Statistical Analysis of Data in the Biotech Industry

Francois Collin

November 04, 2025

Contents

Preliminaries - Leo Breiman

Before machine learning became a household name for engineers and computer scientists interested in making predictions, before AI had any developed applications, and before the name “Data Science” was popularized to describe an area of statistical application which most statisticians ignored, Leo Breiman (**??**) was calling for the field of statistics to modernize or become irrelevant.

Breiman, L. (2001). Statistical modeling: The two cultures. Statistical Science 16, 199–215 is a must-read.

My recent experience working in the context of specifying and validating an omic pipeline used to produce a molecular diagnostic test made it clear that even without AI, DS, and ML, statistical methods needed modernizing: the usual approach of using a named statistical procedure off the shelf will not yield accurate results in most cases - the esoteric error distributions which characterize modern datasets must be accounted for when analyzing these data. See Where’s the Randomness? for a description of this context.[1]

Inspired and encouraged by Leo Breiman, Bin Yu has modernized the concept of variability in data to reflect the character of modern day data collection and analysis. The PQRS framework (Yu and Kumbier, Front Inform Technol Electron Eng 2018 19(1):6-9)) highlight 4 elements of statistical practice which remain crucially important for the successful analysis of modern day datasets: population (P), question of interest (Q), representativeness of training data (R), and scrutiny of results (S). These concepts have always been important, but very often neglected in practice.. These concepts have always been important, but very often neglected in practice.

1 Statistics Training

1.1 Training Program for Statisticians - a collaboration with claude.ai

As an exercise working with claude.ai, we design a training program for experienced statisticians. The professional development training is meant to bring statisticians up to speed with novel statistical methods and techniques which may be outsite of their expertise or sector of industry.

- stats-pd-plan-final-v35.html - final version of the professional development program.
- stats-pd-plan-change-log.html - detailed change log documenting the development process
- ai-assistant-guide - outline principles which apply broadly to working with AI assistants on complex projects.

[1] Why the use of canned procedures persists is easily explained in the context of a client who is uncritical of the form of any proposed solution while being unflinchingly insistent on the timely release of the said solution.

1.2 Specific Training Resources

Generative AI

Resources - online courses, published papers, books, slide presentations - to learn about the areas of applications and development of AI that are most important to statisticians. These should include a basic set of learnings which all professionals interested in harnessing the powers of AI should know. In addition to this basic set, learnings that are most relevant for the integration of AI with statistics.

Yu and Kumbier (Front Inform Technol Electron Eng 2018 19(1):6-9) (**?**) provide a framework for integrating statistical ideas in AI work which they term the PQRS Workflow. In this framework, the statistical concepts of population (P), question of interest (Q), representativeness of training data (R), and scrutiny of results (S) remain critically important in the application of AI to data analysis. We should include any training which can help clarify the framework to non-statisticians, or help statisticians explain the framework to non-statisticians.

If it helps to limit resources or choose among similar options, we can assume that the statisticians work with bio-tech companies which market molecular diagnostics devices which use omic-wide[2] profiels as input.

- Report: Comprehensive Introduction to Generative AI Resources

- Resources:
  - Gen AI - ChangeLog
  - Gen AI - Instructions

Deep Learning Models

Resources - online courses, published papers, books, slide presentations - to learn about the development deep learning theiry ands applications that are most important to statisticians. These should include a basic set of learnings which all professionals interested in harnessing the powers of Deep Leaning should know. In addition to this basic set, learnings that are most relevant for the integration of Deep Learning with statistics.

If it helps to limit resources or choose among similar options, we can assume that the statisticians work with bio-tech companies which market molecular diagnostics devices which use omic-wide[3] profiels as input.

- Report: Comprehensive Resources for Deep Learning in Statistics & Biotech

- Resources:
  - ChangeLog
  - Repository Setup
  - Instructions for Rendering Reports and Changelog
  - Generating HTML Reports from Markdown
  - Markdown Template

2 Data (Re-)Analysis by Topic

2.1 Reference-Free Genomic Inference with SPLASH and OASIS

- Presentation

[2] transcriptomic, genomic, proteomic, methylomic
[3] transcriptomic, genomic, proteomic, methylomic

  - Summarizes the key points made in the SPLASH and OASIS papers
  - Includes comments on alternative approaches or interpretations **but no additional analyses**

- SPLASH - a detailed look at the SPLASH article.

- OASIS - a detailed look at the OASIS article.
  - INCOMPLETE

2.2 scRNAseq UMI count normalization

- slide deck: OneDrive folder of Speed Lab Meeting: https://outlook.office365.com/mail/group/wehi.edu.au/speedlabmeeting/files
- Reference:
  - Ahlmann-Eltze and Huber (2023) (**?**)
- Other:
  - Melms et. al. (2021) (**?**)

2.3 RUV

Normalization Lit Review

- To Be Completed with an appropriate assistant.

The latest RUV

- Link to Ramyar Molania presentations.

Beyond RUV

Can the methodology which enables the basic RUV analysis be adapted to other applications?

- IUV - identification of unwanted variability
  - the automated high throughput processing of samples necessary to get the omic based readouts used to diagnose samples is carried out by a pipeline which strings together numerous steps (typically hundreds).
    - eg. Cell-Free DNA Methylation Profiling Analysis
  - As samples proceed through the pipeline they will be organized in various configurations giving rise to various groupings:
    - plasma isolation
    - cfDNA extraction
    - amplification, conversion, …
  - measurements are recorded as samples travel through the pipeline, each measurement potentially having its own shared variability or dependency structure
  - a real problem is to identify the factors giving rise to excessive variability in the downstream read-outs

- QUV - quantification of unwanted variability
  - can RUV methodology be adapted to provide quantification of reproducibility?

2.4 Miscellanious

- IUV
- Digital Driplet PCR - soon
- Randomness in sample calls - soon
- interimSampleSizeAnalysis - soon
    - [_M3A_study_design_CB_notes.pdf]
- nanopore_talk/microbiome_slides.html
    - nanopore_talk/nanopore_background.html
    - nanopore_talk/umap_explained.html
    - nanopore_talk/pdf_image_extraction.html

3 References

4 Appendix:

Challenges in Statistical Analysis

‘

Types of errors:

- The question is wrong or inadequately posed
    - this is not that infrequent and is largely caused by folks phrasing the question in a form that anticipates the solution. eg, client expresses desire to know what a particular estimated regression coefficient is
    - solution is easy: subject matter experts should think hard about the question which is then phrased using appropriate subject matter specific language.

- An inappropriate model is used as the analysis framework to address a question
    - this is either due to common practice or inadequate statistical training
    - David Freedman discussed the inappropriate use of statistical models in the social sciences (**????**)
    - Inappropriate modeling is not limited to the social sciences
        - in biotech, the unchecked use of logistic regression whenever the response is binary is ubiquitous.
            - the results of such analyses could be misleading due to the biases illustrated in Freedman (2008) (**?**)
    - Leo Breiman pointed to inadequate training as the cause of the reliance on **standard textbook methods** which have a limited range of applicability - Breiman: (**??**)
    - as regression models are used to answer almost every question that come up in the internal analyses conducted by companies in industry, many are going to be flawed.
        - For a review of how regression models go wrong see (**???????**)
    - clinical validation studies for complex omic Dx instruments that are designed according to FDA guidelines will make use of the classical Neyman-Pearson hypothesis testing context which is often invalid:
        - the samples in the study may not be representative of the target population.
            - there is rarely any explicit randomness in the selection of the study samples so inference to any superset of the study samples is iffy at best.