ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Six Statistical Senses

Radu V. Craiu,[1] Ruobin Gong,[2] and Xiao-Li Meng[3]

[1]Department of Statistical Sciences, University of Toronto, Toronto, Canada; email: radu.craiu@utoronto.ca

[2]Department of Statistics, Rutgers University, Piscataway, New Jersey, USA; email: ruobin.gong@rutgers.edu

[3]Department of Statistics, Harvard University, Cambridge, Massachusetts, USA; email: meng@stat.harvard.edu

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

ANNUAL REVIEWS CONNECT

www.annualreviews.org

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Annu. Rev. Stat. Appl. 2023. 10:699–725

First published as a Review in Advance on October 21, 2022

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040220-015348

Copyright © 2023 by the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See credit lines of images or other third-party material in this article for license information.

Keywords

bootstrap, data augmentation, exchangeability, likelihood, propensity score, randomized replication, probabilistic sampling, shrinkage estimation

Abstract

This article proposes a set of categories, each one representing a particular distillation of important statistical ideas. Each category is labeled a "sense" because we think of these as essential in helping every statistical mind connect in constructive and insightful ways with statistical theory, methodologies, and computation, toward the ultimate goal of building statistical phronesis. The illustration of each sense with statistical principles and methods provides a sensical tour of the conceptual landscape of statistics, as a leading discipline in the data science ecosystem.

OPEN ACCESS CC BY

1. WHAT ARE STATISTICAL SENSES?

1.1. Statistical Sentience

Distribution: a mathematical bookkeeping of individual states by their relative abundance or salience (or the lack thereof)

Sentience: the ability to feel or perceive things

Born at the intersection of the empirical and mathematical universes, statistics needed time to develop its own set of principles and tools, intuitions, blunders, and near misses—or, in other words, its character. This article selectively reviews some of the unique features that give statistics its disciplinary identity and strength, and statisticians a toolbox they can lean on in times of creative need.

When writing about important ideas in almost any discipline, one can choose different paths. There is the genesis trail, which considers the emergence of statistics as a field and the ideas that propelled it forward (Hacking 2006, Agresti & Meng 2013; Shafer 2019, 2022; Agresti 2021). There is also the historical perspective, excellently illustrated by Stigler (1986, 2002), in which chronology plays an important role in explaining the evolution of a field. Reading through the history, one cannot avoid the thought that statistics's existence largely preceded the revelation of its essence. The latter concept itself is vulnerable to subjectivity, and had this article been written by a different group, it would have likely resulted in a different creature with different emphases and supported by different interpretations.

To us, the overarching principle of statistics is linked with the idea of a probability distribution—known, hypothesized, or latent. Inexorably connected with a distribution is the concept of variation which, once any effort toward understanding or measuring it is undertaken, leads us to data. Variations permeate our lives because they create simultaneously information and uncertainty; one we love and the other we hate, with our sentiments being also affected by context (Meng 2020). Consequently, variations occupy the attention of the discipline of statistics—and much of the broader data science ecosystem—by perpetually challenging statisticians and data scientists' ability to separate signals from noise, or even define them. More broadly, for humans to successfully navigate the world of variations and its associated hazards, we have developed our senses to handle the physical reality but also to learn from similar occurrences so as to anticipate the future (prediction) or find rational explanations of the real world (inference).

For instance, we humans understand that anything can happen tomorrow, but some events are more likely than others. This knowledge is derived from many tomorrows throughout human history. Tomorrow is always different from today, but it never fails to resemble today to some degree. Such different-yet-similar repetitions continually shape and enhance our cognitive abilities to recognize patterns, contemplate consequences, cope with uncertainties, and ultimately keep ourselves alive as a species and as individuals. The statistical calculus is fueled by this fluid tension between similarity and difference. Both depend on one another to possess meaning, and both require nontrivial formulations in a rigorous sense. A probability distribution prescribes similarities among individuals in a population by describing their differences. An overarching aim for statistical learning is to pin down, from vastly incomplete information, the most suitable set of distributions that can adequately capture the similarity and difference expressed in available data sets for addressing particular substantive inquiries, confirmatory or exploratory, or a mixture of the two.

In their attempt to systematically advance this learning ability, statisticians have developed over time an arsenal of tools that led to a kind of statistical sentience that inherently defies delineation. Yet its formation and evolution are undeniable and rely on a number of principles and techniques that, altogether, have proven elemental in developing the tools we need to navigate the world of variations. The task of a statistical learning can seem overwhelming at first. Based on a sample whose size is often tiny compared with that of the population at large, one is asked to identify complex mechanisms and/or predict future outcomes. Being able to address this enormous problem

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

in all its manifestations would be equivalent to getting a cosmic free lunch. In the absence of the latter, we have to cook ourselves a number of affordable ones to sustain us in different situations.

This aim is served best by developing fundamental statistical ideas that are time honored and timely, with their values to human inquiries particularly important to emphasize and to realize. This article is devoted to them. We focus on statistical ideas that are the building blocks and milestones for establishing and sustaining statistics as a scientific discipline. We connect and highlight these ideas by revealing their roles in forming statistical senses, with each one contributing toward statistical sentience. A precise definition for each sense is neither possible nor desirable, since forcing such a framing would necessarily reduce their rich complexity or interconnectivity. Nor should we partition a statistician's arsenal into distinct categories, because a given technique or principle can appear under different guises or present different merits, and each of them could be serving or enhancing a different sense. Instead, we choose a few distinct concepts and ideas to illustrate each sense.

This gives us the opportunity to revisit some cornerstones of our discipline and, more interestingly, to form new links between classical and modern statistics, between ideas and principles that might look disparate at first. By doing so, we hope to facilitate those who are interested in enhancing their statistical sentience professionally or personally, an ability that may come in handy when dealing and coping with uncertainties and risks in our increasingly multifaceted and volatile societies, with the COVID-19 pandemic serving as a painful and prolonged reminder.

1.2. Statistical Senses

In a nutshell, statistical senses are nothing but common senses developed and guided by statistical insights and probabilistic reasoning. However, some seemingly obvious intuitions, such as the expectation that more data must lead to better results, turn out to be statistical fallacies. This is because data contain both signal and noise. Without properly parsing the two, more data may well inflate the noise instead of enhancing the signals. The development of an individual's statistical senses, therefore, is a maturing process that requires time and experience. The approximate alignment of the statistician's senses with the natural five senses is intentional, and it is meant to reify this process. Just as our natural senses help us navigate and master the physical world, so do the statistical senses help in the world of variations. Indeed, the statistical discipline, as an intellectual body, evolves and thrives because of external stimuli.

The first sense has to do with the informative ignorance statisticians embrace, which leads us to characterize it as selective hearing. The latter allows us to get to the heart of a problem without getting lost in details that are less relevant and might greatly complicate the job. For instance, when setting up a regression model, we must decide which features should be included as predictors or explanatory variables and which ones should be considered noise. Using automated machine learning algorithms to choose features does not release us from making informed judgement. To the contrary, it only increases the demand on our hearing sensitivity and sensibility, in order to prevent us from being deafened by the machine noise.

The second sense codes our ability to "smell" when a problem is too complicated to solve without first making it seemingly even more complicated—that is, by introducing randomness. Matching two groups of patients on all factors that can affect treatment efficacy is an impossible task. Yet by flipping a fair coin to randomly assign each patient to a treatment or a placebo, we ensure the balance statistically between the two groups on all known, unknown, or unknowable factors, from which a causal inference of the treatment efficacy becomes possible. Fully protecting data privacy is another impossible task. Yet by injecting properly designed randomness into data before releasing, we can control the level of privacy loss while maintaining reasonable utility of the data.

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

The third sense is a form of enhanced vision, where experienced statisticians are able to see through data that are not present or, more broadly, dark data. Competent detectives do not reason only from what items are present at a crime scene. A missing portrait can be a smoking gun more than a gun smoking. Most times, data are not missing randomly. The many individuals who did not respond to surveys during the 2016 US presidential election did not make the decision based on flipping a fair coin. We all have seen consequences of such survey data being analyzed and reported without the benefit of a "seeing through" sense.

The fourth sense, which we term the "magic touch," concerns statistical principles and insights that can convert ordinary estimators into extraordinary ones, literally and figuratively. A shining example is the so-called Rao–Blackwellization, which turns a noisy estimator into an optimal one by a form of deep stratification, or more broadly, conditioning. Conditioning, in layperson's terms, means to take into account more detailed information that is judged to be relevant, just as a patient wants an effective individualized treatment instead of one that works on average. The magic touch of conditioning requires both experience and careful thinking. The famous Monty Hall problem, regarding whether one should switch a door or not in order to win a prize behind the door, demonstrates how human minds can easily be misled by a lack of training on probabilistic conditioning.

Extracting meaningful information from data is perhaps the most important part of a statistician's raison d'être, so it is not surprising that the fifth sense is defined by an ability to extract a wealth of information from the frugality of a sample (relative to the much larger population), or "just a taste," the fifth statistical sense. In Occam's razor, statisticians discovered a principle that has guided and served them for many years. The richness that comes from frugality shines through other illustrious instances that range from bootstrap to propensity matching. But there is no free lunch. Bootstrap is made possible by the hidden assumption that there are inherent replications within a single data set, and propensity matching works well when there is sufficient built-in compatibility between the samples being matched. A key part of statistical sentience is to have a good sense of the limitations of each method, no matter how almighty it may appear.

This leads to the sixth (statistical) sense, which is meant to express exactly that: an ability to transcend pure rationality or technical prowess and to create a qualitative jump that is as potentially fallible as it is path-breaking. As we argue and illustrate, all attempts to transition from known to unknown require a leap of faith, explicit or concealed, since statistical contemplation by its very nature involves building shaky bridges that come with structural risks. Much statistical training is needed to build a sixth sense, which would make one daring enough to spearhead new methods despite the inherent risks. Upon completing our sensical tour, one sense per section, we explain in a final section the origin of this tour, how our senses have evolved, and invite our readers to help promote sensical statistics and data science as we venture deeper into the digital age.

2. SELECTIVE HEARING: INFORMATIVE IGNORANCE

Statisticians must be able to hear the essential and ignore the incidental in their data so that they can develop statistical models with the relevant inferential aims. The latter could be the essential effects of a treatment in randomized studies or disentangling the persistent similarities from spurious differences using de Finetti's elegant exchangeability result.

2.1. Randomized Replications

To provide a concrete example of how randomized replications and probability distributions work in practice, consider the problem of comparing the resilience of K different brands of tires. A total of M drivers are selected at random and are asked to test-drive all brands for a number of

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

weeks. Let Y_{ij} denote the wear recorded for brand i by driver j. The Y_{ij}s differ from each other, but intuitively those sharing the same driver or the same brand should be more similar than those which share neither. We may capture such similarities and differences by positing the following distributional model for Y_{ij}:

$$Y_{ij} = a + b_i + d_j + \epsilon_{ij}, \qquad 1.$$

where a represents a common baseline, and b_i and d_j aim to capture the effects of using brand i and driver j, respectively. The so-called error term ϵ_{ij} describes the idiosyncratic variabilities, which are often assumed to form an independent and identically distributed (i.i.d.) sample from $N(0, \sigma^2)$, with σ^2 to be estimated from the data.

