Statistics: expanding our reach by reviewing our roots

James Garrett

October 20, 2025

Introduction

About me

- PhD University of Minnesota
- BD Diagnostic Systems
  - Experimental design, optimization
  - Validations, QC questions
  - Machine-learning, automated decision-making
- Novartis
  - Assay development
  - Biomarker exploratory analysis
    - Shoulder to shoulder with data scientists
  - Set incoming material specifications (submitted to FDA) for first marketed gene therapy, Kymriah
- Recently I've started my own consulting business.
  - Replicate! Statistical analysis and planning

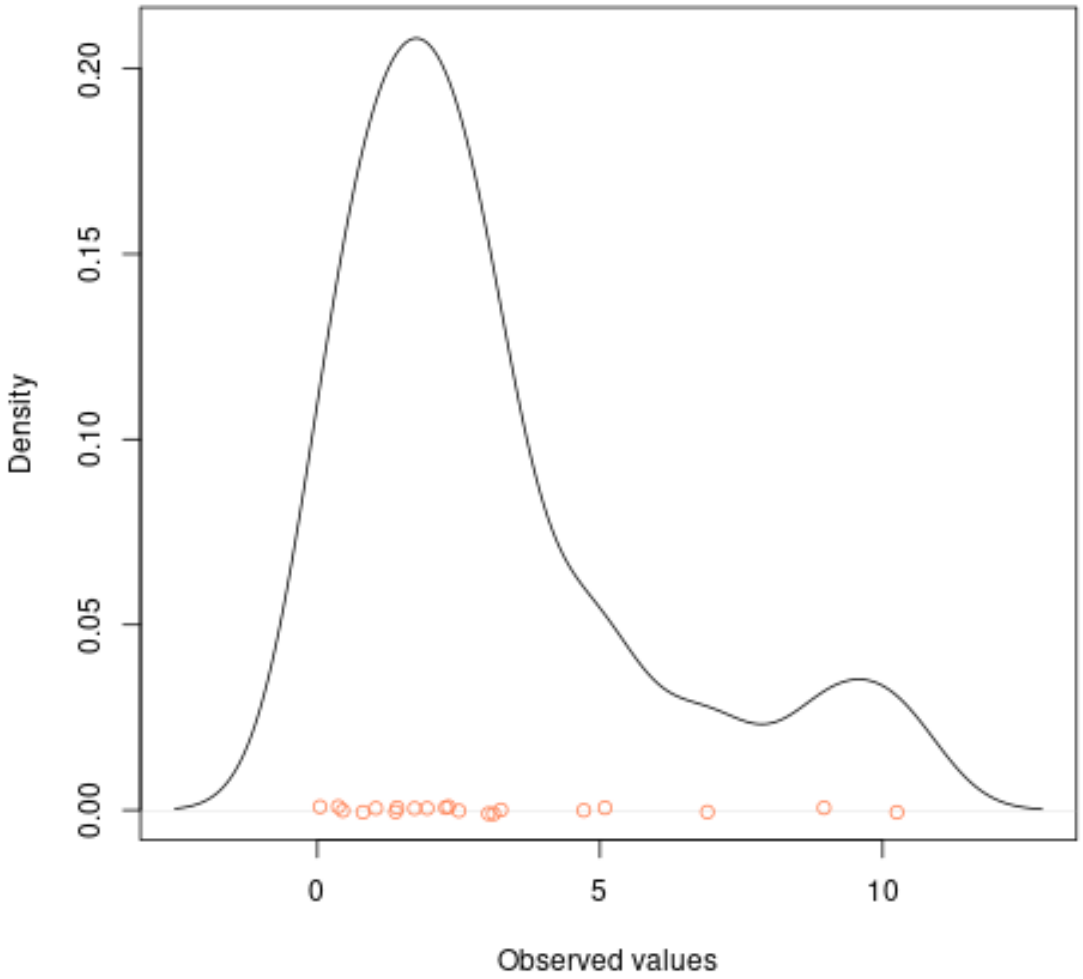
Grounding

What's in a name?

- Statistics is the art and science of picking a decision rule (noisy data $\rightarrow$ decision) that is most replicable and useful, and discovering principles that lead to such rules.
  - And gathering data efficiently.
  - I.e., simply using a data-analysis method doesn't make you a statistician. Using a method for good reasons makes you a statistician.
- Anyone can do anything, if given enough time.
- It's possible that someone is doing statistics without realizing it.

