Statistics: expanding our reach by reviewing our roots (handout)

Jim Garrett

October 20, 2025

Abstract

This document is a handout accompanying Jim Garrett's presentation for the IABS conference, titled "Statistics: expanding our reach by reviewing our roots."

Contents

1 Introduction **2**
1.1 A little about me 3

2 Groundwork **3**
2.1 What is statistics? 3
2.2 What's in a name? 4
2.3 Aside: a contrast between statistics and data science 4

3 Statistical principles make a difference **5**
3.1 Motivating example 1 6
3.2 Motivating example 2 10
3.3 History of statistical contributions to bioinformatics 10
3.3.1 CART 10
3.3.2 Random Forest 13
3.3.3 Normalization of 2-color DNA microarrays 14
3.3.4 Affymetrix gene-expression arrays 14
3.3.5 LIMMA 15
3.3.6 Gradient boosting 15
3.3.7 Support Vector Machines 15
3.3.8 Smoothing splines 16
3.3.9 Summary of history: Statistical principles 16

4 **Statistics cedes ground** 17
4.1 Statistical leadership cedes ground 17
4.2 Individuals cede ground 18
4.2.1 Discomfort with uncertain environment 18
4.2.2 Lack of familiarity with technology 19
4.2.3 Inference referee uncomfortable with judgment 19
4.2.4 Inference referee in "impure" context 20

5 **Axes of orientation** 20
5.1 Simplicity 20
5.1.1 Value of technically simple tools 22
5.2 Articulating statistical value 22
5.2.1 What does effective statistics provide? 24
5.2.2 Small data set? I'll show you a small data set 24
5.3 Inference referee vs. modeling tactician 25

6 **Technical differentiators** 27
6.1 Transformations 27
6.2 Model multiple contributors 28
6.3 Analyze continuous data 28

7 **Work practice differentiators** 29
7.1 Statisticians think differently 29
7.2 Be the problem solver 29
7.3 Build a relevant toolbox 30
7.4 Step up if you're the best qualified 31

8 **Collective movement forward** 31

1 Introduction

This document provides additional detail for a presentation for IABS. The presentation falls within one theme of the conference, "Contribution & Differentiation." This is a timely topic since "statistics" is no longer operating in isolation, if in fact it ever was. We now have data scientists carrying out analyses, as well as managing data. Where is the border between the two? Is there one? Meanwhile we have quality engineers who are actively using basic designed experiment principles, even as many statistics graduate programs have reduced or eliminated requirements for students to receive training in this area.

I've done quite a bit of work shoulder-to-shoulder with data scientists in the pharmaceutical industry, and I've noticed some differences in orientation. In this presentation I focus on the comparison between statisticians and data scientists. I imagine many similar comments could be made in the experimental-design context.

1.1 A little about me

I earned my PhD in statistics from the University of Minnesota in 1996. At that time, all graduate students were required to take a year-long sequence in experimental design methods, as well as stat theory, some specific topics, and some time in supporting the university's consulting clinic.

My first job was with a division of Becton Dickinson that specialized in infectious disease diagnostic devices. Our management at the time gave us support to make ourselves useful in whatever ways we saw opportunities to contribute, and so we helped scientists with system-optimization experiments; deployed QC lot-disposition plans; developed and analyzed system validation plans; analyzed clinical trial data; and so on.

Because some of the assay systems were highly automated, I became interested in the problem of decision-making based on observed data, without a human in the loop. I began to train myself on all manner of machine learning and other means of data-based decision-making, by reading books and attending conferences. I could generally attend one conference per year, and I generally did *not* attend one focused on diagnostic devices, or health care, but rather sought conferences that demonstrated automated decision-making in action. Most of what I know now I learned *after* graduate school, but nevertheless my graduate training provided a solid foundation.

Later I moved to Novartis to support the development of companion diagnostic assays by an in-house team. After a few years the company decided to rely exclusively on diagnostic partners, and this team was dissolved; I then spent a number of years doing exploratory biomarker analysis of clinical data (Phase II and Phase III).

Novartis was the first pharmaceutical company to bring a gene therapy to market, and through FDA approval, under the marketing name "Kymriah." Due to my experience supporting manufacturing (even if only of devices), I was informally enlisted to look at very early data and to specify incoming material specifications for viral vectors. These specifications were part of the package submitted to the FDA. I'll mention that occasionally below.

A little over a year ago I founded an independent consulting business, "Replicate! Statistical planning and analysis".

If you'd like to discuss any of these ideas, please drop me a line at james@replicate-stats.com, or as a backup, jamesgarrettllc@posteo.com.

2 Groundwork

We need to establish a few terms.

2.1 What is statistics?

If we're comparing statistics to data science, we really need to know what statistics is anyway. I have my own working statement:

> Statistics is the art and science of picking a decision rule (noisy data $\rightarrow$ decision) that is most replicable and useful, based on statistical properties and the nature of the data-generating process.

I intend this to include the art of *gathering* the data too, so experimental design and study design practices fall within this.

2.2 What's in a name?

Note that simply analyzing data doesn't make one a statistician. Rather, considering multiple data-analysis strategies and recommending one based on its statistical performance is what makes you a statistician.

Note also that one can become a statistician by learning and deploying wisdom and knowledge about the statistical properties of data-analysis strategies. People who didn't get a graduate degree from a statistics program may find themselves doing this. Of course, getting formal training can certainly help!

Conversely, someone with a statistics degree may find themselves analyzing data according to rules prescribed by others; they may not be exercising statistical judgement very often. They are *not* acting like statisticians, and are not providing their full value as statisticians.

Really, anyone can do anything if they take the time and effort to learn the ways. Data scientists can become statisticians, and statisticians can become data scientists.

This is similar to the practice of law; a student gets her law degree, and then has to decide what sub-field to pursue. Patent law? Corporate law? Divorce law? Real-estate law? All are possible, and change from one to another is possible, given enough training and experience. Yet the switching cost is high, so we don't do it very often.

2.3 Aside: a contrast between statistics and data science

As someone who has worked shoulder-to-shoulder with data scientists at a large pharmaceutical company, I'd like to offer some comments based on my own experience about how statisticians and data scientists tend to differ. Again, these points don't necessarily have to hold; there are no doubt people out there who deviate from these patterns. Also, I'm noting observations concerning how (clinical) statistics and (R&D) data science were organized at one corporation.

- Data scientists tend to take the craft of coding seriously, and are, in the main, better programmers than statisticians are. That is, if I picked a statistician at random and a data scientist at random...you can fill in the rest.
  - They tend to take ownership of "pipelines," or computing sequences that process data from raw to ready-to-analyze.
- Data scientists tend to learn more molecular biology concerning the problem area than statisticians.

- Most of the data scientists I've worked with tended to approach multivariate problems (e.g., gene expression for many genes) by applying simple methods (t-tests, boxplots) many times and aggregating the results. Nothing wrong with this, but it leaves on the table the opportunity for shrinkage or borrowing of information among the multiples.
  - However, a small proportion became proficient in using machine-learning tools and multivariate clustering tools.
- Statisticians have always owned the problem of inference.
- How does one handle potential batch effects? To statisticians they're just another parameter in a model, which in no way threatens conclusions or analysis flow. I've noticed data scientists deploy a process that "removes" batch effects, and then proceed as if that batch-effect removal is truth.
  - Come to think of it, I've also seen this among engineers.
  - In fact, if the ability to handle multiple effects simultaneously via a model is not a natural human approach. I think statisticians share this with other disciplines that describe systems involving multiple factors via models, such as physicists and economists. But there aren't many.

