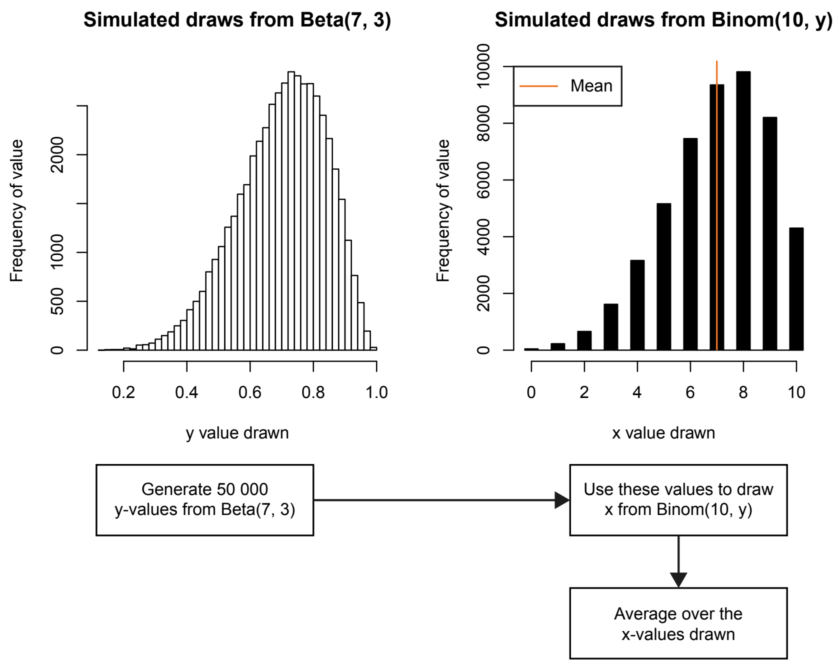

Figure 7: Process diagram for a Monte Carlo simulation.

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 8. The effectiveness of this use-case is evaluated in Table 3.

Bar charts are also used fairly frequently, and a particular format of scatterplot is used to present the results of a tolerance limit analysis. The latter is shown in Figure 9.

The tolerance intervals in the preceding section are, by protocol, plotted

Table 3: 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.

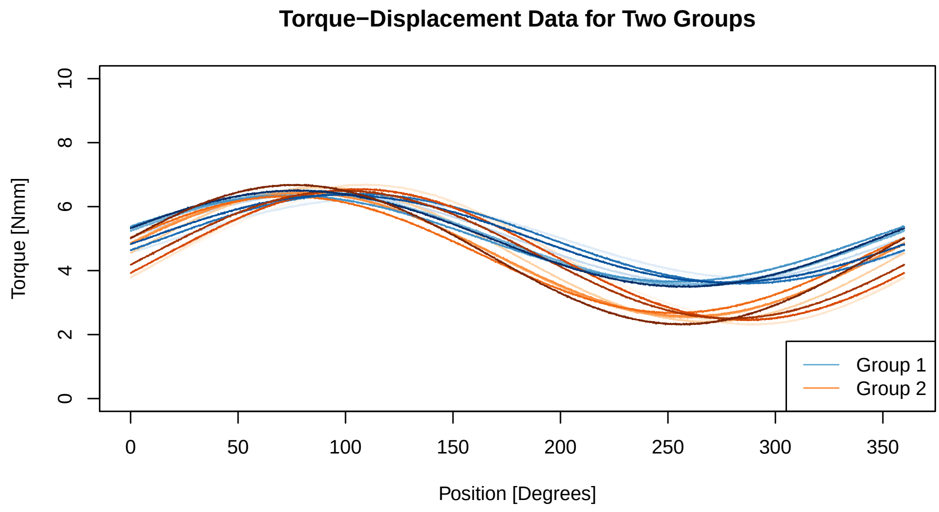

Figure 8: Sample summary line chart from a DCA test report.

