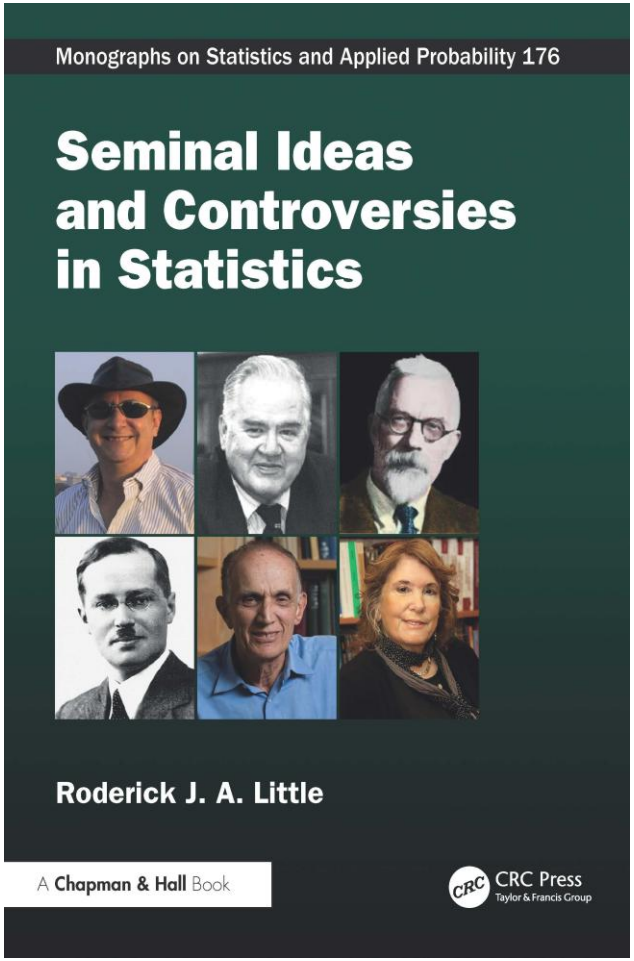
Seminal Ideas and Controversies in Statistics

Rod Little

The book

Little, R.J. (2025). *Seminal Ideas and Controversies in Statistics*. Chapman Hall/CRC Press



Statistics has developed as a field through seminal papers and fascinating controversies. The book covers a wide-ranging set of fifteen statistical topics, grouped into three parts:

Part I, chapters 1-6. Philosophies of statistical inference
Part II, chapters 7-12. Advances in statistical methodology
Part III, chapters 13-15. Randomization in statistical design

25% discount code

SICS26 at checkout here on any version at:

https://www.routledge.com/Seminal-Ideas-and-Controversies-in-Statistics/Little/p/book/9781032493565

Eight Major Themes {corresponding chapters}

The 15 chapters can be arranged around 8 themes:

1. Estimating equations vs likelihood-based methods {1,10}
2. Frequentist vs Bayesian inference {2-6}
   - Conditionality principle/role of ancillary statistics
   - Caibrated Bayes: a good frequentist/Bayes compromize
3. Models with random effects {7,8}
4. Multiple comparisons {9}
5. Impact of advances in computing power {11}
   - Bootstrap, EM algorithm, Bayesian Monte Carlo methods
6. Exploratory Data Analysis and Data Science{12}
   - Parametric inference vs prediction
7. Random Sampling {13}
8. Causal inference and Randomized Clinical Trials {14,15}

Omit here (No time!)