Independent and identically distributed (i.i.d.): describes a set of random variables that are mutually independent of each other but share the same probability distribution

This model alone does not capture similarities among drivers or brands. The fact that they are all persons or tires implies that they have similarities. We can again use distributions to model their similarities, such as

$$b_1, \ldots, b_K \overset{\text{i.i.d.}}{\sim} N(0, \tau_b^2), \quad d_1, \ldots, d_M \overset{\text{i.i.d.}}{\sim} N(0, \tau_d^2), \qquad 2.$$

where $\{b_1, \ldots, b_K\}$ and $\{d_1, \ldots, d_M\}$ are independent of each other, and the variances τ_b^2 and τ_d^2 are to be estimated from the data. The specification given by Equation 1 and Expression 2 is a special case of the so-called random-effects model, whose essence is to describe a randomized replication study via a probabilistic distribution. Fitting and checking such models constitutes a large portion of statistical endeavors, yet setting up the appropriate randomized replications is always the most critical step, be it a thought experiment or a real one. Without a (thought) mechanism that permits us to ignore in order to inform, we would drown in the data tsunami, or even a data stream.

If, instead of wear, we record the time, T_{ij}, of the first failure of a brand j tire driven by driver i, then studying the dependence between the response and, say, a covariate vector $\mathbf{x}_{ij}$ using the model in Equation 1 is inadequate. Estimates of regression coefficients are difficult to interpret and could be biased because the model ignores important features of the data structure, e.g., censoring. For instance, the latter occurs if some of the drivers do not experience a tire failure before the study is completed. The proportional hazards (PH) model of Cox (1972) (see the sidebar titled Cox's Proportional Hazards Model), developed precisely for studying the covariate effects on a time-to-event response variable, is another example of a sentient statistician's selective hearing. Cox's model assumes that the ratio of hazards associated to two different brand/driver pairs depends solely on the covariate vector and is constant in time. An added benefit is that the modeler is freed from the essentially impossible task of correctly specifying the entire baseline hazard function. This ingenious construction has turned the PH model into one of the most widely used statistical models due to its versatility, which allows, among other things, time-dependent covariates (Fisher & Lin 1999) and the integration of random effects (Vaida & Xu 2000).

COX'S PROPORTIONAL HAZARDS MODEL

The PH model (Cox 1972) allows the estimation of covariate effects on time-to-event response variables. It has found many applications in biomedical studies of survival and engineering studies of reliability. The hazard function at time t for a survival time variable T, $\lambda(t)$, can be interpreted as proportional to the instantaneous probability of the event (death, failure, etc.) at time t. In the presence of a covariate vector $\mathbf{x}$ that is expected to influence the distribution of T, the PH model posits that the hazard is the product between a baseline hazard $\lambda_0(t)$, independent of $\mathbf{x}$, and a positive function of the linear predictor $\beta^\top \mathbf{x}$ [usually $\exp(\beta^\top \mathbf{x})$, but other forms are possible].

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

2.2. Exchangeability

Exchangeability: a set of random variables are exchangeable if their joint probability stays the same regardless of how we permute their labels

Statistics offers expectations and predictions about occurrences that are not observed, based on some that are. What makes an inferential claim statistical is that, typically, the observed evidence on which the inference is based consists of a collection of variables that are amenable to a probabilistic description about their similarities and differences. Any reader of statistical textbooks would be familiar with the i.i.d. assumption introduced in the previous section as a foundational element for proving validity of various probabilistic methods. A collection of i.i.d. random variables exists only in an idealized form, under the assumption that the distribution of outcomes for the repeated events is known to be stable (in time, across subpopulations, etc.), or when the i.i.d. concept is used as a theoretical tool to prescribe differences and similarities, as in Equation 1 and Expression 2. In reality, the i.i.d. assumption rarely withstands close scrutiny.

Not surprisingly, statisticians and probabilists have been greatly bothered by this compromise. The great insight offered by the exchangeability concept allows statistical analyses to retain many of the advantages offered by the i.i.d. setup while capturing realistically a much wider range of repeatable random phenomena. Mathematically, the key assumption we invoke to model randomized replications is that they share the same probabilistic distribution—or rather, that we have no evidence to claim that they do not. Such failure to distinguish the random quantities from one another beyond their shared probabilistic distribution grants them the status of being replications to one another. De Finetti (2017) used the term exchangeability to replace what he thought of as a misnomer description of "independent events of unknown probability." A finite number of random variables are called exchangeable if their joint distribution is invariant to permutations. By definition, an infinite collection is exchangeable as long as so are any of its finite subsets.

De Finetti showed that an infinite sequence of Bernoulli random variables $\{X_n; n \geq 1\}$ is exchangeable if and only if there exists a distribution F such that $\{X_n; n \geq 1\}$ are conditionally independent and $\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) = \int \theta^{S_n}(1-\theta)^{n-S_n}\mathrm{d}F(\theta)$, where $S_n = \sum_{k=1}^{n} x_i$ for all $n \geq 1$. Moreover, θ is the limiting frequency $\theta = \lim_{n\to\infty} S_n/n$. Prior to collecting data, one could model a priori the distribution F using the available knowledge about the long range behavior of the Bernoulli trials, and the conditioning variable θ emerges naturally as the unknown probability of success. Therefore, the de Finetti theorem provides coherence and theoretical conceptualization for all the ingredients required in a Bayesian analysis. Exchangeability also builds connections with Fisher's concept of subpopulations, as noticed by Lindley & Novick (1981), and allows a principled treatment of underlying heterogeneous structures in a population.

In practice, finding the right variable to condition on provides clear benefits because it allows us to think of the data as i.i.d., which, in turn, makes it possible to predict the unseen from the seen. Using temporal data for illustration, if one believes that the future is exchangeable with the past, then the knowledge accumulated from measuring and understanding the past can inform, in probabilistic ways, the future. However, this needs careful qualifications, since tomorrow is clearly not exchangeable with today in a strict sense, minimally because the time is not reversible. What we typically mean is that those aspects of the future that we cannot (reasonably) describe using known characteristics will present uncertainties that are exchangeable with those recorded in the past.

3. FOLLOWING OUR NOSE: DETERMINACY THROUGH RANDOMNESS

Insertion of randomness in a system is a general principle that, when mastered properly, can overcome complexity in computation or in inference. There are many rewarding uses of this seemingly counterintuitive sense, with the aforementioned randomization being an obvious one. Here we further illustrate its versatility via Monte Carlo and statistical privacy.

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

3.1. Monte Carlo Integration

Statistical practice often requires investigating the properties of a function $f(\theta)$ when θ varies in some state space Θ. Consider the common problem of computing the expectation (i.e., average) of f with respect to a density $\pi(\theta)$: $I = \int_\Theta f(\theta)\pi(\theta)\mathrm{d}\theta$. Analytical calculation or deterministic numerical approximation (e.g., Owen 2003) often is impractical for many problems, especially when the dimension of θ is high. The Monte Carlo method (Metropolis & Ulam 1949) then becomes the only choice, by randomly sampling points $\{\theta_1, \ldots, \theta_n\}$ from $\pi(\theta)$, and then taking average of $\{f(\theta_i), i = 1, \ldots, n\}$, denoted by $\hat{I}_n$, to form the Monte Carlo estimator for I.

Obviously, the quality of the sampled points $\{\theta_1, \ldots, \theta_n\}$ is critical in determining the statistical accuracy of the $\hat{I}_n$ as an estimator for I. An i.i.d. sample from π generally is considered ideal but typically is very difficult to obtain even for reasonable-looking π. However, the unbiasedness of $\hat{I}_n$ is preserved as long as each θ_i is drawn from π, regardless of how statistically $\{\theta_1, \ldots, \theta_n\}$ may depend on each other. Consequently, we can give up the independence requirement in i.i.d. and construct the so-called Markov chain in the form of $\theta_{t+1} = \psi(\theta_t, U_{t+1})$, where t indexes the iteration, ψ is a deterministic updating function, and U_{t+1} is a random vector independent of all its predecessors $\{U_0, U_1, \ldots, U_t\}$. The central idea here is that by carefully choosing the function ψ, the resulting chain will generate a sequence of θ_t whose statistical properties approach those that are generated directly from π, as t increases. This construction is known as Markov chain Monte Carlo (MCMC), which has revolutionized Bayesian computation since the 1990s (Geman & Geman 1984, Tanner & Wong 1987, Gelfand & Smith 1992, Gilks et al. 1994, Brooks et al. 2011).

MCMC is also an example of a statistical idea that was not an idea in statistics—that is, it was neither invented by statisticians nor first published in the statistical literature—because it was developed in the early 1950s by physicists (Metropolis et al. 1953). Later, building on an analogy between images and lattice physical systems, Geman & Geman (1984) introduced the Gibbs sampler (see the sidebar titled The Gibbs Sampler), which was soon thereafter used in Bayesian methods for image reconstructions (Besag 1986, Besag & Green 1993), and Bayesian analysis for general statistical models (Gelfand & Smith 1992), including those formulated by augmenting the observed data with missing or latent ones (Tanner & Wong 1987).

The well-known algorithm by Metropolis et al. (1953) demonstrates how to construct ψ to sample from a target π. We start with an initial θ_0 and then sample θ^* (independently) from a symmetric proposal distribution p, typically chosen as a decent approximation of π but much easier to sample from. We then compute $r = \pi(\theta^*)/\pi(\theta_0)$. [The generalization by Hastings (1970) allows nonsymmetric proposals, at the expense of a proposal dependent r.] If $r \geq 1$, then we know that under π the value θ^* should appear more often in our sample than θ_0. Since θ_0 is already in our sample, we will need to include θ^* as well, and hence we let $\theta_1 = \theta^*$. If $r < 1$, say $r = 0.25$, then the value θ_0 should appear in our sample four times more frequently than θ^*. We can achieve

THE GIBBS SAMPLER

The Gibbs sampler is an MCMC algorithm that is used to sample from a d-dimensional posterior distribution $\pi(\theta)$ (Geman & Geman 1984, Liu et al. 1995). After initializing the chain at $\theta^{(0)}$, the chain updates at any iteration $t \geq 1$ by cycling through all the components of $\theta^{(t)}$ and sampling a new value for $\theta_j^{(t)}$ from $\pi(\theta_j^{(t)} \mid \theta_{1:(j-1)}^{(t)}, \theta_{(j+1):d}^{(t-1)})$, the conditional distribution of the jth component given the values of the remaining components at the current iteration. An important variation is the block Gibbs sampler, which cycles and updates groups of components instead of individual ones.

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

this by flipping a biased coin that lands on heads 25% of the time and let $\theta_1 = \theta^*$ if it is heads, and $\theta_1 = \theta_0$ if it is tails. We then repeat this process, with θ_1 replacing θ_0, to determine θ_2, and subsequently θ_t, $t = 3, 4, \ldots$.

Theoretical guarantees require a set of regularity conditions, under which the distribution of θ_t from the Metropolis–Hastings, Gibbs, or many other MCMC algorithms would approach the target π (Meyn & Tweedie 1994). This means that when t is sufficiently large—say, larger than an integer B, which is referred to as the burn-in and can be estimated in various ways (Brooks & Gelman 1998, Rosenthal 2002)—we can treat $\{\theta_t, t \geq B\}$ as an approximate sample from π and hence use them to form the Monte Carlo estimate I_n.