Statistical principles make a difference

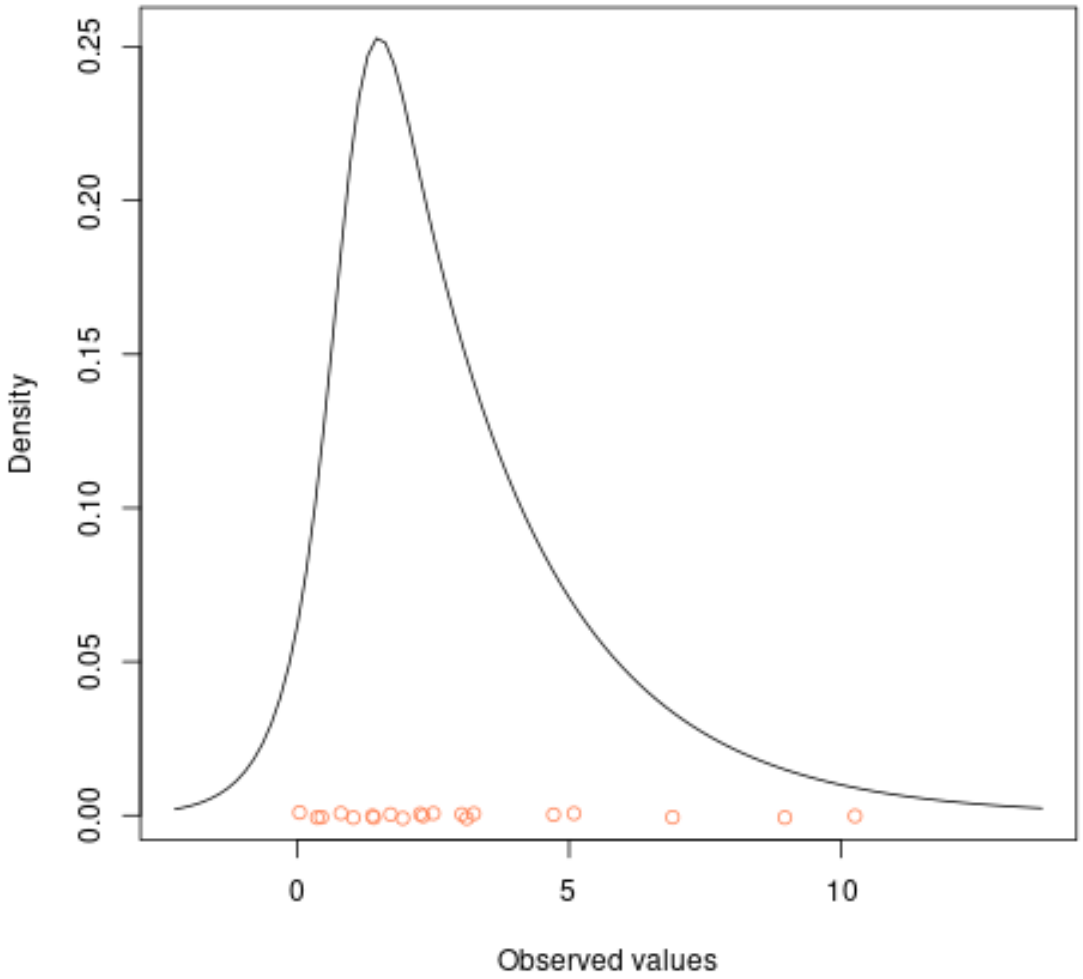
Principles make improvements: kernel density estimate