3 Statistical principles make a difference

Data-analytic methods can be:

Ad-hoc Something seems to work, for reasons perhaps not well understood.

Algorithmic One obtains the answer by processing available data using some rules.

Model-based One assumes an error distribution and possibly some model components, which are sometimes flexible.

It's been my experience that the more modeling and probability principles we bring to a problem, approaching point 3, the better the method works.

While it is valuable to add model-based structure, this also raises the possibility that that structure uses assumptions that the data don't support. Therefore there is also a benefit to adding flexibility. Flexibility often comes in the form of

- Using splines or other smooth curve methodologies to relax linearity assumptions;
- Using mixtures to allow for multiple parametric components;
- Using covariance structures to allow for dependence in some limited ways.

Here I provide some illustrative examples.

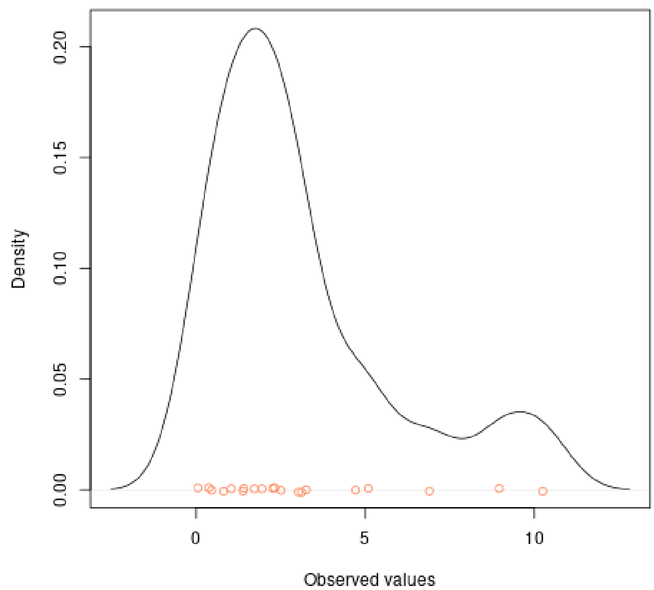

Figure 1: Kernel density estimate.

3.1 Motivating example 1

If you have a set of continuous observations, a popular visualization tool is the kernel density estimate. This is an algorithmic method; to calculate the density value at a point x, one must keep the data set and carry out an operation relative to the observe data.

Figure 1 gives a kernel density estimate for some simulated data. The data comes from a non-symmetric distribution with no support below zero.

This is a fine visualization. It is plain that the data distribution is skewed; we see what appears to be a second mode near the value 10, which calls our eye to the asymmetry. But what if we were a bit more demanding, for instance, what if we were to use a density estimate to establish a normal range? Is it plausible that the “bump” near 10 represents a replicable second mode? No, we don’t have any data to support such a representation. Still, there it is, our estimate has done it. Thus it is a nice visualization, but it is not a “serious” estimate.

Let’s add some probabilistic structure. A univariate density method I’ve found to behave very well is the `logspline` approach, implemented in the `polspline` R package. It works as follows:

- Suppose that the logarithm of the density function can be represented as a cubic spline.

- Given a number of "knot" locations we can generate a basis for cubic splines. Then the model needs only coefficients to be fitted.
- Given a number of knots, we can apply a heuristic to determine their locations (quantiles of the data, I believe).
- Therefore for a given number of knots, we can create the basis elements, and then use a numerical search algorithm to find maximum likelihood estimates of the coefficients.
- This provides deviance as well. Using the number of basis elements as degrees of freedom, we can calculate the BIC model-selection criterion.
- We can repeat this for a sequence of knot numbers, and choose the knot number that optimizes BIC.
- Furthermore, if data observations are censored, this can be accommodated, because censored data still contributes to the likelihood in a known way.
- Similarly, if the range of the data is physically bounded, we can accommodate this in the likelihood definition.

If we apply this approach to the same data, we get Figure 2. The addition of a likelihood as well as model-building machinery has added a great deal of restraint to the estimate; we now have only one mode, yet the asymmetry of the distribution is clear.

We can take this one more step. We know that data values cannot be below zero, and the fitted distribution gives these values positive probability. We can incorporate this information in the likelihood definition, yielding Figure 3. This looks a little odd, because we're not used to seeing density estimates with lower bounds; yet if we think correctly, this estimate is enforcing the lower gound as expected; it's also indicating a bolus of observations between zero and 2.5, and meanwhile there is one mode, and the distribution is skewed.

Incidentally, if some observations are interval censored, this can also be incorporated in the likelihood, and can inform this density estimate.

In this example, adding probabilistic structure and model-building processes improve performance, even if the estimate remains very flexible.

In my experience, this is almost always the case.

This technology was first published in 1992. I think it might be about time to consider it tried-and-true.

I used this technology to estimate quantiles as limits for incoming material, in some cases where data could not quite be transformed to a normal distribution. I carried out some simulations which demonstrated, to my own satisfaction, that this approach generated reasonable results for normally-distributed data. We received no questions from the FDA concerning this approach.

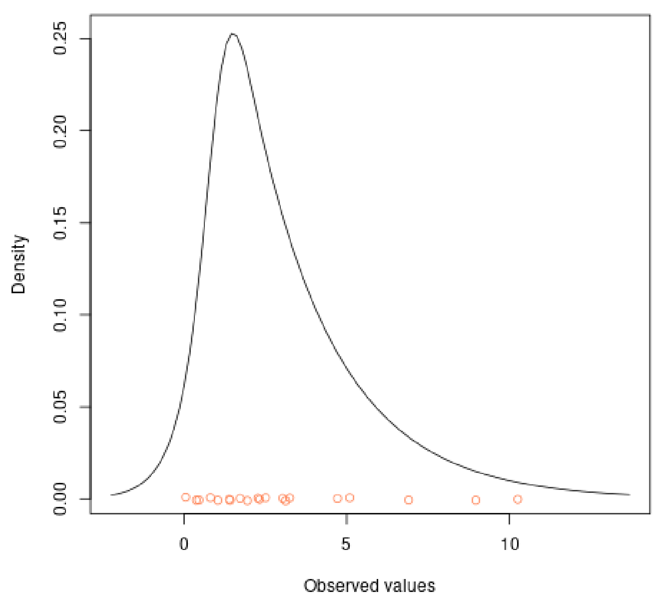

Figure 2: Vanilla logspline density estimate.

