Research Article

Winston Lin*, Sandrine Dudoit, Deborah Nolan, and Terence P. Speed

From urn models to box models: Making Neyman's (1923) insights accessible

https://doi.org/10.1515/jci-2023-0073
received October 29, 2023; accepted June 28, 2024

Abstract: Neyman's 1923 paper introduced the potential outcomes framework and the foundations of randomization-based inference. We discuss the influence of Neyman's paper on four introductory to intermediate-level textbooks by Berkeley faculty members (Scheffé; Hodges and Lehmann; Freedman, Pisani, and Purves; and Dunning). These examples illustrate that Neyman's key insights can be explained in intuitive and interesting ways to audiences at all levels, including undergraduates in introductory statistics courses. We have found Freedman, Pisani, and Purves's box-of-tickets model to be a valuable expository tool, and we also find their intuitive explanation of Neyman's variance result helpful: It is a "minor miracle" that in randomized experiments, the two-sample z-test is conservative because of "two mistakes that cancel." All four books take a more positive view of Neyman's results than Neyman himself did. We encourage educators and researchers to explore ways to communicate Neyman's ideas that are helpful for their own audiences.

Keywords: statistics education, potential outcomes, design-based inference, two-sample test, standard error

MSC 2020: 97K70, 62A01, 62D20

1 Introduction

One hundred years ago, Neyman [1] introduced the concept of potential outcomes and laid the foundations for randomization-based inference about average treatment effects. In the context of an agricultural experiment, Neyman assumes that the assignment of treatments (crop varieties) to units (plots of land) is analogous to randomly drawing balls from urns without replacement, so that each treatment group is a simple random sample from the finite population of experimental units. Although he does not propose randomization in a literal sense [2,3], his urn model is equivalent to a completely randomized experiment, and he considers the problem of estimating the average effect of treatment i relative to treatment j. In his framework for statistical inference, the units in the experiment are not assumed to be a random sample from a larger target population. Instead, the experimental units *are* the target population, and random assignment of treatment is the sole source of randomness.

Neyman's first key result follows immediately from the properties of simple random samples: the difference in mean outcomes between treatment groups i and j is an unbiased estimator of the average treatment effect. However, his next result is far from obvious. Consider the usual estimator of the variance of the

*** Corresponding author: Winston Lin,** Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, United States of America, e-mail: linston@wharton.upenn.edu

Sandrine Dudoit: Department of Statistics and Division of Biostatistics/School of Public Health, University of California, Berkeley, CA 94720, United States of America, e-mail: sandrine@stat.berkeley.edu

Deborah Nolan: Department of Statistics, University of California, Berkeley, CA 94720, United States of America, e-mail: nolan@stat.berkeley.edu

Terence P. Speed: Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia, e-mail: terry@wehi.edu.au

difference-in-means estimator: $s_i^2/n_i + s_j^2/n_j$, where s_i^2 and s_j^2 are the sample variances of the outcome in the two groups and n_i and n_j are the sample sizes. A traditional justification for this formula assumes two independent random samples from two infinite populations. Neyman's framework violates these assumptions (the population is finite and the treatment groups are mutually exclusive, not independent), but surprisingly, he finds that the usual variance estimator is unbiased if there is a constant additive treatment effect and conservatively biased if there are heterogeneous treatment effects.

Other articles in this special issue examine the influence of Neyman's paper on research in causal inference. Our aim is complementary: to discuss the paper's influence on four introductory to intermediate-level textbooks [4–7] in statistics and causal inference by faculty members from Berkeley (where Neyman founded the Department of Statistics). Although the original paper is difficult for modern readers, we think its key insights can and should be explained in intuitive and interesting ways. This can be helpful for audiences at all levels, from high school students and undergraduates in introductory statistics classes to Ph.D. students and scholars reading recent papers on design-based uncertainty in econometrics and causal inference [8–13].

The books we discuss are the exceptions, not the rule. Our impression is that most statistics textbooks (including one that two of us wrote [14]) still do not cover potential outcomes, and very few discuss Neyman's variance estimation result. When students learn about the two-sample t- and z-tests and the corresponding standard errors and confidence intervals, the derivations usually assume two independent random samples from two infinite populations. Textbooks often assert that the same procedures are appropriate for randomized experiments without clearly explaining why or conveying that this is surprising. This approach reminds us of Somers's [15] comment on his high school biology classes: "In the textbooks, astonishing facts were presented without astonishment." We prefer Freedman, Pisani, and Purves's (FPP's) [6] explanation: Neyman's result is a "minor miracle" where two wrongs make a right.

Sections 2–5 discuss how Scheffé [4], Hodges and Lehmann [5], Freedman et al. [6], and Dunning [7] depict Neyman's contributions; we order the four books chronologically because the later ones build on the insights of the earlier ones. Section 6 offers concluding thoughts, and the appendix provides mathematical details.

2 Scheffé (1959), *The Analysis of Variance*

Scheffé's classic textbook [4] does not discuss Neyman's 1923 paper in any detail, but appears to have played a role in keeping its flame alive. At the time of Scheffé's writing, the paper was not well known and was only available in Polish with a German summary. Neyman's better-known contribution to randomization-based inference was his 1935 paper [16], which focuses on the randomized-blocks and Latin-square designs and uses a more complicated model in which the potential outcomes are stochastic because of random "technical errors." Scheffé's chapter on randomization models ([4], ch. 9) similarly introduces these complexities from the outset, but mentions Neyman's 1923 paper very briefly in the first footnote, crediting it for the first formulation of a randomization model for the completely randomized design. (In the final footnote of a 1956 paper, Scheffé [17] gives a one-paragraph summary of Neyman [1].)

