# Statistics, Overview

Brett Stoudt
John Jay College, The City University of New York, New York, NY, USA

## Introduction

Quantitative research is the dominant paradigm in psychology and, thus, the primary way the discipline judges "truth" and creates new knowledge. Like any privileged standpoint, the use of aggregated numbers to understand individual psychological processes, attitudes, or behaviors is rarely questioned. Debates and critiques about statistics are often held at the level of what techniques are most appropriate to use technically and mathematically. Rarely are theoretical discussions had about why, when, or even *if* quantification can accurately model human experience. Thus, while psychology students get trained in statistics, often very sophisticated statistics, they are seldom offered an opportunity to approach the subject from a critical perspective; nevertheless, a critical quantitative tradition does exist in the social sciences.

## Definition

Though "critical statistics" is an uncommon topic in psychology, the occasional article, chapter, or book has accumulated over the years into a respectable body of work. Collectively, five themes emerge from this literature.

1. A "critical statistics" is not an outcome or an achievement but a way to approach the entire research process involved with quantifying psychological and social experience. Applying a critical perspective to statistics includes a critical awareness at all the stages along the way: the questions, the research designs, the types of instruments, the strategies for measurement, the questions that are asked, the sampling, how the data are "cleaned" in a dataset, how the variables are disaggregated and aggregated, what statistical procedures are used, how the data are explored, how and with whom the analyses are discussed, how the analyses are visualized, how the findings are presented/written, where and with whom the analyses are presented/written, and many more small and large pivot points that evolve as research unfolds.
2. A "critical statistics" acknowledges that quantitative methods can both distort and enlighten, has strengths and weaknesses, and therefore is in conversation with not in opposition to qualitative methods. The divide between qualitative and quantitative techniques is an unproductive debate that can prevent researchers from using methods that more fully address their research questions as well as divert researchers from much more important ontological and epistemological discussions (i.e., what exists of human psychology and behavior to be studied and what evidence will we trust as "truth," "fact," or "knowledge"). Methodological pluralism, also known as mixed methods or triangulated methods, has garnered significant momentum and support. Therefore, quantitative research is *a* useful method for understanding social experience, but it is not *the only* useful method for understanding social experience.
3. A "critical statistics" is a principled and action-oriented approach to the social sciences in the larger pursuit of democratic participation, equality, and justice. It is in service of marginalized communities to reveal oppressive systems, institutions, and policies. A critical approach to quantitative methods is connected with critical theories including critical race theory, feminist theory, indigenous theory, queer theory, and any other theories that help reveal social inequality and help promote activism, emancipation, and justice. A quantitative criticalist, as described by Stage (2007), has two primary responsibilities. The first is to use quantitative data to represent "processes and outcomes on a large scale to reveal inequalities and to identify social or institutional perpetuation of systematic inequalities in such processes and outcomes" (p. 96).

The second is to "question models, measures, and analytic practices of quantitative research in order to offer competing models, measures and analytic practices that better describe experiences of those who have not been adequately represented" (p. 96).

4. A "critical statistics" understands quantitative approaches as a sociopolitical practice that is historically and contextually situated. Professional statisticians and researchers are not objective and nor are their decisions as to what should be researched, what questions to ask, how the data are interpreted, and what should be presented. Counting is a fundamentally exclusionary human activity. To count is to make a choice about what is included and what is excluded: not only what to count and how to count but who to count. Statistical methods and analyses, while potentially informative, are tools to convey a sense of authority and persuasiveness. This is true of an academic publishing numbers to garner support for a statistical model or politicians using numbers to garner support for proposed legislation. Therefore, the production of knowledge through quantitative research cannot be separated from the distribution of power at multiple layers of social experience.
5. A "critical statistics" promotes statistical literacy and critical participation throughout the quantitative process. Quantitative methods are gatekeepers to participation, separating expert from layperson. The necessary critical conversations the numbers should facilitate are too often reserved for "professionals." And even among professional researchers, the complexity of mathematics can inhibit fruitful dialogue. A critical approach to quantitative methods should promote democratic participation among colleagues and among all citizens, particularly those who are most affected by unjust policies and oppressive systems, by demystifying the too technical complexity of quantification while at the same time insisting on troubling the too simplistic interpretations of human experience. A movement called "Barefoot Statisticians" embodies this ethic – inspired by China's "Barefoot Doctors" where locally trained medical intermediaries within poor and rural communities are trained to handle the community's basic health needs; barefoot statisticians are trained to serve the local community's basic quantitative needs for the purposes of critical democratic engagement and activism.