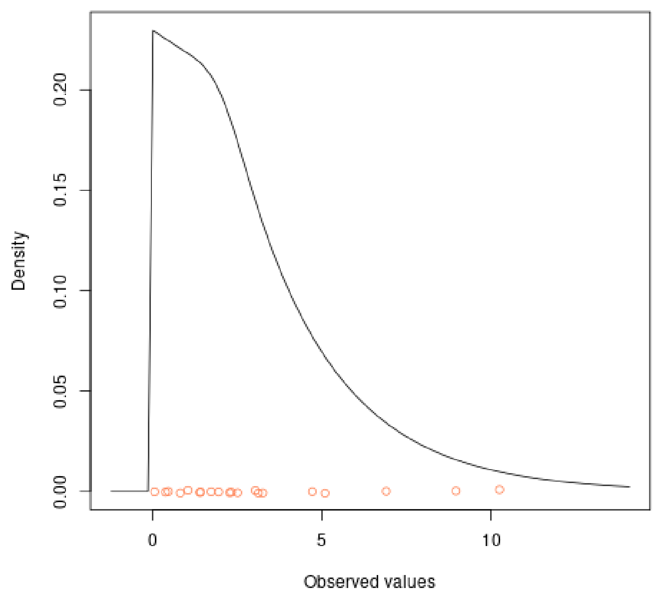

Figure 3: Bounded logspline density estimate.

3.2 Motivating example 2

The previous example was illustrative, I hope, yet contrived. The next example comes from a real application with solid value.

I came across this article evaluating clustering methods; it makes a distinction between algorithmic and model-based methods, and finds that model-based ones tend to do well (concurrent with my experience). One in particular, implemented in the `VarSelLCM` R package, uses finite mixture models and can can accommodate both continuous and categorical variables. It can also apply a criterion to exclude variables that apparently vary little from one mixture component to another.

Have hammer, will find nail: shortly after I investigated this package, an application landed on my desk in the form of a clinical biomarker exploratory analysis. I advocate for including clinical covariates in exploratory biomarker modeling, and the team had obligingly suggested over 10 clinical variables that could reasonably be relevant to the outcome of interest. Yet they steadfastly refused to suggest that one, two, or three might be of paramount interest.

After checking for suitable transformations of continuous variables, I applied the `VarSelLCM` package. It identified two populations, the most important variables, and initial estimates of which trial subjects belonged to which population. We presented this information to the team and they found it very reasonable from a clinical and biological perspective.

We added putative subject component membership as an additional clinical covariate and it proved to be a compact summary of clinical condition, and added substantially to exploratory models. Thus it served as a useful data reduction tool. (Note that arbitrarily much exploration of predictor variables does not induce bias in modeling, as long as one doesn't include the model response variable in the exploration.)

Figure 4 shows an example generated in the documentation for the the function `VarSelCluster`.

3.3 History of statistical contributions to bioinformatics

Statisticians, applying statistical insights, have provided enormous contributions to bioinformatics associated with the exploratory assay technology as it evolved. Table 1 presents contributions, who made them, and what statistical principles made the contributions work well.

3.3.1 CART

Tree-structured predictive models have been discussed for a long time, but one technology has become dominant. This involves:

- Searching predictor variables for the best possible split;
- Applying a criterion to score the usefulness of a split;

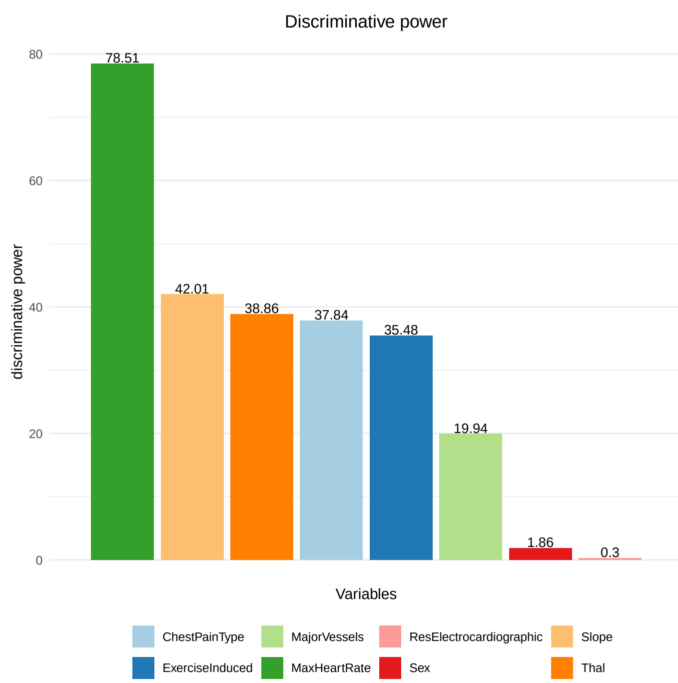

Figure 4: Example from documentation of the `VarSelCluster` function in the `VarSelLCM` package, showing apparent discriminatory power of different variables.

Achievement	Purpose	Who	Statistical ideas
CART	Prediction	Leo Breiman Jerome Friedman Richard Olshen Charles Stone	Variability measures
Random Forest	Prediction	Leo Breiman	Law of large numbers Bias Correlation
2-color DNA microarray normalization	Normalize cDNA spotted arrays	Gordon Smyth Terrence Speed	Smoothing Multiple perturbation factors
Affymetrix	Normalize Affy gene-expression arrays	Raphael Irizarry	Defined objective assessment protocol
LIMMA	Differential gene expression	Gordon Smyth et. al.	Multivariable model Empirical Bayes Pooled variance Shrinkage estimation
Gradient boosting	Prediction	Jerome Friedman	“How does it work?”
SVM	Classification	Vladimir Vapnik	VC dimension (model criterion)
Smoothing splines, GCV	Curve fitting	Grace Wahba	model selection

Table 1: Statistical contributions to bioinformatics.

- Applying "pruning" to simplify a tree model after the splitting process has been carried out until a stopping rule is triggered.

This paradigm was proposed by Breiman, Friedman, Olshen, and Stone. The key statistical insights are:

- The splitting criterion: reduction in mean squared error for continuous response, and reduction in Gini index for categorical response.

- A complexity measure that applies to a branch; this weighs the total benefit of all of the leaves of the branch against the number of splits.

3.3.2 Random Forest

Random Forest is one of the most widely-used machine-learning methods. It is rarely optimal, but it is almost always close to optimal, and it has no critical tuning parameters. It is nearly invariant to transformations of predictor variables.

Random Forest was created by Leo Breiman, who contributed significantly to CART (Section 3.3.1).

By the way, to get a sense of how a brilliant statistical thinker took a broad view of how statistical thinking can be applied, read this conversation with Leo Breiman.

The principles that make Random Forest tick are uniquely statistical, and ingenious. I highly recommend reading for yourself at the Random Forest website curated by Adele Cutler. Briefly, though:

- We're going to make a prediction of y at a multivariate location x. And we have training data.

- Suppose we have a large number of weakly-correlated predictive models $f_i(x)$. At each point x, f_i is nearly unbiased, yet highly variable.

- By taking the average prediction across the models, $\hat{y} = f(\bar{x})$, we get a prediction that is also nearly unbiased (owing to the linear nature of averaging) but also much smaller variance.

Put simply, Random Forest leverages the Law of Large Numbers. The trick is constructing a large number of weakly-correlated, nearly-unbaised models. If one can contrive this, there's little that can go wrong.