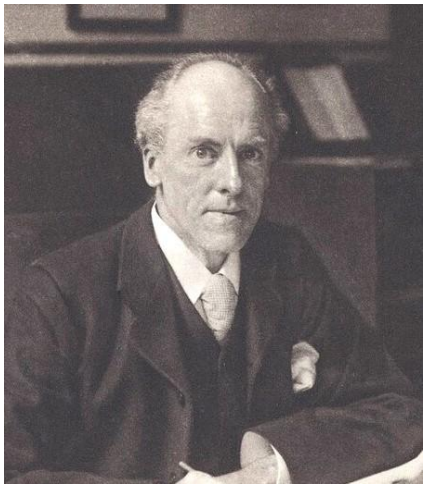
Theme 1: Estimating equations vs likelihood methods

- Fisher, R.A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A,* Vol. 222, pp. 309-368
  - The topic of Chapter 1
- Liang, K-Y. & Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73, 1, 13-22.
  - The topic of Chapter 10

Fisher (1922)

R.A. Fisher

Until Fisher's landmark paper, the main general approach to estimation was Karl Pearson's method of moments (MM), based on estimating equations that equate sample moments (sample mean, sample variance, etc.) to their expectations under an assumed model



Karl Pearson

Fisher proposed maximum likelihood (ML) as an alternative. He showed that ML satisfies three important properties of statistical procedures – Consistency, Efficiency and Sufficiency. He also introduced the Fisher information, a key measure of the information supplied by a sample, and illustrated the application of ML to a variety of problems.

Fisher (1922)

Fisher describes estimating the population median from a sample from the Cauchy distribution as an example where ML works and the method of moments fails

> … because the Cauchy distribution doesn't have any moments!

As Stigler (2005 *Statist. Sci.*) describes, Fisher's paper created some bad feelings between these two statistical giants.

More on ML vs GEE

- The GEE methodology in Liang and Zeger (1986) can be viewed as a descendant of the Pearson MM approaches to estimation discussed in Fisher (1922).

  Pearson $\rightarrow$ method of moments $\rightarrow$ GEE

  Fisher $\rightarrow$ likelihood $\rightarrow$ maximum likelihood, Bayes

- These lineages characterize two major alternative schools of thought in statistics:
  - GEE: less specification and limiting explicit assumptions; can trade efficiency for *robustness*
  - Likelihood / full probability modeling: more specification, but makes assumptions explicit
  - A recurring question here is: are assumptions good or bad?

Fisher and Bayesian methods

Fisher (1922) railed against Bayesian methods, noting the lack of uniqueness, and inability to capture "ignorance" in the prior

However, I think Fisher's philosophy of statistics is quite Bayesian in spirit:

1. Both Bayes and ML involve the data through the likelihood function, and ML can be viewed as a form of large sample Bayes

2. Fisher recognized the need to condition on ancillary statistics, a strength of Bayesian methods – the posterior distribution conditions on all the data. I return to this point in Theme 2.

3. Fisher was above all a scientist and statistical modeler

4. He later attempted to "make the Bayesian omelet without breaking the Bayesian egg" with the method he called fiducial inference – ultimately a failure, as I discuss in Chapter 4.

ML is marvellous, but Bayes is better!

- The theory of ML assumes large samples, and Bayesian inference under suitable priors is better than ML in small samples:
  - Bayes under Jeffreys' priors yield degrees of freedom and student t corrections for normal models, and improve on the standard Wald confidence intervals for proportions
  - For the Fisher-Behrens problem, namely inference for the difference in means in two independent normal samples with different variances, Ghosh and Kim's (2001 *Can. J. Statist.*) Bayesian model yields better frequentist answers than the usual approaches proposed by Fisher and Welch, and should be applied more often (see Chapter 4)

ML is marvellous, but Bayes is better!