## Keywords

Critical statistics; radical statistics; quantitative criticalist

## History

Before the twentieth century, the earliest applied materialization of quantitative social research began in service of governments (i.e., England, Germany, France) and the political elite. "Political arithmetic," as it was called, collected social facts about the population for the purposes of governing, though occasionally in the name of democracy, largely in the name of state control (Desrosieres, 1999). Since then, quantitative social science research, especially psychology, boomed and became part of the public's consciousness. For example, the modern social scientific surveys of the twentieth century had profound influence on American society. The use of statistical procedures such as frequency counts, aggregated majorities, and averages to represent the survey findings informed and continues to inform what is considered normality, morality, identity, and democratic participation in our country (e.g., through the United States Census to measure the population, Gallup Polls to capture public opinion, the Kinsey Report to learn about private sexual activity, Myers-Briggs to reveal internal personality structures). Igo (2007) noted, "In the concrete techniques of the questionnaire and the interview, in public debates over survey findings, and in encounters between researchers and the researched, a new mode of knowing "ourselves" took shape in the twentieth–century United States" (p. 282).

The establishment of statistics and psychology as academic fields was closely related. One particularly important development in the history of quantitative psychological research was the move from individual to statistical aggregate. The origins of modern experimental psychology started in Wilhelm Wundt's laboratory studies of self-observation. Psychologists were looking to distinguish the field of psychology from philosophy and as a "real" experimental science like physics. Early attempts at psychological experimentation by Wundt and his early followers were heavily focused on repeated trials of a small sample of individuals rather than the aggregation of many individuals. The experimenter and the subject were often colleagues in the same lab and thus interchanged their roles. The individual was so clearly present in the data that even the names of the experimenter and the subjects were included in the published analysis. To many early psychologists, learning about the individual from the group seemed fundamentally in opposition to their goals. However, through the pioneering work of Francis Galton, quantitative aggregation would soon become the more dominant approach in psychology (Danzinger, 1990).

Galton and those influenced by his approach used large samples to examine aggregated psychological processes. This approach allowed claims to be made about the individual from numerical patterns found in the collective. Thus, learning about the individual involved comparing the extent to which personal responses deviated ("error") from statistical norms derived from large samples of responses. Through probability, general inferences were then made from the statistical aggregation to the theoretical population of interest. The influence of this new logic based on the statistical regularities of large numbers and probability could eventually be found across the field of psychology from the aggregated scores of experimental and control groups to the psychological attributes measured from multiple survey items then collapsed into a single score through factor analysis. The statistical approaches influenced by Galton were increasingly seen as useful to advancing practical questions and in retrospect were also, like political arithmetic of the nineteenth century, useful in social control and administration (e.g., IQ testing to manage the influx of immigrants entering public school or managing the distribution of soldiers into the military hierarchy of WWII). Ultimately this approach became and has remained the dominant research paradigm in psychology (Danzinger, 1990; Stigler, 1999).

Statistics has profoundly influenced the logic of psychology. Throughout the twentieth century, inferential statistical procedures became firmly understood, mainstream, increasingly sophisticated, associated with "real" science, and seen as objective. Statistical conventions such as statistical significance were adopted and routinized, binding psychologists together in a standardized quantitative language spoken across the field. Much of modern statistics taught to psychology researchers today has a lineage from such pioneers as Quetelet (the Newton of statistics) and emerged directly from the efforts *and assumptions* of Galton, Yule, Pearson, and Fisher, among others (e.g., ANOVA, correlation, regression, factor analysis). Their approaches were not derived from neutral space. They were developed with applied purposes and informed by their sociopolitical theories, especially eugenics, social evolution, and biology (Dorling & Simpson, 1999; Yu, 2006). Therefore, the critical question to ask is: what did psychologists lose by applying to psychological problems a set of quantitative methods and statistical procedures that were not, in their fundamental assumptions, developed to specifically address the theoretical concerns of psychology?

