Chapter 1: What is Statistics?

Christopher J. Wild, Jessica M. Utts, and Nicholas J. Horton

Christopher J. Wild
University of Auckland, New Zealand

President, American Statistical Association

Jessica M. Utts
University of California, Irvine, United States

Chair, Statistical Education Section, ASA

Nicholas J. Horton
Amherst College, United States

e-mail: c.wild@auckland.ac.nz, jutts@uci.edu, nhorton@amherst.edu

Abstract

What is Statistics? We attempt to answer this question as it relates to grounding research in statistics education. We discuss the nature of statistics (the science of learning from data), its history and traditions, what characterises statistical thinking and how it differs from mathematics, connections with computing and data science, why learning statistics is essential, and what is most important. Finally, we attempt to gaze into the future, drawing upon what is known about the fast-growing demand for statistical skills and the portents of where the discipline is heading, especially those arising from data science and the promises and problems of big data.

Key Words

Discipline of statistics; Statistical thinking; Value of statistics; Statistical fundamentals; Decision making; Trends in statistical practice; Data science; Computational thinking

1.1 Introduction

In this, the opening chapter of the *International Handbook of Research in Statistics Education*, we ask the question, "What is statistics?" This question is not considered in isolation, however, but in the context of grounding research in statistics education. Educational endeavour in statistics can be divided very broadly into "What?" and "How?" The "What?" deals with the nature and extent of the discipline to be conveyed, whereas any consideration of "How?" (including "When?") brings into play a host of additional factors, such as cognition and learning theories, audience and readiness, attitudes, social interactions with teachers and other students, and learning and assessment strategies and systems.

This chapter discusses the nature of statistics as it relates to the teaching of statistics. The chapter has three main sections. Section 2 discusses the nature of statistics, including a brief historical overview and discussion of statistical thinking and differences with mathematics. Section 3 follows this up with a discussion of why learning statistics is important. Section 4 concludes the chapter with a discussion of the growing demand for statistical skills and a look at where the discipline is heading. A thread pervading the chapter is changing conceptions of the nature of

statistics over time with an increasing emphasis recently on broad, as opposed to narrow, conceptions of what statistics is. We emphasize broader conceptions because we believe they best address emerging areas of need and because we do not want researchers to feel constrained when it comes to deciding what constitutes fair game as targets for research.

1.2 The Nature of Statistics

"Statistics" **–** as defined by the American Statistical Association (ASA) **–** "***is the science of learning from data, and of measuring, controlling and communicating uncertainty."*** Although not every statistician would agree with this description, it is an inclusive starting point with a solid pedigree. It encompasses and concisely encapsulates the "wider view" of Marquardt (1987) and Wild (1994), the "greater statistics" of Chambers (1993), the "wider field" of Bartholomew (1995), the broader vision advocated by Brown and Kass (2009), and the sets of definitions given in opening pages of Hahn and Doganaksoy (2012) and Fienberg (2014). It also encompasses the narrower views.

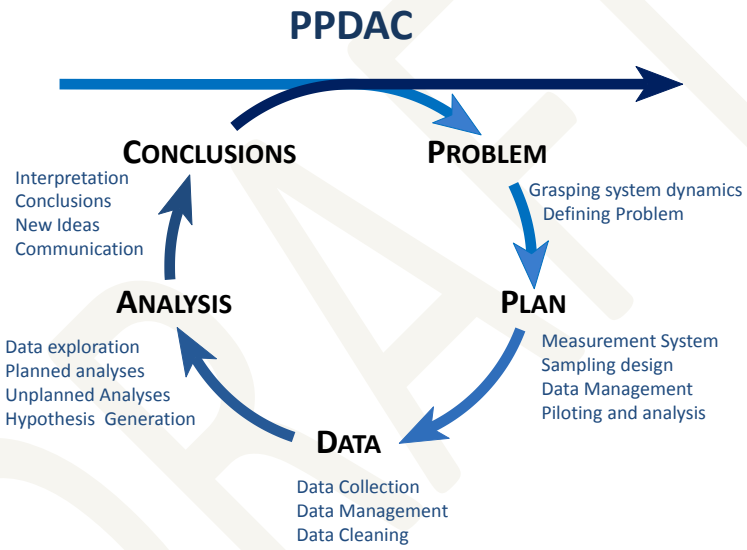

Fig. 1.1: The statistical inquiry cycle

Figure 1.1 gives a model of the statistical inquiry cycle from Wild and Pfannkuch (1999). This partial, rudimentary "map" hints at the diversity of domains that contribute to "learning from data." The ASA description of statistics given above covers all elements seen in this diagram and more. Although statisticians have wrestled with every aspect of this cycle, particular attention has been given by statistical theory-and-methods thinkers and researchers to different elements at different times. For at least the last half century, the main focus has been on the use of probabilistic models in the Analysis and Conclusions stages and to a lesser extent, on sampling designs and experimental designs in the Plan stage. But a wider view is needed to chart the way of statistics education into the future.

The disciplines of statistics and, more specifically, statistics education are, by their very nature, in the "future" business. The mission of statistical education is to provide conceptual frameworks (structured ways of thinking) and practical skills to better equip our students for their future lives in a fast-changing world. Because the data-universe is expanding and changing so fast, educators need to focus more on looking forward than looking back. We must also look back, of course, but predominantly so that we can plunder our history's storehouses of wisdom to better chart pathways

into the future. For educational purposes, statistics needs to be defined by the ends it pursues rather than the means statisticians have most often used to pursue them in the past. Changing capabilities, like those provided by advancing technology, can change the preferred means for pursuing goals over time but the fundamental goals themselves will remain the same. The big-picture definition that we opened with "keeps our eyes on the ball" by placing at the centre of our universe the fundamental human need to be able to learn about how our world operates using data, all the while acknowledging sources and levels of uncertainty.

"Statisticians develop new methodologies in the context of a specific substantive problem," Fienberg (2014) says, "but they also step back and integrate what they have learned into a more general framework using statistical principles and thinking. Then, they can carry their ideas into new areas and apply variations in innovative ways." At their core, most disciplines think and learn about some particular aspects of life and the world, be it the physical nature of the universe, living organisms, or how economies or societies function. Statistics is a meta-discipline in that it *thinks about how to think* about turning data into real-world insights. Statistics as a meta-discipline advances when the methodological lessons and principles from a particular piece of work are abstracted and incorporated into a theoretical scaffold that enables them to be used on many other problems in many other places.

1.2.1 History of Statistics

Although the collection of forms of census data goes back into antiquity, rulers "were interested in keeping track of their people, money and key events (such as wars and the flooding of the Nile) but little else in the way of quantitative assessment of the world at large" (Scheaffer, 2001, para. 3). The statistical analysis of data is usually traced back to the work of John Graunt (e.g., his 1662 book *Natural and Political Observations*). For example, Graunt concluded that the plague was caused by person-to-person infection rather than the competing theory of "infectious air" based on the pattern of infections through time. Graunt and other "political arithmeticians" from across Western Europe were influenced during the Renaissance by the rise of science based on observation of the natural world. And they "thought as we think today … they reasoned about their data" (Kendall, 1960, p. 448). They estimated, predicted, and learned from the data – they did not just describe or collect facts – and they promoted the notion that state policy should be informed by the use of data rather than by the authority of church and nobility (Porter, 1986). But the political arithmetician's uses of statistics lacked formal methodological techniques for gathering and analysing data. Methods for sample surveys and census taking were in their infancy well into the nineteenth century (Fienberg, 2014).

Another thread in the development of modern statistics was the foundations of probability, with its origins in games of chance, as laid down by Pascal (1623–1662) and later Bernoulli (1654–1705). The big conceptual steps towards the application of probability to quantitative inference were taken by Bayes in 1764 and Laplace (1749-1827) by inverting probability analyses.

> The science that held sway above all others around 1800 was astronomy, and the great mathematicians of the day made their scientific contributions in that area. Legendre (least squares), Gauss (normal theory of errors), and Laplace (least squares and the central limit theorem) all were motivated by problems in astronomy. (Scheaffer, 2001, para. 6)

These ideas were later applied to social data by Quetelet (1796–1874), who was trying to infer general laws governing human action. This was after the French Revolution when there was a subtle shift in thinking of statistics as a science of the state with the statists, as they were known,

conducting surveys of trade, industrial progress, labour, poverty, education, sanitation, and crime (Porter, 1986).

A third thread in the development of statistics involves statistical graphics. The first major figure is William Playfair (1759–1823), credited with inventing the line and bar charts for economic data and the pie chart. Friendly (2008) characterises the period from 1850 to 1900 as the "golden age of statistical graphics" (p. 2). This is the era of John Snow's dot map of cholera data and the Broad Street pump, of Minard's famous graph showing losses of soldiers in Napoleon's march on Moscow and subsequent retreat, of Florence Nightingale's coxcomb plot used to persuade of the need for better military field hospitals, of the advent of most of the graphic forms we still use for conveying geographically-linked information on maps, including such things as flow diagrams of traffic patterns, of grids of related graphs, of contour plots of 3-dimensional tables, population pyramids, scatterplots and many more.