- The inclusion of a proper prior distribution increases modeling flexibility
- Adding a suitable proper prior distribution can improve inferences for problems with a large number of parameters and a relatively small sample size.
- See for example, empirical Bayes shrinkage (Chapter 7) or the ridge prior for multiple regression (Chapter 8)
- These modifications improve small sample inferences, but do not affect the properties of Bayes inferences in large samples, which parallel those of ML

GEE for longitudinal data

- ML inference for two particularly useful classes of models was available in widely-distributed software in the 1980's:
  - Normal model for repeated measures $y_i = (y_{i1}, ..., y_{ik})$
    $$(y_i \mid X_i, \beta, \phi) \sim_{\text{ind}} N_K\left(X_i\beta, \Sigma(\phi)\right)$$
  - Generalized linear models (GLIM, e.g. Baker and Nelder 1978) extensions of normal linear models for non-normal independent data.
    $$f(y_i \mid x_i, \beta, \phi) = \exp\left[\left(y_i\theta_i - a\left(\theta_i\right) + b(y_i)\right)\phi\right]$$
  - ML for non-normal (e.g. binary, count) repeated-measures data was limited to particular cases.
- Liang and Zeger (1986) described a useful extension of the GLIM system to non-normal repeated-measures data
  - Estimation by Generalized Estimating Equations (GEE) not ML.

Robustness property

- Ignoring dependence and simply assuming iid residuals leads to consistent estimates, and se's can still be computed by (for example) applying the bootstrap to whole cases (not just the individual repeated measures)
- Liang and Zeger point out that estimating under a working covariance structure can increase efficiency of estimates
- Their method has a form of robustness, in that the working covariance structure does not have to be correctly specified to get valid estimates, and estimates of uncertainty

Liang and Zeger (1986)

- With increased computing power, we now have ML and Bayesian methods for correlated normal data, through generalized linear mixed models (GLMM, e.g. PROC MLMIXED in SAS).
- These approaches make more explicit distributional assumptions, but have some advantages, e.g.
  - do not assume missing data are missing completely at random (though weighted GEE also relaxes this assumption
  - Bayes propagates error better in small samples – GEE is basically asymptotic.
  - GLMM's have more potential problems with misspecification (Hubbard et al. 2010 *Epidemiology*)

Time-varying covariates

- Be clear about conditioning!
- Pepe and Anderson (1994 *Comm. Statist. Sim. and Comp.*) warn that Liang and Zeger (1986) implicitly assume that if $x_i = (x_{i1}, ..., x_{ik})$,

$$f(y_{it} \mid x_i, \beta, \phi) = f(y_{it} \mid x_{it}, \beta, \phi)$$

- This is a strong assumption which limits the utility of the model for causal inference

Theme 2: Frequentist vs Bayes

- When I was a student at Imperial College in the 1970s, debates raged over frequentist vs Bayesian inference
  - The debate was somewhat theoretical, because at that time Bayes was limited in practice by computational limitations
- Bayes bases inference on the posterior distribution, which by Bayes' rule is proportional to the likelihood multiplied by the prior distribution
- Frequentist inference is based on properties of statistics in repeated sampling – hypothesis tests, confidence intervals

A key issue in the frequentist-Bayes debate: whether to condition on ancillary statistics

- Frequentist inference is based on repeated-sampling properties of statistics:
  - Hypothesis testing: Goals are close to nominal size, good power against realistic alternatives
  - Confidence intervals: goals are narrow intervals with nominal or conservative coverage
- Question: what is the appropriate set of repeated samples … the *reference set* … for frequentist calculations?
- In particular, should this set *condition on ancillary statistics*?
- Note that it's not an issue with Bayesian methods because the posterior distribution conditions on all of the data

Ancillarity and conditionality

See e.g. Ghosh, Reid and Fraser (2010 *Statist. Sinica*)

1. (y, a) minimal sufficient for θ: a is ancillary if

$$f(y, a \mid \theta) = f_A(a) f_{Y|A}(y \mid a, \theta)$$

2. (y, a) minimal sufficient for θ, ϕ: a is S-ancillary for θ if

$$f(y, a \mid \theta, \phi) = f_A(a \mid \phi) f_{Y|A}(y \mid a, \theta)$$

Conditionality principle states that inferences about θ should condition on (S)-ancillary statistics a

Fisher's flowering plants

- Fisher invoked the following example to justify conditioning on ancillary statistics:
- A germinated flower has two colors, purple or red