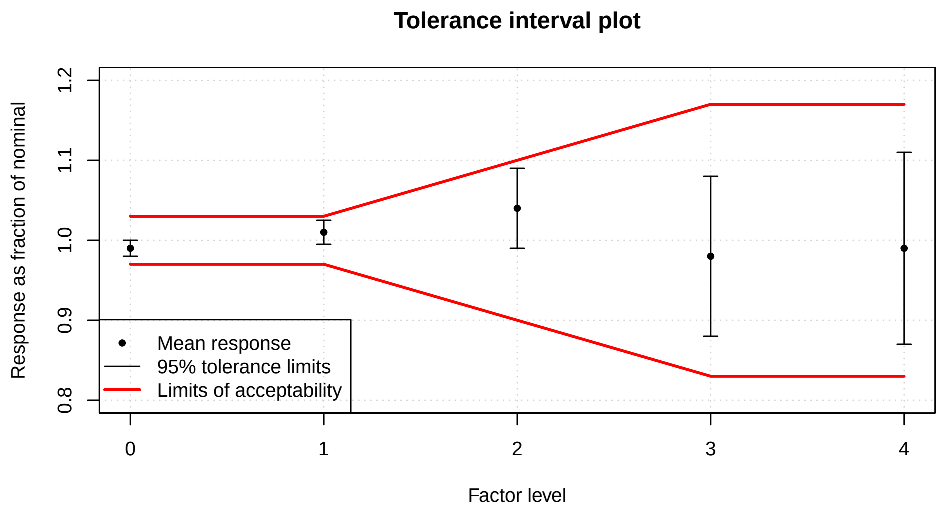

Figure 9: Tolerance interval plot.

according to the graph in Figure 9.

+ Something nice

2.5 Other Activities

3 Suggested Methods

3.1 Experiment design

Experimental pre-planning

Fractional factorial designs

Taguchi designs

3.2 Analysis

Regression Models

Regression models are a flexible tool that allow us to

- Estimate the effects of discrete and continuous factors on a unit's response - even if they interact or have a nonlinear influence.
- Identify differences between units that aren't immediately apparent from raw data
- Measure how well our understanding of a product lines up with its reality

Bayesian Inference

Summary statistics focus attention on a particular possible value of a distribution's parameter. In reality a sample will suggest a range of plausible values, each with an associated probability. In Bayesian inference, this probability distribution is known as the parameter's *posterior*. It's the result of combining knowledge available before data is collected and the information in the data itself. Estimating the posterior, rather than the 'most likely' parameter, has several benefits:

- Its spread makes it clear how much uncertainty there is about what the true parameter value is.
- The distribution directly describes the probability of a particular parameter value, making it much easier to communicate to the uninitiated than hypothesis tests or interval estimates.
- The assumptions being made in Bayesian inference are clearer than in hypothesis testing.

To retiterate, Bayesian inference combines prior knowledge with test data to estimate a distribution's parameters. Knowing what parameters are plausible allows us to place bounds on the behaviour of a population.

An example of where this would be applicable in DCA is when estimating the proportion of units that will pass a given test. $p(\theta, y)$ is the probability that this proportion is θ and that y units in a sample of n pass. Bayes' rule relates the joint distribution, $p(\theta, y)$, to the probability that θ is a particular value, given the data:

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

Qualitatively, this expression states that the probability of a parameter given the data (the *posterior*) is directly proportional to the data's probability, given that parameter value (the *likelihood*). The denominator of this expression is constant w.r.t. θ, therefore we can reduce the above to

$$p(\theta|y) \propto p(y|\theta) \cdot p(\theta)$$

To estimate $p(\theta|y)$ then, we need two things: * $p(y|\theta)$ - the probability of the data for a given probability of success (the *likelihood*). * $p(\theta)$ - the probability of a parameter value according to prior knowledge (the *prior*).

The probability of y successes in a fixed number of trials for a given probability of success θ is described by the binomial distribution

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

The prior distribution, $p(\theta)$, is a way to encode our knowledge that of whether a unit will pass into the model. If all probabilities of success seem equally plausible, then it's reasonable to assume that $p(\theta)$ is uniform over $\theta \in [0,1]$:

$$\theta \sim \text{Unif}(0,1) : p(\theta) = 1, \theta \in [0,1]$$

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

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

It might be easier to understand this process as one of simulation.

Markov Chain Simulation

Decision Trees

3.3 Presentation & Visualization

3.3.1 The Psychology of a Plot

3.3.2 Scatter plots

3.3.3 Histograms

3.3.4 Box plots

3.3.5 Separation Plots

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

3.4 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:

$$\mathrm{Var}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \mathrm{Var} X_i \tag{4}$$

$$\mathrm{Var} n\bar{X} = n \mathrm{Var} X \tag{5}$$

$$\implies \sqrt{\mathrm{Var}\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{6}$$