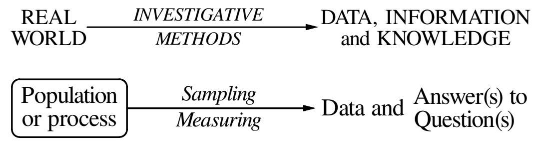
Figure 1.1. DATA-BASED INVESTIGATING: Introductory Overview

1. What is Statistics About?

As summarized in the two schemas at the right, statistics is concerned with *data-based investigating* (or empirical problem solving) of the real world, which means investigating some population or process on the basis of *data* to *answer* one or more *questions*. For this introductory discussion, only the investigative methods of *sampling* and *measuring* are shown in the lower schema.



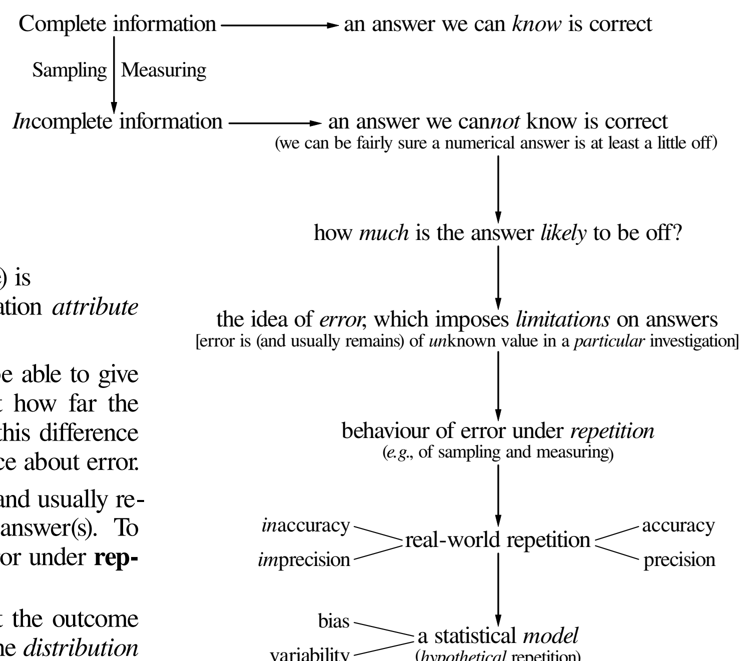
This introduction presupposes initially that data-based investigating is concerned with answer(s) to question(s) about a collection of *elements* which comprise a *population*. For example:

- a computer manufacturer may want to assess warranty claims for one of its products – an *element* would be one item of the product and the *population* is all such items sold over some specified period;
- a government department may want to know the proportion of Canadian adults who have concerns about the level of funding of the health care system – an *element* would be a Canadian adult and the *population* is all such Canadians;
- a financial institution may wish to find people whose profile makes them more likely to accept an unsolicited offer of a credit card – an *element* would again be a person but the *population* is harder to specify; it could be people whose profile puts them at or above an acceptance level deemed likely to make the credit card offering profitable for the financial institution.

The question(s) to be answered may *sometimes* be concerned with an *individual* to be selected in future from the population.

2. Key Ideas: Uncertainty, Error and Repetition

A central issue in data-based investigating is that external constraints and the type of information we require *impose* sampling and measuring processes on most investigating; we then need to assess the likely 'correctness' of the answer from the investigating in light of the *uncertainty* introduced by these two processes. In essence, this is the problem of *induction*. An overview of this matter is given in the schema at the right below; the main ideas are summarized at the left.



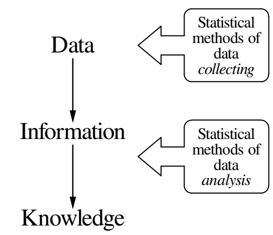
* If we have *complete* information, we can obtain a *certain* answer; that is, an answer we can *know* is correct.
* If we have *in*complete information, we can*not* know an answer is correct (an *un*certain answer) – in fact, we can be *fairly sure* a numerical answer is at least a little off;
  - sampling and measuring yield data (and, hence, information) that are *inherently* incomplete.
* An answer which is a *number* (like a sample average) is called a *point estimate* of the corresponding population *attribute* (the population average).
* To make such an answer more useful, we want to be able to give an *interval estimate*, a quantitative statement about how far the point estimate is *likely* to be from the true value – this difference is the **error** of the estimate. **Uncertainty** is ignorance about error.
* In a *particular* investigation, the size of the error is (and usually remains) *un*known; this is reflected in *limitations* on answer(s). To quantify uncertainty, we turn to the behaviour of error under **repetition** of the sampling and measuring processes;
  - an analogy is tossing a coin – we cannot predict the outcome of a *particular* toss but *probability* can describe the *distribution* of outcomes under *repetition* of the process of tossing the coin.
* In the classroom, we can do *actual* repetition (repeating over and over) to demonstrate, for example, the statistical behaviour of the values of sample attributes (*e.g.*, averages) and of the values which arise from measuring processes;
  - the two characteristics of *sign* and *magnitude* of (numerical) error lead, under repetition, to what we call *inaccuracy* and *imprecision*; the *inverses* of these two ideas – accuracy and precision – provide more familiar terminology, but we must realize that statistical methods (try to) manage the (undesirable) *former* to achieve needed levels of their (desirable) inverses.
* *Out*side the classroom, we recognize that *actual* repetition is usually not a viable option – we use instead *hypothetical* repetition based on an appropriate statistical *model* (for the selecting and measuring processes, for example);
  - we help maintain the distinction between the real world and the model by using bias and variability as the *model* quantities which represent inaccuracy and imprecision in the real world.