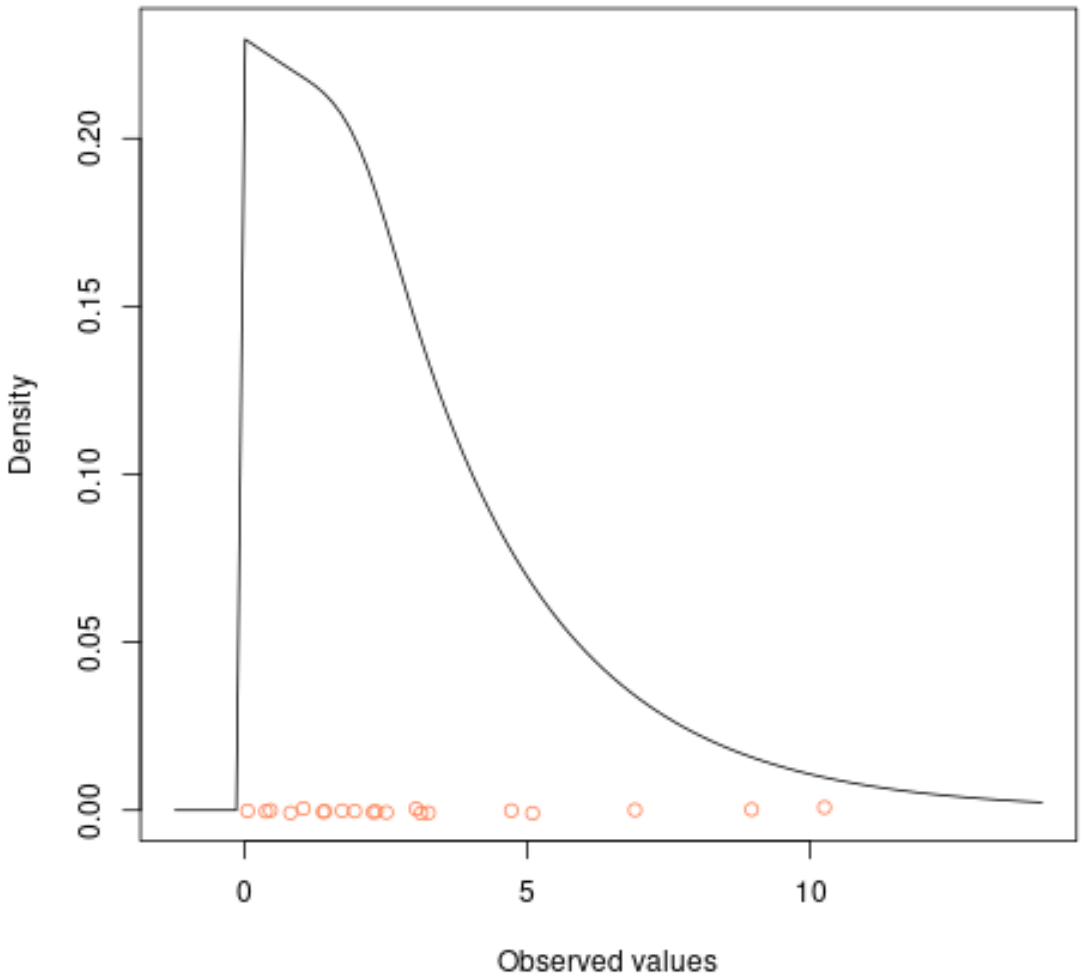
Principles make improvements: "logspline" density estimate

- `logspline` function in R's `polspline` package
- Assume the log density is a cubic spline
- Use a heuristic to place knots sensibly, for a given number of knots.
- For a set of knots, fit coefficients via maximum likelihood.
- Try a range of numbers of knots and choose the one that minimizes BIC.
- In other words, use established statistical modeling principles.

Principles make improvements: "logspline" density estimate



Logspline, but we know that $X > 0$



Logspline

- logspline was created in 1992.
- I used it to set incoming-material acceptable ranges for parameters that could not be transformed to normal, and submitted results to the FDA.
  - I conducted a small simulation study to convince myself that its answers were reasonable for normally-distributed data.
  - They did not respond with any questions about it.

Principles make improvements: Multivariate clustering

Context	Exploratory biomarker modeling
Problem	Many potential clinical covariates of uncertain value
Innovation	R package `VarSelLCM` carries out model-based clustering for mixed data, with variable selection
Contribution	Identified patient clusters that made physical sense and improved biomarker analysis
Reference	"Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark", Preud'homme et. al. (2021), *Nature Scientific Reports*.

Statistical influence in bioinformatics and machine learning

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, variance 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

Statistical principles in action

- Statistical principles "make things work better"
- Among those principles are:
  - Critical thinking concerning performance metrics
  - Curve smoothing
  - Flexible basis elements
  - Shrinkage across multiple instances; (Empirical) Bayesian information sharing
  - "Deconvoluting" multiple additive simultaneous effects
  - Model selection criteria

Statistics cedes ground

Statistics cedes ground

- We ourselves self-censor
  - Collectively
  - Personally
- Application areas
- Tools

Statistical leadership cedes ground

"Data science handles high-dimensional data, statistics handles low-dimensional data and inference"