In practice, each sub-model is a greatly-overfitted CART model fitted to its own bootstrap data set, with predictor variables randomly selected. Random Forest leverages the following statistical ideas:

- Law of Large Numbers

- Bias

- Variance

- Correlation (how can the correlation between models be reduced?)
- Bootstrap
- Random selection

3.3.3 Normalization of 2-color DNA microarrays

This assay technology is obsolete now, but just a few decades ago it was a primary tool in the fast-developing world of molecular biology. Fragments of capture DNA were sprayed onto very small spots on a glass slide. The manufacturing as well as the reading processes could give rise to spatial effects. Statisticians Gordon Smyth, Terrence Speed, and others developed an algorithm to "normalize" the readings for increased accuracy, and this algorithm was apparently highly respected and widely used. The normalization algorithm developed linear models for assay readings in each color that included smooth continuous spatial contributions, fitted by lowess (or loess; spellings vary). Thus they used principles of smoothing and fitting models with multiple simultaneous effects.

3.3.4 Affymetrix gene-expression arrays

Some years after two-color spotting, Affymetrix deployed a much higher-resolution approach which still involved DNA fragment capture at many points on a surface, this time a silicon chip. Again, spatial effects were possible. The Affymetrix company provided a normalization algorithm, but they also allowed users to access the raw data. Not surprisingly, users began to try out their own algorithms.

Raphael Irizarry, then a professor in the statistics department at Johns Hopkins University, made an enormous contribution to the field, but he didn't add his own algorithm. Rather, he developed a protocol to evaluate algorithms, and fashioned two reference data sets. He then created a web site. Individuals wishing to have their algorithm evaluated and scored publicly would download blinded reference data, apply their algorithm, and upload generated results in a standardized format. Then the evaluation protocol was applied, and the method would be scored according to multiple criteria. (I may be misrepresenting some details; if so, I apologize!)

Eventually over 20 algorithms were publicly scored. No algorithm was ideal across all criteria, but two methods generally became generally recommended: RMA, and g-RMA (a variant of the RMA algorithm that adjusted for different nucleic acid binding strength).

Dr. Irizarry's development of a fair and informative evaluation process led to the effective crowdsourcing of Affymetrix normalization methods. Interestingly, Affymetrix's own original algorithm didn't look horrible, but did not come out in the top tier. Since, initially, this was the only algorithm available, the crowdsourcing proved to be very effective.

3.3.5 LIMMA

LIMMA is an R package developed to estimate contrasts in gene expression (such as from Affymetrix chips) while adjusting for background noise sources such as reagent lots and batched runs.

It uses an Empirical Bayes framework to pool estimates of noise variance, as well as multivariable models to deconvolute contrast effects and noise effects. Therefore it leveraged the statistical principles of shrinkage estimation, Bayesian methodology, and multivariable modeling.

3.3.6 Gradient boosting

This example powerfully exemplifies how statistical principles can improve analytical methods.

"Boosting" was an emerging classification technique in the computer science machine-learning community. It consisted of:

1. Fitting a model;

2. Calculating residuals;

3. Applying a specific transformation to the residuals;

4. Fitting another model to the transformed residuals;

5. Upon reaching a stopping criterion, adding the models together.

Computer scientists found that this "trick" yielded good results, but no one was altogether sure why.

Statistician Jerome Friedman looked closely at the transformation and deduced that fitting a model to approximate it was taking a step towards optimizing binomial likelihood. What, then, would optimize an *actual* binomial likelihood? Friedman figured out what this would look like, and found that it worked better than the ad-hoc boosting, and "gradient boosting" was born.

And if we can develop gradient boosting for a binomial likelihood, what of a normal likelihood? Or a Poisson likelihood, or a Cox partial likelihood? The idea generalized.

Today gradient boosting is a highly effective tool in the machine learning toolbox. It is typically among the top performers for predictive accuracy. Furthermore, it tends to select a sparse set of predictor variables, and it admits variable importance scores.

All because a statistician looked at an ad-hoc method and sought statistical rigor. This exemplifies the definition of statistical activity: wondering *why* analytical methods work.

3.3.7 Support Vector Machines

Vladimir Vapnik's The Nature of Statistical Learning Theory comes across like mathematics from an alient civilization: irrefutable logic, but beginning from different propositions and asking different questions. Vapnik mines a Russian tradition of probabilists asking: if I see n observations in an interval I, can I assert bounds on the probability of future observations falling into I? Vapnik develops a convincing criterion for predictive quality, the Vapnik-Chervonenkis Dimension, and developed Support Vector Machines to optimize it.

I'll admit that I didn't read through every theorem carefully, but the point I'll make is that SVM's are based on a disciplined development, and they work well.

Again, careful thought with mathematical discipline yields improved performance.

3.3.8 Smoothing splines

A smooth curve can be fitted by an excessively rich set of basis elements that would cause gross overfitting, if not constrained by a roughness penalty. But what should that roughness penalty be? It could be determined empirically by cross validation, but that can be expensive. Grace Wahba looked at the arithmetic and found that leave-one-out cross validation can be closely approximated by a cheaply-calculated metric, "Generalized Cross Validation" (GCV). Now GCV turns up almost anywhere curve smoothing does, not least being Generalized Additive Models (GAM's). Dr. Wahba exploited research into model selection criteria and spline methodology.

Dr. Wahba has also advocated for kernel-based modeling methods, of which SVM's are one type.

3.3.9 Summary of history: Statistical principles

Here's a summary of the statistical principles that emerged in reviewing the history above:

- Measures of variation (SD, variance, Gini index)
- Expectation, bias
- Correlation
- Simultaneous additive effects
- Deriving performance measures; system validation
- Likelihood
- Model selection criteria
- Smoothing
- Shrinkage estimation

4 Statistics cedes ground

In spite of the capabilities of statistical principles, statisticians are ceding perfectly fertile ground, both collectively (through our leaders) and individually.

4.1 Statistical leadership cedes ground

I hope I've made a convincing case that statistical principles make decision-making based on noisy data work better. Since I've defined statistics as the study of principles for optimal decision-making based on noisy data, the fact that statistical principles do in fact make decision-making based on noisy data work better is a bit circular, and shouldn't be surprising.

What is the appropriate domain for statistics, then? Where does humanity practice decision-making based on noisy data? And also, where are the stakes high enough to pay someone to ensure that decision-making based on noisy data is done well?

However, it seems that a sense has been emerging that statistics operates in a certain domain, a domain that is quite a bit smaller than the domain of decision-making based on noisy data.

In a large pharmaceutical company within which I worked, sometimes statisticians and data scientists were assigned to the same projects. Data scientists reported in through R&D and were responsible for parsing and interpreting high-throughput assays (e.g., sequencing, Affymetrix). Statisticians reported in through the clinical statistics function and were responsible for inference and clinical application (e.g., Cox models assessing biomarker influence on relative risk). Between these poles, there was a lot of gray area. How should we determine who was to do what?

Finally, the respective organizational leaders (clinical statistics on the one hand, R&D on the other) gave us clarification:

- Statisticians worked with low-dimensional data (usually clinical)
- Data scientists worked with high-dimensional data (usually bioinformatics information coming from high-throughput platforms)

This flies directly in the face of the history of significant statistical contributions I describe above. I don't know whom to blame harder: the clinical statistics leadership, which of course was focused on clinical applications and for whom this research was a distraction from their main clinical mission, or the R&D leadership that was ignorant of this history and perhaps didn't realize that statistics had the potential to contribute to high-dimensional assays.

Incidentally, sitting as I was within the clinical organization, I had no awareness of CMC statisticians. In this organization, clinical operations truly dominated. We all had "biostatistician" in our title and, frankly, I think many of my statistical colleagues were ignorant of our history of contributions.