The magic of statistical estimation is that it not only provides an estimator with known theoretical properties but also assesses the statistical uncertainty about the estimator itself. A central limit theorem for Markov chains (Roberts & Tweedie 1996, Jones & Hobert 2001) ensures that the distribution of I_n will be well approximated by the Normal distribution $N(I, \hat{\sigma}_f^2/n)$ when n is sufficiently large. Here, $\hat{\sigma}_f^2$ is computed from $\{f(\theta_t), t \geq B\}$ (Geyer 1992, Flegal et al. 2008, Vats et al. 2019) using

$$\hat{\sigma}_f^2 = \hat{V}_f \left[1 + \frac{2}{n} \sum_{k=1+B}^{n+B} (n-k)\hat{\rho}_k \right], \qquad 3.$$

where $\hat{V}_f$ and $\hat{\rho}_k$ are the sampling variance and lag-k auto-correlation from $\{f(\theta_t), t \geq B\}$, respectively. This would allow us to provide an approximated 95% confidence interval estimator for I in the form of $\hat{I}_n \pm \hat{\tau}$, where $\hat{\tau}$ is an estimated margin of error given by $\hat{\tau} = 2\hat{\sigma}_f/\sqrt{n}$.

From Expression 3, we see Monte Carlo integration avoids the curse of dimension once we know how to sample correctly from the target π. Expression 3 also makes it apparent that reducing the within-chain correlations ρ_k will reduce the MCMC variance. Interestingly enough, randomness can be again adapted to serve our goals, for instance by introducing dependencies in the design of MCMC samplers to yield negative ρs (Rubinstein & Samorodnitsky 1985, Frigessi et al. 2000, Craiu & Meng 2005, Craiu & Lemieux 2007) or even to dissolve the dependence altogether and generate i.i.d. samples, the so-called perfect or exact sampling (e.g., see Propp & Wilson 1996, Craiu & Meng 2011). The construction of a perfect sampler relies on a clever use of randomness via a coupling strategy of multiple Markov chains, which often is not an easy task. Fortunately, similar coupling techniques have led to unbiased estimates of I for general MCMC algorithms (Heng & Jacob 2019, Jacob et al. 2020), which can be viewed as a sensible compromise between being perfect and being practical (Craiu & Meng 2022).

3.2. Randomized Responses and Differential Privacy

A challenge to the elicitation of survey response is when the question being surveyed is sensitive in nature. When respondents, especially those with perceived negative attributes (e.g., health condition, smoking, substance abuse), evade responses or lie about them, they induce systematic biases in the resulting statistical inference, and the power of randomization cannot be exercised fully.

The randomized response mechanism (Warner 1965) is an ingenious idea that can alleviate bias due to evasive answers in surveys. It uses a surprising double application of randomization within survey sampling. The mechanism, as originally constructed by Warner, is described in the sidebar titled Randomized Response. Using a random device—a biased coin with a known probability $p \in (1/2, 1)$ of turning up heads—to elicit answers, no individual has to report their sensitive feature verbatim, yet these randomized answers allow the interviewer to leverage statistical methods to unbiasedly infer the underlying true proportion and quantify uncertainty.

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

RANDOMIZED RESPONSE

The randomized response strategy of Warner (1965) is the earliest known mechanism that satisfies differential privacy (Dwork et al. 2006). Let $X_i \in \{0, 1\}$ be a binary attribute for individual i, and π the population proportion of 1s, which we wish to estimate. A total of n individuals are sampled with equal probability, and each of them is given a random device R_i (e.g., a biased coin) that simulates a Bernoulli random variable with known probability $p \in (1/2, 1)$. The individual reports $Y_i = 1$ if $R_i = X_i$, and $Y_i = 0$ otherwise, but does not disclose R_i or X_i. Consequently, the probability of observing $Y = 1$ is $\pi p + (1 - \pi)(1 - p)$. This allows us to estimate π unbiasedly—without knowing anyone's actual data X—via $\hat{\pi} = \left(p - 1 + \bar{Y}_n\right) / (2p - 1)$, where $\bar{Y}_n$ is the average of the observed $\{Y_1, \ldots, Y_n\}$.

The randomized response mechanism is a differentially private mechanism (Dwork et al. 2006), a formal privacy concept proposed decades later. Differential privacy has become the state-of-the-art standard for disclosure limitation for major corporations and statistical agencies, as seen in its high-profile adoption by the US Census Bureau for protecting the 2020 Decennial Census, including both the Public Law 94-171 redistricting data released in August 2021 (Abowd et al. 2022) and the Demographics and Housing Characteristics data scheduled for release in late 2023 (Abowd & Hawes 2023).

To say that a random function Y is ϵ-differentially private means that, for all neighboring values (x, x') (typically with unit metric in the neighborhood definition), the input data X can take, and for all possible states y for Y (here we assume Y is discrete for simplicity),

$$\frac{\Pr(Y(X) = y \mid X = x)}{\Pr(Y(X) = y \mid X = x')} \leq \exp(\epsilon) \qquad 4.$$

for a chosen $\epsilon > 0$, known as privacy loss budget. The larger the ϵ, the lesser the privacy protection, with a trade-off of greater preservation of information in the data. Differential privacy amounts to requiring that the probability distribution of Y does not change too much upon small changes in X. We see that the randomized response mechanism is an ϵ-differentially private mechanism with $\epsilon = \text{logit}(p) = \log(p/(1 - p))$, because the conditional probability ratios that Equation 4 requires, namely $\Pr(Y_i = 1 \mid X_i = 1)/\Pr(Y_i = 1 \mid X_i = 0)$ and $\Pr(Y_i = 0 \mid X_i = 0)/\Pr(Y_i = 0 \mid X_i = 1)$, are both equal to $p/(1 - p)$ by design. This also explains why $p = 1/2$ will make $\hat{\pi}$ take the useless value of ∞, because consequently $\epsilon = 0$, which means that the data are completely private, and hence no information for estimating π may come from the data alone.

However, when p approaches 1, ϵ approaches infinity, which means that we lose all privacy protection. But as a trade-off, we can have the fully efficient estimate $\hat{\pi} = \bar{Y}_n$. Of course, this full efficiency is not achievable in practice, because without some degrees of privacy protection, some individuals would refuse to respond. Worse, their decisions to not respond likely correlate with answers to sensitive question being asked. This would lead not only to reduced sample size but, most critically, also to a biased sample (see the general discussions of these problems in the next section), the very reason that we want to use a randomized mechanism in the first place.

The randomized response mechanism, and the more general differentially private mechanisms that come after it, are great examples of how a calculated randomness can be used to our advantage to probe difficult and confidential questions that do not otherwise succumb to direct and deterministic conquering. The transparent specification of these mechanisms allows for the principled decoding of statistical information contained in the veiled responses elicited from individuals (Gong 2022a,b). For an in-depth review of differential privacy and statistical disclosure limitation, readers are directed to Slavković & Seeman (2023).

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

4. SEEING THROUGH: ENLIGHTENMENT FROM DARK DATA

Data are never completely observed in reality, and their hidden parts have generating mechanisms that can greatly complicate statistical analyses because of the selection biases they create. The statistician's ability to decipher the latter and conjure ways to complete the former, seemingly out of thin air, has led to computational shortcuts and to more meaningful analyses. We illustrate the benefits of this enhanced vision with topics in sample survey and Bayesian computation.

Selection bias: systematic distortions created by selection processes in data collection that destroy their representativeness

Dark data: unobserved or unobservable data whose contemplation may lead to better models or computations

Finite-population statistics: all individual attributes are treated as fixed, and all statistics are formed by enumerating over individual indices in the (finite) population

Missing-data mechanism: the process that prevents data from being fully observed; often is described by a probability model

4.1. Quantifying Missing-Data Mechanisms and Data Defects

Dark data is a general term for any kind of data that are lost or distorted before analysis, as well as for unobserved or unobservable data, such as latent variables, that are constructed or conceptualized because they serve modeling and computational purposes (e.g., Hand 2020). The need to deal with missing or incomplete data, the most common form of dark data, is a rule rather than an exception. All of us, for example, have ignored survey questionnaires or provided only partial answers multiple times in our lives. Missing or incomplete data generally create at least three problems for analysis: (*a*) deteriorating data quality, (*b*) reducing data quantity, and (*c*) impeding the use of standard methods and software (Meng 2012). For example, when the reason for a nonresponse is correlated with the answers we are seeking, the survey data we observe are no longer representative of our target population. The consequence of this data distortion is more devastating than commonly realized.

To see this clearly, Meng (2018) shows that the actual error induced when using a sample mean ($\bar{Y}_n$) to estimate a population mean ($\bar{Y}_N$) can be decomposed into three factors:

$$\underbrace{\bar{Y}_n - \bar{Y}_N}_{\text{Actual error}} = \underbrace{\rho}_{\text{Data quality}} \times \underbrace{\sqrt{\frac{1-f}{f}}}_{\text{Data quantity}} \times \underbrace{\sigma}_{\text{Problem difficulty}} . \qquad 5.$$

Here, ρ is the finite-population correlation between Y_i and the binary recording indicator variable R_i (i.e., $R_i = 1$ if Y_i is recorded in the sample, and $R_i = 0$ otherwise), and σ is the standard deviation of the finite population $\{Y_1, \ldots, Y_N\}$. The second factor is entirely determined by the relative sample size $f = n/N$, and the third measures the finite population heterogeneity and hence the difficulty of estimating its mean. The first factor, the data defect correlation, ρ, is a useful measure for data quality because it captures the selection bias created by the dependence of R_i on Y_i; for a truly probabilistic sample, ρ is zero on average since whether $R_i = 1$ or not is uninfluenced by Y_i.

When ρ is not negligible (compared with N^{-1}), the resulting mean-squared error of $\bar{X}_n$ (averaging over possible randomness in R) is the same as that of a simple random sample with size $n_{\text{eff}} \approx f/(1-f)\rho^{-2}$. For the 2016 US presidential election, $\rho \approx -0.005$ pertains to voting for Trump based on survey data used by Meng (2018). This means that the sample mean from 2.3 million potential voters—about 1% of the eligible US voters—would not be more accurate for estimating Trump's vote share than that from a simple random sampling of about 400 voters who respond fully and truthfully, a 99.98% reduction from the apparent big data. Hence, the importance of data quality cannot be overemphasized.

The seemingly trivial adoption of the recording indicator R reflects a great advance in statistical analysis because it introduces a probabilistic framework for quantifying and modeling the selection biases created by any kind of missing-data mechanism. Specifically, Rubin (1976) used the conditional probability $\Pr(R = 1|Y_{\text{obs}}, Y_{\text{mis}})$ to model the mechanism, where Y_{obs} denotes the observed data points and Y_{mis} the missing ones. Selection bias exists when $\Pr(R = 1|Y_{\text{obs}}, Y_{\text{mis}})$ varies with Y_{mis}, a situation typically described as suffering from a nonignorable mechanism. When $\Pr(R = 1|Y_{\text{obs}}, Y_{\text{mis}}) \equiv \Pr(R = 1|Y_{\text{obs}})$, the so-called missing at random, we can avoid selection bias

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

by incorporating $\Pr(R = 1|Y_{\text{obs}})$ in the analysis. The easiest case is when $\Pr(R = 1|Y_{\text{obs}}, Y_{\text{mis}})$ is not influenced by either Y_{obs} or Y_{mis}, in which case Y_{obs} is simply a random subsample of the random sample we intended to collect, and hence Y_{obs} can be analyzed using standard methods. But this is a very rare occurrence in practice, so many more advanced methods have been developed and applied to address the issue of missing data (Rubin 1987, Little & Rubin 2019, Enders 2022) and more generally the challenges associated with a nonprobabilistic sample (Elliott & Valliant 2017, Zhang 2019, Wu 2022).

4.2. Intentionally Constructed Dark Data

Rather than regretting their unfortunate occurrence, one may choose to intentionally construct dark data either to aid the expression of the statistical model or to facilitate computation. As reviewed by Meng (2000), hypothetical data constructions come in many forms and shapes, such as latent variables (Loehlin 2004), auxiliary variables (Pollock 2002), and hidden states (Elliott et al. 2008).