Scheffé's book is also noteworthy for the spirit in which he discusses randomization-based inference. He writes ([4], p. 106):

> The logical reason for randomizing is that it is possible on the basis of a model reflecting the randomization to draw sound statistical inferences, the probability basis of the model being provided not by wishful thinking but by the actual process of randomization which is part of the experiment. It is fortunate that in situations where the randomization models are more appropriate, statistical inferences from the corresponding "normal-theory" models usually are fair approximations in the more realistic randomization models. However, to profit from this happy relationship of the two models *randomization must be incorporated into the experiment.*

The italics are Scheffé's, but we also want to highlight his immediately preceding phrase, "this happy relationship of the two models." Neyman ([1], p. 471) had taken a gloomier view of his own results, describing the usual variance estimator as a method that was necessary "for the time being" but "has to be considered inaccurate" because of its conservative bias. For Scheffé, the glass is half full: it is a happy coincidence [18] that statistical inferences from the usual formulas can often be interpreted as approximate randomization-based inferences.

Neyman's 1923 paper and its results are not mentioned in any of the eleven other textbooks from the 1950s on statistics or experimental design that we have examined [19–29], including Neyman's own books [19,20]. The widely used statistics textbooks by Yule and Kendall [21] and Snedecor [22] rely on normal-theory justifications for two-sample tests and the analysis of variance; the groups being compared are assumed to be independent random samples from infinite populations. Kempthorne [23] includes detailed sections on randomization-based inference, but his discussion of Neyman's work only mentions the 1935 paper [16] and describes its null hypothesis as "artificial" because "a series of repetitions is envisaged, the experimental conditions remaining the same but the technical errors being different" ([23], p. 133).

3 Hodges and Lehmann (1970), *Basic Concepts of Probability and Statistics*, 2nd edition

Hodges and Lehmann's textbook [5] is dedicated to Neyman. As described in Lehmann's preface to the 2005 reissue, "It filled the need for an introduction to the fundamental ideas of modern statistics that was mathematically rigorous but did not require calculus." In section 9.4, after using examples (such as a cloud-seeding experiment) to introduce the concepts of treatments, subjects, and responses, Hodges and Lehmann write, "Let us now specify just what is meant by the 'effect' of such a treatment." The definition that follows is helpful and precise, using potential outcomes without referring to them by that name:

> Consider one particular subject, and suppose that his response would be equal to w if he were given the treatment, while it would be equal to v if he were not given the treatment. Then the difference $w - v$ is the *additional* response elicited by the treatment, above what it would have been without the treatment. We shall define this difference to be the *effect* of the treatment *on that subject*, and denote it by Δ; that is,
>
> $$\Delta = w - v.$$
>
> For example, if a particular storm would produce 2.4 inches of rain without seeding, and $w = 2.7$ inches with seeding, then the effect of seeding on that storm is $\Delta = 2.7 - 2.4 = .3$ inch.

Hodges and Lehmann then define the average treatment effect $\overline{\Delta}$ in a population of N subjects, writing that $\overline{\Delta}$ "is usually of primary interest" but making its limitations clear: "If a drug will speed up the recovery of half the patients by two months, while slowing down the recovery of the other half by two months, $\overline{\Delta}$ is 0; but the important question would be to identify the patients whom the drug will benefit." (Similar concerns motivate recent research on estimating heterogeneous treatment effects for personalized medicine and other applications [30–32].) They give examples where it seems possible or impossible to measure both w and v for the same subject. In the latter case, they propose a completely randomized experiment that assigns subjects to a treatment group of size t and a control group of size s. The experiment may use all the available subjects ($s + t = N$), or it may use only a random sample ($s + t < N$). Thus, Hodges and Lehmann simplify Neyman's model (which allows multiple treatment groups) by assuming a two-group experiment, but their framework is general enough to allow designs where the experimental subjects are a random sample from a larger population. FPP [6] uses the same framework.

Hodges and Lehmann explain that the usual formula for the variance of the difference-in-means estimator is approximately correct if the sampling fractions s/N and t/N are small (in other words, if the experimental subjects are a random sample from a much larger population) or if there is a constant additive treatment effect. How should the variance be estimated if the sampling fractions are not small and it is not reasonable to assume a constant additive treatment effect? Hodges and Lehmann leave this question unaddressed, but in other ways, they give very helpful advice about the design and analysis of experiments.

4 Freedman, Pisani, and Purves (2007), *Statistics*, 4th edition

Freedman et al. [6] is an introductory statistics textbook that emphasizes intuitive understanding and critical thinking. It is also dedicated to Neyman. In the preface, FPP write:

> We are going to tell you about some interesting problems which have been studied with the help of statistical methods, and show you how to use these methods yourself. We will try to explain why the methods work, and what to watch out for when others use them. Mathematical notation only seems to confuse things for many people, so this book relies on words, charts, and tables; there are hardly any x's or y's. As a matter of fact, even when professional mathematicians read technical books, their eyes tend to skip over the equations. What they really want is a sympathetic friend who will explain the ideas and draw the pictures behind the equations. We will try to be that friend, for those who read our book.

FPP give a remarkably accessible and interesting three-page distillation of Neyman [1] in their section on "Experiments" ([6], pp. 508–11). First, they use a hypothetical randomized clinical trial on vitamin C and colds to illustrate the mechanics of a two-sample z-test, writing, "Just pretend that you have two independent samples drawn at random with replacement." They then explain that this assumption makes two mistakes: (1) the draws were actually made without replacement and (2) the two samples' average outcomes are actually dependent (e.g., if we randomly assign the most cold-prone subject to the vitamin C group, that subject cannot be in the placebo group). The question they pose and answer in the remainder of the section is: "Why does the SE come out right, despite these problems?"