θ = Pr(plant is purple). $H_0 : \theta = 0.5, H_a : \theta > 0.5$

ϕ = Pr(plant germinates) [Suppose ϕ is known to be 3/4]

n = number of plants

a = number of plants that germinate, y = number that are red

Data: $n = a = y = 4$

a is then ancillary for θ

Flowering plants

- The P-value conditioning on a is
  $\Pr(y = 4 \mid a = 4, n = 4) = 1/2^4 = 1/16$
- The P-value averaging over a is

$$\Pr(y = 4, a = 4 \mid n = 4) = \Pr(y = 4 \mid a = 4, n = 4) \times \Pr(a = 4)$$
$$= (1/16)(3/4)^4 = (1/16)(81/256)$$

because all other combinations of y and a are less extreme. Thus, including the probability that all four plants germinate reduces the P-value by 81/256; but Fisher argued:

"what if someone else had discovered how to get his plants to flower every time? He would, surely, justifiably complain if he, getting the same result, had it judged non-significant at 5 per cent, just because of his skill in horticulture?"

Flowering plants

- Fisher concluded that the P-value for the reference set that conditions on *a* is the appropriate one. George Barnard “after long meditation” was forced to agree, leading him to abandon his CSM test independence in a 2x2 table. Fisher asserted that "Barnard is the only statistician who has ever admitted he was wrong.“

Did that resolve the issue? Apparently not! According to latest Wikipedia

- “Under specious pressure from Fisher, Barnard retracted his test in a published paper, however many researchers prefer Barnard’s exact test over Fisher's exact test for analyzing 2 × 2 contingency tables, since its statistics are more powerful for the vast majority of experimental designs, whereas Fisher’s exact test statistics are conservative.”

Tests of independence in a 2x2 contingency table

One-sided test in 2x2 table, $\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \text{odds ratio}$

$H_0 : \theta = 1, \quad H_a : \theta > 1$

$P = \Pr(T \geq t_{\text{obs}} \mid H_0)$

The Pearson chi-squared test (C) and the Fisher exact test (F) or its approximation, the Yates continuity-corrected chi-squared test (Y), give similar results in large samples, but differ in small samples

Yates (1984 *J. Roy. Statist. Soc. A*) discussed these alternative tests, 50 years after his paper on the continuity-corrected chi-squared test

The choice remains contentious, over 40 years later

Issue of ancillarity plays a key role…

Independence in 2x2 tables

- F and Y are conservative when one margin is fixed (as is common in many practical designs):
  - The test statistic for F or Y is "less extreme" than for P and B. In particular, Yates' continuity correction makes the chi-squared statistic for Y smaller than the chi-squared statistic for P.
  - With discrete data, hypothesis tests for a fixed nominal size are inherently conservative. This conservatism is more pronounced for F and Y than for P and B.
- On the other hand, F is an exact test if both margins are fixed
- Should we condition on second margin or not?
  - It's not clear – the second margin is approximately, but not exactly, S-ancillary for odds ratio

Bayes posterior prob vs p-value

Adopting a Bayesian approach for the odds ratio avoids the conditioning issue, because the posterior distribution of the odds ratio conditions on all the data.

But … the Bayesian answer depends on the choice of prior

It can be shown that the posterior probability under an independent Jeffreys' prior for the two proportions is very close to the p-value obtained from the standard Pearson chi-square statistic C

However, Howard (1998 *Statist. Sci.*) argues persuasively that the priors should be dependent, and this leads to a Bayesian answer closer to the P-value from F or Y (!)

So he ends up favoring F for frequentist inference, though not for Fisher's reasons.