It is also important to remember the forgotten possibilities of past psychology, like Wundt's individualized experiments, *or* the radical strands of early social psychology, like the use of social statistics to pursue social justice. For example, consider the nineteenth-century amateur and localized quantitative efforts that began to flourish independent of the official statistical collection of the state. Now known as the Social Survey Movement, these multi-methods studies (e.g., community surveys, mapping, interviews) tended to be explicitly conducted for the purposes of social justice, reform, or human rights.

Usually with the help of local volunteers and grassroots organizations, extensive data on many layers of social and economic factors were collected within a relatively defined community. In the early twentieth century, a distinction was made between "Sociological Survey" and "Social Survey," the former for the purposes of advancing the objective pursuit of knowledge through sophisticated statistics and probability sampling, while the latter a nonscientific-biased pursuit of activism. The amateur-driven Social Survey Movement fell out of favor as the definition of social scientific expertise matured (Bulmer, Bales, & Sklar, 2001). Though the commitments of the Social Survey Movement still continues in some areas of psychology today through such organizations as the Society for the Psychological Study of Social Issues (SPSSI), these early amateur pursuits force us to ask important questions of the dominant assumptions underneath modern research: why is expertise (e.g., quantitative analysis) defined narrowly within the university-educated academic researcher, why are community members not included in the entire research process (including the statistical analysis), and why must "good" science exclude public and political engagement?

## Critical Debates

The appearance of precision and the illusion of neutrality can make the products of statistics seem somehow magical: facts above critique. The complexity of mathematics makes the discipline of statistics seem settled: ancient proofs solved long ago. It may surprise many that even the most common practices in quantitative methods and statistics were once fiercely debated and many are still sources of tension. There are several debates or, rather, tensions that exist in the critical statistics literature. Three are briefly described below.

### CFA Versus EDA

John Tukey was one of the most influential statisticians and mathematicians of the twentieth century. Tukey drew a distinction between what he called confirmatory data analysis (CDA) and exploratory data analysis (EDA). Traditional introductory statistics courses devote most of the time to CDA. CDA tends to be a deductive, hypothesis-driven approach with heavy guidance from predetermined theories. At its most conservative, all of the analyses are preplanned so as to not capitalize on chance or random fluctuations (i.e., type I error, familywise error). CDA models also tend to be interested in using a sample of data in order to estimate (infer or generalize to) the population of interest using statistical probability. For example, the sample mean is used as a proxy to estimate the true population mean. Tukey contributed heavily to advancing CDA and acknowledged its importance. However, he worried that the myopic pursuit of CDA facilitated a "specific mental rigidity" that can come from fitting complex data into a set of very restrictive assumptions.

Tukey's critique of CDA was especially focused on social science's use of statistics. More recently, other prominent statisticians have also cautioned the social sciences at their overreliance on complex modeling, controlling for covariates, and unwarranted causal claims. The general concern is that the complexity of social and psychological experience is often inadequately captured by quantitative approaches. The proposed remedies often involve what David Freedman called "shoe leather research": greater awareness of the boundaries of quantification, more varied use of research methodologies, uncovering patterns across multiple studies over time, and, as Tukey suggested, a greater willingness to explore one's data (Freedman, 2010; Lieberson, 1985). Tukey and his colleagues developed an approach to statistical analysis in contrast though complimentary to CDA: what they called exploratory data analysis (EDA). EDA, as most clearly articulated by Tukey and others in the mid-1970s to early 1980s, was an iterative, descriptive, graphical approach to statistics: one that was less concerned with using the data to statistically generalize to a population but instead took seriously the data for what those individual responses might reveal

collectively through exploratory probing. Tukey explained, "Exploratory data analysis is detective work—numerical detective work—or counting detective work—or graphical detective work" (1977, p. 1). He was "Looking at what data seems to say" (Tukey, p. v) rather than confirming or testing previously stated hypotheses from predetermined theories.