Like Hodges and Lehmann, FPP introduce potential outcomes without calling them by that name. Unlike Hodges and Lehmann, FPP explain the concept with a diagram of a "box model" – a device used throughout the book to discuss chance variability and sampling distributions. Box models make analogies between chance processes and randomly drawing numbered tickets from a box. (Sun and Alfredo [33] give a thoughtful discussion of the pedagogical benefits and limitations of FPP's box model approach and introduce a simulation-based enhancement.) In FPP's box model diagram for experiments (recreated in our Figure 1), the box has a ticket for each subject, and each ticket has *two* numbers, A and B, representing the subject's responses (potential outcomes) to the two treatments. But when we draw a random sample of tickets from the box and assign those subjects to one of the two treatments, we can only observe the numbers representing their responses to the assigned treatment. The other numbers are shaded out.

As FPP explain again, the usual standard error calculation for the difference-in-means estimator makes two mistakes: (1) "the SEs are computed as if drawing with replacement," and (2) "the SEs are combined as if the averages were independent." (The first mistake ignores finite-population correction factors, and the second

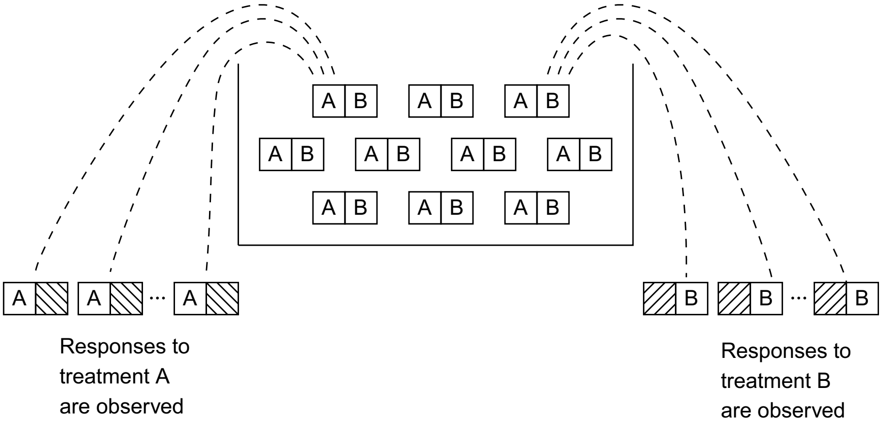
A B
A B
A B
A B
A B
A B
A B
A B
A B
A B
A ▨
A ▨
···
A ▨
▨ B
▨ B
···
▨ B
Responses to treatment A are observed
Responses to treatment B are observed

Figure 1: Redrawn from Freedman et al. ([6], p. 509). The box of tickets represents a population of subjects available for an experiment. On each ticket, the two numbers "A" and "B" (which may vary from one ticket to another) represent a subject's potential outcomes (responses to treatments A and B). We draw two distinct simple random samples of subjects without replacement and assign them to treatments A and B. For each subject, we can only observe the potential outcome corresponding to the assigned treatment; the other potential outcome is shaded out.

mistake ignores the covariance between the two samples' average outcomes.) They summarize the consequences intuitively:

> It is a lucky break that when applied to randomized experiments, the procedure of section 2 [the two-sample z-test with unequal variances] is conservative, tending to overestimate the SE by a small amount. That is because the two mistakes offset each other.
> - The first mistake inflates the SE.
> - The second mistake cuts the SE back down.

At the end of this section, they write, "In summary: when the data come from a randomized experiment ... the procedure of section 2 can be used even though there is dependence." The chapter's concluding section describes this result as "a minor miracle – two mistakes that cancel" ([6], p. 517).

FPP give mathematical derivations, references, and discussions of technical issues in the endnotes ([6], pp. A32–4), which are helpful for instructors and advanced readers. One note refers to "combinatorial calculations" that are omitted; calculations can be found in [34] (pp. 237–238), [35] (pp. 187–188), and our Appendix. For readers who want to understand the details and implications of Neyman's results but find the original paper difficult, we also recommend Reichardt and Gollob's exposition and discussion [36].

5 Dunning (2012), *Natural Experiments in the Social Sciences*

Dunning's textbook [7] focuses on causal inference and research design in the social sciences. His emphasis is on "natural experiments," but as he writes in the introduction, "The evaluative framework developed in this book is intentionally broad, and it may apply to other kinds of research designs – including true experiments as well as conventional observational studies." (It is a pleasure to associate Dunning's book with Berkeley even though he wrote it at Yale. Dunning is now a professor of political science at Berkeley, where he earned his Ph.D. The book is dedicated to the memory of Freedman and acknowledges Freedman's influence.)

After introducing the potential outcomes framework with examples from the social sciences, Dunning discusses how to estimate average treatment effects when treatment assignment is random or as-if random ([7], pp. 107–115). He writes, "For true experiments and strong natural experiments, the Neyman approach lends itself naturally to a sensible statistical model – that of sampling potential outcomes at random from an urn." Like FPP, he uses a box of tickets instead of an urn to make the model more accessible to students. Dunning's box model diagram ([7], p. 113) is adapted from FPP's, now labeling the potential outcomes under treatment and control as T_i and C_i. He explains intuitively that the difference in mean outcomes is an unbiased estimator of the average treatment effect because "the treatment group and control group are both random samples of the tickets in the box."

Dunning gives a thoughtful discussion ([7], pp. 119–120) of the noninterference assumption (potential outcomes for one unit do not depend on another unit's treatment assignment), which was left implicit in Neyman's model and made explicit by Cox [28] (ch. 2) and Rubin [37]. In a later section on instrumental-variables designs ([7], pp. 135–153), he explains how the potential outcomes framework can be used to study randomized experiments and natural experiments where there is noncompliance with the assigned treatment. He uses box model diagrams to illustrate the concepts of compliers, always-takers, and never-takers that were introduced by Angrist et al. [38].