A daring yet productive construction of dark data is the potential outcome notation, to refer to the complete set of possible values of a unit under all arms of treatment in an experiment (Neyman 1990). Suppose a clinical trial for a vaccine is designed such that a representative sample from the population was randomly assigned to two treatment arms, with $Z_i = 1$ indicating receipt of the vaccine by individual i and $Z_i = 0$ receipt of the placebo. All individuals complied with the study, and their immunity effects were measured and recorded accurately as $Y_i = Y(Z_i)$. However, even with such practically unachievable idealization, there is an inherent missing data structure that prevents us from even defining the treatment efficacy if we do not recognize it. In a typical experiment, an individual can receive either the vaccine or its placebo, but not both. But it is exactly the difference of the outcomes from both that defines the efficacy for the individual. In notation, we are interested in assessing $e_i = Y_i(1) - Y_i(0)$ for individual i, which is never directly observable. Nevertheless, the formulation via the potential outcome $\{Y_i(1), Y_i(0)\}$ is critical in clearly defining the scientifically meaningful estimand. It also clarifies what is and is not estimable.

For example, whereas it is impossible to assess individual efficacy e_i, for a finite population of size N, we can estimate the average causal effect, $A_N = \sum_{i=1}^{N} e_i/N$, by $\hat{A}_n = \bar{Y}(1) - \bar{Y}(0)$, where $\bar{Y}(Z)$ is the sample average of the outcomes from all individuals who have received the treatment $Z(= 0, 1)$. When the assignment to the treatment is done randomly, which is the case for most clinical trials, $\hat{A}_n$ is an unbiased estimate of A_N, despite the fact that we cannot estimate any individual e_i unbiasedly without making model assumptions (e.g., all individual efficiencies are the same).

Potential outcomes follow naturally from our imagination and permit the entertainment of counterfactuals and possible worlds beyond the observable one (Lewis 1974, Menzel 2017). With or without Neyman's notation, the idea permeates the study of causal inference from randomized experiments in statistics, economics and the social sciences (e.g., Mill 1906, book III, chapter 5; Fisher 1919; Tinbergen 1930; Haavelmo 1943; Cochran & Cox 1957; Cox 1958a). Rubin (1974, 1978) established the use of potential outcomes in observational studies, providing a formal framework for the analysis of causal effects that extends beyond classical randomized experiments.

Another kind of intentionally constructed missing data appears in computational algorithms, known as data augmentation, a term coined by Tanner & Wong (1987) in the context of MCMC, as discussed in Section 3.1. One of the most useful illustrations of this principle is represented by the EM algorithm for computing the maximum likelihood estimator (MLE) with missing data. Suppose that we are interested in computing the MLE for θ from a log-likelihood $\ell(\theta|X)$, where X denotes the data we observe. Here X may be only part of an intended larger data set, such as

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

EXPECTATION–MAXIMIZATION ALGORITHM

The EM algorithm is used to find the MLE of a parameter, θ, when a portion of the data is missing. Initialized at θ_0, the algorithm proceeds in an iterative manner over two steps (until convergence under a prespecified criterion):

- **E-step (expectation step)**: Compute $Q(\theta|\theta^{(t)})$, the conditional expectation of the complete-data log likelihood given the observed data X and $\theta = \theta^{(t)}$, the parameter estimate at the current iteration.
- **M-step (maximization step)**: Find the θ value that maximizes $Q(\theta|\theta^{(t)})$ and set it to $\theta^{(t+1)}$.

Emerging first in the works of Baum et al. (1970) and Sundberg (1976, 1974), the EM algorithm took shape and became widely popular through Dempster et al. (1977). It remains one of the most influential algorithms of all time.

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

$X = Y_{\text{obs}}$ using the notation from before, or it is the entire intended data set. Nevertheless, we can always augment X to $\{X, Z\}$, since bringing the augmented data Z can make finding MLE from the augmented data likelihood $\ell(\theta|Z, X)$ easier than from our original $\ell(\theta|X)$. However, once we have an estimate for θ, we can use the conditional model $f(Z|X, \theta)$ to impute/predict the $\ell(\theta|Z, X)$ function (since it depends on the unknown Z), which in turn can lead to a better estimate of θ. The expectation–maximization (EM) algorithm formalizes this intuitive notion of iterative improvement with an iterative scheme. Its popularity is due to both its simplicity and its stability as captured by the ascent property (Dempster et al. 1977, Wu 1983), which guarantees the iterative sequence $\{\theta^{(t)}, t = 0, 1, \ldots\}$ (see the sidebar titled Expectation–Maximization Algorithm) will never lower the likelihood being maximized, that is, $\ell(\theta^{(t+1)} \mid X) \geq \ell(\theta^{(t)} \mid X)$, for all $t \geq 0$. The simplicity here is both computational and conceptual, because statisticians' familiarity with many standard complete data models make it easier for us to identify effective data augmentation schemes.

The idea of easier or more feasible optimization via introducing hypothetical data translates to Monte Carlo sampling problems. The data augmentation (DA) algorithm of Tanner & Wong (1987) epitomizes this idea, which has stimulated many of the subsequent refinements and generalizations (e.g., Liu & Wu 1999, Van Dyk & Meng 2001, Pal et al. 2015; see Hobert 2011 for a review). With a DA algorithm, sampling from a posterior density $\pi(\theta \mid X) \propto p(\theta)p(X \mid \theta)$ is augmented to sampling from $\pi(\theta \mid X, Z) \propto p(\theta)p(X, Z \mid \theta)$. One then implements the Gibbs sampler described in Section 3.1 by iterating between sampling from $\pi(\theta \mid X, Z)$ and from $p(Z \mid X, \theta)$, which are the sampling counterparts of the M-step (maximization step) and E-step (expectation step), respectively. DA works whenever the introduction of Z allows the construction of a Markov chain with stationary density $\pi(\theta, Z \mid X)$ that is considerably easier to run than the one with stationary density $\pi(\theta \mid X)$. Not surprisingly, there is a large class of statistical models for which both EM and DA are effective, for likelihood and Bayesian computation, respectively. There are also many parallels between EM-type algorithms (e.g., the expectation and conditional maximization algorithm; Meng & Rubin 1993) and MCMC algorithms in terms of both theoretical properties and implementation strategies, as detailed by Van Dyk & Meng (2010).

5. THE MAGIC TOUCH: REFINEMENT BY CONFINEMENT

This sense brings to the fore the advantage of creating ingenious model constraints and links, be they conceptual or functional, between different parameters. Bayesian pooling and shrinkage tether the parameters in the model and allow the transfer of information between groups of observations. This results in more information being available for each parameter estimate. Similarly, the general concept of conditioning shrinks the sample space to relevant subspaces and allows the insertion of subject-matter knowledge into the mathematical aspects of the statistical analysis.

5.1. Shrinkage Estimation and Bayesian Pooling

Bayesian ideas have a long history that is interwoven with frequentist and fiducial insights (e.g., Fienberg 2006). Their importance is rooted in a series of attractive features that are available almost automatically and are backed by principles with firm theoretical support. In their modern form, Bayesian analysis and inference are fundamental to statistical modeling and to practical data analysis.

One powerful expression of Bayesian modeling is the flexible pooling of information. In the simplest terms, pooling is the act of borrowing strengths from individual constituents to achieve something superior than would be possible by each individually. The James–Stein estimator (Stein 1956, James & Stein 1961), and shrinkage estimators in general, emphasizes the importance of pooling in the simultaneous estimation of multiple means. Suppose that there are M groups, each with population mean $\{d_j : j = 1, \ldots, M\}$. One measurement is taken per group with standard Normal error: $Y_j \sim N(d_j, 1)$. The MLE in this case is just $\hat{d}_j^{\text{MLE}} = Y_j$. As much as we expect it to hold many good properties, it turns out that whenever $M \geq 3$, the MLE is inadmissible. The latter term signifies that the MLE can be easily dominated in terms of the risk incurred in the estimation: $R_L(\mathbf{d}, \hat{\mathbf{d}}(\mathbf{Y})) = \mathbb{E}(L(\mathbf{d}, \hat{\mathbf{d}}(\mathbf{Y})))$, where expectation is taken with respect to the sampling distribution of the data $\mathbf{Y}$. The loss function L can be measured in terms of the usual average squared error (James & Stein 1961) but also under a more general class of convex functions (Brown 1966). The James–Stein estimator,

$$\hat{d}_j^{JS} = \left(1 - \frac{M-2}{\sum_{j=1}^{M} Y_j^2}\right) Y_j, \qquad 6.$$

improves the estimation risk everywhere and, in particular, has a lower risk than the MLE.

The surprising result that MLE is not admissible can be explained rather intuitively from a data augmentation perspective by considering a regression setting with $\{(Y_i, d_i) : i = 1, \ldots, M\}$ as augmented data (since d_i is not observed). From this perspective, Stigler (1990) argued that the MLE corresponds to regressing Y on d because $\mathbb{E}(Y_i \mid d_i) = d_i$. But this is the wrong regression, since what we really want is to predict d_i from Y_i. Hence, we should regress d_i on Y_i, which is in the form of βY_i. Since d_i is unobserved, the multiplier in front of Y_i in Equation 6 is an unbiased estimator of β based on the Y_is alone (Stigler 1990, Meng 2005).

The James–Stein estimator itself is still not an admissible estimator for the population means. It also cannot be derived as a classic Bayes estimator. However, it can be motivated as an empirical Bayes solution to the estimation problem and serves as a basis for estimators with improved risk. Strawderman (1971) derived the class of minimax and admissible estimators, assuming the dimension $M \geq 5$, as the proper posterior mean under a class of Normal scale mixture priors. The powerful result of Brown (1971) established that proper and generalized Bayesian solutions constitute a complete class for the estimation of multivariate Normal means under quadratic loss, further strengthening the connection between minimax shrinkage estimation and Bayesian procedures.

Shrinkage estimators achieve uniformly superior estimation risk via an act of pooling the observations toward a point or a subspace. The hierarchical Bayes models, also known as multilevel models, allow for the flexible pooling of information about subpopulation-specific parameters. Through the specification of the hyperpriors, the modeler controls for the extent of pooling performed on the constituents. Consider a simplified version of the tire example from Section 2, where we assume there is no brand effect, and hence $b_i = 0, i = 1, \ldots, K$. We also absorb the baseline term a into the driver effects $d_j, j = 1, \ldots, K$, so Equation 1 becomes $Y_{ij} = d_j + \epsilon_{ij}$ with $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, and accordingly the prior from Expression 2 becomes $d_1, \ldots, d_M \overset{\text{i.i.d.}}{\sim} N(\mu_d, \tau_d^2)$. Here the value μ_d can be viewed as our prior knowledge of the average driver effects, and τ_d^2 expresses the similarity between these effects.