The Royal Statistical Society began in 1834 as the London Statistical Society (LSS), and the American Statistical Association was formed in 1839 by five men interested in improving the U.S. census (Horton, 2015; Utts, 2015). Influential founders of the LSS (Pullinger, 2014, 825–827) included Adolphe Quetelet, Charles Babbage (inventor of the computer), and Thomas Malthus (famous for his theories about population growth). The first female LSS member was Florence Nightingale, who joined in 1858 (she also became a member of the ASA, as did Alexander Graham Bell, Herman Hollerith, Andrew Carnegie, and Martin Van Buren). These early members of LSS and ASA were remarkable for representing such a very wide variety of real-world areas of activity (scientific, economic, political, and social), and their influence in society.

> Near the end of the nineteenth century, the roots of a theory of statistics emerge from the work of Francis Galton and Francis Ysidro Edgeworth and from that of Karl Pearson and George Udny Yule somewhat later. These scientists came to statistics from biology, economics, and social science more broadly, and they developed more formal statistical methods that could be used not just within their fields of interest but across the spectrum of the sciences. (Fienberg, 2014)

Another wave of activity into the 1920s was initiated by the concerns of William Gosset, reaching its culmination in the insights of Ronald Fisher with the development of experimental design, analysis of variance, maximum likelihood estimation, and refinement of significance testing. This was followed by the collaboration of Egon Pearson and Jerzy Neyman in the 1930s, giving rise to hypothesis testing and confidence intervals. At about the same time came Bruno de Finetti's seminal work on subjective Bayesian inference and Harold Jeffreys's work on "objective" Bayesian inference so that by 1940 we had most of the basics of the theories of the "modern statistics" of the twentieth century. World War II was also a time of great progress as a result of drafting many young, mathematically gifted people into positions where they had to find timely answers to problems related to the war effort. Many of them stayed in the field of statistics swelling the profession. We also draw particular attention to John Tukey's introduction of "exploratory data analysis" in the 1970s.

Short histories of statistics include Fienberg (2014, Section 3); Scheaffer (2001), who emphasized how mathematicians were funded or employed and the influence this had on what they thought about and developed; and Pfannkuch and Wild (2004) who described the development of statistical thinking. Lengthier accounts are given by Fienberg (1992), and the books by Porter (1986), Stigler (1986, 2016), and Hacking (1990). Key references about the history of statistics education include Vere-Jones (1995), Scheaffer (2001), Holmes (2003), and Forbes (2014); see also Chapter 2.

1.2.2 Statistical Thinking

Statisticians need to be able to think in several ways: statistically, mathematically, and computationally. The thinking modes used in data analysis differ from those used in working with mathematical derivations, which in turn differ from those used for writing computational code. Although there are very strong internal connections within each of these thinking modes, there are relatively weak connections among them. Here we will concentrate on "statistical thinking" in the sense of the most distinctively statistical parts of the thinking that goes on in solving real-world problems using data.

In statistics, however, we sometimes talk about "solving real world (or practical) problems" far too loosely. For the general public, "solving a real-world problem" involves taking action so that the problem either goes away or is at least reduced (e.g., unemployment levels are reduced). We need to better distinguish between satisfying "a need to act" and "a need to know." Figuring out how to act to solve a problem will typically require acquiring more knowledge. This is where statistical inquiry can be useful. It addresses "a need to know." So when statisticians talk about solving a real-world problem, we are generally talking about solving a (real-world) *knowledge-deficit* or *understanding-deficit* problem.

1.2.2.1 Dimension 1 - PPDAC Wild and Pfannkuch (1999) investigated the nature of statistical thinking in this sense using available literature, interviews with practicing statisticians, and interviews with students performing statistical-enquiry activities; and presented models for different "dimensions" of statistical thinking. Dimension 1 of their work was the PPDAC model (Fig. 1.1) of the inquiry cycle. The basic PPDAC model was due to and later published by MacKay and Oldford (2000). There are also other essentially equivalent descriptions of the statistical inquiry cycle. The inquiry cycle has connections with standard descriptions of the "scientific method" (cf. https://en.wikipedia.org/wiki/Scientific_method) but is more flexible, omitting the latter's strong emphasis on being hypothesis driven and having (scientific) theory-formulation as its ultimate objective.

The PPDAC inquiry cycle reminds us of the major steps involved in carrying out a statistical inquiry. It is the setting in which statistical thinking takes place. The initial "P" in PPDAC spotlights the problem (or question) crystallisation phase. In the early stages, the problem is typically poorly defined. People start with very vague ideas about what their problems are, what they need to understand, and why. The *Problem* step is about trying to turn these vague feelings into much more precise informational goals, some very specific questions that should be able to be answered using data. Arriving at useful questions that can realistically be answered using statistical data always involves a lot of hard thinking and often a lot of hard preparatory work. Statistics education research says little about this but the PhD thesis of Arnold (2013) makes a very good start.

The *Plan* step is then about deciding what people/objects/entities to obtain data on, what things we should "measure," and how we are going to do all of this. The *Data* step is about obtaining the data, storing it, and "whipping it into shape" (data wrangling and cleaning). The *Analysis* step which follows, and the *Conclusions* step, are about making sense of it all and then abstracting and communicating what has been learned. There is always a back-and-forth involving doing analysis, tentatively forming conclusions, and doing more analysis. In fact there is back-and-forth between

the major steps whenever something new gets learned in a subsequent step that leads to modifying an earlier decision.

Any substantial learning from data involves extrapolating from what you can see in the data you have to how it might relate to some wider universe. PPDAC focuses on data gathered for a purpose using planned processes, processes that are chosen on statistical grounds to justify certain types of extrapolation. Much of the current "big-data" buzz relates to exploiting opportunistic (happenstance or "found") data – data that just happen to be available in electronic form because they have accumulated for other reasons, such as the result of the administrative processes of business or government, audit trails of internet activity, or billing data from medical procedures.

> In a very real sense, we have walked into the theatre half way through the movie and have then to pick up the story. … For opportunistic data there is no extrapolation that is justified by a data-collection process specifically designed to facilitate that extrapolation. The best we can do is to try to forensically reconstruct what this data is and how it came to be (its 'provenance'). What entities were 'measures' taken on? What measures have been employed and how? By what processes did some things get to be recorded and others not? What distortions might this cause? It is all about trying to gauge the extent to which we can generalize from patterns in the data to the way we think it will be in populations or processes that we care about. (Wild, 2017)

In particular, we are on the lookout for biases that could lead us to false conclusions.

1.2.2.2 Other Dimensions Dimension 2 of Wild and Pfannkuch's model lists Types of Thinking, broken out into General Types and Types Fundamental to Statistics. The General Types are *strategic*, *seeking explanations*, constructing and *using models*, and *applying techniques* (solving problems by mapping them on to problem archetypes). The Types Fundamental to Statistics listed are *recognition of the need for data*, *transnumeration* (changing data representations in search of those that trigger understanding), *consideration of variation* and its sources, *reasoning using statistical models*, and *integrating the statistical and the contextual* (information, knowledge, conceptions). Something that is not highlighted here is the inductive nature of statistical inference – extrapolation from data on a part to reach conclusions about a whole (wider reality).

Dimension 3 is the *interrogative cycle*, a continually-operating high-frequency cycle of *Generating* (possible informational requirements, explanations, or plans of attack), *Seeking* (information and ideas), *Interpreting* these, *Criticizing* them against reference points and *Judging* whether to accept, reject, or tentatively entertain them. Grolemund and Wickham (2014) dig much deeper into this dimension bringing in important ideas from the cognitive literature such as *schemas* (mental models in which knowledge is stored). In very oversimplified terms, when new information "arrives" a relevant schema is searched for internally to which it is compared. If discrepancies are detected ("insights") the schema is updated or the information is dismissed as non-credible. Grolemund and Wickham explore this in the context of types of information we seek and obtain in the process of analysing data.

Dimension 4 consists of a list of personal qualities, or *dispositions*, successful practitioners bring to their problem solving: *scepticism*; *imagination, curiosity and awareness*, a *propensity to seek deeper meaning, being logical*, *engagement*, and *perseverance*. This is amplified in Hahn and Doganaksoy's chapter "Characteristics of Successful Statisticians" (2012, Chapter 6).

1.2.2.3 Statistical Thinking for Beginners Although it only scratches the surface, the above still underscores the richness and complexity of thinking involved in real-world statistical problem solving and provides a useful set of reference points against which researchers and teachers can triangulate educational experiences ("Where is … being addressed?"). It is, however, far too complex for most students, particularly beginners. In discussing Wild and Pfannkuch, Moore (1999) asked, "What Shall We Teach Beginners?" He suggested "… we can start by mapping more detailed structures for the 'Data, Analysis, Conclusions' portion of the investigative cycle, that is, for conceptual content currently central to elementary instruction. Here is an example of such a structure:

> When you first examine a set of data, (1) begin by graphing the data and interpreting what you see; (2) look for overall patterns and for striking deviations from those patterns, and seek explanations in the problem context; (3) based on examination of the data, choose appropriate numerical descriptions of specific aspects; (4) if the overall pattern is sufficiently regular, seek a compact mathematical model for that pattern. (p. 251)

Moore (1998) offered the following for basic critique, which complements his 1999 list of strategies with "Data beat anecdotes" and the largely metacognitive questions, "Is this the right question? Does the answer make sense? Can you read a graph? Do you have filters for quantitative nonsense?" (p. 1258).