Calibrated Bayes

- Much of the frequentist and Bayes literature is like an estranged brother and sister who don't talk to one another
- As discussed in Chapter 6, two notable papers by Box (1980 *J. Roy. Statist. Soc. A*) and Rubin (1984 *Ann. Statist.*) present unified approaches that attempt to capture the best features of both philosophies

Bayesians should be frequentist

- Bayes is optimal if the model and prior are well specified, but
- "all models are wrong"…
- … and Bayes may give very poor inferences if the model is misspecified
- Thus Bayesian inferences, such as posterior credible intervals, should be *well calibrated*, in the sense of having approximately the nominal coverage when regarded as confidence intervals

Frequentists should be Bayesian

- Good frequentist properties are useful…
- … but do not necessary imply that an inference is appropriate *for the data set actually observed*.
  - Rubin calls this "scientifically justifiability"
  - Frequentist analysis does not necessarily condition on all the relevant information, such as ancillary statistics

Calibrated Bayes

Activity	Bayes	Frequentist
Inference under assumed model	strong	weak
Model formulation / assessment	weak	strong

Bayes for best for the inference

Frequentist properties are best for model development and assessment (enriched by Bayesian ideas)

Seek Bayesian models that yield inferences with good frequentist properties

Summary of Theme 2

- The calibrated Bayes philosophy is a useful perspective on statistical modeling that brings out the best features of frequentist and Bayesian paradigms
- The methodological issues in Parts 2 and 3 of the book – repeated measures models, random effects models, regression alternatives, survey inference, data science – all benefit from applying calibrated Bayes ideas.

Theme 3: Random Effects Models

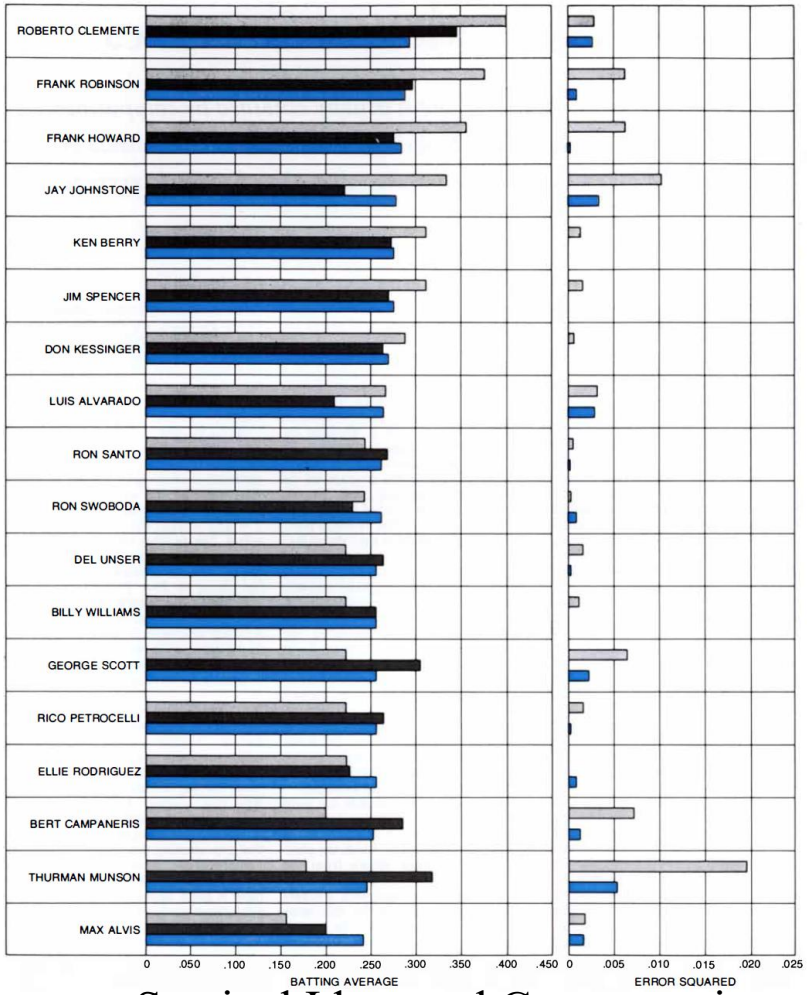
Efron, B. & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 1977, 119-127.