Admissibility: a statistical procedure is admissible if it is not completely dominated by another with respect to a given criterion

Empirical Bayes: the practice of Bayesian inference where aspects of the prior distribution are determined by the data being analyzed

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

Under this setup, after observing data $\vec{Y}_j = \{Y_{1j}, \ldots, Y_{Kj}\}$, the posterior distribution for d_j is

$$d_j \mid \vec{Y}_j \sim N((1-\lambda_j)\bar{Y}_{\cdot j} + \lambda_j \mu_d,\ \tilde{\tau}_d^2), \quad j = 1, \ldots, M, \tag{7.}$$

Ancillarity: a statistic is ancillary to a parameter if it is part of the minimum sufficient statistics but its distribution is parameter-free

where $\bar{Y}_{\cdot j}$ is the average of $\{Y_{1j}, \ldots, Y_{Kj}\}$, and λ_j is the pooling weight, given by $\lambda_j = \tilde{\tau}_d^2/\tau_d^2$, which is the ratio of prior precision to posterior precision, where precision is defined as the reciprocal of the variance. Here, the posterior precision $\tilde{\tau}_d^{-2}$ itself has a very intuitive expression:

$$\underbrace{\tilde{\tau}_d^{-2}}_{\text{Posterior precision}} = \underbrace{K\sigma^{-2}}_{\text{Data precision}} + \underbrace{\tau_d^{-2}}_{\text{Prior precision}}. \tag{8.}$$

Also, the term $K\sigma^{-2}$ is the data precision because it is the reciprocal of the variance of $\bar{Y}_{\cdot j}$, our estimate of d_j from data alone.

The pooling attribute of the multilevel model (Expression 7) is apparent, since the posterior mean of d_j is a linear combination between the no pooling mean (corresponding to $\lambda_j = 0$, or $\tau_d^2 = \infty$) and the complete pooling one (corresponding to $\lambda_j = 1$, or $\tau_d^2 = 0$). In the former case, the prior average μ_d is useless for estimating d_j because by setting $\tau_d^2 = \infty$, we declare that there is no similarity whatsoever among the drivers' effects. Hence there is no information to borrow from other drivers (via the mean μ_d), as captured by the prior precision of μ_d for estimating d_j being zero. Consequently, all our information about d_j comes from the data produced by the jth driver. In the latter case, by setting $\tau_d = 0$, we impose the assumption that all driver effects are the same, and since we have already specified that their average is μ_d, each of them must be μ_d, and hence no data would be needed or can help. In general, the value of λ_j determines how much pooling there is toward the common μ_d, depending on how similar we consider the driver effects to be: the more similarity, the smaller is τ_d^2, and hence the larger λ_j, resulting in more pooling.

In general, the posterior variance is smaller than the variance of the sample mean since it is shrunk by our prior knowledge/assumption of similarity among the individual effects. The borrowing of information phenomenon described in the simple model above is a cornerstone of hierarchical Bayes models (Gelman et al. 2013) that is useful in moderating the effect of extreme observations. One of its most lauded features is the improvement it brings to predictions of cluster-level effects (e.g., Gelman 2006), which, in our example, are the driver-specific effects d_j, for all $1 \le j \le K$.

5.2. Conditioning

As one of our favorite teachers of statistics likes to say, conditioning is the soul of statistics (Blitzstein & Hwang 2015, p. 42). Conditioning is an important operation that permeates methods under all schools of statistical inference: frequentist, likelihood, Bayesian, and fiducial.

To condition on a random variable is to reduce the realm of possible outcomes to only those that yield the same value of the conditioning variable. By imposing such a confinement, the variability is on average reduced. Broadly speaking, we perform conditioning in order to sharpen the focus and provide an inferential conclusion with better relevance to the question at hand (Liu & Meng 2016). By way of analogy, when asked a question (especially a trick question), the sassy response would always open with "it depends." By asserting that the answer to the question is a function of the scenario under which we envision the answer, we are effectively performing conditioning. But precisely what kind of things should the answer depend on?

Bayesian inference is always conditioned on the entirety of the observed data. Since the data contribute to the Bayesian posterior through the likelihood, the same thing can be said about likelihood inference. Often, there exists a sufficient reduction to the data—that is, a function $t(X)$—such that the likelihood function depends on the data X only through the sufficient statistic $t(X)$.

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

Operating outside of the likelihood and Bayesian inference paradigms, however, the importance of conditioning on the sufficient statistics is most elegantly captured by the Rao–Blackwell theorem (Rao 1945, Blackwell 1947). The theorem establishes that, if $y(X)$ is an estimator for the parameter of interest θ, the conditional expectation of $y(X)$ given a sufficient statistic of the parameter, say $t(X)$, will improve under any convex loss function. In other words, through conditioning, the sufficient statistic is capable of turning any estimator into a better one. The improved estimator, $\mathbb{E}(y(X) \mid t(X))$, is unbiased if and only if the original estimator $y(X)$ is unbiased. If the sufficient statistic $t(X)$ is, furthermore, complete, to Rao–Blackwellize the unbiased estimator results in a uniformly minimal variance unbiased estimator (Lehmann & Scheffé 1950, 1955), implying that it cannot be further improved upon under squared loss. The Rao–Blackwell theorem has important implications in the design of MCMC methods (Robert & Roberts 2021, Kong et al. 2007), among many other applications. It provides a guideline and inspires methods to construct estimators with reduced Monte Carlo variability (e.g., Liu et al. 1995).

In frequentist hypothesis testing, conditioning on ancillary statistics helps achieve another goal, to derive relevant conclusions based on the data at hand. Specifically, we wish to ensure that the conclusion enjoys the same quality—as measured by the test's statistical power—conditional on aspects that in themselves do not provide discriminative evidence toward the hypothesis. The idea is captured by the principle of conditionality (Cox & Hinkley 1974, Reid 1995), which requires that the inference be drawn conditional on an ancillary statistic when one exists. Consider the classic example given by Cox (1958b) and recapitulated by Fraser (2004). A noisy measurement about an unknown parameter θ was taken under one of the two equally possible scenarios. Under the first scenario, we have $X \sim N(\theta, \sigma^2)$, and under the second scenario, $X \sim N(\theta, (100\sigma)^2)$, making it much noisier than the first. Notably, the indicator for the scenario under which the measurement was taken is ancillary to the parameter of interest. For testing the null hypothesis of $\theta = 0$, the most powerful 95% hypothesis test would call for a rejection rule that is $|x| > 5\sigma$ if the experiment were conducted under the first scenario and $|x| > 164\sigma$ if under the second scenario. While this rejection region provides the greatest unconditional statistical power for all $\theta' \neq \theta$, the merits of the conclusions under the two scenarios are vastly different. In particular, it has excellent power under the first scenario but is only mediocre under the second. In contrast, the conditional test that would achieve the same statistical power under both scenarios would reject the null hypothesis when $|x| > 1.96\sigma$ if the measurement were taken under the first scenario and $|x| > 196\sigma$ if under the second. Even though the conditional test does not achieve the maximal unconditional power, the analyst would have the comfort of knowing that, regardless of which scenario they are working under, the quality of the resulting statistical conclusion is the same.

Conditioning can also be used to simplify more complex testing problems, particularly by getting rid of nuisance parameters. Because the benefit of dimensionality reduction is too great, it is sometimes performed at the expense of discriminative information about the parameter of interest. In the testing of independence from two-by-two contingency tables, Fisher's exact test is constructed conditional on the row and column margins. The row and column margins are not ancillary to the parameter of interest—here, the cross-product ratio of the population proportion parameters of the four cells. Conditioning on them is justified on the grounds that table marginals contain very little information about the cross-product ratio (Yates 1984) and because it turns an otherwise three-dimensional problem into a unidimensional one (for a more detailed discussion, see Little 1989). Conditioning as a guiding principle is reflected in practical techniques, such as stratification in survey sampling and blocking in experimental design. The sidebar titled Stratification and Blocking reviews these concepts.

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

STRATIFICATION AND BLOCKING

In survey sampling and observational studies, stratification refers to the method of obtaining samples from explicitly specified partitions of the intended population (Kish 1965). These partitions are usually determined by important features of the population, such as people of the same age, sex, race, and ethnicity, or other aspects that matter in the context of the study. We employ stratification to make sure that each defined subpopulation is represented in the sample precisely according to a defined proportion (usually the population proportion), without deviation due to randomness. A similar idea, called blocking, is encountered in experimental design (Box et al. 1978). Blocking is the deliberate effort on the experimenter's part to limit the random assignment of treatments in such a way that, again, each treatment arm comprises the defined subpopulations precisely according to a defined proportion. Both stratification and blocking are applications of the idea of conditioning.

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

6. JUST A TASTE: RICHNESS IN FRUGALITY

The title of this section may evoke the principle of parsimony, widely embraced and assiduously practiced by statisticians. However, the sense of frugality goes beyond that. As hinted in Section 1, the statistician is constrained to frugality by the very nature of the discipline and its methodology. Bootstrap demonstrates brilliantly that within the sample lies a wealth of information. The propensity score exemplifies frugality in the form of a single matching score that balances many observed attributes and empowers the study's conclusions to go beyond the mere detection of association.

6.1. Bootstrap

Bootstrap (Efron 1979) is a great illustration of the essence of randomized replications, as well as the caution needed to implement them correctly. A set of observations $D_n = \{X_1, \ldots, X_n\}$ does not automatically imply replications of any kind without further description of how they arrived at our desk or disk, regardless of how large n is or how many different values the observations may take. For example, they could be just a single deterministic sequence out of a mathematical textbook. But when we know, or more likely we assume, that it comes from a simple random sample of size n, we can obtain many (hidden) randomized replications. For example, there are n simple random samples of size $n-1$, denoted by $D_{n,-i}$, that are obtained by removing X_i from D_n. They are highly related to each other, but nevertheless, they are authentic randomized replications of each other because they are created by the same sampling process.

Now suppose we want to evaluate the variability of a statistic $g(D_n)$, such as the sample mean, $\sum X_i/n$. If we could have many replications, say $D_n^1, \ldots, D_n^m$, we could estimate the variability by using the variability among $g(D_n^1), \ldots, g(D_n^m)$. In practice, we only observe one D_n. Nevertheless, if we permit ourselves to approximate $g(D_n)$ by $g(D_{n-1})$, then we have at our disposal n replications of $g(D_{n-1})$ via $g(D_{n,-i})$, $i = 1, \ldots, n$, with which we can assess the variability and other properties (e.g., bias) of $g(D_{n-1})$. Historically, this is called the jackknife method (Quenouille 1956, Miller 1964), and the bootstrap is an improvement that avoids the need to approximate $g(D_n)$ by $g(D_{n-1})$. By resampling from $\{X_1, \ldots, X_n\}$ n times with replacement, bootstrap creates n (approximate) randomized replications of D_n.

The intention of bootstrap strategy is the same as the jackknife: using the internal replicability of D_n to create approximate randomized replications for itself. In that regard, bootstrap accomplishes a seemingly impossible task, "pulling ourselves up by our bootstraps," by concocting apparently new data from the old. On the surface, therefore, bootstrap might look like an act

of double dipping: From one sample, we create a large collection of resamples that allows one to build inference about yet a larger entity (the population). However, the nuanced theoretical justification of bootstrap would make clear that what is being leveraged are the hidden (assumed) replications within the observed D_n, information that is not fully utilized by the task of estimating $g(D_n)$ itself. The same idea has been put to good use in deriving other statistical methods, such as high-dimensional linear models (Zhao et al. 2021).

But there is no free lunch. Bootstrap is possible only if there indeed are hidden randomized replications in the sample. Suppose the original $D = \{X_1, \ldots, X_n\}$ actually comes from a time series—that is, the index represents a time order. In such cases, removing one observation or sampling with replacement would destroy this dependence structure, and hence there is little reason to expect that the resulting bootstrap sample, if naïvely constructed as before, would provide meaningful results. However, if we know that the time series is stationary—that is, any continuous segment of the same size shares the same probabilistic behavior regardless of their locations—then effectively we have hidden randomized replications (in units of segments) to build upon. This leads to the idea of using block bootstrap to sample (continuous) segments (i.e., blocks) of time series. The theoretical justification of block bootstrap is more involved—Bühlmann (2002) provides an excellent overview of block bootstrap and other bootstrap methods for time series. But the fundamental idea is the same: using internal replicability to approximate randomized replication.