There are great advantages in short, snappy lists as starting points. Chance's (2002) seven habits (p. 4) bring in much of Moore's lists, and the section headings are even "snappier": "Start from the beginning. Understand the statistical process as a whole. Always be sceptical. Think about the variables involved. Always relate the data to the context. Understand (and believe) the relevance of statistics. Think beyond the textbook." Grolemund and Wickham (2014, Section 5) give similar lists for more advanced students. Brown and Kass (2009) state, "when faced with a problem statement and a set of data, naïve students immediately tried to find a suitable statistical technique (e.g., chi-squared, *t*-test), whereas the experts began by identifying the scientific question" (p. 123). They highlighted three "principles of statistical thinking":

1. Statistical models of regularity and variability in data may be used to express knowledge and uncertainty about a signal in the presence of noise, via inductive reasoning. (p. 109)
2. Statistical methods may be analyzed to determine how well they are likely to perform. (p. 109)
3. Computational considerations help determine the way statistical problems are formalized. (p. 122)

We conclude with the very specialised definition of Snee (1990), which is widely used in quality improvement for business and organisations,

> I define statistical thinking as thought processes, which recognize that variation is all around us and present in everything we do, all work is a series of interconnected processes, and identifying, characterizing, quantifying, controlling, and reducing variation provide opportunities for improvement. (p. 118)

1.2.3 Relationship with Mathematics

Although definitions that characterise statistics as a branch of mathematics still linger in some dictionaries, the separate and distinct nature of statistics as a discipline is now established.

"Statistical thinking," as Moore (1998) said, "is a general, fundamental, and independent mode of reasoning about data, variation, and chance" (p. 1257). "Statistics at its best provides methodology for dealing empirically with complicated and uncertain information, in a way that is both useful and scientifically valid" (Chambers, 1993).

"Statistics is a methodological discipline. It exists not for itself but rather to offer to other fields of study a coherent set of ideas and tools for dealing with data" (Cobb and Moore, 1997, p. 801). To accomplish those ends it presses into service any tools that are of help. Mathematics contains many very useful tools (as does computing). Just as physics attempts to understand the physical universe and presses mathematics into service wherever it can help, so too statistics attempts to turn data into real-world insights and presses mathematics into service wherever it can help. And whereas in mathematics, mathematical structures can exist and be of enormous interest for their own sake, in statistics, mathematical structures are merely a means to an end (see also Box, 1990, paragraph 2; De Veaux and Velleman, 2008). A consequence is that whereas a mathematician prefers an exact answer to an approximate question, an applied statistician prefers an approximate answer to an exact question.

1.2.3.1 Role of Context The focus of the discipline, and in particular the role of *context*, is also distinct. "Statistics is not just about the methodology in a particular application domain; it also focuses on how to go from the particular to the general and back to the particular again" (Fienberg, 2014).

> Although mathematicians often rely on applied context both for motivation and as a source of problems for research, the ultimate focus in mathematical thinking is on abstract patterns: the context is part of the irrelevant detail that must be boiled off over the flame of abstraction in order to reveal the previously hidden crystal of pure structure. *In mathematics, context obscures structure*. Like mathematicians, data analysts also look for patterns, but ultimately, in data analysis, whether the patterns have meaning, and whether they have any value, depends on how the threads of those patterns interweave with the complementary threads of the story line. *In data analysis, context provides meaning*. (Cobb & Moore, 1997, p. 803; our emphasis)

There is a constant "interplay between pattern and context" (Cobb & Moore, 1997). As for statistical investigations for real-world problems, the ultimate learning is new knowledge about the context domain – we have gone "from the particular to the general" (to enable us to use methods stored in our statistical repository) "and back to the particular again" (to extract the real-world learnings).

1.2.3.2 Role of Theory When most statisticians speak of "statistical theory" they are thinking of mathematical theories comprising "statistical models" and principled ways of reasoning with and drawing conclusions using such models. "Statistical models," which play a core role in most analyses, are mathematical models that include chance or random elements incorporated in probability-theory terms. Perhaps the simplest example of such a model is $y = \mu + \varepsilon$ where we think in terms of y being an attempt to measure a quantity of interest μ in the process of which we incur a random error ε (which might be modelled as having a normal distribution, say). In the simple linear model: $y = \beta_0 + \beta_1 x + \varepsilon$, the mean value of y depends linearly on the value of an explanatory variable x rather than being a constant. Random terms (cf. ε) model the unpredictable part of a process and give us ways of incorporating and working with uncertainties. "Statistical theory" in this sense is largely synonymous with "mathematical statistics."

"Probability theory" is a body of mathematical theory, originally motivated by games of chance but latterly much more by the needs of statistical modelling, which takes abstracted ideas about randomness, forms mathematical structures that encode these ideas and makes deductions about the behaviour of these structures. Statistical modellers use these structures as some of the building blocks that they can use in constructing their models, as with the random-error term in the very simple model above. Recent work by Pfannkuch et al. (2016) draws on interviews with stochastic-modelling practitioners to explore probability modelling from a statistical education perspective. The paper offers a new set of (conceptual) models of this activity. Their SWAMTU model is basically a cycle (with some feedback). It has nodes *problem Situation* $\rightarrow$ *Want (to know)* $\rightarrow$ *Assumptions* $\rightarrow$ *Model* $\rightarrow$ *Test* $\rightarrow$ *Use*. Models are always derived from a set of mathematical assumptions so that assumption checking against data is, or should be, a core part of their construction and use. As well as being used in statistical analysis, they are commonly used to try to answer "what-if" questions (e.g., "What would happen if the supermarket added another checkout operator?"). Although there is much that is distinct about statistical problem solving, there is also much that is in common with mathematical problem solving so that statistics education researchers can learn a lot from work in mathematics education research, and classic works such as Schoenfeld (1985).

When most statisticians think "theory," they are thinking of mathematical theories that underpin many important practices in the Analysis and Plan stages of PPDAC. But "theory" is also applicable *whenever* we form abstracted or generalized explanations of how things work. Consequently, there is also theory about other elements of PPDAC, often described using tools like process diagrams (cf. Fig. 1.1). Grolemund and Wickham (2014) propose a theoretical model for the data analysis process by comparing it to the cognitive process of the human mind called "sensemaking."

In recent years there has also been a shift in the "balance of power" from overtly mathematical approaches to data analysis towards computationally-intensive approaches (e.g., using computer-simulation-based approaches including bootstrapping and randomisation tests, flexible trend smoothers, and classification algorithms). Here the underlying models make much weaker assumptions and cannot be described in terms of simple equations. So, although "the practice of statistics requires mathematics for the development of its underlying theory, statistics is distinct from mathematics and requires many nonmathematical skills" (American Statistical Association Undergraduate Guidelines Workgroup, 2014, p. 8). These skills (required also by many other disciplines) include basic scientific thinking, computational/algorithmic thinking, graphical/visualisation thinking, and communication skills.

So, "… how is it then that statistics came to be seen as a branch of mathematics? It makes no more sense to us than considering chemical engineering as a branch of mathematics" (Madigan and Gelman's discussion of Brown and Kass, 2009, p. 114). The large majority of senior statisticians of the last half century began their academic careers as mathematics majors. Originally, computational capabilities were extremely limited and mathematical solutions to simplified problems and mathematical approximations were hugely important (Cobb, 2015). The academic statisticians also worked in environments where the reward systems overwhelmingly favoured mathematical developments. The wake-up call from "big data" and "data science" is helping nudge statistics back toward its earlier, and much more holistic, roots in broad scientific inference. Though the elements of all of this, and their champions, have always been there, in universities the broader practical dimensions became overshadowed and underappreciated.

1.3 Why Learning Statistics Is More Important Than Ever

In today's data-rich world, all educated people need to understand statistical ideas and conclusions, to enrich both their professional and personal lives. The widespread availability of interesting and complex data sets, and increasingly easy access to user-friendly visualization and analysis software mean that anyone can play with data to ask and answer interesting questions. For example, Wild's Visual Inference Tools (https://www.stat.auckland.ac.nz/~wild/VIT/) and iNZight software (https://www.stat.auckland.ac.nz/~wild/iNZight) allow anyone to explore data sets of their own choosing. The CODAP (Common Online Data Analysis Platform, https://concord.org/projects/codap) provides a straightforward platform for web-based data analysis, as does iNZight Lite (http://lite.docker.stat.auckland.ac.nz/).

Statistical methods are used in almost all knowledge areas and increasingly are used by businesses, governments, health practitioners, other professionals, and individuals to make better decisions. Conclusions and advice based on statistical methods abound in the media. Some of the thinking used for decision-making based on quantitative data carries over into decision-making involving uncertainty in daily life even when quantitative data are not available. For these reasons, probably no academic subject is more useful to both working professionals and informed citizens on a daily basis than statistics.

The rapid development of data science and expansion of choices for what to teach in statistics courses provides challenges for statistics educators in determining learning goals, and opportunities for statistics education researchers to explore what instructional methods can best achieve those goals. For example, in articles that appeared almost simultaneously, both Cobb (2015) and Ridgway (2015) argue that we need a major overhaul of the statistics curriculum, particularly the introductory course and the undergraduate curriculum. To ignore the impact of the widespread availability of data and user-friendly software to play with data would lead to marginalization of statistics within the expanding world of data science.