Dempster, A.P., Schatzoff, M. & Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares (with discussion). *J. Amer. Statist. Assoc.*, 72, 357, 77-106.

Empirical Bayes methods

- Unbiasedness is not always a good property
- Empirical Bayes (or Bayes) estimates are biased but yield better mean squared error.
- Efron and Morris were primary proponents of empirical Bayes methods, and their 1977 paper explains the idea in nontechnical terms to readers of *Scientific American*.
- Their famous baseball example is a simple and intuitive application

Figure 7.1. Batting averages of 18 baseball players. The upper bars are sample means after 45 at bats, the middle bars are the averages at the end of the season, and the lower bars are the James-Stein estimates. The right panel indicates that the James-Stein estimates are closer to the averages at the end of the season for 16 of the 18 players.



Empirical Bayes methods

- The mathematics of Stein's result seems to allow shrinkage of means from unconnected problems, in particular adding an additional data point, the "percentage of foreign cars in Chicago" to the baseball players.
- Efron and Morris describe this as a paradox. But from a modelling perspective, the implied assumption, that baseball averages and percent foreign cars are exchangeable, makes no sense, so I do not view this as a paradox.
- Statistics is modelling, not math; a major theme of Chapter 7.

Empirical Bayes methods

- The Bayesian perspective on random effects models:
  - Unknowns are all random and assigned a distribution to reflect uncertainty
  - Fixed effects have flat priors, random effects have proper priors
- The frequentist perspective is that some parameters are random, some are fixed – I find this confusing (not to say confused).
- See Chapter 7 for examples in ANOVA

Alternatives to Least Squares in Regression

- Multiple linear regression estimated by ordinary least squares does not work well when the sample size is small compared to the number of predictors, and the predictors are highly associated.
- The paper "57 varieties paper" by Dempster, Schatzoff & Wermuth (1977) compares a large number (yes, 57!) of alternatives to ordinary least squares in this situation, by a way of a simulation study that systematically varies the important factors.
- Ridge regression performs the best, although commenters argue that this finding is dependent on the choice of simulation parameters.
- Other alternatives to least squares emerged after this work, notably the Lasso introduced by Tibshirani (1996 *J. Roy. Statist. Soc. B,*), and other "large p, small n" methods that are assessed in the simulation study by Chaibub Neto, Bare & Margolin (2014 *PLoS ONE*).
- Bayesian and frequentist perspectives on these methods are briefly described in Chapter 8.

Theme 7. Random sampling

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. Roy. Statist.Soc.*, 97, 4, 558-625.

Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council investigation. *Brit. Med. J.*, 2, 769-782.

Neyman (1934): Probability Sampling versus "Purposive Sampling"

- Definition of probability sampling:
  - every sample has a known probability of being selected
  - every individual in the population has a positive probability of being selected
- Initially, probability sampling was equated with its basic form, simple random sampling (SRS)
  - Every sample of size *n* has *equal* chance of being selected, hence an equal probability of selection method (*epsem*)
  - Samples of size other than *n* have no chance of being selected
  - With and without replacement

"Purposive Sampling"

- "Non-probability sampling" – but hard to define a negative.
- Units are picked so that sample matches distribution of a characteristic known for the population.
- E.g. if we know distribution of age and gender in population, choose sample cases to match this distribution.
- A common form is *quota sampling*: interviewers are given a quota for each age group and gender and interview individuals until this quota is met
- Becoming increasingly common with increased barriers to probability sampling

The Controversy

- Under simple random sampling, distribution of a known characteristic in the sample can deviate considerably from its (known) distribution in the population, purely by chance
- This “lack of representativeness” led some to prefer purposively picking the sample to match the population distribution

Neyman's "Resolution"