6.2. Propensity Matching

Central to the analysis of randomized experiments and observational studies is the assignment mechanism, that is, the procedure through which units are allocated to different treatment groups. The key is to ensure that such an assignment will not introduce imbalances between the groups, which would make it difficult, if not impossible, to attribute differences in outcome to the treatments only. When the assignments are not made probabilistically, as in virtually all observational studies, we make attempts to rebalance the two groups. For example, we may choose a subset of those who received an existing treatment by matching their pretreatment attributes to those of participants who received an experimental treatment according to some matching criteria. Propensity matching (Rosenbaum & Rubin 1983) is perhaps the most popular rebalancing method, partly because it offers a surprisingly practical and effective method for the seemingly impossible task of maintaining balance statistically by matching only on a univariate score, the propensity score.

The propensity score of a unit i is its probability to appear in the treatment group, where every unit is identified by its known covariates X_i as well as its pair of potential outcomes $\{Y_i(1), Y_i(0)\}$, as introduced in Section 4.2. When the assignment mechanism, $Z_i \in \{0, 1\}$, is (conditionally) unconfounded (Dawid 1979), i.e., the assignment received by unit i is conditionally independent of its potential outcomes given the covariates, the propensity score can simply be written as (Rosenbaum & Rubin 1983, Imbens & Rubin 2015)

$$e(x) = \Pr(Z_i = 1 \mid X_i = x, Y_i(1), Y_i(0)) = \Pr(Z_i = 1 \mid X_i = x).$$

The propensity score plays a crucial role in the practical extension of causal analysis to observational studies. In classical randomized experiments, similarity among units (in the form of partial exchangeability) between the treatment arms can be established, more or less objectively, using assignment mechanisms under the control of the experimenter. Such luxury does not exist in observational studies, so to be able to conduct causal analysis, one must be able to decide whether two units are deemed similar based on available information. Thus, the criterion must be computable from the unit's available information, that is, their covariate information. This criterion that we need is the balancing score, a function b of the covariates X_i with the property that if we condition

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

on it, all differences, if any, in the covariates no longer bias the assignment in any way: $Z_i \perp X_i \mid b(X_i)$. Trivially, the totality of the covariate itself is a balancing score: $b(X_i) = X_i$. But matching on X_i, although desirable, is not practical, especially when X_i is of high dimension, because we would quickly run out of matching sample (e.g., to form a control group) since each person is unique.

Fiducial: a trusted standard for comparison, coming from a Latin word meaning "trusted"

The propensity score is the coarsest balancing score, in the sense that it is a function of all other balancing scores. This analogy is precisely that, in the effort to debias assignment probabilities of units, the balancing score is a sufficient statistic of the unit's characteristics, whereas the propensity score is a univariate minimal sufficient statistic. That is, we need only to match on a univariate quantity $e(X)$ in order to balance on the entire X for estimating the treatment effect, a rather remarkable achievement. The propensity score allows for the estimation of the causal effect in observational studies through identifying comparable units (Rosenbaum & Rubin 1984, Rubin & Thomas 1992) when we do not have the luxury of controlled randomized trials that enforce such comparability.

7. SENSE AND SENSIBILITY

We intend for this sense to be a culmination of our categories, in that it relies on all senses but allows the total ability to be larger than the sum. Anything that we wish to predict or infer remains unknowable to us unless we build a bridge between that and what we already know. Yet, such bridges can never be fully tested before we embark on them, precisely because we rely on them to find out where they would lead us. This bridge-building is the paramount challenge behind the practice of inference and prediction, and a leap of faith is necessary for such statistical endeavors. Of course, there are different bridges one can build, and which one is better for which journey requires the ultimate statistical sense and sensibility.

When we are asked to solve a deterministic problem, say, to find the value of θ when we know

$$X = \theta + U, \tag{9.}$$

there is only one bridge: We solve it by first expressing θ as

$$\theta = X - U \tag{10.}$$

and then use whatever information we have about X and U to determine θ. If $X = 6$ and $U = 0.1$, then θ must be 5.9. If $X = 6$, but we do not know U other than $|U| \leq 0.1$, then θ will be restricted to an interval $[5.9, 6.1]$. In either case, the link (Equation 9) permits us to transfer our deterministic knowledge, expressed as an equality or inequality, about X and U into either full or partial knowledge about θ.

Suppose now that our knowledge about U is stochastic, mathematically described by a probabilistic distribution: U is uniformly distributed on $[-0.1, 0.1]$. How can we transfer such stochastic knowledge to that about θ? This is where all the fun and frustration take place. We perhaps all have some vague sense that this distributional knowledge must impose some restrictions on θ. The question, then, is how should we proceed sensibly? At first sight, this seems to be a rather trivial problem, at least for this toy example. Since $X = 6$, and U is uniform on $[-0.1, 0.1]$, should then θ be uniform on $[X - 0.1, X + 0.1] = [5.9, 6.1]$ as Equation 10 suggests?

7.1. Fiducial Inference

Above is the answer provided by so-called fiducial inference, put forward by R.A. Fisher (1935), the founder of likelihood theory and many other statistical methods that are in use today. Fiducial inference is generally regarded as Fisher's biggest blunder because the seemingly natural operation of mimicking the deterministic equation-solving step turned out to be rather problematic

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

with stochastic quantities (see Zabell 1992, Dawid 2022). Even so, it influenced generations of statisticians and inspired fascinating topics such as structural inference (Fraser 1968), functional inference (Dawid & Stone 1982), Dempster–Shafer theory (Dempster 1966, Shafer 1976), generalized fiducial inference (Hannig et al. 2016), and inferential models (Martin & Liu 2015). A fundamental issue is that, whereas the distributional assumption on U induces a distribution for X for any given value of θ in Equation 9, it says nothing about whether θ can be described by a distribution or not. If θ is not even endowed with a distributional structure, then what does it mean to say θ is distributed uniformly on $[X - 0.1, X + 0.1]$? If it is given a distributional description, e.g., the so-called prior distribution π, then how θ is distributed after knowing X will depend on how we specify π. For example, if π is uniform on [6, 7], then it is impossible for θ to be uniformly distributed on [5.9, 6.1] after observing $X = 6$ because any value below 6 has already been eliminated by the prior π.

Objective Bayes: a somewhat paradoxical term describing efforts to invoke a prior while minimizing its impact

The fiducial argument relies on an explicit leap of faith, in that it continues to regard (Dempster 1963) the distribution of the auxiliary variable as unchanged after the observation. In the toy example, this means to steadfastly treat U as uniform on $[-0.1, 0.1]$ even after observing $X = 6$. The argument seems to be sensible intuitively. If we have not been given any prior knowledge or constraints about θ, then whatever information is generated by X via the unknown prior π about U is not quantifiable; hence, we can ignore it and persistently use the predata distribution information about U. Unfortunately, this argument cannot be coherently rationalized and operationalized in terms of probabilistic distributional calculations. This is because there is no probability distribution to describe ignorance: Any probability distribution specification about θ encodes restrictions on how likely one possible state of θ is versus another, and hence it cannot represent ignorance (see Section 7.2). The situation is a bit like doing arithmetic without the number zero, and using some other (small) numbers to represent zero, which might be acceptable in some practical cases, but logical inconsistencies and paradoxes would inevitably ensue (e.g., only a true zero remains zero after multiplication by any number).

7.2. Bayesian Inference

The Bayesian paradigm invokes a different kind of leap of faith, which is to require a prior distribution for θ. In reality, we may have very good knowledge to impose, say, $5 < \theta < 6$ but almost never have full confidence to say how θ is distributed over the interval (5, 6). This is even the case for those of us who are comfortable with the idea of using a probability distribution as a mathematical tool to describe our uncertainties (and hence avoid the issue of considering θ to be random). It is extremely rare, if not impossible, to know precisely the relative uncertainty about one possible state of θ versus another, and for all such pairs. Nevertheless, once a prior for θ is posed, it allows us to create the most relevant replications (θ', X) for drawing inference about θ that corresponds to our actual data X, as depicted in **Figure 1*b***. It is the most relevant because replications (θ', X) all share the same data X as observed, and hence the posterior distribution $p(\theta' \mid X)$ directly addresses our inference question: how likely is each possible value θ, denoted by θ', given our data X?

The concern of the unreliability of arbitrary priors has led to the quest for formal rules to guide their choice (Kass & Wasserman 1996). The quest gave rise to the literature of objective Bayes priors (e.g., Berger & Sun 2008; Ghosh 2011; Berger et al. 2012, 2015), or noninformative priors, as they are generally known. A specific class of objective priors are the so-called reference priors, which encompass priors that maximize some chosen distance between the prior and posterior distributions (Berger & Bernardo 1989, Ye 1993, Berger et al. 2009). The rationale of this maximization is to ensure that the likelihood function has the most impact, or to let the data speak as loudly as possible. Historically, the concern of not being able to choose a sensible prior was

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

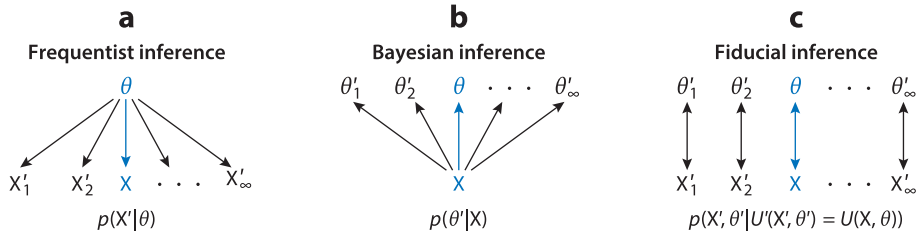

Figure 1

Replications underlying the frequentist, Bayesian, and fiducial inferential paradigms. X is the data, θ is the estimand, and p is the respective conditional distribution.

addressed by requiring a statistical procedure to be reliable regardless of the choice of the prior (Neyman 1934), which is the same as requiring reliability when the prior is a point mass at any plausible value of θ, as described in the next section.

7.3. Frequentist Inference

The frequentist paradigm corresponds to a top-down replication postulate, which considers all possible data sets that share the same θ value as the actual observed data set, as illustrated by **Figure 1*a***. The frequentist approach then investigates the long-run frequency properties of various methods over such hypothetical data sets, denoted by X', and chooses one with the desirable properties, such as being unbiased. A key feature of this approach is that it entirely bypasses the need to specify a distribution for θ. To some, this represents scientifically a more objective paradigm for inference. However, the no-free-lunch principle tells us that there must be a catch. The long-run frequency properties are assessed by considering how our procedures perform on these hypothetical X's, such as coverage probability of a confidence procedure or power of a hypothesis testing procedure. Whether such properties can guarantee anything about the procedures' reliability on the actual X we observe is fundamentally a matter of transferring faith about their collective merits into individualized performance. To some, this is an even more dangerous leap of faith than either the Bayesian one or the fiducial one, since the latter can be viewed as creating replications on the joint space (X', θ'), subject to the constraint that their noise U' must be the same as the U corresponding to the one underlying our actual data, as depicted in **Figure 1*c***.

But whichever replications we prefer, there is no escaping the leap of faith. Put differently, ultimately it is our sixth sense that completes our journey from the known to the unknowns. The differences between paradigms are more a matter of what we consider sensible, a question whose universal answer is no easier than establishing a theory of everything.