In this section we provide some motivation for why everyone should study statistics, and some examples of what could be useful for various constituencies. Statistics educators should keep these ideas in mind, and make sure to emphasize the usefulness of statistics when they teach, especially for students who are studying statistics for the first (and possibly only) time. Statistics education researchers should study how we can train students to use statistical reasoning throughout their lives to ask and answer questions relevant to them.

The ideal type and amount of statistical knowledge needed by an individual depends on whether the person will eventually be working with data as a professional researcher (a *producer* of statistical studies), interpreting statistical results for others (a *professional user* of statistics), or simply needing to understand how to use data and interpret statistical information in life (an *educated consumer* of data and statistics).

Professional users include health workers (who need to understand results of medical studies and translate them into information for patients), financial advisors (who need to understand trends and variability in economic data), and politicians (who need to understand scientific data as it relates to public policy, as well as how to conduct and understand surveys and polls). Producers of statistical

studies are likely to take several courses in statistical methods, and will not be the focus of this chapter (see Section 4 for more discussion).

Educated consumers include pretty much everyone else in a modern society. They need to understand how and what valid conclusions can be made from statistical studies and how statistical thinking can be used as a tool for answering questions and making decisions, with or without quantitative data.

1.3.1 What Professional Users of Statistics Need to Know

Many professionals do not need to know how to carry out their own research studies, but they do need to know how to interpret results of statistical studies and explain them to patients and customers. In business applications, professionals such as marketing managers may need to understand the results generated by statisticians within their own companies. In this section we provide a few examples of why professional users of statistics need to understand basic statistical ideas beyond what is needed for the general consumer.

Commonly used statistical methods differ somewhat across disciplines, but there are some basic ideas that apply to almost all of them. One of the most fundamental concepts is the importance of variability and a distribution of values. The first example illustrates how that concept is important for financial advisors and their clients.

> *Example 1: How Much Should You Save? Income and Expense Volatility*
> In 2015 the financial giant JP Morgan Chase announced the establishment of a research institute to utilize the massive and proprietary set of financial data it owns to answer questions about consumer finances. The inaugural report, published in May 2015 (Farrell and Greig, 2015), examined financial data for 100,000 individuals randomly selected from a 2.5 million person subset of Chase customers who met specific criteria for bank and credit card use. One of the most important and publicized findings (e.g., Applebaum, 2015) was that household income and expenditures both vary widely from month to month, and not necessarily in the same direction. For instance, the report stated that "41% of individuals experienced fluctuations in income of more than 30% on a month-to-month basis" (p. 8) and "a full 60% of people experienced average monthly changes in consumption of greater than 30%" (p. 9). Additionally, the report found that changes in income and consumption don't occur in tandem, so it isn't that consumers are spending more in months when they earn more. Why is this important information? Financial planners routinely advise clients to have liquid savings equivalent to three to six months of income. But in some cases, that may not be enough because of the volatility in both income and expenditures, and the possibility that they can occur in opposite directions. One of three main findings of the report was "The typical individual did not have a sufficient financial buffer to weather the degree of income and consumption volatility that we observed in our data." (p. 15)

A concept related to variability is that individuals are not all "typical." In the previous example, individual consumers should know whether the warnings in the report are likely to apply to them based on how secure their income is, what major expenditures they are likely to encounter, and what resources they have to weather financial storms. Knowing that the "typical" or "average" consumer needs to stash away three to six months of income as savings should lead to questions

about how each individual's circumstances might lead to recommendations that differ from that advice.

Baldi and Utts (2015) discuss topics and examples that are important for future health practitioners to learn in their introductory, and possibly only, statistics course. One central concept is that of natural variability and the role it plays in defining disease and "abnormal" health measurements. The next example, adapted from Baldi and Utts, illustrates how knowledge about variability and a distribution of values can help physicians and their patients put medical test results in perspective.

Example 2: High Cholesterol for (Almost) All

Medical guidelines routinely change based on research results, and statistical studies often lead pharmaceutical companies and regulatory agencies to change their advice to medical practitioners about what constitutes pathology. According to the United States National Institutes of Health, high cholesterol is defined as having total blood cholesterol of 240 mg/dl or above, and elevated or borderline high cholesterol is defined as between 200 and 240 mg/dl. (http://www.nhlbi.nih.gov/health/health-topics/topics/hbc). Suppose you are diagnosed with high or borderline high cholesterol and your doctor recommends that you take statin drugs to lower it. You might be interested in knowing what percentage of the population is in the same situation. Using data from the World Health Organization (Lawes et al, 2004), we can model the total cholesterol levels of women aged 45 to 59 years old in the United States using a normal distribution with mean of about 216 mg/dl and standard deviation of about 42 mg/dl. Figure 1.2 illustrates this distribution. As shown in the figure, about 35% of women in this age group have high cholesterol, and an additional 36.5% have borderline high values. That means that only about 28.5% do not have this problem! Should more than 70% of middle-aged women be taking statin drugs? Given that there are risk associated with high cholesterol and side effects associated with taking statin drugs, this is a discussion for individuals to have with their doctors. But it would be helpful for both of them to understand that more than a majority of the population fall into these cholesterol risk categories. Additional statistical reasoning tools (beyond the scope of this example) are needed to help physicians and consumers understand the tradeoff in risks associated with taking statin drugs or not. But this example illustrates the importance of understanding the concept of a distribution of values, and how it relates to decisions individuals need to make.

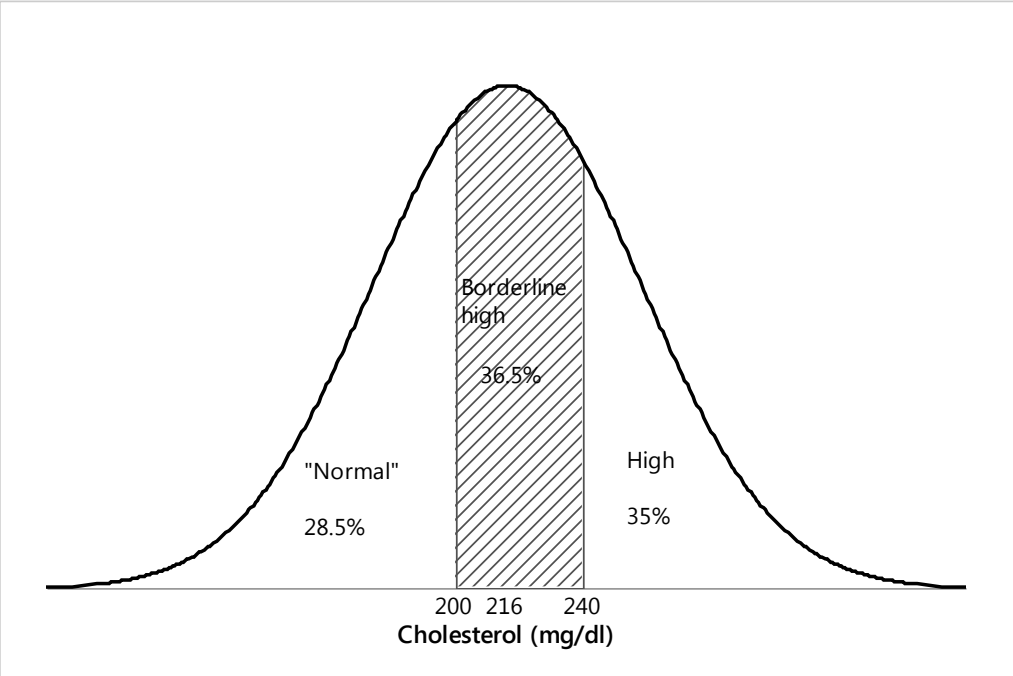

Fig. 1.2: Cholesterol values for women aged 45 to 59; mean ≈ 216 mg/dl and s.d. ≈ 42 mg/dl

1.3.2 What Educated Consumers of Data and Statistics Need to Know

The majority of students who take a statistics course will never use statistics in their professional lives, but quite often they will encounter situations for which they could utilize data and statistical information to make informed decisions. Teachers of introductory statistics should provide instruction that will help people utilize that information. In the papers by Ridgway (2015), Baldi and Utts (2015), Utts (2003), and Utts (2010), more than a dozen important topics are described, with multiple examples. A few of them were covered in the previous section on what professional users of statistics need to know. Here, we list more of them and explain why they are important for consumers. Examples to illustrate each of these can be found in the references mentioned here, as well as in textbooks with a focus on statistical literacy such as Cohn and Cope (2011), Hand (2014), Moore and Notz (2012), and Utts (2014). A resource for current examples is the website http://www.stats.org, a joint venture of the American Statistical Association and Sense About Science USA.