- Neyman (1934) showed that we can get the best of both worlds by <u>stratified sampling</u>:
  - Create strata by the classifying population according to the known characteristics
  - Select a simple random sample of known size n_j from population of size N_j in stratum j
- If $f_j = n_j/N_j$ = const., results in epsem sample, retains probabilistic selection, and sample matches distribution of strata in population
- Also one can vary f_j and weight sample cases by $1/f_j$: Neyman's optimal allocation

More Complex Designs

- Neyman's paper helped to set the stage for other complex design features: cluster sampling, multistage sampling, etc., greatly extending the practical feasibility and utility of probability sampling in practice
- E.g. simple random sampling of people in the US is not feasible – we do not have a complete list of everyone in the population from which to sample
- Work of Mahalanobis, Hansen, Cochran, Kish, ….

Design-based vs Model-based inference

- *Design-based* inference: population values are fixed, inference is based on probability distribution of sample selection. Obviously this assumes that we have a probability sample (or "quasi-randomization", where we pretend that we have one)
- *Model-based* inference: survey variables are assumed to come from a statistical model
  - probability sampling is not the basis for inference, but randomization ensures that the sample selection is *ignorable*.
- *Calibrated Bayes*: build Bayesian models that incorporate features of the sample design, and yield posterior credible intervals that have good confidence coverage
- I have argued that Calibrated Bayes is the best framework

Theme 8. Causal inference

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.*, 6, 1, 34-58.

Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 1, 41-55.

Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council investigation. *Brit. Med. J.*, 2, 769-782.

Polack, F.P. et al. (2020) for the C4591001 Clinical Trial Group. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *New Engl. J. Med.* 383, 2603-2615.

Some key Ideas

- Neyman/Rubin causal model
  - causal effect is a measure comparing outcomes under two treatments
  - only one is observed, corresponding to the assigned treatment
- Ignorable selection, assignment and recording mechanisms
  - Randomized Clinical Trials – assignment is ignorable, a key to insuring internal validity (Chapter 14)
- Propensity score
  - Analog of randomization for observational data (Chapter 15)

A seminal paper

Rubin, D.B. (1978). Bayesian Inference for Causal Effects: the Role of Randomization. *Annals Statist.*, 6, 34-58

- Neyman/Rubin causal model: causal effects as the comparison of outcomes under different treatments
- Formulates causal inference as a missing data problem: only see outcome under the treatment assigned
- Includes selection, assignment and recording mechanisms as random variables in full statistical model
- Role of randomization: makes selection, assignment, or recording mechanisms *ignorable*: if nonignorable, mechanism needs to be modeled

All values in a study of T treatments (Rubin 1978)