Many of Tukey's exploratory strategies stayed close to the original data, minimizing statistical abstraction by using techniques that were largely descriptive. He wrote, "We ... regard simple descriptions as good in themselves" (Tukey, 1977, p. vi). As part of these descriptions, he worried about using statistical techniques like the common average that were very sensitive to outlier (i.e., unusual) values and therefore susceptible to distorted interpretations. He in fact, worried about any single value, like the average, used to describe a set of data points without also exploring the entire variability of those data points (i.e., the distribution). In other words, instead of using probabilistic standards of $p = <.05$ to determine "if he had something significant," he sought strategies that would most clearly and with the least distortion allow the stories within the data to emerge. His approach was particularly well suited to discovering insightful questions one should ask of the data and the topic of interest. Indeed he frequently wrote, "Finding the question is often more important than finding the answer" (Tukey, 1980, p. 24).

Therefore, EDA was not only a set of techniques; *it was most importantly* a state of mind, an attitude, and a way of perceiving the data. Tukey (1980) argued, "No catalog of techniques can convey a willingness to look for what can be seen, whether or not anticipated. Yet this is the heart of exploratory data analysis" (Tukey, p. 24). This attitude was an open-ended approach that stressed iteration and flexibility – seeing what the data revealed rather than imposing rigid assumptions onto the numbers. At the time, EDA was a radical shift as both a technique and an epistemological stance.

Of particular importance to the field of psychology is the co-optation of some techniques in traditional introductory statistics texts but an absence or removal of the philosophy. Hoaglin (2003) explained of EDA, "within a few years, the basic techniques, particularly displays, were available in statistical software. By now a number of those techniques have become part of statistical instruction at all levels. So, at the level of tools, the impact of EDA has been broad and lasting. I am not sure about the attitudes, which require more effort to teach and more reflection..." (p. 313). It is this exploratory attitude towards statistics using a set of techniques designed to find the right questions that is a forgotten alternative deserving a second look by current quantitative and critically oriented psychologists.

## NHST

The renowned statistician Karl Fisher invented the null hypothesis and set the socially constructed p value (cautiously) at $p = <.05$, though by his own admission, there is nothing sacred about that number. The $p$ value is used to test the extent to which the results derived from the observed data are consistent with the null hypotheses. The null hypothesis, as it is most commonly practiced, assumes that *in the population* there is no difference or relationship between your variables of interest. When the relationship produced from the data is very unlikely (e.g., a small $p$ value), given the assumption of no relationship in the population is held as true, the null hypothesis is then rejected. However, for Fisher, it stopped there. He did not invent a mechanism to thereby accept the alternative hypothesis: in other words, the hypothesis that *in the population* there likely *is* a difference or relationship between your variables of interest. Though this is the current convention in psychology and why we say something is "statistically significant," Jerzy Neyman and Egon Pearson (the son of Karl Pearson) did not make an argument for the alternative hypothesis until years after Fisher developed the null hypothesis. Fisher was vehemently opposed to the use of the alternative hypothesis, and it ignited a bitter feud with Neyman and Pearson. Yet, these procedures, what in

combination are called Null Hypothesis Statistical Testing (NHST), have become cemented into the conventional practices of research psychologists. They are one of the most important indicators by which knowledge is defined in psychology and the social sciences in general. However, the procedures and assumptions attached to NHST most notably signified through the p value (i.e., $p = <.05$) continue to be highly critiqued (Morrison & Henkel, 1970).

Indeed, Cohen argued, "NHST has not only failed to support the advance of psychology as a science but also has seriously impeded it" (1994, p. 997). Rozeboom (1997) was less kind; he argued that "The Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students" (p. 335). The problems with NHST are multiple. It facilitates dichotomous, true-false decision-making around a socially constructed cutoff point ($p = <.05$) that should be more appropriately determined depending on the context. The $p$ value indicates the likelihood of collecting data that produced the relationship of interest, given the assumption that no relationship in the population actually exists. However, the p value is often misinterpreted to mean the opposite: the probability that the null hypothesis is *true* given the data collected and relationship found. The null hypothesis further assumes *no relationship at all*. Thus, even miniscule and irrelevant deviation from zero in the population will be statistically significant with a large enough sample size. Further, the null hypothesis does not indicate strength of relationship, though small p values are often misinterpreted as numerical estimations of how strong the relationship between the variables are (i.e., effect sizes). What can p values do? They can simply indicate how confident one is that there is enough power to detect whatever difference (whether small or large) inevitably exists in the population. Thus, given appropriate sampling, NHST *can* lend confidence to what direction the population relationship is in (Morrison & Henkel, 1970).

