Chapter 1: What is Statistics?

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.

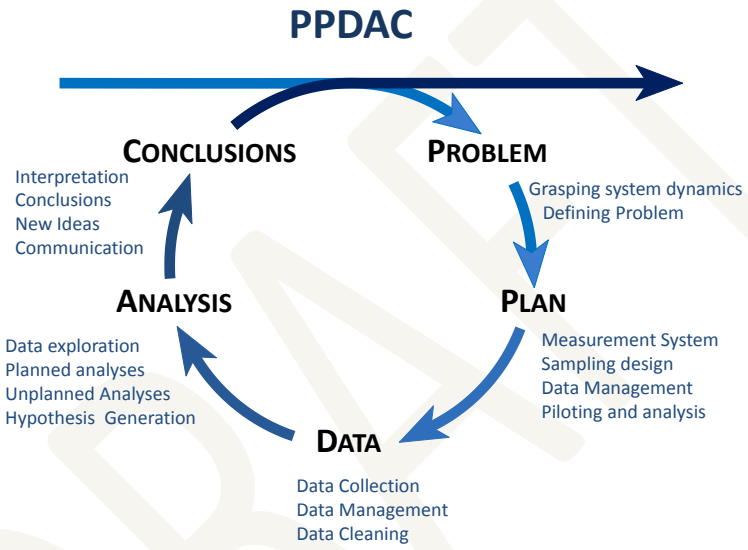

Fig. 1.1: The statistical inquiry cycle

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.