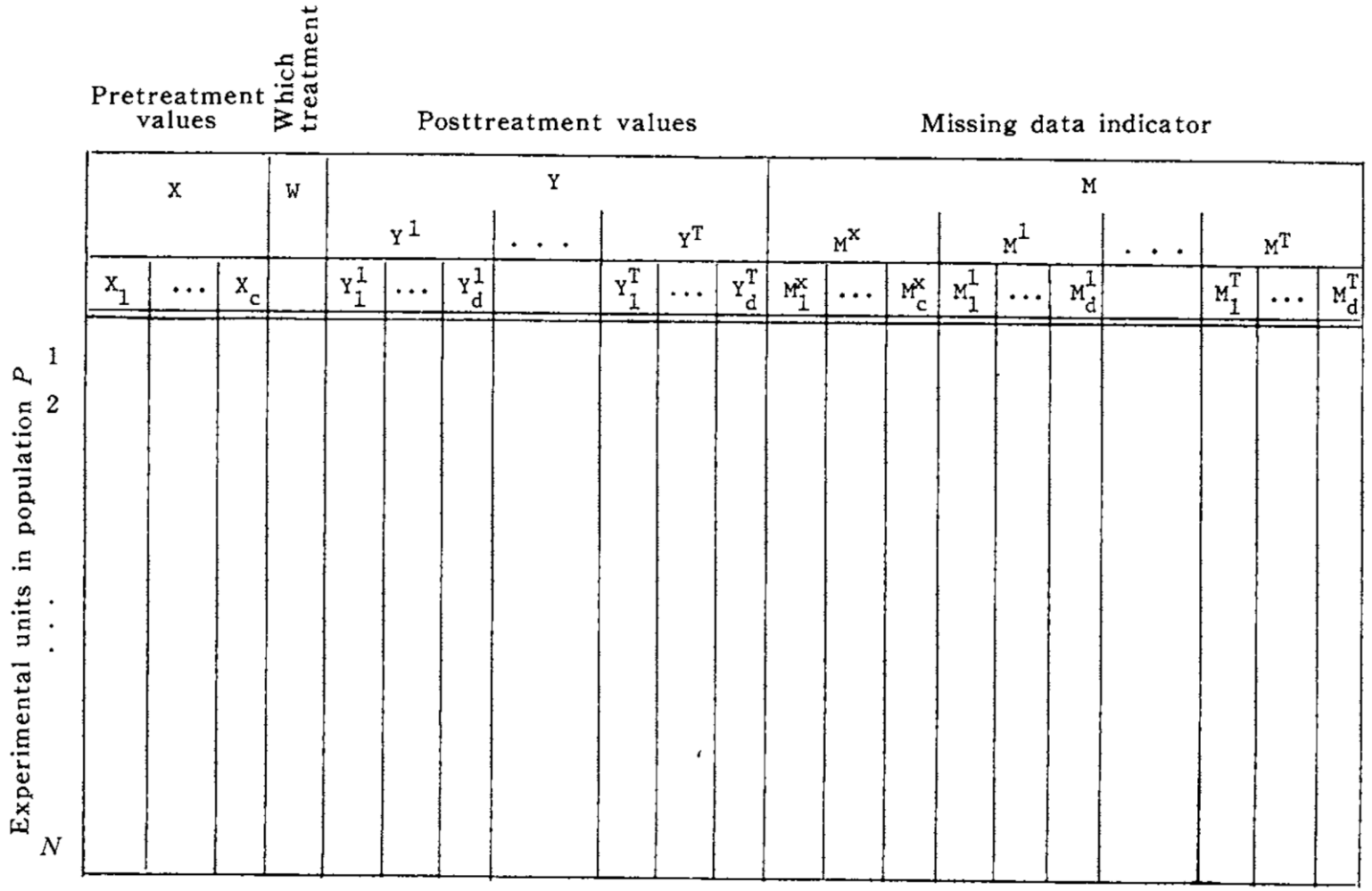


FIG. 1. All values in a study of T treatments.

Randomization eases the modeling task

$X =$ pre-treatment values

$Y = (Y^1, \ldots Y^T), Y^t =$ post-treatment values given treatment t

$S =$ selection indicator

$A =$ treatment assignment (if selected)

$M =$ missingness indicator

Full model:

$$f(X, Y, S, A, M \mid \theta)$$
$$= f_{XY}(X, Y \mid \theta) f_{S|XY}(S \mid X, Y, \theta) f_{A|SXY}(A \mid S, X, Y, \theta) f_{M|SAXY}(M \mid S, A, X, Y, \theta)$$

- Random selection: don't need to model S
- Random treatment allocation: don't need to model A
- Ignorable missingness: don't need to model M
- This is why randomization matters to Bayesians

Conclusion

Sharon Lohr (2025 J Appl. Stats.) writes:

“It is incredibly valuable for students to read the papers in which researchers first struggled with a concept…

Although the book features controversies in statistics, it also has the implicit theme that the principles that unite statisticians are far greater than the controversies that divide them. The statisticians featured in the book may have disagreed on specific issues or philosophical frameworks, but all dedicated themselves to improving the practice of statistics. Little argues that diverse perspectives spur new methodology, and advocates drawing on the strengths of both Bayesian and frequentist approaches to inference.”