In a chapter on sampling processes and standard errors ([7], ch. 6), Dunning draws on FPP's discussion of the standard error of the difference-in-means estimator, giving intuitive explanations in the main text and mathematical derivations in an appendix. He helpfully explains ([7], p. 171):

> Assuming independent sampling typically leads to standard errors that are if anything a little conservative (that is, slightly too big), but that's a good thing if we want to be cautious about our causal claims. Thus, we can treat the treatment and control samples as independent for purposes of calculating standard errors.

Thus, like Scheffé [4], Hodges and Lehmann [5], and FPP [6], Dunning takes a more positive view of Neyman's results than Neyman himself did. He also uses a box model diagram with cluster sampling ([7], p. 177) to help readers understand issues in the analysis of cluster-randomized experiments and as-if cluster-randomized natural experiments.

6 Discussion

Most introductory statistics textbooks do not use Neyman's [1] framework for inference from randomized experiments. Discussions of two-sample tests and confidence intervals assume that the groups being compared are two independent random samples from two larger populations; random assignment of treatment is listed as one of the ways to satisfy these assumptions [39,40]. The "populations" are usually hypothetical since most experiments do not randomly select subjects from a larger population. For example, in a careful discussion of a randomized trial of diet versus exercise, Utts and Heckard ([40], p. 446) write:

> The difference between the sample means estimates $\mu_1 - \mu_2$, the difference in the means of two hypothetical populations defined by the two weight-loss strategies in the study. The first hypothetical population is the population of weight losses for all sedentary men like the ones in this study if they were to be placed on this diet for a year. The second hypothetical population is the population of weight losses for those same men if they were to follow the exercise routine for a year.

In contrast, Neyman's framework for statistical inference does not ask readers to imagine random sampling from hypothetical superpopulations that are similar to but much larger than the groups in the experiment. As Scheffé [4] writes, the basis for inference is "provided not by wishful thinking but by the actual process of randomization which is part of the experiment." The actual subjects in the experiment are the target population for inference. Even if we have outcome data for the entire target population, there is still uncertainty about causal effects because we only observe one potential outcome for each subject.

We do not mean to suggest that randomization-based inference should always be preferred to superpopulation inference in experiments (for a balanced discussion, see Reichardt and Gollob [36], pp. 125–127). Both frameworks are important to study, but the randomization-based approach is underemphasized in statistics education. Many users of statistics are unaware that potential outcomes and randomization-based inference provide another way to think about frequently asked questions such as "how do we interpret standard errors ... when we have data on the entire population" [41] and whether it is "wrong" to report *p*-values from a randomized experiment with a convenience sample [42].

We have found FPP's box model especially valuable for teaching randomization-based causal inference concepts in an introductory statistics course. For alternative views, see Sun and Alfredo [33] and Rossman and De Veaux [43]. Our aim here is not to prescribe specific pedagogical tools, but to show that Neyman's insights on potential outcomes and standard errors *can* be expressed intuitively, and to encourage educators and researchers to explore ways to do this. For example, Lau et al.'s introductory data science textbook [44] uses urns filled with colored or labeled marbles as models for survey sampling, randomized experiments, and measurement error. Sampling distributions are discovered by running simulation studies, and causal inference is introduced as a simple extension following FPP, with two numbers on each marble. Imbens and Rubin's graduate-level causal inference textbook ([45], ch. 1) gives a thoughtful discussion of potential outcomes with an example about aspirin and headaches and an anecdote from the movie *It's a Wonderful Life*.

Potential outcomes are now covered in a growing number of textbooks and other expository books, mostly outside of the core statistics curriculum. Ding's new causal inference textbook [46] is based on undergraduate and graduate courses at Berkeley and provides helpful, detailed expositions of basic and advanced topics, including Neyman's [1] results. His preface recommends other valuable textbooks on causal inference. At a more elementary level, Rosenbaum's books for general readers [47,48] and several introductory textbooks by social scientists [49–52] cover potential outcomes and causal inference concepts with minimal mathematics and without assuming prior knowledge of statistics.

In an intermediate-level statistical inference course, teaching a derivation of Neyman's variance result and its "minor miracle" can be a useful way to give students practice with concepts such as conditional expectations, variances, covariances, standard error estimation, and confidence interval coverage. (We thank an anonymous reviewer for this suggestion.) Our appendix gives one derivation, and we recommend Ding's textbook [46] for alternative derivations, asymptotic results, and simulation examples.

Neyman's 1923 paper is now justly regarded as a pathbreaking contribution to causal inference, but its main ideas are still unfamiliar to many statisticians and students of statistics. In our view, Neyman's paper has not yet had the influence on statistics education that it ought to have. As Hill [53] writes:

> In terms of education, we need to move beyond the catch-all "correlation is not causation" admonition and help create a deeper understanding about the broader populace about what it means to make a causal claim and what kind of research strongly supports such claims. That would mean starting to teach about counterfactuals and a wider range of designs even in introductory statistics courses (I would like some of these concepts to be taught in grade school).

We agree, and we hope it won't take another 100 years.

Acknowledgments: We are grateful to Philip Dawid (the associate editor), Thad Dunning, Don Green, Philip Stark, and two anonymous reviewers for very thoughtful suggestions and to Mauricio Daros Andrade for kind help with Figure 1.

Funding information: The authors state no funding involved.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results, and approved the final version of the manuscript. SD, DN, and TPS conceived the topic. WL prepared the manuscript with contributions from all co-authors.

Conflict of interest: The authors state no conflict of interest.