How is your organization oriented? Does it employ statisticians to get the best possible analysis from talented statisticians?

I think we as individuals also can step away from applications, if we view ourselves as inference referees and the application involves a degree of ambiguity. See Section 5.3 for more discussion.

At any rate, I suggest that the orientation of our community, from our leadership to each of us as individuals, contributes to a narrowing of perceived value and application scope.

4.2 Individuals cede ground

I've seen many individual statisticians step away from meaningful contribution for a variety of reasons. Some patterns I've seen are the following:

- Discomfort with uncertain environment.
- Lack of familiarity with technology, coupled with lack of willingness to gain that familiarity.
- Statistician habituated to inference referee role feeling discomfort with use of professional judgment.
- Statistician habituated to inference referee role feeling concerned that context is not pure enough.

4.2.1 Discomfort with uncertain environment

Here I offer Exhibit A: myself.

In Section 5.3 I discussed how we are sensitized in graduate school to the principle "First Say Nothing Wrong." I recall feeling very much this way when I first began working in industry.

During my first year or two working in industry, someone stopped by my desk to ask if I thought statistics might have something to offer regarding the problem of detecting pathogens growing in a liquid broth incubation tube; one of the company's major product lines was blood culture. Pathogens in liquid growth media exhibit exponential growth, while human blood cells don't reproduce. Therefore detection amounts to finding positive curvature in estimated metabolism activity.

I had a few ideas, but for each idea I could also imagine scenarios in which the idea would fail. And because at the time I was still sensitized to First Say Nothing Wrong, I said I couldn't think of anything in particular.

This is a pattern in which individual statisticians can cede territory unnecessarily: by being concerned about an uncontrolled environment in which something might go wrong.

Some years later, I learned about the methods the team was using at the time. The methods were ad-hoc, complex, and required many man-hours to retune if any modifications were needed.

My initial fears that my ideas might not work were probably justified. That is, most initial ideas require some tweaks, and don't work optimally the first time we try them. However, the statistical tools I would have employed would have been conceptually simpler and easier to tune.

The point is not to get it right the first time, but to get the organization directed on the best path with the best framework with which to address problems that are certain to arise.

4.2.2 Lack of familiarity with technology

I once suggested to a statistician who reported to me that a Generalized Additive Model could be a good approach for a particular problem. He responded that he wasn't familiar with this method, and that he didn't want to use any method unless he was very thoroughly familiar with it. It was implied that becoming so thoroughly familiar could take significant time–months?–and moreover that he wasn't likely to spend that time because he didn't really care that much about the method in the first place. (Maybe it was just a passive-aggressive way of asking me not to tell him what technology he needed to learn.)

This is a matter of degree, of course. None of us can do everything. We have to choose where we spend our time. We should learn continually, but we have to make choices. Still, I'm of the opinion that if we want to differentiate and contribute as statisticians (a theme of this conference!) then we must stretch technically. We must constantly search for technology that will apply to our work area.

4.2.3 Inference referee uncomfortable with judgment

This is the first of two points concerning the "inference referee" role I described in Section 5.3, in contrast with the statistical tactician. These two points arise when a statistician habituated to an "inference referee" role are put into a different context and fail to adapt appropriately.

First, some statistical methods require some professional judgment. Using judgment can appear unseemly; we are supposed to let the data speak without our ideas coloring it–especially if we are inference referees. However, we use judgment all the time, such as when we design studies.

It often happens that a statistical method having a bit of ambiguity (which may require some choice, either arbitrary or expert) will outperform another method that lacks ambiguity. For example, fitting a smooth curve requires some sort of specification of smoothness, but over a wide range of choices, will improve on a straight line fit, if there is non-linearity. Often Bayesian methods with weakly informative priors will work very well, even if "weakly informative" means (say) the variance is probably between 1 and 1 million, or between 1 and 2 million. It's in the nature of solving infinite-dimensional problems with finite data that a range of parameters may suffice, and some choices simply have to be made.

Again, context and degree matters. For most statisticians this is not a serious issue, but I have indeed seen a few individuals who expressed this pattern to a detrimental extent, out of connection to the problem's context.

4.2.4 Inference referee in "impure" context

Similarly, the habituated inference referee, when put into a messy research context, may complain that sample size is too small, model assumptions are not guaranteed, etc., etc., and may insist that a self-respecting statistician would not undertake an analysis. I maintain, alternatively, that statisticians can make valuable contributions in messy contexts. I think again of the statisticians supporting Eisenhower on the German Tank Problem, as in Section 5.3.

I would note also a hierarchy of statistical knowledge and expertise. First we are taught the assumptions; with experience or exploration we find out which assumptions are critical and which are not. Then we attain a sense of how *much* deviation from the ideal is going to cause how much of a problem. Case in point: the t-test. First we are taught that the error distribution should be normal. Then we learn that this test is actually robust because, due to averaging and the central limit theorem, the distribution of the statistic will be more normal than the underlying data distribution. The next stage of knowledge is looking at the data histogram and having a sense for what sample size will make the mean follow approximately a normal distribution. Someone who insists on purity in all cases is simply not getting past the first rung of the ladder.

Once again, the statistician who strongly resists contributing in any area except where all matters are nominal is rare, but I have seen this pattern; it does exist.

5 Axes of orientation

Most of the remainder of this document concerns how statisticians can differentiate themselves and make unique contributions to their organizations. The following sections consider:

- "Axes of orientation:" dimensions or principles that inform our strategies and our work.
- Technical differentiators: technical statistical practices that can contribute to unique impact.
- Work practice differentiators: practices concerning how statisticians organize their work and interact with collaborators in order to maximize unique impact.

Therefore, first, axes of orientation:

5.1 Simplicity

One principle that has become clear over the course of my career is this: everyone loves simplicity. However, they usually have in mind technical simplicity. Let's fit a straight line using Ordinary Least Squares; let's use ANOVA.

However, there is another attribute of simplicity: conceptual, or decision-making. Is a decision cut-and-drived, evaluated in a single step? Or are there numerous layers of questions to work through in order to arrive at a conclusion?

Very frequently, almost axiomatically, technical simplicity leads to decision-making or conceptual complexity. Complexity arises when there are multiple aspects to deal with when interpreting data:

- Are there interactions among additive effects?
- Are there outliers?
- Might a continuous variable contribute non-linearly?

Typically we need to resolve these questions with additional decision-making steps, and suddenly we find we have a layering of simple methods whose overall statistical performance isn't clear. What's the probability of making an incorrect decision at any early stage? How does such an error impact subsequent stages, or the final decision? Suddenly the simple becomes complex.

This can be merely annoying until a case arrives in which an upstream decision concerning a nuisance factor happens to lead to a different conclusion. A colleague once asked me something along the following lines (not exact, but you'll get the idea): there was a question of whether Vendor A's or Vendor B's products behaved differently, so they set up an experiment. The process is run at two labs, so the experiment was divided between the two labs.

It turned out that the A vs. B comparison looked a little different between the labs, i.e., there was some indication of an interaction between vendor and lab. If the labs are analyzed separately, one of the labs failed the acceptance criteria, hence A cannot be equivalent to B. However, if the two labs are pooled, the difference between A and B meets acceptance criteria.