8. HOW DID WE FORM OUR SENSES?

Our sensical tour started with a pleasantly surprising invitation from the editors of this journal to write about "The Top X Big Ideas in Statistics." Given both the long history and the accelerating evolution of statistics, where does one even begin with such a task?

Being all trained quantitatively, but with appreciation for qualitative thinking, we first pondered on the meaning of almost every word in the ask. In this context, "top X" seems to be a sure invitation for controversy because it demands a univariate ordering of a high-dimensional enterprise. "Big" is also big trouble: big on what? Ingenuity, insight, influence, impact, or all of these? And by "ideas," do we mean products of human intelligence, not discovery, or as secrets of God (or nature) revealed by humans? Two central pillars of statistics and probability are the law of large

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

numbers and the central limit theorem. Neither is an idea. Whether or not human beings find its mathematical expression, the bell curve would still reveal itself everywhere from Galton's boards to scoreboards.

Phronesis: a form of practical wisdom and the exercise of sound judgment in defining aims and devising ways to attain them

To further complicate the matter, "ideas in statistics" are not the same as "statistical ideas," which need not be in statistics. A recent example is the use of data augmentation for incorporating prior knowledge or model considerations by creating synthetic training samples that reflect them (see, e.g., Taylor & Nitschke 2018, Shorten & Khoshgoftaar 2019). This is a fine statistical idea and a good example of using randomized replications to operationalize probabilistic modeling. But, because it is proposed in the machine learning literature, it is unclear whether there has been an awareness in that literature of the terminology clash with the statisticians' own data augmentation, reviewed in Section 4, let alone any discussion of the similarities and differences between the two usages of the same term [e.g., no mention of the statistical usage in the review article by Shorten & Khoshgoftaar (2019)].

Ultimately we were inspired by the wisdom behind Stigler's (2016) *The Seven Pillars of Statistical Wisdom* to accentuate the essence of statistics not by top or big ideas or discovery, but by its broad conceptual framing and its intellectual axes. We chose to categorize our framing into the six senses of statistics for reasons listed in Section 1. Evidently, our own intellectual trajectories led us to the particular set of examples used to illustrate each sense. Space constraints have limited the scope of illustrations even further. We would have loved to include other topics that have colored our relationship with statistics. Indeed, we only touched upon a few of the seven pillars of Stigler (2016): aggregation, information measurement, likelihood, intercomparison, regression, experimental design, and residual. The overlap between our example topics and those listed in "What Are the Most Important Statistical Ideas of the Past 50 Years?" (Gelman & Vehtari 2021) is of a similar order: counterfactual causal inference, bootstrap and simulation-based inference, overparameterized models and regularization, Bayesian multilevel models, generic computation algorithms, adaptive decision analysis, robust inference, and exploratory data analysis.

Such comparisons should make clear that, whereas we believe the six senses are essential for the statistical enterprise, there are many other important methods and ideas than those mentioned in this article (or any single article). The more mathematically inclined reader may have also noticed that we have steered clear of probability concepts (e.g., central limit theorems, ergodic theory, asymptotics) that are fundamental to statistical derivations and justifications. They are, in our opinion, the tools we need to advance once our senses are pointing us in the right direction, but they do not define a statistician's phronesis. Nevertheless, we hope that the examples used will facilitate the reader's ability to sharpen their statistical senses and guide discovery of new links with their favorite methods.

We are writing this at a time when the discipline of statistics is expanding rapidly under important stimulants, such as the emergence of data science as an ecosystem (Meng 2019a,b). Statistical ideas and principles are meshing with computational algorithms to solve scientific and societal problems that carry the burden of enormous, messy, and often confidential data; the responsibility of dealing with complex relationships; the increased darkness of the black boxes promoted by advances in machine learning; and the societal demands of transparency and interpretability due to fairness considerations, to name only a few. New senses therefore are likely to emerge, but those discussed in this article are time honored and will continue to play an important role in the way new problems are tackled.

For example, the time-honored bias–variance trade-off manifests itself in the relevance–robustness trade-off (Liu & Meng 2016) in the context of accumulating statistical evidence for assessing the effectiveness of individualized treatments, an increasingly common desire due to the availability of the (seemingly) big data. Since no two individuals are identical, whether as biological

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

or social beings, evidence from proxy individuals is approximate in nature. The more resemblance between the proxy individuals (or training sample) and the target individual, the more relevant are the approximating assessments. But this relevance comes at the expense of fewer available proxies, resulting in assessments that are less robust. Conversely, we can assemble many proxy individuals if we relax the proxy criterion and hence have more stable statistical assessment. But the results may not be that relevant for the target individual because of the loose criterion employed in choosing the proxies.

Whereas the optimal construction of the training sample is a holy grail in data science, the kind of statistical phronesis that is discussed in this article helps us make a sensible treatment decision. As an extreme example, insisting on having fully robust evidence is similar to insisting on 100% coverage, which would then lead to a useless—tautological—confidence interval. But by giving up a small amount of robustness, we can achieve much better relevance, just as we obtain a more meaningful confidence interval by giving up a small amount of confidence (e.g., 5%).

Those with good statistical senses are likely to possess more confidence in their abilities to handle whatever the future may bring. Our belief is built on the observation (or our sixth sense?) that statistical principles have demonstrated enduring importance and influence in every scientific revolution, be it big data mining, machine learning, data science, or the *n*th Spring of Artificial Intelligence.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Nancy Reid and David Madigan for the invitation to write this article. This work has been supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada (R.V.C.) and National Science Foundation (NSF) (R.G. and X-L.M.).

LITERATURE CITED

Abowd J, Ashmead R, Cumings-Menon R, Garfinkel S, Heineck M, et al. 2022. The 2020 Census Disclosure Avoidance System TopDown algorithm. *Harv. Data Sci. Rev.* **https://doi.org/10.1162/99608f92.529e3cb9**

Abowd JM, Hawes MB. 2023. Confidentiality protection in the 2020 US Census of Population and Housing. *Annu. Rev. Stat. Appl.* 10:119–44

Agresti A. 2021. The foundations of statistical science: a history of textbook presentations. *Braz. J. Probab. Stat.* 35(4):657–98

Agresti A, Meng X-L. 2013. *Strength in Numbers: The Rising of Academic Statistics Departments in the US*. New York: Springer

Baum LE, Petrie T, Soules G, Weiss N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* 41(1):164–71

Berger JO, Bernardo JM. 1989. Estimating a product of means: Bayesian analysis with reference priors. *J. Am. Stat. Assoc.* 84(405):200–7

Berger JO, Bernardo JM, Sun D. 2009. The formal definition of reference priors. *Ann. Stat.* 37(2):905–38

Berger JO, Bernardo JM, Sun D. 2012. Objective priors for discrete parameter spaces. *J. Am. Stat. Assoc.* 107(498):636–48

Berger JO, Bernardo JM, Sun D. 2015. Overall objective priors. *Bayesian Anal.* 10(1):189–221

Berger JO, Sun D. 2008. Objective priors for the bivariate normal model. *Ann. Stat.* 36(2):963–82

Besag J. 1986. On the statistical analysis of dirty pictures. *J. R. Stat. Soc. Ser. B* 48(3):259–79

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

Besag J, Green PJ. 1993. Spatial statistics and Bayesian computation. *J. R. Stat. Soc. Ser. B* 55(1):25–37
Blackwell D. 1947. Conditional expectation and unbiased sequential estimation. *Ann. Math. Stat.* 18:105–10
Blitzstein JK, Hwang J. 2015. *Introduction to Probability*. Boca Raton, FL: Chapman & Hall/CRC
Box GE, Hunter WH, Hunter S. 1978. *Statistics for Experimenters*. New York: Wiley
Brooks S, Gelman A, Jones GL, Meng X-L, eds. 2011. *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman & Hall/CRC
Brooks SP, Gelman A. 1998. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 7(4):434–55
Brown LD. 1966. On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Stat.* 37(5):1087–136
Brown LD. 1971. Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Stat.* 42(3):855–903
Bühlmann P. 2002. Bootstraps for time series. *Stat. Sci.* 17:52–72
Cochran WG, Cox GM. 1957. *Experimental Designs*. New York: Wiley
Cox DR. 1958a. *Planning of Experiments*. New York: Wiley
Cox DR. 1958b. Some problems connected with statistical inference. *Ann. Math. Stat.* 29(2):357–72
Cox DR. 1972. Regression models and life-tables. *J. R. Stat. Soc. Ser. B* 34(2):187–202
Cox DR, Hinkley DV. 1974. *Theoretical Statistics*. Boca Raton, FL: Chapman & Hall/CRC
Craiu RV, Lemieux C. 2007. Acceleration of the multiple-try Metropolis algorithm using antithetic and stratified sampling. *Stat. Comput.* 17(2):109
Craiu RV, Meng X-L. 2005. Multiprocess parallel antithetic coupling for backward and forward Markov chain Monte Carlo. *Ann. Stat.* 33(2):661–97
Craiu RV, Meng X-L. 2011. Perfection within reach: exact MCMC sampling. In *Handbook of Markov Chain Monte Carlo*, ed. S Brooks, A Gelman, G Jones, X-L Meng, pp. 199–226. Boca Raton, FL: Chapman & Hall/CRC
Craiu RV, Meng X-L. 2022. Double happiness: enhancing the coupled gains of L-lag coupling via control variates. *Stat. Sin.* 32(4):1745–66
Dawid AP. 1979. Conditional independence in statistical theory. *J. R. Stat. Soc. Ser. B* 41(1):1–15
Dawid AP. 2022. Fiducial inference then and now. In *Handbook on Bayesian, Fiducial and Frequentist (BFF) Inferences*, ed. J Berger, X-L Meng, N Reid, M Xie. Boca Raton, FL: Chapman & Hall/CRC. In press
Dawid AP, Stone M. 1982. The functional-model basis of fiducial inference. *Ann. Stat.* 10(4):1054–67
De Finetti B. 2017. *Theory of Probability: A Critical Introductory Treatment*. New York: Wiley
Dempster AP. 1963. On direct probabilities. *J. R. Stat. Soc. Ser. B* 25(1):100–10
Dempster AP. 1966. New methods for reasoning toward posterior distributions based on sample data. *Ann. Math. Stat.* 37(2):355–74
Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* 39(1):1–22
Dwork C, McSherry F, Nissim K, Smith A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, ed. S Halevi, T Rabin, pp. 265–84. New York: Springer
Efron B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7:1–26
Elliott MR, Valliant R. 2017. Inference for nonprobability samples. *Stat. Sci.* 32(2):249–64
Elliott RJ, Aggoun L, Moore JB. 2008. *Hidden Markov Models: Estimation and Control*. New York: Springer
Enders CK. 2022. *Applied Missing Data Analysis*. New York: Guilford. 2nd ed.
Fienberg SE. 2006. When did Bayesian inference become "Bayesian"? *Bayesian Anal.* 1(1):1–40
Fisher LD, Lin DY. 1999. Time-dependent covariates in the Cox proportional-hazards regression model. *Annu. Rev. Public Health* 20:145–57
Fisher RA. 1919. The causes of human variability. *Eugen. Rev.* 10(4):213
Fisher RA. 1935. The fiducial argument in statistical inference. *Ann. Eugen.* 6(4):391–98
Flegal J, Haran M, Jones G. 2008. Markov chain Monte Marlo: Can we trust the third significant figure? *Stat. Sci.* 23(2):250–60
Fraser DA. 1968. *Structural Inference*. New York: Wiley
Fraser DA. 2004. Ancillaries and conditional inference. *Stat. Sci.* 19(2):333–69