References

[1] Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated and edited by DM Dabrowska and TP Speed from the 1923 Polish original. Statist Sci. 1990;5(4):465–72. doi: https://doi.org/10.1214/ss/1177012031.

[2] Speed TP. Introductory remarks on Neyman (1923). Statist Sci. 1990;5(4):463–4. doi: https://doi.org/10.1214/ss/1177012030.

[3] Rubin DB. Comment: Neyman (1923) and causal inference in experiments and observational studies. Statist Sci. 1990;5(4):472–80. doi: https://doi.org/10.1214/ss/1177012032.

[4] Scheffé H. The analysis of variance. New York: Wiley; 1959.

[5] Hodges JL Jr, Lehmann EL. Basic concepts of probability and statistics. 2nd ed. Philadelphia: Society for Industrial and Applied Mathematics; 2005. Unabridged republication of the work first published by Holden-Day, San Francisco, 1970.

[6] Freedman D, Pisani R, Purves R. Statistics. 4th ed. New York: Norton; 2007.

[7] Dunning T. Natural experiments in the social sciences: a design-based approach. Cambridge (UK): Cambridge University Press; 2012.

[8] Abadie A, Athey S, Imbens GW, Wooldridge JM. Sampling-based versus design-based uncertainty in regression analysis. Econometrica. 2020;88(1):265–96. doi: https://doi.org/10.3982/ECTA12675.

[9] Abadie A, Athey S, Imbens GW, Wooldridge JM. When should you adjust standard errors for clustering? Q J Econ. 2023;138(1):1–35. doi: https://doi.org/10.1093/qje/qjac038.

[10] Wooldridge JM. What is a standard error? (And how should we compute it?) J Econometrics. 2023;237(2 Pt A):Article 105517, doi: https://doi.org/10.1016/j.jeconom.2023.105517.

[11] Xu R. Potential outcomes and finite-population inference for M-estimators. Econom J. 2021;24(1):162–76. doi: https://doi.org/10.1093/ectj/utaa022.

[12] Rambachan A, Roth J. Design-based uncertainty for quasi-experiments. 2024. arXiv: https://arxiv.org/abs/2008.00602.

[13] Bai Y, Shaikh AM, Tabord-Meehan M. A primer on the analysis of randomized experiments and a survey of some recent advances. 2024. arXiv: https://arxiv.org/abs/2405.03910.

[14] Nolan D, Speed T. Stat labs: mathematical statistics through applications. New York: Springer; 2000.

[15] Somers J. I should have loved biology. 2020 Nov 18 [cited 2024 Jul 2]. In: James Somers [blog on the internet]. https://jsomers.net/i-should-have-loved-biology/.

[16] Neyman J. with co-operation of Iwaszkiewicz K, Kolodziejczyk S. Statistical problems in agricultural experimentation. Suppl J Roy Statist Soc. 1935;2(2):107–54; discussion 154–80. doi: https://doi.org/10.2307/2983637.

[17] Scheffé H. Alternative models for the analysis of variance. Ann Math Statist. 1956;27(2):251–71. doi: https://doi.org/10.1214/aoms/1177728258.

[18] Samii C. Should you use frequentist standard errors with causal estimates on population data? Yes. 2014 Feb 3 [cited 2024 Jul 5]. In: Cyrus Samii [blog on the internet]. https://cyrussamii.com/?p=1622.

[19] Neyman J. First course in probability and statistics. New York: Holt; 1950.

[20] Neyman J. Lectures and conferences on mathematical statistics and probability. 2nd ed. Washington: Graduate School, U.S. Department of Agriculture; 1952.

[21] Yule GU, Kendall MG. An introduction to the theory of statistics. 14th ed. New York: Hafner; 1950.

[22] Snedecor GW. Statistical methods applied to experiments in agriculture and biology. 5th ed. Ames (IA): Iowa State University Press; 1956.

[23] Kempthorne O. The design and analysis of experiments. New York: Wiley; 1952.

[24] Davies OL. The design and analysis of industrial experiments. New York: Hafner; 1954.

[25] Federer WT. Experimental design: theory and application. New York: Macmillan; 1955.

[26] Finney DJ. Experimental design and its statistical basis. Chicago: University of Chicago Press; 1955.

[27] Cochran WG, Cox GM. Experimental designs. 2nd ed. New York: Wiley; 1957.

[28] Cox DR. Planning of experiments. New York: Wiley; 1958.

[29] Lehmann EL. Testing statistical hypotheses. New York: Wiley; 1959.

[30] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Amer Statist Assoc. 2018;113(523):1228–42. doi: https://doi.org/10.1080/01621459.2017.1319839.

[31] Xu Y, Ignatiadis N, Sverdrup E, Fleming S, Wager S, Shah N. Treatment heterogeneity with survival outcomes. In: Zubizarreta JR, Stuart EA, Small DS, Rosenbaum PR, editors. Handbook of matching and weighting adjustments for causal inference. Boca Raton (FL): Chapman & Hall/CRC; 2023. p. 445–82.

[32] Boileau P, Qi NT, van der Laan MJ, Dudoit S, Leng N. A flexible approach for predictive biomarker discovery. Biostatistics. 2023;24(4):1085–105. doi: https://doi.org/10.1093/biostatistics/kxac029.

[33] Sun DL, Alfredo J. A modern look at Freedman's box model. Tech Innovat Stat Educ. 2019;12(1):15p. doi: https://doi.org/10.5070/T5121044395.

[34] Cornfield J. On samples from finite populations. J Amer Statist Assoc. 1944;39(226):236–9. doi: https://doi.org/10.1080/01621459.1944.10500680.