Their biggest error may have been failing to specify in advance how to handle the labs. I've contributed to many analysis plans and study protocols, and my rule of thumb is that the statistician should make provision in those prospective documents for the lion's share of unexpected eventualities: interactions, nonlinearities, outliers, etc. If something occurs that makes a study plan in a protocol moot, we can create a "deviation." However, the presence of many deviations can erode trust by regulatory authorities. If a protocol study plan includes contingencies for not-unlikely eventualities, then the protocol can be executed with no deviation. It's impossible to foresee all possible eventualities; also, analysis plans should not be as thick as a novel. Nevertheless, I advocate for study plans covering a good number of the contingencies a statistician can imagine.

If you write a few study plans and try to address a good number of contingencies, you'll begin to appreciate complexity. "Simple" statistical routines do not insulate you from complex decision-making protocols, especially if they deal with only one factor or depend on narrow assumptions. In fact, statistical methods that expand assumptions reduce the chance of violating the necessary assumptions. Such statistical methods are usually "complex" from an operational perspective, though they can be simple from a conceptual perspective (e.g., "fitting a smooth curve" needs no explanation to non-statisticians, but requires math that is more involved than simple linear regression).

Technically sophisticated but conceptually simple methods can reduce or eliminate multiple steps.

- Concerned that a continuous variable could contribute non-linearly? Add a smooth spline basis to the model.
- Concerned about non-constant variance? Consider fitting a model in which variance is estimated as a power of the mean, or related to a particular factor if scientifically warranted.
- Screen outliers based on outlier-resistant yet efficient estimators, such as M-estimators. Complement with a standard policy for number of outliers allowed (e.g., the process fails its acceptance criteria if more than two outliers are detected).
- Hierarchical Bayesian models can automatically include interaction terms, and still return inference with integrity.
  - Bootstrapping offers a similar capability, but a parametric bootstrap may be required if an experiment is sparse.

Usually in these technically more complex models, there are more parameters that can influence outcomes, so acceptance criteria may need to be based on predictions. If we fit a smooth curve, we don't have a single regression slope to evaluate, but we have an expected value and, if necessary, a confidence envelope. If we're interested in a difference but we entertain the possibility of an interaction, then we may specify a criterion based on the maximum difference across possibly-interacting levels.

Bayesian methods, especially hierarchical ones that allow for data-dependent shrinkage towards zero of higher-order effects, make it straightforward to produce predicted values and prediction intervals to use in acceptance criteria, even with large models. Therefore this approach has large potential to promote decision-making simplicity.

I recommend that, whenever possible, you strive for decision-making simplicity.

5.1.1 Value of technically simple tools

I must call out, though, that there is real value in having everyone in an organization trained in technically simple tools, and to understand variance components. Statisticians should be able to discuss projects in terms location and scale, even if they deploy more sophisticated evaluation tools.

5.2 Articulating statistical value

The appropriate domain of a field of study dedicated to optimizing decision-making based on noisy data is huge. However, statistics takes effort, which

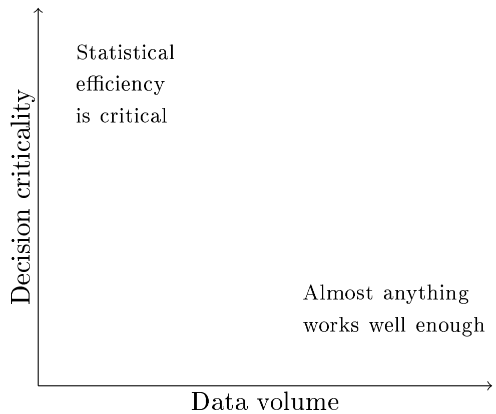

Figure 5: Diagram showing the importance of statistical efficiency depending on the amount of information and the criticality of the decision.

requires people who are paid, so we have to focus on cases where statistical involvement can yield high value.

We must consider Figure 5. For example:

- Evaluating a medical product whose manufacturing process is very expensive to run, leading to a small data set: This would be in the top left.

- Data on web site visits in order to determine whether a particular sales strategy is effective at recruiting customers: bottom right. The analyst has lots of data and the outcome isn't life-or-death, so it's not critical whether highly efficient statistical methods are used.

Most of the data you're thinking about (either analyzing or planning) is probably high on the vertical axis (high decision-making importance), otherwise your employer wouldn't spend money having you work on it. The real question is whether you have so much data that an efficient method isn't necessary.

If you don't have more than enough data, I suggest you should be thinking about statistical efficiency.

Also, we need to generally advance the proposition that not all analytical approaches are equally efficient. What can be the consequence when an inefficient approach is used?

Of course, statistics being about variability, it's possible that in a given instance, for a given data set, a shrewdly efficient approach and a cheap dumb-as-rocks approach will reach the same conclusion. But we can know this only after the data is available. When planning, we need to make provision for the possibility of a wrong conclusion, and consider what it's worth to decrease the probability of such an outcome.

5.2.1 What does effective statistics provide?

What value does a good statistician provide? What happens when a statistician does a poor job?

Remember that statistics is about contrasts. Consider a project on which an effective statistician is present. To contrast, consider that same project lacking an effective statistical voice. What would be different between the two?

Clarity.

With an effective statistician present, data is collected and then it's analyzed with a view towards informing important decisions facing the team. The analysis is decisive and so the decision is obvious and quick and straightforward. Results are repeatable, so the team isn't called back together to revisit issues that re-emerge later. Data-analytic complexities such as discussed in Section 5.1 are handled gracefully and quickly; the team doesn't spend a lot of time discussing them, and the analyst doesn't spend a lot of time re-analyzing due to them, and doesn't show the team multiple versions of results.

In contrast, a team lacking an effective statistical voice spends a lot of time arguing over suggestive but non-convincing data. They may make a poor decision. They may have to revisit it later. Lots of FTE time is wasted. People work hard but things don't seem to move ahead, or rather, nothing is ever finished; the team comes back to revisit issues. A Six Sigma practitioner would call it "rework."

5.2.2 Small data set? I'll show you a small data set

This story is not critical for this discussion but it's fun. How valuable is data? How expensive is it to attain? Is there such a thing as "too small for statistical analysis?"

Here's an interesting example of a small and critical data set. The cost per observational unit was astronomical. It concerns the shape of the earth.

During my undergraduate education at the University of Chicago, I was fortunate to take a regression class from Stephen Stigler, whose specialty is the history of statistics. I recommend his book, "The History of Statistics: The Measurement of Uncertainty before 1900." As an example of the use of linear regression, Dr. Stigler presented the class with the early consideration of the shape of the earth:

- Is it spherical?
- Is it extended at the poles ("prolate")?
- Or is it extended along the equator ("oblate")?

The question could be resolved experimentally by measuring distances along the Earth's surface and comparing the measured distance to differences in latitude as measured astronomically. After some geometry, trigonometry, and algebra, we find that the question hinges on how the discrepancy changes with latitude (the equator has latitude $0°$) in a linear relationship:

- If the slope of the line is zero, the earth is spherical.
- If the slope is positive, the earth is prolate (extended at the poles).
- If the slope is negative, the earth is oblate (extended at the equator).

To gather "data," the French Academy of Science sent an expedition to Ecuador to carry out extensive surveying to measure the distance between two latitudes differing by 1°. This is about 69 miles. Through tropical mountains and jungle. This took months, or years. One team member became acclimated to the tropics, married a local, and took an academic position in Columbia. How much did *that* data point cost?