## Categorization and Measurement

The politics of categorization is a hotly debated topic. Categories such a race/ethnicity or gender are politically charged social constructions, despite appearing as naturally occurring groups. Though consistently being resisted and queered, the process of categorization aided substantially by quantification can make rigid and standardize individual and group identities in a way that looks objective. Making the decision to collapse a race/ethnicity variable into "white or nonwhite" or examining gender as a male versus female without including the category of transgendered has critical implications for reflecting accurate statistical representations of lived experience produced by researchers. Quantitative methods communicate a false sense of unbiased precision that can erase the politics and assumptions that are inevitably attached to any pursuit of knowledge. This is particularly true of the quantitative researcher's pursuit of measurement (Porter, 1995; Saetnam, Lomell, & Hammer, 2010).

Rating scales are one of the most common tools in the quantitative method toolbox. Rating scales are used to measure everything from depression to intelligence to personality traits. The historical development of scales coincides with the pressure and desire of the emerging psychological field to be seen as a "real" science, one that could objectively quantify its subjects on par with the natural sciences. Scales, such as asking "to what extend do you agree or disagree with the following" using five options (i.e., strongly agree, agree, neither agree nor disagree, disagree, strongly disagree), are so settled as a psychological instrument that their usefulness is seldom questioned. However, the assumptions that researchers must make of their respondents when using scales are quite lofty. Rosenbaum and Valsiner (2011) outlined this list: "During the rating process, the respondents are assumed to (a) have direct access to their personal and stable meanings of the given scale endpoints and (b) accept the assumption of the continuous nature of the linear space between the points. Perhaps most importantly, researchers then assume that (c) the different respondents' personal understanding of the questions to be

similar to those of all other respondents, making it possible to aggregate the ratings from a single participant to a sample of participants" (p. 51). This is a tall order.

Furthermore, multiple questions in the form of rating scales are often used to measure complex psychological constructs (e.g., intelligence) not able to be captured with a single item. It is hoped that the series of questions aggregated together measure all of the theorized qualities that make up the construct of interest. The series of measurements produce a score, and that score is a more or less flawed representation of something thought to be real and meaningful. However, it is seldom considered if this hypothesized psychological construct is actually a quantity or are we just forcing quantification onto it through artificial rating scales. In statistical speak, does it have interval or ratio properties such as temperature or height (i.e., a one-point difference between 32 and 33° or 60 and 61 in. are the same one point difference throughout the temperature and height scales) or does it have ordinal properties such as personal perception of social class (i.e., one's perceived difference between middle class and upper class has rank order in that one has more money than the other but lacks continuous precision in that we do not know how much the difference is). This is not an issue of how it is measured, but an ontological argument about what exists to be measured. Michell (1999) argues that when statistics are used to measure quantifiable things in the world, it provides a set of highly powerful and predictive procedures. His proof is that there are bridges that have not crumbled and rockets that landed on the moon. This is because things that are quantifiable have a structure that can be accurately described by statistics. However, he argues that most socially constructed psychological concepts are at best ordinal. If true, this has huge implications for the field of psychology since most commonly used statistical techniques require the dependent variable to have interval or ratio properties. Furthermore and even more pervasive, averages are not appropriate calculations with ordinal variables.

## International Relevance

If critical statistics has an active home in the field of psychology or the broader social sciences, it is not in the United States. If it lies anywhere, it lies in England with the Radical Statistics Group (i.e., Radstats). The Journal of Radical Statistics was developed in 1975 as part of the establishment of the British Society for Social Responsibility in Science. The society has since become defunct; however, Radstats continues to function as an organization, release its journal, hold yearly conferences, and publish books. They are a diverse group with varied political perspectives united by a commitment to building a freer, democratic society, and they believe that the critical use of statistics can contribute to this effort. An early policy statement explained that they sought to provide "free access to, and free discussion of, the information, political and commercial criteria, and procedures used in decision-making, by all those affected by the decision" (Thomas, 2001, pp. 66–67). Radstats concerns itself with the use and misuse of statistics in the service of hegemony, government power, and privileged groups. Their early policy states that "Although statistics sometimes helps to create the conditions of change, it is usually used to protect the status quo" (Thomas, pp. 66–67). As a result, they are interested in "the production and publication of statistics needed by the disadvantaged groups in society, e.g. on wealth, income, prices, housing, social services, education" (Thomas, pp. 66–67). This group continues to inspire new generations of critically minded quantitative researchers and activists.