[35] Freedman DA. On regression adjustments in experiments with several treatments. Ann Appl Stat. 2008;2(1):176–96. doi: https://doi.org/10.1214/07-AOAS143

[36] Reichardt CS, Gollob HF. Justifying the use and increasing the power of a *t* test for a randomized experiment with a convenience sample. Psychol Methods. 1999;4(1):117–28. doi: https://doi.org/10.1037/1082-989X.4.1.117.

[37] Rubin DB. Comment. J Amer Statist Assoc. 1980;75(371):591–3. doi: https://doi.org/10.1080/01621459.1980.10477517.

[38] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Amer Statist Assoc. 1996;91(434):444–55; discussion 456–72. doi: https://doi.org/10.1080/01621459.1996.10476902.

[39] Moore DS, McCabe GP, Craig BA. Introduction to the practice of statistics. 10th ed. Austin (TX): Macmillan Learning; 2021.

[40] Utts JM, Heckard RF. Mind on statistics. 6th ed. Boston: Cengage; 2022.

[41] Angrist JD. Pop quiz. 2013 Nov 28 [cited 2024 Jul 2]. In: *MHE* Blog [internet]. https://www.mostlyharmlesseconometrics.com/2013/11/pop-quiz-2.

[42] Gelman A. How to interpret inferential statistics when your data aren't a random sample. 2021 Jul 22 [cited 2024 Jul 2]. In: Statistical Modeling, Causal Inference, and Social Science [blog on the internet]. https://statmodeling.stat.columbia.edu/2021/07/22/how-to-interpret-inferential-statistics-when-your-data-arent-a-random-sample/.

[43] Rossman A, De Veaux R. Interview with Richard De Veaux. J Stat Educ. 2016;24(3):157–68. doi: https://doi.org/10.1080/10691898.2016.1263493.

[44] Lau S, Gonzalez J, Nolan D. Learning data science: data wrangling, exploration, visualization, and modeling with Python. Sebastopol (CA): O'Reilly; 2023.

[45] Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences: an introduction. New York: Cambridge University Press; 2015.

[46] Ding P. A first course in causal inference. Boca Raton (FL): Chapman & Hall/CRC; 2024.

[47] Rosenbaum PR. Observation and experiment: an introduction to causal inference. Cambridge (MA): Harvard University Press; 2017.

[48] Rosenbaum PR. Causal inference. Cambridge (MA): MIT Press; 2023.

[49] Angrist JD, Pischke JS. Mastering 'metrics: the path from cause to effect. Princeton (NJ): Princeton University Press; 2014.

[50] Bueno de Mesquita E, Fowler A. Thinking clearly with data: a guide to quantitative reasoning and analysis. Princeton (NJ): Princeton University Press; 2021.

[51] Green DP. Social science experiments: a hands-on introduction. Cambridge (UK): Cambridge University Press; 2022.

[52] Llaudet E, Imai K. Data analysis for social science: a friendly and practical introduction. Princeton (NJ): Princeton University Press; 2022.

[53] Hill JL. Lessons we are still learning. Obs Stud. 2015;1(1):196–9. doi: https://doi.org/10.1353/obs.2015.0019.

[54] Ding P, Li X, Miratrix LW. Bridging finite and super population causal inference. J Causal Inference. 2017;5(2):20160027. doi: https://doi.org/10.1515/jci-2016-0027.

[55] Cochran WG. Sampling techniques. 3rd ed. New York: Wiley; 1977.

[56] Rice JA. Mathematical statistics and data analysis. 3rd ed. Belmont (CA): Thomson Brooks/Cole; 2007.

[57] Freedman DA. On regression adjustments to experimental data. Adv in Appl Math. 2008;40(2):180–93. doi: https://doi.org/10.1016/j.aam.2006.12.003.

Appendix

This appendix gives some of the details of Neyman's results in modern notation. Our derivations are adapted from FPP ([6], pp. A32–4), Freedman ([35], pp. 187–8), and Reichardt and Gollob [36].

Ding et al. [54] give an elegant alternative derivation, using a variance decomposition and a "completeness-style" argument to connect finite- and infinite-population inference.

A.1 Setup

We use Hodges and Lehmann's ([5], p. 277) and FPP's ([6], p. A33) simplification of Neyman's [1] setup. Suppose there are N subjects available, and we assign two disjoint simple random samples of fixed sizes n_1 and n_0 to treatment and control, respectively, where $n_1 + n_0 \leq N$.

Let $y_i(1)$ and $y_i(0)$ denote subject i's potential outcomes under treatment and control, for $i = 1, \ldots, N$. Also, let $T_i = 1$ if the subject is assigned to treatment and $T_i = 0$ otherwise, and similarly let C_i be an indicator for assignment to control. In this framework, T_i and C_i are random variables, but $y_i(1)$ and $y_i(0)$ are constants.

The true average treatment effect is $\text{ATE} = \overline{y}(1) - \overline{y}(0)$, where $\overline{y}(1) = \frac{1}{N}\sum_{i=1}^{N} y_i(1)$ and $\overline{y}(0) = \frac{1}{N}\sum_{i=1}^{N} y_i(0)$ are the average potential outcomes. The difference-in-means estimator of ATE is defined as follows:

$$\widehat{\text{ATE}} = \widehat{y}(1) - \widehat{y}(0),$$

where $\widehat{y}(1) = \frac{1}{n_1}\sum_{i=1}^{N} T_i y_i(1)$ and $\widehat{y}(0) = \frac{1}{n_0}\sum_{i=1}^{N} C_i y_i(0)$ are the average observed outcomes in the treatment group and control group.