A second expedition did similar work in Lapland (Finland). Other data points were collected from Paris and Rome. Another measurement was taken from the Cape of Good Hope. Altogether, discrepancy measurements based on 1° latitude difference (astronomical) were collected at five locations.

Find interesting description at these links concerning the French Geodesic Mission to the Equator and the French Geodesic Mission to Lapland.

Certainly the researchers carrying out these measurements at great cost, effort, and time did not feel that this data set of five observations was too small to analyze. Indeed, substantial discussion went into the best avenue of analysis because ordinary least squares was not yet ensconced as a standard regression method. This data set provided evidence that the Earth is oblate, i.e., extended at the equator and flattened at the poles. Scientific opinion hasn't changed since, indeed, this has since been confirmed in great detail. The scientific finding has been replicated, so the data set was effective.

5.3 Inference referee vs. modeling tactician

In graduate school we're penalized heavily for getting anything wrong. On a test, if we violate assumptions, we'll have points taken off. We arrive in industry sensitized to First Say Nothing Wrong. At work, if we are generating results to present to an adversarial or skeptical entity, such as a government regulator, we want to have the "i"'s dotted and the "t"'s crossed: don't give any gratuitous opportunity for your adversary to find fault with your work.

I believe that in such a setting, we can become habituated to not generating any results, by any method, where assumptions are not clearly met and where optimal sample size is not assured.

One can forget about the many sources of complexity described in Section 5.1. One could imagine that for any problem, there is a single best, most fair representation of data.

Alternatively, consider post-hoc exploratory modeling of data. I've heard statisticians habituated to the referee mindset above say that we shouldn't even do such modeling. However, it's a valuable activity that can be done well, or poorly.

Note that conducting a modeling exercise multiple time degrades the quality of the result. Consider the following cases:

1. Fit multiple models and report the best one (only).

2. Fit a single model, evaluate it, find fault with it, and refine it. Iterate until no clear fault is found.

3. Fit a number of models and report all of them without indicating any preference.

4. Explore predictor variables thoroughly and conduct "feature engineering" with subject-matter experts. Make provision for interactions and nonlinear predictor contributions. Do all of this without consideration of the response variable. Finally, fit a model to a derived set of predictors using penalties, where penalties are chosen to optimize a model-selection criterion such as BIC.

Cases 1, 2, and 3 can induce severe over-optimism concerning the model result. Case 4, inspired by Frank Harrell, Jr.'s methodology, is much more trustworthy. See "Regression Modeling Strategies" for more information.

Clearly, the more models one fits, the less reliable or trustworthy is the final model.

An important related thought: if ideally one fits a model once, and one makes analysis decisions prior to fitting a model, then one needs to make good analysis decisions!

In other words, if one can't just go back and refit models to explore a different analysis path, then analysis choices matter. One needs to try to get it right the first time. (And statisticians who make good choices should be valued. And paid accordingly.)

In the context of exploratory modeling, the notion of statistician as referee or arbiter of fair representation of data is replaced with the statistician as valued tactician.

It's important to recognize the axis or dimension running between inference referee at one extreme and data-analysis tactician at the other. There are right times and contexts for each. Just recognize the appropriate context at hand and adapt accordingly, if you can. (I think some cannot, or perhaps they haven't tried.) Serious problems arise when statisticians fail to adapt to the appropriate context.

When I think of statisticians opting out because they want to be inference referees but they're placed into a tactician context, I think of statisticians on Eisenhower's staff during World War II. They were charged to address the German tank problm. During WWII campaigns against Nazi Germany and Fascist Italy in North Africa and Italy, US Sherman tanks had fought successfully against German Panzer III and IV tanks, but Germany had just introduced the more powerful and dangerous Panzer V. As allies planned for the invasion of Normandy, they wanted to know whether a large number of Panzer V tanks were situated in France. How many tanks had Germany manufactured? Information from spies on the ground suggested a number in the thousands.

Allies also collected serial numbers from parts recovered from destroyed German tanks. With this, statisticians applied the Pareto distributions, maximum likelihood, and combining multiple sources to arrive at a number of tanks likely manufactured: 270.

That was a tolerable number, and the allies went ahead with the invasion. You know the rest of the story.

The actual number turned out to be 276. This is a great moment in the history of statistics.

Imagine if a statistician on staff said, seeing himself as an inference referee, said, "There are too many assumptions that we can't be sure about. We can't be confident in our answer. We think it's best to not make an attempt." Fortunately, the statisticians on Eisenhower's staff didn't say this, but instead stepped up to serve their countries.

If we self-censor, it's no wonder that the rest of the organization pigeonholes us, and does not expose us to the full breadth of problems that we could usefully address.

6 Technical differentiators

Here are technical data-analysis practices with which statisticians can maximize their contributions. They are also practices that are growing less common outside of statistics, and so they can also be important differentiators.

6.1 Transformations

My standard practice in exploratory modeling is to transform all continuous variables towards normal. I've found that this exhibits numerous benefits.

So many times, in a biomarker context, I've heard, "Look at change from baseline." "Should that change be additive, or percentage change?" No answer from the data scientists.

An aside: percentage change is never ideal, and it should be replaced with difference on log scale. But that's another talk.

But the meat of the answer is: The scale on which the central limit theorem operates, on which many unobserved but presumed perturbations accumulate to a normally-distributed observation, is the right scale for difference. If a log transformation works, then the right difference is an additive difference on log scale. If data are proportions, then a difference of logits is log odds.

I've noticed that there is a growing unease with transformations. Increasingly, non-statisticians view transformation as a suspicious tinkering with data. I think this is because they lack statistical training. This is an area where statisticians need to push, and can differentiate themselves.

Certainly, if there is a clinical standard for transformation or lack of it, we should report results accordingly, but I recommend applying transformations to get the most effective possible analysis and then back-transform results if necessary.

6.2 Model multiple contributors

I've actually heard from more than one data scientist, in the context of biomarker exploratory analysis, "We're interested in the biomarker. We should calculate a test for the effect, ignoring all other factors, even if there's another factor that may contribute."

This is wrong. Deeply. This is logically equivalent to running an experiment in which you knowingly fail to block on factors known to be important.

In biomarker exploratory analysis, I've consistently pushed to include important clinical covariates—at the very minimum, an assessment of how sick the subject is. Including such information almost always sharpens results.

Fitting models with multiple factors, even just a few, may sound so basic as to be not worth mentioning, yet I find that it is a novel concept for some non-statisticians, and so is an important differentiator.

6.3 Analyze continuous data

I wrote a long blog piece on this, which you can find here.

You've heard that binning continuous data loses information, and that therefore a data analysis with continuous data will be more efficient than the analysis with binned data.

However, that isn't the only aspect of the story. In fact, it's not the main aspect of the story. In fact, it may not even always be true. If you bin a continuous variable into tertiles, you add two d.f. to a model; if you include a continuous variable and add a spline basis expansion, you'll add two, three, or four d.f. The two approaches may in fact give similar power.

The real benefit of analyzing continuous data is that it is closer to Nature. If you are brought into a problem and want a data analysis to uncover, as quickly and expeditiously as possible, "What is going on," binning data is adding an insulating layer between your eyes and the phenomenon.