1. Unwarranted conclusions based on observational studies

The media have improved in the interpretation of observational studies in recent years, but reports implying causation based on observational studies are still quite common. Students should learn to recognize observational studies and know that cause and effect conclusions cannot be made based on them. Here are some recent examples of misleading headlines based on observational studies or meta-analyses of them:

- "6 cups a day? Coffee lovers less likely to die, study finds" (NBC News, May 16, 2012; http://vitals.nbcnews.com/_news/2012/05/16/11704493-6-cups-a-day-coffee-lovers-less-likely-to-die-study-finds?lite)
- "Citrus fruits lower women's stroke risk" (Live Science, February 22, 2012; http://www.livescience.com/18608-citrus-fruits-stroke-risk.html)

- "Walk faster and you just might live longer" (NBC News, January 4, 2011; http://www.nbcnews.com/id/40914372/ns/health-fitness/t/walk-faster-you-just-might-live-longer/#.Vc-yHvlViko)

In many cases those conducting the research cautioned against making a causal conclusion, but the headlines are what people read and remember. Students should always question whether a causal conclusion is warranted, and should know how to answer that question.

2. Statistical significance versus practical importance

Ideally, students would learn the correct interpretation of *p*-values. (See Nuzzo, 2014, for a non-technical explanation of *p*-values.) But the most common misinterpretation, that the *p*-value measures the probability that chance alone can explain observed results, is difficult to overcome. . So students at least should learn the importance of distinguishing between statistical significance and practical importance. The easiest way to illustrate the difference is to look at a variety of studies that have small *p*-values, and then look at a confidence interval for the population parameter in each case. Although *p*-values are often used in situations that don't involve an easily interpretable parameter (for which a confidence interval could be computed), showing examples for which there is an interpretable parameter will make the point about statistical versus practical significance, and especially about the importance of sample size.

In the previously referenced article, Nuzzo (2014) provides an example of a study claiming that those who meet online have happier marriages ($p < 0.001$) and lower divorce rates ($p < 0.002$) than those who meet offline. The *p*-values of less than 0.001 and 0.002 look impressive, but on a 7-point scale the average "happiness" ratings for the two groups were 5.48 and 5.64, and the divorce rates were 7.67% and 5.96%. The impressive *p*-values resulted from the large sample size of over 19,000 people.

Aspects such as the size of the study and how that impacts the *p*-value are important if students are to understand the distinction between statistical significance and practical importance. Introducing the concept of "effect size" can help make this point, as illustrated in the next section.

3. The difference between no effect and no statistically significant effect

Studies that find no statistically significant effect may never make it into publication because they are not thought to have found anything of interest and are not newsworthy. But often media reports will mention an "unsuccessful replication" of an earlier study, and report it as if it contradicts the earlier result. Consumers should understand that there are multiple reasons for this apparent contradiction, and they do not all imply a real contradiction.

Consider the following scenario, called the "hypothesis testing paradox" (Utts & Heckard, 2015). A researcher conducts a *t*-test for a population mean based on a sample of $n = 100$ observations, obtaining a result of $t = 2.50$ and $p = 0.014$, so the null hypothesis is rejected. The experimenter decides to repeat the experiment with $n = 25$ observations to verify the result, but finds disappointingly that the result is $t = 1.25$, $p = 0.22$, and so she cannot reject the null hypothesis. The effect seems to have disappeared. To salvage the situation, she decides to combine the data, so now $n = 125$. Based on the combined data, $t = 2.80$, *p*-value = 0.006! How could a second study

that seemed to diminish the statistically significant result of the first study somehow make the result stronger when combined with the first study?

The paradox is that the second study alone did not replicate the finding of statistical significance, but when combined with the first study the effect seems even stronger than the first study alone, with the *p*-value going from 0.014 to 0.006. The problem of course is that the test statistic and *p*-value depend on the sample size. In fact in this example, the effect size (measured as $(\bar{x} - \mu_0)/s$) is the same in both studies. It is the sample size that creates the difference in *t* and the *p*-value. See Table 1.1 for a numerical explanation.

Table 1.1: Hypothetical example of the relationships among sample size, test statistic and p-value. The same effect size yields a statistically significant result for a larger study.

Study	*n*	$(\bar{x} - \mu_0)/s$	Test statistic *t*	*p*-value
1	100	0.25	2.50	.014
2	25	0.25	1.25	.22
Combined	125	0.25	2.80	.006

4. Sources of potential bias in studies and surveys, and the population to which results apply

As fewer households maintain landline telephones and caller ID makes it easy to ignore calls from unfamiliar numbers, it is becoming increasingly difficult to get representative samples for surveys and other statistical studies. Consequently, in many cases the results of surveys and other studies may not reflect the population of interest. According to Silver (2014), "Even polls that make every effort to contact a representative sample of voters now get no more than 10 percent to complete their surveys — down from about 35 percent in the 1990s" (para. 1).

Lack of response is just one of many sources of bias that can enter into surveys and other studies. Other sources of bias include poor and/or intentionally biased wording of questions, the order in which questions are asked, who is asking, and whether the topic is one for which people are inclined to lie (for instance, to appear to conform to social norms). When reading results of surveys it's important to know exactly what was asked, who asked, whether questions were asked in person, by mail, phone or online, and whether special interest groups were involved in any way that could affect the results. Cohn and Cope (2011) provide a detailed list of questions journalists should ask when covering these kinds of studies, but the list is relevant to anyone interested in learning more about how to detect bias in survey results.

5. Multiple testing and selective reporting

Almost all studies measure multiple explanatory and/or response variables, and therefore conduct multiple statistical tests to find out which variables are related. It is very common for the media to report the findings that happen to be statistically significant without mentioning that they were part of a larger study. If the original research report did not correct the reported results for multiple testing, the statistically significant findings could easily be spurious. Given that 20 independent tests with true null hypotheses are expected to yield one with statistical significance, it is not surprising that false claims make their way into the media. Students should learn to ask how many different tests were conducted in a study, and whether the statistical results were adjusted for

multiple testing. Ioannidis (2005) illustrates this problem and related issues with many examples in his article "Why most published research findings are false."

6. Interpreting relative risk, absolute risk, personal risk, and risk tradeoffs

Knowing how to think about risk can help consumers in many ways. Education research has shown that teaching risk and relative risk in terms of frequencies instead of probabilities will make them easier for most students to understand, and giving baseline risks in addition to relative risks will allow people to make more informed decisions (Gigerenzer et al., 2008). For instance, students understand the idea that 3 out of 1000 people may die from a certain treatment more readily than they understand that 0.003 or 0.3% of those treated may die. Saying that a certain behavior doubles your probability (or risk) of cancer from 0.003 to 0.006 is not easy for most people to understand, but saying that the behavior increases the *number* of cases of cancer in people similar to them from 3 in 1000 to 6 in 1000 is much easier to understand. And reporting frequencies instead of proportions makes the role of baseline risk much more clear. Most people can understand that an increase from 3 in 1000 to 6 in 1000 is different than an increase from 30 in 1000 to 60 in 1000, and will immediately recognize what the baseline risk is in each case.

Another important feature of risk to explain to students is that changing one behavior to avoid risk may lead to increasing risk of a different outcome. For instance, taking drugs to reduce blood pressure or cholesterol may increase the risk of other medical problems. Having a mammogram to reduce the risk of undetected breast cancer may increase the risk of the effects of radiation, or add a psychological risk of having a false positive result, and the accompanying stress. A striking example by Gigerenzer et al. (2008), described by Utts (2010), illustrated how a media scare associated with birth control pills in the United Kingdom resulted in reduced use of the pills, but led to large increases in abortions and teen pregnancies, which had much higher risks than the use of the pills would have had. Students should learn to view behavior changes that reduce risk in the broader context of risk trade-offs.

7. Conditional probability and "confusion of the inverse"

Psychologists know that people have very poor intuition about probability. One example is called "confusion of the inverse," in which people confuse conditional probability in one direction with conditional probability in the other direction. A classic example is confusing the probability of a positive test result given that you have a disease with the probability of having the disease, given a positive test result. The two probabilities can of course be vastly different, and this confusion has undoubtedly led to much unnecessary angst in people who receive a false positive medical test result. In one scenario, Eddy (1982) showed that physicians gave estimates of the probability of having breast cancer given a positive mammogram that were 10 times too high, close to 0.75 when the actual probability was 0.075.

Another example of this confusion is in the courtroom, where the probability of guilt given particular evidence is not the same as the probability of that evidence, given that the person is guilty. Again these can be vastly different, and juries need to recognize the difference. Similar to explanations of relative risk, conditional probabilities are easier to understand using frequencies instead of proportions or probabilities. One of the easiest ways to illustrate conditional probabilities in both directions, and how they differ, is through the use of a "hypothetical hundred thousand" table. See Utts (2010) for an example, as well as for other examples of how psychological influences can affect probability estimates and interpretation.

1.3.3 What Decision Makers Need to Know

Statistical ideas and methods provide many tools for making decisions in life, especially decisions that involve trade-offs. Previously we discussed how competing risks often need to be taken into account when making decisions. We now discuss some other ways in which statistical ideas can help with making decisions when trade-offs are involved.