A.2 Unbiased estimation of the average treatment effect

The properties of simple random samples imply that $\widehat{y}(1)$ and $\widehat{y}(0)$ are unbiased estimators of $\overline{y}(1)$ and $\overline{y}(0)$, and therefore, $\widehat{\text{ATE}}$ is unbiased for ATE. For example, by linearity of expectation,

$$E[\widehat{y}(1)] = \frac{1}{n_1}\sum_{i=1}^{N} E[T_i] y_i(1) = \frac{1}{n_1}\sum_{i=1}^{N} \frac{n_1}{N} y_i(1) = \overline{y}(1).$$

(The second equality uses the fact that each subject's probability of assignment to treatment is n_1/N. One way to prove that result is to start with $\sum_{i=1}^{N} T_i = n_1$ and then use linearity of expectation and symmetry. Another way is to note that there are $\binom{N}{n_1}$ possible samples of size n_1, all equally likely to become the treatment group, and $\binom{N-1}{n_1-1}$ of these include subject i.)

A.3 True variance of the difference-in-means estimator

The variance of the difference-in-means estimator is

$$\text{Var}(\widehat{\text{ATE}}) = \text{Var}(\widehat{y}(1)) + \text{Var}(\widehat{y}(0)) - 2\text{Cov}(\widehat{y}(1), \widehat{y}(0)).$$

Using well-known formulas for the variance of the sample mean under simple random sampling without replacement ([34]; [55], p. 23; [56], p. 208), the variance of the treatment group's average observed outcome is

$$\text{Var}(\widehat{y}(1)) = \left(1 - \frac{n_1 - 1}{N - 1}\right)\frac{\sigma_1^2}{n_1} = \left(1 - \frac{n_1}{N}\right)\frac{\tilde{\sigma}_1^2}{n_1},$$

where $\sigma_1^2 = \frac{1}{N}\sum_{i=1}^{N}[y_i(1) - \overline{y}(1)]^2$ and $\tilde{\sigma}_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}[y_i(1) - \overline{y}(1)]^2$. Similarly, the variance of the control group's average observed outcome is

$$\mathrm{Var}(\widehat{\overline{y}}(0)) = \left(1 - \frac{n_0 - 1}{N - 1}\right)\frac{\sigma_0^2}{n_0} = \left(1 - \frac{n_0}{N}\right)\frac{\tilde{\sigma}_0^2}{n_0},$$

where $\sigma_0^2 = \frac{1}{N}\sum_{i=1}^{N}[y_i(0) - \overline{y}(0)]^2$ and $\tilde{\sigma}_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}[y_i(0) - \overline{y}(0)]^2$. The denominator $(N - 1)$ in $\tilde{\sigma}_1^2$ and $\tilde{\sigma}_0^2$ helps simplify results in finite-population sampling theory ([36], p. 119; [55], p. 23).

The covariance between the two groups' average observed outcomes is

$$\mathrm{Cov}(\widehat{\overline{y}}(1), \widehat{\overline{y}}(0)) = -\frac{1}{N-1}\sigma_{1,0} = -\frac{1}{N}\tilde{\sigma}_{1,0},$$

where $\sigma_{1,0} = \frac{1}{N}\sum_{i=1}^{N}[y_i(1) - \overline{y}(1)][y_i(0) - \overline{y}(0)]$ and $\tilde{\sigma}_{1,0} = \frac{1}{N-1}\sum_{i=1}^{N}[y_i(1) - \overline{y}(1)][y_i(0) - \overline{y}(0)]$. To prove this result, we adapt Freedman's derivation ([35], pp. 187–8) and fill in omitted steps. Define the mean-centered potential outcomes $y_i^*(1) = y_i(1) - \overline{y}(1)$ and $y_i^*(0) = y_i(0) - \overline{y}(0)$. Since $\widehat{\overline{y}}(1)$ and $\widehat{\overline{y}}(0)$ are unbiased estimators of $\overline{y}(1)$ and $\overline{y}(0)$,

$$\begin{aligned}
\mathrm{Cov}(\widehat{\overline{y}}(1), \widehat{\overline{y}}(0)) &= E[(\widehat{\overline{y}}(1) - E[\widehat{\overline{y}}(1)])(\widehat{\overline{y}}(0) - E[\widehat{\overline{y}}(0)])] \\
&= E\left(\left[\frac{1}{n_1}\sum_{i=1}^{N} T_i y_i(1) - \overline{y}(1)\right]\left[\frac{1}{n_0}\sum_{j=1}^{N} C_j y_j(0) - \overline{y}(0)\right]\right) \\
&= E\left(\left[\frac{1}{n_1}\sum_{i=1}^{N} T_i y_i(1) - \frac{1}{n_1}\sum_{i=1}^{N} T_i \overline{y}(1)\right]\left[\frac{1}{n_0}\sum_{j=1}^{N} C_j y_j(0) - \frac{1}{n_0}\sum_{j=1}^{N} C_j \overline{y}(0)\right]\right) \\
&= E\left(\left[\frac{1}{n_1}\sum_{i=1}^{N} T_i y_i^*(1)\right]\left[\frac{1}{n_0}\sum_{j=1}^{N} C_j y_j^*(0)\right]\right) \\
&= \frac{1}{n_1}\frac{1}{n_0}E\left(\sum_{i \neq j} T_i C_j y_i^*(1) y_j^*(0) + \sum_{i=1}^{N} T_i C_i y_i^*(1) y_i^*(0)\right).
\end{aligned}$$

We have $T_iC_i = 0$ for all i, since a subject cannot be assigned to both treatment and control. Also, the mean-centered potential outcomes are constants. So the covariance simplifies to

$$\mathrm{Cov}(\widehat{\overline{y}}(1), \widehat{\overline{y}}(0)) = \frac{1}{n_1}\frac{1}{n_0}\sum_{i \neq j} y_i^*(1) y_j^*(0) E(T_i C_j).$$