## Practice Relevance

Faith in statistical evidence not only continues to grow in psychology but throughout our culture as well. Terms such as "business analytics," "big data," "infographics," "predictive modeling," "political forecasting," and "Wall Street Quants" are commonplace in our public consciousness. Public institutions like police

departments and private institutions like Google are increasingly numbers driven. And the 24 h news cycle is saturated with social media polling, economic markers, and government budgets. Griffiths, Irvine, and Miles (1989) argued that "Radical statisticians may succeed in quantifying the world in new ways, but what really counts is whether they succeed in helping to change it" (p. 367). If our dependence on statistics remains as pervasive into the future as it was throughout the twentieth century and into the first decade of the twenty-first century, then an organized countermovement of critical statisticians, critical researchers, and citizens with critical quantitative perspectives will be equally important to radically question what impact the use of numbers has on our lives.

## Future Directions

There are many future paths for critical statistics. One particularly fruitful and fast developing area is called "participatory statistics." It is often an illusion that statistics is an individual process. At each point in the quantitative process, the potential for social engagement is possible and common. And at each moment along the way, important conceptual and theoretical decisions are made that ultimately effect what knowledge is produced: in other words, what questions to ask in a survey, how to categorize or combine variable responses, who to include in the sample, what questions to ask, how to interpret the numbers, and which findings should be presented as "the story." Who is in the room when it comes time to make these numerous choices represent the dynamic processes of quantitative social research (though largely invisible to outside audiences and rarely written about). Given how important these decision points are, it is important to find ways to fill those moments with a diverse group of experts – particularly experts who are most closely connected to the research topic (e.g., community members). Participatory action researchers consider this an issue of validity, epistemology, and ethics. The quality of the research is thought to be hindered to the extent that those engaging in the research design and analyses are not able to fully and intelligently participate in critical thinking and discussion because of, for example, the technical aspects of developing a survey or running statistics.

"Stats-n-Action" bridges the epistemological and methodological commitments of participatory action research with the framework of exploratory data analysis. It is a growing set of collaborative techniques and strategies designed to take seriously the quantitative process as democratic group work. Stats-n-Action is an iterative, flexible, participatory, and critical approach applied to four explicitly quantitative moments throughout the PAR process: the development of quantitative instruments, the discussion of who and how to sample, the analysis and interpretation of data, and the communication of quantitative stories to the public. "Stats-n-Action" seeks to apply a principled approach to quantitative and mixed methods research with the larger goal of activism for social equality.

## References

Bulmer, M., Bales, K., & Sklar, K. K. (Eds.). (2001). *The social survey in historical perspective (1880–1940)*. Cambridge, UK: Cambridge University Press.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 49*(12), 997–1003.

Danzinger, K. (1990). *Constructing the subject: Historical origins of psychological research*. Cambridge, UK: Cambridge University Press.

Desrosieres, A. (1999). *The politics of large numbers: A history of statistical reasoning*. Paris, France: La Decouverte. (Original work published 1993).

Dorling, D., & Simpson, S. (1999). *Statistics in society: The arithmetic of politics*. London, England: Arnold Press.

Freedman, D. A. (2010). *Statistical models and causal inference: A dialogue with the social sciences*. Cambridge, UK: Cambridge University Press.

Griffiths, D., Irvine, J., & Miles, I. (1989). Social statistics: Towards a radical science. In J. Irvine, I, Miles, & J. Evans (Eds.), *Demystifying social statistics*. London, England: Pluto Press.

Hoaglin, D. C. (2003). John W. Tukey and data analysis. *Statistical Science, 18*(3), 311–318.