1.3.3.1 Using expected values to make decisions Insurance companies, casinos, lottery agencies and sellers of extended warranties all rely on expected values, and can exploit consumers who do not understand them. In all of those cases consumers overall are the losers, but sometimes the protection (as with insurance and extended warranties) is worth the loss. The important point for consumers is to understand how to figure out when that is true. As an example, if you buy an appliance, should you buy the extended warranty? If you have sufficient income or financial reserves to have it fixed or replaced, the answer is probably not. In the long run, you will lose money if you often buy insurance and extended warranties. But an individual consumer does not get the benefit of "the long run" in something like the purchase of a house or car, so in those cases it may be worth having the insurance in the (small probability) event of a disaster. Also, if you are a real klutz and tend to break things, you might actually come out ahead with certain extended warranties. Students should understand these issues so they can make informed choices.

Example 3: Should you Pay in Advance?

Here is a simple example of where knowledge of expected value could be useful in making a decision (Utts, 2014, Exercise 17.29). Suppose you are planning to stay at a hotel a month from now but are not 100% sure you will take the trip. The hotel offers two payment choices. You can pay an "advance purchase" price of $85 now, nonrefundable. Or, you can reserve a room and pay $100 when you go, but not pay anything if you decide not to go. Which choice should you make? With the advance purchase, the "expected value" of what you will pay is $85, because the probability is 1.0 that you will pay that amount. Define p to be the probability that you will take the trip. If you don't use the advance purchase option, the expected value for what you will pay is $(\$100)(p) + (\$0)(1 - p) = \$100p$. Note that $\$100p$ is less than $85 if p is less than 0.85. So if you think the probability of taking the trip is less than 0.85, the advance purchase is not a good idea, but if you think the probability is higher than 0.85, the expected value is lower by taking advantage of the advance purchase.

1.3.3.2 Using the Hypothesis Testing Framework to Make Decisions In addition to the technical aspects of statistical studies, the reasoning used for hypothesis testing can be useful in making decisions even without quantitative data. Consider the following example from the life of one of the authors.

Example 4: Making a Tough Decision
The empty wrapper for a chocolate bar that someone in your household remembers partially eating is sitting on a table, carelessly left where your dog could find it. The human involved cannot remember whether he ate the whole chocolate bar or left half of it exposed on the table. You fear that the dog consumed the remainder of the chocolate, an indulgence that could be fatal to the dog. Should you rush to the veterinarian to have your dog's stomach pumped?

We can think about this decision in the framework of hypothesis testing, and look at the equivalent of Type 1 and Type 2 errors when considering the decision, as shown in Table 1.2. The actual decision will depend on how likely you think the two hypotheses are, but illustrating the possible choices and their consequences can be informative and helpful in making a decision.

Table 1.2: Did the dog eat the chocolate?

	DECISION	
HYPOTHESIS	*Dog did not eat the chocolate*	*Dog did eat the chocolate*
Null: Dog did not eat the chocolate	No trip to veterinarian; OK	Type 1 error: Go to vet; dog has stomach pumped needlessly
Alternative: Dog did eat the chocolate	Type 2 error: Do not go to vet. Dog could die	Go to vet; thank goodness you had stomach pumped

In this example, if there was even a relatively small chance that the dog ate the chocolate most dog-owners would be likely to take the dog to the vet for an evaluation. In general, the decision would be based on the seriousness of the consequences of the two types of errors. Laying the choices out in this kind of table makes those consequences more clear.

1.3.4 Final Remarks about the Importance of Statistics

It is hard to predict the future of data science and statistics, as resources become available that allow easy access to data and methods for visualizing and analyzing them. As eminent statistician Brad Efron noted, "Those who ignore statistics are condemned to reinvent it" (attributed to Brad Efron by Friedman, 2001, p. 6), and as Wild (2015) notes, "their ignorance can do real damage in the meantime" (p. 1). Statistics educators have a grave responsibility and an exciting opportunity to make sure that everyone learns how useful statistics can be. Statistics education researchers have their work cut out for them in figuring out how best to convey these ideas in ways that are useful and that will allow students to continue to make informed decisions throughout life.

1.4 Where Statistics is Heading

1.4.1 An Exciting Time to be a Statistician

This is an exciting time to be a statistician. Interest in the discipline of statistics and the analysis of data is booming. The amount of information collected in our increasingly data-centered society is staggering. Statistical expertise is more valuable than ever, with society and employers clamoring for graduates with the ability to blend knowledge of statistics, data management, computing, and visualization to help make better decisions. But with the opportunities afforded by this rich information come threats for statistics as a discipline. What does the future hold? What do we need to be addressing to ensure that students are developing the statistical skills necessary?

Speaking of the recent availability of a vast flood of data is not hyperbole. George Lee of Goldman Sachs estimates that 90% of the world's data have been created in the last two years (http://www.goldmansachs.com/our-thinking/trends-in-our-business). The 2013 Future of Statistics (London) report (http://bit.ly/londonreport) enumerates examples such as astronomy, where new telescopes will generate a petabyte of data each day and commercial databases at social media

companies such as Facebook, which generate 500 terabytes per day. United States President Barack Obama signed an Open Data Executive Order in 2013 (https://www.whitehouse.gov/sites/default/files/microsites/ostp/2013opendata.pdf) that called for data on health, energy, education, safety, finance, and global development to be made machine accessible to "generate new products and services, build businesses, and create jobs," and this has led to increased access to sophisticated and detailed information.

These increasingly diverse data are being used to make decisions in all realms of society. Consider the theme for the AAAS annual meeting in 2015 (Innovations, Information, and Imaging): "Science and technology are being transformed by new ways to collect and use information. Progress in all fields is increasingly driven by the ability to organize, visualize, and analyze data" (http://meetings.aaas.org/program/meeting-theme).

Planning a trip to New York City (NYC)? It's straightforward to download and analyze data on all commercial flights in the United States since 1987 (180 million records, http://www.amherst.edu/~nhorton/precursors), fourteen million taxi rides in 2013 (http://www.andresmh.com/nyctaxitrips/) and over a billion records for 2009-2015 (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml), millions of Citi Bike rentals (http://www.citibikenyc.com/system-data), and restaurant violations (http://www.nyc.gov/html/doh/html/services/restaurant-inspection.shtml). Useful information of this type is widely available for many other large cities, governments, and domains.

It's worth noting that although much is said about "big data," none of these NYC examples qualify. A typical definition of "big data" requires datasets with sizes that are difficult to process in a timely fashion using a typical workflow. The NYC examples have a modest *Volume* (the first of the 3V's model for describing big data (http://web.archive.org/web/20110710043533/http://www.gartner.com/it/page.jsp?id=1731916) but low *Velocity* and modest *Variety* (forms of data not easily stored in a rectangular array). Though issues of "big data" are important, many more challenges and opportunities are available for smaller scale information.

The development of new and easier to use computational tools (many of which are open-source with less barrier to adoption) has spurred the analysis of these new data sources. Recent efforts include general frameworks for data manipulation (Wickham, 2014), improved access to high performance database systems (e.g., http://cran.rstudio.com/web/packages/dplyr/vignettes/databases.html), and sophisticated interfaces for data scraping and related web technologies (e.g., Nolan & Temple Lang, 2014). A number of instructors have taken heed of the advice of those working to incorporate data wrangling and management skills early in the curriculum (Carver & Stevens, 2014).

1.4.2 A Challenging Time for Statistics

Although this is undoubtedly an exciting time to be a statistician, there are a number of challenges that loom. The demand for quantitative skills is clearly there. The widely cited McKinsey report (http://tinyurl.com/mckinsey-nextfrontier) described the potential shortage of hundreds of thousands of workers with the skills to make sense of the enormous amount of information now available. But where will the hundreds of thousands of new workers anticipated by the McKinsey report and others come from? Graduates of undergraduate statistics programs will be a small fraction (even if the growth seen in the recent decade continues or accelerates). Increased supply

is unlikely to be solved by an influx of new statistics doctoral students; while the number of doctoral graduates is slowly increasing, growth is insufficient to meet demand for new positions in industry, government, and academia.

Where else can these skilled graduates be found? If they aren't produced by statistics programs, where will they come from? The London report describes the need for *data scientists*, the exact definition of which is elusive and very much a matter of debate-and raises important questions about the identity and role of statisticians (Horton, 2015; Wasserstein, 2015). What is meant by data science? What skills are required? What training is needed to be able to function in these new positions? What role does statistics have in this new arena? How can it be ensured that critical statistical messages be transmitted to students educated in other types of program that feed the data-science shortfall?

A widely read Computing Research Association white paper on the challenges and opportunities with "Big Data" starts in an encouraging manner: "The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of 'Big Data'" (http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf). But it is disconcerting that the first mention of statistics is on the sixth page of the report: "Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples." The remaining references include statistics in passing as a bag of tricks (but not central to the use of data to inform decision-making).

In his video introduction to the keynote for the Strata+Hadoop Big Data Conference in 2015, United States President Barack Obama stated that "understanding and innovating with data has the potential to change how we do almost anything for the better." We applaud these sentiments. However, the fact that "statistics" was not mentioned in the presentation (or in many media depictions of this new and growing field) is a serious concern.

As another example of the challenges for the role of statistics in this era of "Big Data," consider the new Advanced Placement Computer Science Principles course, offered for the first time in the Fall of 2016 (https://advancesinap.collegeboard.org/stem/computer-science-principles). This course focuses on the development of foundational computing skills, programming literacy, and an understanding of the impact of computing applications. It includes "creativity" and "data and information" as two of seven "big ideas" that underlie the curriculum. The description of creativity includes discussion of how "computing facilitates exploration and the creation of computational artifacts and new knowledge that help people solve personal, societal, and global problems."