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

Frigessi A, Gåsemyr J, Rue H. 2000. Antithetic coupling of two Gibbs sampler chains. *Ann. Stat.* 28:1128–49
Gelfand AE, Smith AFM. 1992. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* 87:523–32
Gelman A. 2006. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics* 48(3):432–35
Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013. *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC
Gelman A, Vehtari A. 2021. What are the most important statistical ideas of the past 50 years? *J. Am. Stat. Assoc.* 116(536):2087–97
Geman S, Geman D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intel.* 6(6):721–41
Geyer CJ. 1992. Practical Markov chain Monte Carlo (with discussion). *Stat. Sci.* 7:473–511
Ghosh M. 2011. Objective priors: an introduction for frequentists. *Stat. Sci.* 26(2):187–202
Gilks WR, Thomas A, Spiegelhalter DJ. 1994. A language and program for complex Bayesian modelling. *J. R. Stat. Soc. Ser. D* 43(1):169–77
Gong R. 2022a. Exact inference with approximate computation for differentially private data via perturbations. *J. Priv. Confid.* 12(2). **https://doi.org/10.29012/jpc.797**
Gong R. 2022b. Transparent privacy is principled privacy. *Harv. Data Sci. Rev.* **https://doi.org/10.1162/99608f92.b5d3faaa**
Haavelmo T. 1943. The statistical implications of a system of simultaneous equations. *Econom. J. Econom. Soc.* 11:1–12
Hacking I. 2006. *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge, UK: Cambridge Univ. Press
Hand DJ. 2020. *Dark Data: Why What You Don't Know Matters*. Princeton, NJ: Princeton Univ. Press
Hannig J, Iyer H, Lai RC, Lee TC. 2016. Generalized fiducial inference: a review and new results. *J. Am. Stat. Assoc.* 111(515):1346–61
Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1):97–109
Heng J, Jacob PE. 2019. Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika* 106(2):287–302
Hobert JP. 2011. The data augmentation algorithm: theory and methodology. In *Handbook of Markov Chain Monte Carlo*, ed. S Brooks, A Gelman, G Jones, X-L Meng, pp. 253–93. Boca Raton, FL: Chapman & Hall/CRC
Imbens GW, Rubin DB. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge Univ. Press
Jacob PE, O'Leary J, Atchadé YF. 2020. Unbiased Markov chain Monte Carlo with couplings (with discussion). *J. R. Stat. Soc. Ser. B* 82(3):543–600
James W, Stein C. 1961. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1: *Contributions to the Theory of Statistics*, ed. J Neyman, pp. 361–79. Berkeley: Univ. Calif. Press
Jones G, Hobert J. 2001. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Stat. Sci.* 16:312–34
Kass RE, Wasserman L. 1996. The selection of prior distributions by formal rules. *J. Am. Stat. Assoc.* 91(435):1343–70
Kish L. 1965. *Survey Sampling*. New York: Wiley
Kong A, McCullagh P, Meng X-L, Nicolae DL. 2007. Further explorations of likelihood theory for Monte Carlo integration. In *Advances in Statistical Modeling and Inference: Essays in Honor of Kjell A. Doksum*, ed. V Nair, pp. 563–92. Singapore: World Sci.
Lehmann E, Scheffé H. 1950. Completeness, similar regions, and unbiased estimation: part I. *Sankhyā Indian J. Stat.* 10:305–40
Lehmann E, Scheffé H. 1955. Completeness, similar regions, and unbiased estimation: part II. *Sankhyā Indian J. Stat.* 15(3):219–36
Lewis D. 1974. Causation. *J. Philos.* 70(17):556–67
Lindley DV, Novick MR. 1981. The role of exchangeability in inference. *Ann. Stat.* 9(1):45–58

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

Little RJ. 1989. Testing the equality of two independent binomial proportions. *Am. Stat.* 43(4):283–88
Little RJ, Rubin DB. 2019. *Statistical Analysis with Missing Data*. New York: Wiley
Liu JS, Wong WH, Kong A. 1995. Covariance structure and convergence rate of the Gibbs sampler with various scans. *J. R. Stat. Soc. Ser. B* 57(1):157–69
Liu JS, Wu YN. 1999. Parameter expansion for data augmentation. *J. Am. Stat. Assoc.* 94(448):1264–74
Liu K, Meng X-L. 2016. There is individualized treatment. Why not individualized inference? *Annu. Rev. Stat. Appl.* 3:79–111
Loehlin JC. 2004. *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis*. London: Psychology
Martin R, Liu C. 2015. *Inferential Models: Reasoning with Uncertainty*. Boca Raton, FL: Chapman & Hall/CRC
Meng X-L. 2000. Missing data: Dial M for??? *J. Am. Stat. Assoc.* 95(452):1325–30
Meng X-L. 2005. Comment: computation, survey and inference. *Stat. Sci.* 20(1):21–28
Meng X-L. 2012. You want me to analyze data I don't have? Are you insane? *Shanghai Arch. Psychiatry* 24(5):297
Meng X-L. 2018. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* 12(2):685–726
Meng X-L. 2019a. Data science: an artificial ecosystem. *Harv. Data Sci. Rev.* 1(1). **https://hdsr.mitpress.mit.edu/pub/jhy4g6eg**
Meng X-L. 2019b. Five immersive 3D surroundings of data science. *Harv. Data Sci. Rev.* 1(2). **https://doi.org/10.1162/99608f92.ab81d0a9**
Meng X-L. 2020. Information and uncertainty: two sides of the same coin. *Harv. Data Sci. Rev.* 2(2). **https://doi.org/10.1162/99608f92.c108a25b**
Meng X-L, Rubin DB. 1993. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 80(2):267–78
Menzel C. 2017. Possible worlds. In *The Stanford Encyclopedia of Philosophy*, ed. EN Zalta. Stanford, CA: Metaphysics Res. Lab, Stanford Univ. **https://plato.stanford.edu/entries/possible-worlds/**
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21(6):1087–92
Metropolis N, Ulam S. 1949. The Monte Carlo method. *J. Am. Stat. Assoc.* 44(247):335–41
Meyn SP, Tweedie RL. 1994. Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.* 4(4):981–1011
Mill JS. 1906. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*. London: Longmans, Green, Reader and Dyer
Miller RG. 1964. A trustworthy jackknife. *Ann. Math. Stat.* 35(4):1594–605
Neyman J. 1934. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. R. Stat. Soc.* 97:558–625
Neyman J. 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat. Sci.* 5(4):465–72
Owen AB. 2003. Quasi–Monte Carlo sampling. In *Monte Carlo Ray Tracing: SIGGRAPH 2003, Course 44*, ed. H Jensen, pp. 69–88. New York: ACM
Pal S, Khare K, Hobert JP. 2015. Improving the data augmentation algorithm in the two-block setup. *J. Comput. Graph. Stat.* 24(4):1114–33
Pollock KH. 2002. The use of auxiliary variables in capture-recapture modelling: an overview. *J. Appl. Stat.* 29(1–4):85–102
Propp JG, Wilson DB. 1996. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struct. Algorithms* 9(1&2):223–52
Quenouille MH. 1956. Notes on bias in estimation. *Biometrika* 43(3/4):353–60
Rao CR. 1945. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* 37(3):81–91
Reid N. 1995. The roles of conditioning in inference. *Stat. Sci.* 10(2):138–57
Robert CP, Roberts G. 2021. Rao–Blackwellisation in the Markov chain Monte Carlo era. *Int. Stat. Rev.* 89(2):237–49
Roberts GO, Tweedie RL. 1996. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* 83(1):95–110

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55
Rosenbaum PR, Rubin DB. 1984. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* 79(387):516–24
Rosenthal JS. 2002. Quantitative convergence rates of Markov chains: a simple account. *Electron. Commun. Probab.* 7:123–28
Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66(5):688–701
Rubin DB. 1976. Inference and missing data. *Biometrika* 63(3):581–92
Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6(1):34–58
Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley
Rubin DB, Thomas N. 1992. Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika* 79(4):797–809
Rubinstein YR, Samorodnitsky G. 1985. Variance reduction by the use of common and antithetic random variables. *J. Stat. Comput. Simul.* 22(2):161–80
Shafer G. 1976. *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton Univ. Press
Shafer G. 2019. Pascal's and Huygens's game-theoretic foundations for probability. *Sartoniana* 32(9):117–45
Shafer G. 2022. "So much data. Who needs probability?" Have we been here before? *Int. J. Approx. Reason.* 141:183–89
Shorten C, Khoshgoftaar TM. 2019. A survey on image data augmentation for deep learning. *J. Big Data* 6(1):1–48
Slavković A, Seeman J. 2023. Statistical data privacy: a song of privacy and utility. *Annu. Rev. Stat. Appl.* 10:189–218
Stein C. 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1: *Contributions to the Theory of Statistics*, ed. J Neyman, pp. 197–206. Berkeley: Univ. Calif. Press
Stigler SM. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: Harvard Univ. Press
Stigler SM. 1990. The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Stat. Sci.* 5(1):147–55
Stigler SM. 2002. *Statistics on the Table: The History of Statistical Concepts and Methods*. Cambridge, MA: Harvard Univ. Press
Stigler SM. 2016. *The Seven Pillars of Statistical Wisdom*. Cambridge, MA: Harvard Univ. Press
Strawderman WE. 1971. Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Stat.* 42(1):385–88
Sundberg R. 1974. Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Stat.* 1(2):49–58
Sundberg R. 1976. An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Commun. Stat. Simul. Comput.* 5(1):55–64
Tanner MA, Wong WH. 1987. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82(398):528–40
Taylor L, Nitschke G. 2018. Improving deep learning with generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1542–47. Piscataway, NJ: IEEE
Tinbergen J. 1930. Bestimmung und Deutung von Angebotskurven: ein Beispiel. *Z. Nationalökonomie* 1(5):669–79
Vaida F, Xu R. 2000. Proportional hazards model with random effects. *Stat. Med.* 19(24):3309–24
Van Dyk DA, Meng X-L. 2001. The art of data augmentation. *J. Comput. Graph. Stat.* 10(1):1–50
Van Dyk DA, Meng X-L. 2010. Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: a graphical guide book. *Stat. Sci.* 25(4):429–49
Vats D, Flegal JM, Jones GL. 2019. Multivariate output analysis for Markov chain Monte Carlo. *Biometrika* 106(2):321–37
Warner SL. 1965. Randomized response: a survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.* 60(309):63–69

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

Wu C. 2022. Statistical inference with non-probability survey samples (with discussions). *Surv. Methodol.* In press
Wu CJ. 1983. On the convergence properties of the EM algorithm. *Ann. Stat.* 11(1):95–103
Yates F. 1984. Tests of significance for 2×2 contingency tables. *J. R. Stat. Soc. Ser. A* 147(3):426–49
Ye K. 1993. Reference priors when the stopping rule depends on the parameter of interest. *J. Am. Stat. Assoc.* 88(421):360–63
Zabell SL. 1992. R.A. Fisher and fiducial argument. *Stat. Sci.* 7(3):369–87
Zhang LC. 2019. On valid descriptive inference from non-probability sample. *Stat. Theory Relat. Fields* 3(2):103–13
Zhao S, Witten D, Shojaie A. 2021. In defense of the indefensible: a very naive approach to high-dimensional inference. *Stat. Sci.* 36(4):562–77

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

Annual Review of Statistics and Its Application

Volume 10, 2023

Contents

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics

Annu. Rev. Stat. Appl. 2023.10:699-725. Downloaded from www.annualreviews.org
Access provided by 93.49.167.13 on 03/11/23. See copyright for approved use.