The story in the blog concerns a diagnostic assay used in a therapeutic drug trial. Examing binned assay results and ignoring the underlying continuous results caused a delay in the trial, cost hundreds of thousands of dollars, and sucked up hours from senior executives who must have had better things to do (one would hope). All for no very good reason, it turned out.

I strongly encourage all statisticians to analyze continuous data whenever possible. However, this leads to complications which must be dealt with:

- Contributions to model functions may not be linear.
- Outliers may perturb estimates.
- Variance may not be constant.
- Distributions may not be normal.

Many statisticians look at that list and go back to binning.

My recommendation to analyze continuous data is therefore a recommendation to learn tools to handle the complexities above. These tools continue to evolve and this is a life-long process.

People working in many topic areas are prone to bin data; in the clinical setting, the practice is almost reflexive. Therefore this can be an important differentiating factor.

7 Work practice differentiators

7.1 Statisticians think differently

One day an engineer I had worked with stopped by my desk with a problem. He was a highly experienced and thoughtful person, you must understand, as you'll see. He was developing an internal tool to measure the fluorescence of liquids. This involved putting liquids in a chamber called a cuvette.

He was finding that his fluorescence measurements were noisy. Being an engineer, he sought a fix, and found an optical filter to place in the path of optical measurement. This did indeed reduce the noise, but he found that every time he placed the filter, it shifted the baseline of the measurement. Did I have any ideas?

"Can you measure replicates?"

"I can get about a thousand per second."

"What if you averaged about three seconds' worth of measurements?"

He nodded and went off to try it. And that was the measurement system.

I'm not the first person to note that engineers try to fix things, while statisticians tend to adapt to things. And fixing things is quite often the best path forward. But do take a moment to consider that your statistical training, and quite probably an inherent inclination towards probabilistic thinking that led you to statistics in the first place, gives you a perspective that's different from that of many other people.

Lean into it. It's a valuable perspective. It will compete with the engineers' perspective, and it won't always win. It shouldn't always win. But by putting it on the table alongside the other perspectives, you're giving your organization the opportunity to adapt, to select the best perspective for the given moment. If your voice isn't there, your perspective will never be available.

7.2 Be the problem solver

Sometimes statisticians can be too eager to please to serve their organization well. Have you seen any of these patterns?

- The statistician gets a highly specific request, and delivers exactly that request, without investigating alternative pathways.
- The statistician carries out multiple permutations of an analysis, such as combinations of subsets of data, in order to address any anticipated question.

- The statistician creates an online RShiny web app that lets her collaborator choose alternative analyses.

Taken alone, none of the above are bad practices. However, they do not entail understanding the one or a few key decisions the team is trying to make.

Fulfilling a specific request turns off your statistical judgement. You're not bringing your professional judgement to the table, you're bringing your coding skills.

Generating lots and lots of statistical reports isn't by itself addressing the key questions. Similarly, building a software tool that lets the collaborator explore aspects of the analysis themselves doesn't provide statistical guidance.

If your team is investigating multiple subsets, for instance, don't just generate results for all the subsets. Also use your statistical expertise to inform the team which subsets they should pay attention to.

Use your statistical expertise to make a recommendation regarding the key decision. What do the reams of statistical results say about the one or two things the team should do with the information?

7.3 Build a relevant toolbox

When I started working in industry, I found myself in an organization that manufactured instrumented assay systems. These took in samples, carried out autonomous processing and interrogation, and returned an answer: positive, negative, susceptible to such-and-such antimicrobials, etc. The idea of a machine that processed noisy data to make a decision (a statistical act!) without any human intervention fascinated me.

I began to read all I could about any applicable statistical models and machine-learning methods. I went to conferences, when allowed, just to learn what others were doing. These conferences were not in disease areas that I worked in, and they had no clear connection to my work area–except that people were discussing data-analytic methods that had potential in our area. I bought and read books (great when I was spending down time waiting for my toddler children to go to sleep!).

I would hear about a new method that purported to do something much better than the standard methods, and had to investigate them.

Importantly, I tried out these methods. Many methods were embodied in R packages, and those packages included examples in their documentation. Trying the method out on real data involved simply installing the package and running the example.

I would create my own examples, too. They didn't have to be complicated. So many times I generated a sine wave with added Gaussian noise and fitted a curve-smoothing algorithm, or such a wave included as one of several additive effects. Even simple examples can give you valuable intuition about how an algorithm behaves.

And more often than not, the new method turned out to be more limited or less effective than advertised. I was gullible. It's a good thing I tested these packages out! Nevertheless, I would find perhaps one truly valuable new method per year. Over ten to fifteen years, that's ten to fifteen solidly useful new tools that my colleagues had not yet tried.

Tool development is important. It's important that it be applicable to your work area. Your expanding toolbox shouldn't be sophisticated for sophistication's sake. It should be ever-focused on solving recurring problems in your environment.

7.4 Step up if you're the best qualified

In Section 4.2.1 I told a story in which I didn't step forward because I feared my ideas wouldn't be foolproof. Taking that negative and phrasing it as a positive: look around at who is working on a problem. At the people sitting at the meeting table, for instance. Who is best-suited to contribute to the task at hand? If it's someone else, give them your support. If it's you, step forward.

The question isn't whether you're right or wrong. The question is who has the best framework with which to move forward and interact with the problem. You don't have to have the answers. You just have to have the tools with which to figure out the answers.

8 Collective movement forward

In sections above I've discussed personal strategies with which we statisticians can differentiate ourselves from data scientists (and perhaps also others with overlapping work, such as quality engineers and computer scientists). What can we do collectively?

First, we must collectively assert that any activity that involves learning or making decisions based on noisy data is in the purview of statistics. And any such activity that is critically important deserves a statistician.

We must advocate to our own leadership that the remit of statistics is broad. And then we must prepare ourselves to service that broad remit, and that means stepping out more than we have in the past.

We can't make a claim to the rest of the world that statistics has the most effective approach and fail to challenge ourselves to find the best approach. If we pigeonhole ourselves with our own restrictions about what statistics can do, or does do, we have to be content with the pigeonholes that society puts us into.

There's an analogy here to science itself. In popular usage, we refer to "science" as a body of facts–the "scientific" facts. However, strictly speaking, science is not a body of facts but rather a *method* for interrogating Nature to find facts; these facts are held provisionally until they are displaced by still-more-complete facts. Newtonian mechanics was displaced by Einstein's general theory of relativity and quantum mechanics. These in turn have unresolved issues and are certain to be displaced in the future by something else as yet unknown. No respectable physicist would argue that Relativity and Quantum Mechanics are inviolate; rather, many physicists are searching for that next better theory. Einstein himself was until the day he died. Science never stops searching for the next thing. Similarly, statistics, if understood as the science of optimal decision-making based on noisy data, can never stop searching for optima.

After directing attention at our leaders and ourselves, we must turn our attention to the larger society and its decision-makers. We must explain to them the value of statistics.

It's an easy sell when regulators are involved. Statistics provides a neutral language in which adversaries can engage productively. However, statistics provides value to any decision-making based on noisy data. How can we convince leadership that money spent on those who optimize decision-making based on noisy data will be more than repaid? I'm open to suggestions! But I might start by showing them the history I summarized in Section 3.3, and the principles that were at work.

I'd be happy to hear your ideas, suggestions, and reactions to this essay.