Big idea 3 is subtitled "Data and information facilitate the creation of knowledge." The Course Description states:

> Computing enables and empowers new methods of information processing, driving monumental change across many disciplines — from art to business to science. Managing and interpreting an overwhelming amount of raw data is part of the foundation of our information society and economy. People use computers and computation to translate, process, and visualize raw data and to create information. Computation and computer science facilitate and enable new understanding of data and information that contributes knowledge to the world. Students in this course work with data using a variety of computational tools and techniques to better understand the many ways in which data is transformed into information and knowledge. (p. 19)

Learning Objective 3.1.1 describes "use of computers to process information, find patterns, and test hypotheses about digitally processed information to gain insight and knowledge." This objective feels more expansive than the entire Advanced Placement Statistics course, where students are expected to "describe patterns and departures from patterns; plan and conduct a study; explore random phenomena using probability and simulation; and estimate population parameters and test hypotheses" (http://media.collegeboard.com/digitalServices/pdf/ap/ap-course-overviews/ap-statistics-course-overview.pdf).

While the success of its implementation remains to be seen, the Advanced Placement Computer Science Principles course promises an expansiveness of vision and pregnant sense of possibility for personal lives and the wider world. This is something that statistics education needs to learn from. The London report warned that unless statisticians engage in related areas such as computation and data-related skills that are perhaps less familiar there is a potential for the discipline to miss out on the important scientific developments of the 21st century.

What are the areas where statistics may need to adapt to be relevant to data science? Technology is one key realm. While the Guidelines for Assessment and Instruction in Statistics Education (GAISE) K-12 and College reports (http://www.amstat.org/education/gaise) encouraged the use of technology (which, on a more positive note, is now widespread in most courses), hundreds of thousands of high school students still use calculators rather than computers for their analyses, limiting their ability to move beyond simple calculations or gain any sense of realistic workflows that they might encounter in the real world. But much worse, it also narrowly constricts their vision of what statistics is and can be, and walls out the huge potential of the visual sense for gaining insights from data. This is certainly not the technology being used by data scientists (or proposed for the new Advanced Placement Computer Science Principles course).

1.4.3 Where Are We Headed?

The growth of data science as a discipline presents both opportunities and challenges for statisticians and statistical educators (Ridgway, 2015). Data scientists are being hired by employers looking for innovative problem solvers with expertise in programming, statistical modeling, machine learning, and strong communication skills (Rodriguez, 2013).

Computer scientists bring useful skills and approaches to tackle the analysis of large, complex datasets. Statisticians bring important expertise in terms of the understanding of variability and bias to help ensure that conclusions are justified. In addition to "big data," increasingly sophisticated probabilistic (stochastic) models are being developed, for example in areas such as genetics, ecology and climate science. Data science is often described as a "team sport." The complementary skills from many historically disparate disciplines need to be blended and augmented to ensure that data science is on a solid footing. But this means that to be relevant in this age of data, statisticians must be better oriented towards data science, lest data science move on without statistics.

The emergence of statistics as a distinct discipline, and not just as an add-on to mathematics for highly educated specialists, is relatively new. The growth of data science has highlighted the importance of computer science, and shifted the ground in terms of connections with other disciplines. Some aspects of statistics are rooted in mathematics. Moving forward, however, the

connections to mathematics may weaken while the highly dynamic and productive interface with computer science is emphasized.

A number of individuals have proposed creative solutions for statisticians to respond to the data science challenge. In his 2012 ASA presidential address, Robert Rodriguez proposed a "big tent" for statistics that included anyone who uses statistics, including related disciplines such as analytics and data science (http://www.tandfonline.com/doi/abs/10.1080/01621459.2013.771010#.VRcolpPF9E8). Brown and Kass (2009) warned that to remain vibrant, statistics needs to open up its view of statistical training. Nolan and Temple Lang (2010) outlined a curriculum to build computational skills as a basis for real world analysis. Finzer proposed a framework to establish "Data Habits of Mind" (2013). Diane Lambert of Google described the need for students to be able to "think with data" (Horton & Hardin, 2015).

The future of statistics is likely to be closely tied to aspects of data science. Success in this realm will require extensive and creative changes to our secondary and tertiary curriculum, along with partnerships with colleagues in related disciplines. Considerable work in the realm of statistics education research is needed to assess approaches and methods that attempt to address these capacities. The American Statistical Association has been proactive in creating reports to address potential curriculum changes for the present; see the GAISE reports and the Curriculum Guidelines for Undergraduate Programs in Statistical Science (ASA Undergraduate Guidelines Working Group, 2014). But statistics and data science are rapidly evolving, and curriculum and pedagogical changes need to evolve as well to remain current.

1.5 Closing Thoughts

There is a growing demand for statistical skills at the same time that the field of statistics and data science is broadening. Although there are many barriers to the adoption of changes in a curriculum that is already bulging with topics and increasingly heterogeneous in terms of approach, the alternative – allowing data science to proceed without statistics – is not attractive. It would not only diminish statistics, it would also diminish "data science" and worsen data-based decision making in society. It would limit the attractiveness of statistics graduates to the employment market, and through that, limit the attractiveness of statistics programs themselves.

Cobb (2015) likened changing curricula to moving a graveyard: never easy in any circumstance. Developing additional capacities in statistics students takes time. This will likely require approaches that provide repeated exposure and a spiraling curriculum that introduces, extends, and then integrates statistical and data-related skills. It includes getting students to come to grips with multidimensional thinking, preparing them to grapple with real-world problems and complex data, and providing them with skills in computation. These are challenging topics to add to the curriculum, but such an approach would help students to tackle more sophisticated problems and facilitate their ability to effectively make decisions using data. Perhaps most importantly, it should also include successfully fostering an expansiveness of vision within students of the potential of statistics for their world and their own future lives.

With so much curricular and pedagogical change under way, this is an exciting time to be involved in statistics education research. To chart our way into an exciting future of teaching and learning

that best benefits our students, there are so many important research questions to be addressed, including determining how best to target, structure, teach, and assess the emerging curricula. There has never been a wider array of interesting and important problems for statistics education researchers to grapple with than there is right now. The insights within this volume should help spark and guide efforts in this realm for many years to come.

References

American Statistical Association Undergraduate Guidelines Workgroup (2014), *Curriculum Guidelines For Undergraduate Programs in Statistical Science*. American Statistical Association, Alexandria, VA. Online: https://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf.

Applebaum, B. (2015), Vague on your monthly spending? You're not alone. *New York Times*, May 21, 2015, page A3.

Arnold, P.A. (2013). *Statistical Investigative Questions: An enquiry into posing and answering investigative questions from existing data*. PhD thesis, Statistics University of Auckland. Online: https://researchspace.auckland.ac.nz/bitstream/handle/2292/21305/whole.pdf?sequence=2.

Baldi, B., & Utts, J. (2015). What your future doctor should know about statistics: Must-include topics for introductory undergraduate biostatistics. *The American Statistician*, 69(3), 231–240.

Bartholomew, D. (1995). What is statistics? *Journal of the Royal Statistical Society, A*, 158, 1–20.

Blair, R., Kirkman, E.E., & Maxwell, J.W. (2010). Statistical abstract of undergraduate programs in the mathematical sciences in the United States: Fall 2010 CBMS survey. Online: http://www.ams.org/profession/data/cbms-survey/cbms2010-Report.pdf.

Box, G.E.P. (1990). Commentary. *Technometrics*, 32, 251–252.

Brown, E.N., & Kass, R.E. (2009). What is statistics? (with discussion). *The American Statistician*, 63(2), 105–123.

Carver, R. H., & Stevens, M. (2014). It is time to include data management in introductory statistics, Ninth International Conference on Teaching Statistics. Online: http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_C134_CARVER.pdf.

Chambers, J.M. (1993). Greater or lesser statistics: A choice for future research. *Statistics and Computing*, 3(4), 182–184.

Chance, B. (2002). Components of statistical thinking and implications for instruction and assessment, *Journal of Statistics Education*, 10(3), Online: http://www.amstat.org/publications/jse/v10n3/chance.html.

Cobb, G.W. (2015). Mere renovation is too little, too late: We need to rethink the undergraduate curriculum from the ground up. *The American Statistician*, 69(4), 266–282.

Cobb, G.W., & Moore, D.S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, 104(9), 801-823.

Cohn, V., & Cope, L. (2011). *News and Numbers: A Writer's Guide to Statistics*, Wiley-Blackwell.

Davidian, M. (2013). Aren't we data science? *Amstat News*, July 1. Online: http://magazine.amstat.org/blog/2013/07/01/datascience/.

De Veaux, R. D., & Velleman, P. (2008). Math is music; statistics is literature. *Amstat News*, Sept 2008, No 375, pp. 54–60.

Eddy, D.M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (Chapter 18). Cambridge University Press, Cambridge, England.

Farrell, D., & Greig, F. (2015). Weathering volatility: Big data on the financial ups and downs of U.S. individuals, J.P. Morgan Chase & Co. Institute Technical Report, May 2015, accessed online at http://www.jpmorganchase.com/corporate/institute/research.htm, August 15, 2015.