(*continued overleaf*)

∗ A statistical model, including its *assumptions*, allows us to achieve our aim of quantifying (some of) the uncertainty in our estimate(s); part of identifying the limitations on an Answer is to assess model error arising from the difference between ('idealized') modelling assumptions and the real-world situation.

∗ We emphasize sample error and measurement error in this introductory overview; other categories of error, which need more background to understand, are (besides model error) study error, non-response error and comparison error.
– There can also be *non*-statistical error, although subject-matter knowledge is generally required to assess it.
Discussion of each of the six categories of (statistical) error is pursued in later Parts of these Course Materials when the relevant statistical ideas have been developed and the context is appropriate.

NOTES: 1. Another diagrammatic summary of what statistics is about is given at the right; it shows a broad division of the focus of statistical methods into data *collecting* and data *analysis*.

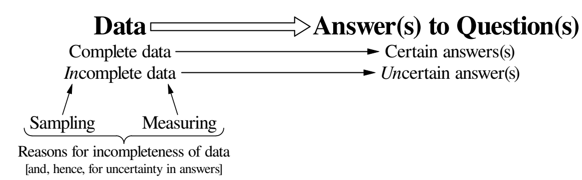


Instances of important questions to be answered (*i.e*., matters needing data-based investigating) are given by P. Calamai: *Why journalists can't add* [Chapter 3 (pages 15-24) in *Statistics, Science, and Public Policy*, A.M. Hertzberg and I. Krupta (editors), Queen's University, Kingston, Ontario K7L 3N6, 1998]:

> What are the big things? They are issues such as silicone breast implants, the downsizing of the middle class, gender pay equity, the consumer price index, violence against women, the poverty line, drinking and pregnant women, race and crime, international debt comparisons, unemployment rates, the disappearance of the Atlantic cod, the tainted blood scandal and, of course, BSE and CJD, two sets of initials known only to a few researchers until Mad Cow Disease hit the headlines in 1996.

2. The following are illustrations of the precept that *certainty* requires *complete information* and, conversely, *in*complete information inescapably yields *un*certainty [ignorance (usually of the size or magnitude) of error]:
   - In the proof of a theorem in mathematics, we have a set of axioms and the rules of logic; within this system of complete information, there can be *certainty* the theorem is true.
   - Games like chess and bridge involve small populations of elements (32 pieces and 52 cards, respectively) and a set of rules; within these systems of complete information, there can be *certainty* which player has won a game.
   - In data-based investigating, information boundaries are seldom as clearly defined as in mathematics, chess and bridge, and practical considerations impose sources of *in*completeness and, hence, of *un*certainty, which include:
     – sampling: although we want Answer(s) to apply to a *population*, we seldom have the resources to investigate more than a *sub*set of the population (a *sample*) – the idea of sample error reminds us a sample seldom agrees *exactly* with the population;
     – measuring: even apparently straight-forward quantities like length and weight have *inherent* uncertainty in the values we obtain (the idea of measurement error), while measuring quantities like the number of cod in the North Atlantic fishing grounds or people's *attitudes* or *opinions* (*e.g*., on mandatory registration of firearms, capital punishment, abortion) may introduce much *greater* measuring uncertainty.

DATA-BASED INVESTIGATING



A diagrammatic summary of these matters is given at the right above.

3. An illustration of *sample* error is polling to estimate a proportion of interest, typically from a national sample of about 1,000 to 1,500 adults. The *sample* proportion is a point estimate of the *population* proportion but is unlikely to be *exactly* equal to it; however, in a properly designed and executed poll, it is likely close enough (*i.e*., to have small enough sample error) to be an Answer with *acceptable* limitations, provided that *measurement* error (*e.g*., arising from the questionnaire, the interviewer and the interview process) has also been adequately managed.

4. Looking ahead, a summary of some **terminology**, and two distinctions (between a particular investigation and repetition, between the real world and the model) it helps us maintain, is given below at the right in Table 1.1.1; individual terms are defined to the left of the Table.
   - **Error**: the *difference* between what is stated [*e.g*., in an Answer] or assumed [*e.g*., in a response model] and the *actual* state of affairs.
   - **Inaccuracy:** *average* error under repetition.
   - **Bias:** the model quantity representing inaccuracy.
   - **Imprecision:** *standard deviation* of error under repetition.
   - **Variability:** the model quantity representing imprecision.
   - **Variation:** (natural) differences in (variate) values among the members of a group; *e.g*., a target population, a study population, a respondent population, a sample, repeated measurements, error values under repetition.

Table 1.1.1:

PARTICULAR INVESTIGATION	(HYPOTHETICAL) REPETITION	
Real World	**Real World**	**Model**
variation		
error	inaccuracy	bias
	imprecision	variability