$\odot$

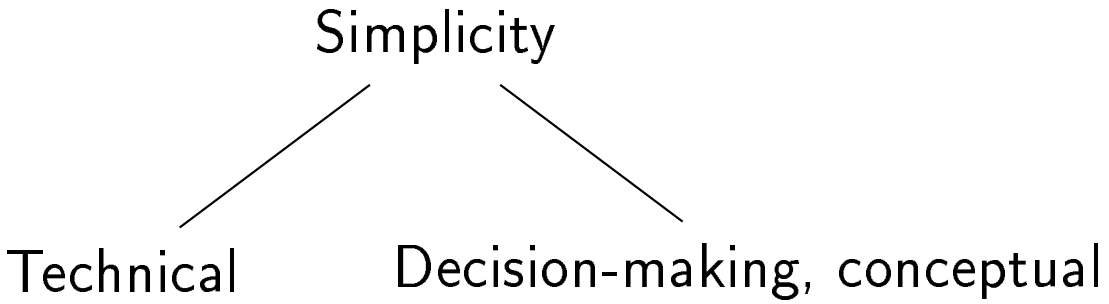
"Data science does machine learning, statistics does simple additive models and inference."

Other ways statisticians cede ground

- "First say nothing wrong." Discomfort in wrestling with complications that could arise.
  - "I'm not familiar with the technology; I won't use it before I study it thoroughly, which takes too long, so I won't."
- We're unaware of our unique perspective; "bystander effect."
- Habituation to "referee" role
  - "If it's not prospective, it has no statistical value."
  - "The sample size is less than what I would prefer, so I recommend we don't do it."
  - "A smoothing parameter is required, which involves some professional judgement, so I won't do it."
- *I don't intend to suggest it's okay for things to be wrong.*

Technical differentiators

"Keep it simple, stupid!"



Simplicity

- Technical and conceptual simplicity are almost always in opposition.
- Decision-making complexity arises from:
  - Nuisance parameters
  - Possibility of nonlinearity
  - Possibility of outliers
  - Possibility of interaction
  - ...
- Technical simplicity $\rightarrow$ Decision-making complexity (often)
- Decision-making simplicity often is based on larger or more sophisticated models (technical complexity)

Aside: Technical simplicity

- It's valuable for knowledge of technically simple tools to be widespread throughout an organization
- If non-statisticians are executing an evaluation, it will need to be technically simple.

Simplicity example: acceptable range specification

- Set incoming material range spec based on 20 historical batches
  - A process change was made after batch 10.
  - Initial visualization indicates no strong evidence of change in variance.
- Options:
  1. Decide whether to combine batch groups before and after change for mean estimation. Then estimate accordingly.
  2. Fit ANOVA model that allows for mean shift. Base conclusions on the mean after change.
     - Bayesian methods allow for mean shift adds no extra difficulty.
     - Optionally, a hierarchical Bayesian approach would allow for data-driven shrinkage between before/after means.
- Option 2 was used to set incoming-material acceptable ranges submitted to the FDA.
  - The FDA eventually asked one question clarifying a detail. They had no issue with it.

Bayesian methods and simplicity

- Conceptually simple, technically complex solutions often *accommodate* $\rightarrow$ Bigger models
- Hierarchical Bayesian models can allow larger models to be well-behaved, and provide inference with integrity.

Simplicity

- Differentiate/contribute by pursuing conceptual / decision-making simplicity.

Transformations

- Transformation is a powerful tool
- My rule of thumb: transform continuous predictor variables towards normal
  - The scale on which the Central Limit Theorem operates is also the scale on which to model, view, and think about data.
  - (You'll probably need to back-transform results.)
- Non-statisticians are growing uncomfortable with transforming data, presumably due to lack of confidence.

Model multiple contributors

"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 an argument I've heard more than once from data scientists.
  - It is very, very, very wrong.
  - If you believe it, you should also not include blocking variables in experimental designs.
- We statisticians are familiar with models that involve multiple factors simultaneously.
  - When does Nature *not* involve multiple factors acting simultaneously?
- You're probably already doing this one; keep it up!