Next, for $i \neq j$, we have

$$E(T_iC_j) = P(T_i = 1, C_j = 1) = P(T_i = 1)P(C_j = 1 \mid T_i = 1) = \frac{n_1}{N}\frac{n_0}{N-1}$$

(since, given that subject i is in the treatment group, the other $N - 1$ subjects are all equally likely to be among the n_0 subjects in the control group). Therefore,

$$\begin{aligned}
\mathrm{Cov}(\widehat{\overline{y}}(1), \widehat{\overline{y}}(0)) &= \frac{1}{N}\frac{1}{N-1}\sum_{i \neq j} y_i^*(1) y_j^*(0) \\
&= \frac{1}{N}\frac{1}{N-1}\left(\left[\sum_{i=1}^{N} y_i^*(1)\right]\left[\sum_{j=1}^{N} y_j^*(0)\right] - \sum_{i=1}^{N} y_i^*(1) y_i^*(0)\right) \\
&= -\frac{1}{N}\frac{1}{N-1}\sum_{i=1}^{N} y_i^*(1) y_i^*(0) = -\frac{1}{N}\tilde{\sigma}_{1,0}.
\end{aligned}$$

Putting it all together,

$$\mathrm{Var}(\widehat{\mathrm{ATE}}) = \left(1 - \frac{n_1}{N}\right)\frac{\tilde{\sigma}_1^2}{n_1} + \left(1 - \frac{n_0}{N}\right)\frac{\tilde{\sigma}_0^2}{n_0} + 2\frac{\tilde{\sigma}_{1,0}}{N}. \tag{A1}$$

Rearranging terms, we have

$$\mathrm{Var}(\widehat{\mathrm{ATE}}) = \frac{\tilde{\sigma}_1^2}{n_1} + \frac{\tilde{\sigma}_0^2}{n_0} - \frac{\tilde{\sigma}_1^2 + \tilde{\sigma}_0^2 - 2\tilde{\sigma}_{1,0}}{N},$$

which simplifies to

$$\mathrm{Var}(\widehat{\mathrm{ATE}}) = \frac{\tilde{\sigma}_1^2}{n_1} + \frac{\tilde{\sigma}_0^2}{n_0} - \frac{\tilde{\sigma}_\Delta^2}{N}, \tag{A2}$$

where $\tilde{\sigma}_\Delta^2 = \frac{1}{N-1}\sum_{i=1}^N [y_i(1) - y_i(0) - \mathrm{ATE}]^2$ is the population variance of the treatment effect (thinking of the N subjects as the finite population, and again using the denominator $N - 1$).

A.4 Unbiased or conservatively biased estimation of the variance of the difference-in-means estimator

The usual (unpooled) estimator of $\mathrm{Var}(\widehat{\mathrm{ATE}})$ is

$$\widehat{\mathrm{Var}}(\widehat{\mathrm{ATE}}) = \frac{s_1^2}{n_1} + \frac{s_0^2}{n_0},$$

where $s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^N T_i[y_i(1) - \widehat{\bar{y}}(1)]^2$ and $s_0^2 = \frac{1}{n_0-1}\sum_{i=1}^N C_i[y_i(0) - \widehat{\bar{y}}(0)]^2$ are the sample variances of the observed outcome in the treatment group and control group. From finite-population sampling theory, s_1^2 and s_0^2 are unbiased estimators of $\tilde{\sigma}_1^2$ and $\tilde{\sigma}_0^2$ ([55], p. 26; [56], p. 211). Thus,

$$E[\widehat{\mathrm{Var}}(\widehat{\mathrm{ATE}})] = \frac{\tilde{\sigma}_1^2}{n_1} + \frac{\tilde{\sigma}_0^2}{n_0}. \tag{A3}$$

Comparing (A3) with (A1), we can see that on average, $\widehat{\mathrm{Var}}(\widehat{\mathrm{ATE}})$ makes "two mistakes." The first mistake is conservative: it ignores the finite-population correction factors $(1 - n_1/N)$ and $(1 - n_0/N)$. The second mistake is anticonservative if the two potential outcomes are positively correlated in the population of N subjects (i.e., if $\tilde{\sigma}_{1,0} > 0$): it ignores the covariance between the two groups' average observed outcomes (which is negative if $\tilde{\sigma}_{1,0} > 0$, because if the luck of the draw assigns many subjects with high potential outcomes to the treatment group, then those same subjects will not be in the control group).

Comparing (A3) with (A2), we can see the net effect of the two mistakes. If there is a constant additive treatment effect, then $\tilde{\sigma}_\Delta^2 = 0$, so $\widehat{\mathrm{Var}}(\widehat{\mathrm{ATE}})$ is unbiased. Otherwise, $\widehat{\mathrm{Var}}(\widehat{\mathrm{ATE}})$ has a conservative bias of magnitude $\tilde{\sigma}_\Delta^2/N$. But if the experimental subjects are a random sample from a much larger population, so that N is much larger than n_1 and n_0, then this bias is negligible, as noted by Hodges and Lehmann ([5], pp. 277–278) and FPP ([6], pp. 510, A33).

Neyman [1] does not study the pooled estimator of $\mathrm{Var}(\widehat{\mathrm{ATE}})$, which is

$$\frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2}\left[\frac{1}{n_1} + \frac{1}{n_0}\right].$$

The pooled estimator is identical to $\widehat{\mathrm{Var}}(\widehat{\mathrm{ATE}})$ if $n_1 = n_0$ and is unbiased if there is a constant additive treatment effect ([36], p. 121). But if $n_1 \neq n_0$ and there are heterogeneous treatment effects, the pooled estimator may be either conservative or anticonservative ([36], p. 124; [57], p. 190).