Fienberg, S.E. (1992). A brief history of statistics in three and one-half chapters: A review essay. *Statistical Science*, 7(2), 208–225.

Fienberg, S.E. (2014). What is statistics? *Annual Review of Statistics and its Applications*, 1, 1–9.

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education*, 7(2). Online: http://escholarship.org/uc/item/7gv0q9dc.

Forbes, S. (2014). The coming of age of statistics education in New Zealand, and its influence internationally. *Journal of Statistics Education*, 22(2). Online: http://www.amstat.org/publications/jse/v22n2/forbes.pdf

Friedman, J.H. (2001). The role of statistics in the data revolution? *Int. Statist. Rev.*, 69, 5–10.

Friendly, M. (2008). The golden age of statistical graphics. *Statistical Science*, 23(4), 502–535.

Futschek, G. (2006). Algorithmic thinking: The key for understanding computer science. In *Lecture Notes in Computer Science 4226* (pp. 159-168), Springer, New York.

GAISE College Report (2016). *Guidelines for Assessment and Instruction in Statistics Education*, American Statistical Association, Alexandria, VA. Online: http://www.amstat.org/education/gaise.

Gentleman, R., & Temple Lang, D. (2004). *Statistical analyses and reproducible research*, Bioconductor Project Working Papers, Working Paper 2. Online: http://biostats.bepress.com/bioconductor/paper2.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L.M., & Woloshin, S. (2008). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8(2), 53–96.

Gould, R. (2010). Statistics and the modern student. *International Statistical Review*, 78(2), 297–315.

Grolemund, G., & Wickham, H. (2014). A cognitive interpretation of data analysis. *International Statistical Review*, 82(2), 184–204.

Hacking I. (1990). *The Taming of Chance*. Cambridge Univ. Press, New York.

Hahn, G. J., & Doganaksoy, N. (2012). *A career in statistics: Beyond the numbers*. Wiley, Hoboken, NJ.

Hand, D. (2014). *The improbability principle: Why coincidences, miracles, and rare events happen every day*, Scientific American, New York.

Holmes, P. (2003). 50 years of statistics teaching in English schools: Some milestones (with discussion). *Journal of the Royal Statistical Society. Series D (The Statistician)*, 52(4), 439–474.

Horton, N.J. (2015). Challenges and opportunities for statistics and statistical education: Looking back, looking forward. *The American Statistician*, 69(2), 138–145.

Horton, N.J., Baumer, B.S., & H. Wickham. (2014). *Teaching precursors to data science in introductory and second courses in statistics*. Online: http://arxiv.org/abs/1401.3269.

Horton N.J., & Hardin, J. (2015). Teaching the next generation of statistics students to "Think with Data": Special issue on statistics and the undergraduate curriculum. *The American Statistician*, 69(4), 258-265. Online: http://amstat.tandfonline.com/doi/full/10.1080/00031305.2015.1094283.

Ioannidis, J. (2005). Why most published research findings are false, *PLoS Medicine*, 2, e124.

Kaplan, D.T. (2011). *Statistical modeling: A fresh approach* (2nd edition), Mosaic Press, Minnesota.

Kendall, M.G. (1960). Studies in the history of probability and statistics. Where shall the history of statistics begin? *Biometrika*, 47(3), 447–449.

Lawes, C. M., Vander Hoorn, S., Law, M. R., & Rodgers, A. (2004). High cholesterol. Chapter 7 in M. Ezzati et al. (Eds.), *Comparative Quantification of Health Risks, Global and Regional Burden of Disease Attributable to Selected Major Risk Factors, Vol. 1* (pp. 391–496). World Health Organization, Geneva.

MacKay, R.J., & Oldford, R.W. (2000). Scientific method, statistical method and the speed of light. *Statistical Science*, 15(3), 254–278.

Marquardt, D.W. (1987). The importance of statisticians. *Journal of the American Statistical Association*, 82(397), 1–7.

Moore, D.S. (1998). Statistics among the Liberal Arts. *Journal of the American Statistical Association*, 93(444), 1253–1259.

Moore, D.S. (1999). Discussion: What shall we teach beginners? *International Statistical Review*, 67(3), 250–252.

Moore, D.S. (2001). Undergraduate programs and the future of academic statistics. *The American Statistician*, 55(1), 1–6.

Moore, D.S., & Notz, W.I. (2012). *Statistics: Concepts and Controversies* (8th edition), Macmillian Learning, New York.

Nature Editorial (2013). Announcement: Reducing our irreproducibility, *Nature*, 496. Online: http://www. nature.com/news/announcement-reducing-our-irreproducibility-1.12852.

Nolan, D., & Perrett, J. (2015). Teaching and learning data visualization: Ideas and assignments, Online: http://arxiv.org/abs/1503.00781.

Nolan, D. & Temple Lang, D. (2007). Dynamic, interactive documents for teaching statistical practice. *International Statistical Review*, 75(3), 295–321.

Nolan, D., & Temple Lang, D. (2010), Computing in the statistics curricula, *The American Statistician*, 64(2), 97–107.

Nolan, D., & Temple Lang, D. (2014). *XML and Web Technologies for Data Sciences with R*, Springer, New York.

Nuzzo, R. (2014). Scientific method: Statistical errors, *Nature*, 506: 150-152, February 13, 2014. Online: http://www.nature.com/news/scientific-method-statistical-errors-1.14700.

Pfannkuch, M., Budget, S., Fewster, R., Fitch, M., Pattenwise, S., Wild, C., & Ziedins, I. (2016). Probability modeling and thinking: What can we learn from practice? *Statistics Education Research Journal*, to appear.

Pfannkuch, M., & Wild, C.J. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking*. Dordrecht, The Netherlands: Kluwer Academic Publishers, Chapter 2.

Porter, T. M. (1986). *The rise of statistical thinking 1820–1900*. Princeton University Press, Princeton, NJ.

Pullinger, J. (2014). Statistics making an impact. *Journal of the Royal Statistical Society, A*, 176(4), 819–839.

Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review*, doi: 10.1111/insr.12110. Online: http://onlinelibrary.wiley.com/doi/10.1111/insr.12110/epdf

Rodriguez, R.N. (2013). The 2012 ASA Presidential Address: Building the big tent for statistics, *Journal of the American Statistical Association*, 108(501), 1–6.

Scheaffer, R. L. (2001). Statistics education: Perusing the past, embracing the present, and charting the future. *Newsletter for the Section on Statistical Education*, 7(1). Online: https://www.amstat.org/sections/educ/newsletter/v7n1/Perusing.html.

Schoenfeld, A. H. (1985). *Mathematical Problem Solving*. Orlando, FL: Academic Press.

Schutt, R., & O'Neil, C. (2013). *Doing data science: Straight talk from the frontline*. O'Reilly Media, Sebastopol, CA.

Silver, N. (2014). Is the polling industry in stasis or in crisis? Published online August 25, 2014 at FiveThirtyEightPolitics, http://fivethirtyeight.com/features/is-the-polling-industry-in-stasis-or-in-crisis, accessed August 15, 2015.

Snee, R. (1990). Statistical thinking and its contribution to quality. *The American Statistician*, 44(2), 116–121.

Stigler, S.M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard Univ. Press, Cambridge, MA.

Stigler, S.M. (2016). *The seven pillars of statistical wisdom*. Harvard Univ. Press, Cambridge, MA.

Utts, J. (2003). What educated citizens should know about statistics and probability. *The American Statistician*, 57(2), 74–79.

Utts, J. (2010). Unintentional lies in the media: Don't blame journalists for what we don't teach, *Proceedings of the Eighth International Conference on Teaching Statistics. Data and Context in Statistics Education*, International Statistical Institute, Voorsburg, The Netherlands.

Utts, J. (2014). *Seeing Through Statistics* (4th ed.), Cengage Learning, Stamford, CT.

Utts, J. (2015). The many facets of statistics education: 175 years of common themes. *The American Statistician*, 69(2), 100–107.

Utts, J., & Heckard, R. (2015), *Mind on Statistics* (5th ed.), Cengage Learning, Stamford, CT.

Vere-Jones, D. (1995). The coming of age of statistical education. *ISI Review*, 63, 3–23.

Wasserstein, R. (2015). Communicating the power and impact of our profession: A heads up for the next Executive Directors of the ASA, *The American Statistician*, 69(2), 96–99.

Wickham, H. (2009). ASA 2009 Data Expo, *Journal of Computational and Graphical Statistics.* 20(2), 281–283.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 10(59). Online: http://www.jstatsoft.org/v59/i10/.

Wild, C.J. (1994). On embracing the 'wider view' of statistics, *The American Statistician*, 48(2), 163–171.

Wild, C.J. (2015). Further, Faster, Wider. *The American Statistician*, online discussion. Online: https://s3-eu-west-1.amazonaws.com/pstorage-tf-iopjsd8797887/2615430/utas_a_1093029_sm0202.pdf.

Wild, C.J. (2017). Statistical literacy as the earth moves. *Statistics Education Research Journal*, to appear.

Wild, C.J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review*, 67(3), 223–265.

Xie, Y. (2014), *Dynamic Documents with R and knitr*, Chapman & Hall/CRC, London.