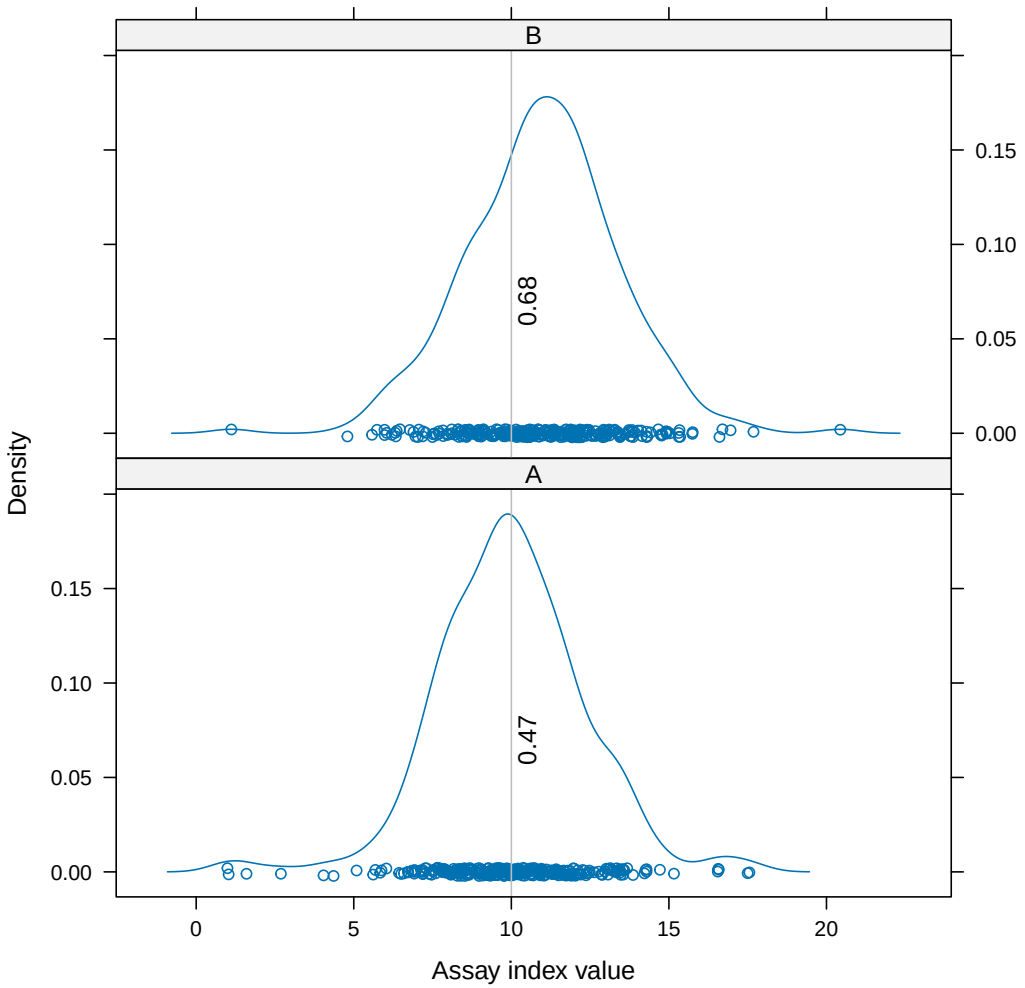
Analyze continuous data

or

A tale of two assay laboratories

(Find narrative here.)

A tale of two assay laboratories



Lesson

- Binning quantitative data is statistically inefficient, yes, but...
- More seriously, it obfuscates data Nature presented, potentially hobbling teams.

Dealing with continuous data requires tools

- Continuous data requires additional assumptions, which may require expanded flexibility.
- Distributional assumptions →
  - Transformations
  - Flexible distributional families (GAM-LSS)
  - Density estimation
  - Resampling methods (bootstrap and permutation testing)
- Linearity / functional form assumptions →
  - Curve smoothing, generalized additive models
- Outliers →
  - Robust estimation
- Simply taking on the challenge of analyzing continuous data will make you a different statistician!

Work practice differentiators

Be aware that statisticians think differently

- A story (see companion handout for written narrative)
- Thinking probabilistically is not universal.
- We don't always realize our perspective is different, much as fish don't perceive water.
- Your unique perspective may be what the team needs, and it isn't going to come from someone else.

Be the problem-solver

- Problem solver, not service worker.
- Understand your team's problem.
- Develop and promote solutions.
  - Rather than delivering requests.
  - Rather than delivering a large package of results without a clear conclusion or recommendation.

Build a relevant toolbox

- Be aware of problems that arise consistently in your work area
- Be on the lookout for solutions
- Read books and attend conferences where solutions to those problems are being proposed
  - These may be *outside* your work area!
  - Avoid "ghettoization"

Collective movement forward

Power and responsibility

- We should be aggressive in reclaiming territory where statistics can contribute usefully.
- Accordingly, we must also aggressively challenge ourselves to find or develop the most flexible and capable tools.
- And we must build the case that it's worth budgeting for statistics in diverse areas.

Thank you!

- Thank you for listening!
- Thanks to conference organizers for the opportunity.
- Thanks to Irina Gershgorin for discussing these slides.