The Four Questions of Data Analysis

Donald J. Wheeler

The four qustions of data analysis are the questions of description, probability, inference, and homogeneity. Any data analyst needs to know how to orgainze and use these four questions in order to obtain meaningful and correct results.

THE DESCRIPTION QUESTION

Given a collection of numbers, are there arithmetic values
that will summarize the information contained in those numbers
in some meaningful way?

The objective is to capture those aspects of the data that are of interest. Intuitive summaries such as totals, averages, and proportions need little explanation. Other summaries that are less commonly used may require some explanation, and even some justification, before they make sense. However, in the end, in order to be effective a descriptive statistic has to make sense—it has to distill some essential characteristic of the data into a value that is both appropriate and understandable. In every case, this distillation takes on the form of some arithmetic operation:

$$Data + Arithmetic = Statistic$$

As soon as we have said this, it becomes apparent that the justification for computing any given statistic must come from the nature of the data themselves—it cannot come from the arithmetic, nor can it come from the statistic. If the data are a meaningless collection of values, then the summary statistics will also be meaningless—no arithmetic operation can magically create meaning out of nonsense. Therefore, the meaning of any statistic has to come from the context for the data, while the appropriateness of any statistic will depend upon the use we intend to make of that statistic.

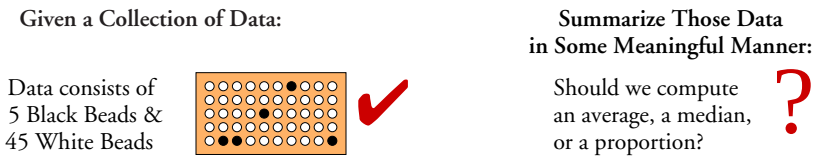

Figure 1: The Question of Description

This means that before we compute the simplest average, range, or proportion, *it has to make sense to do so.* Thus we have to know the *context* for any collection of values before we can select appropriate summary statistics. Among other things, this means that we will need to be careful to avoid mixing up apples, pineapples, and watermelons prior to computing the average weight per piece.

THE PROBABILITY QUESTION

Given a known universe,
what can we say about samples drawn from this universe?

Here we enter the world of deductive logic, the enumeration of possible outcomes, and mathematical models. For simplicity we usually begin with a universe that consists of a bowl filled with known numbers of black and white beads. We then consider the likelihoods of various sample outcomes that might be drawn from this bowl. This is illustrated in Figure 2.

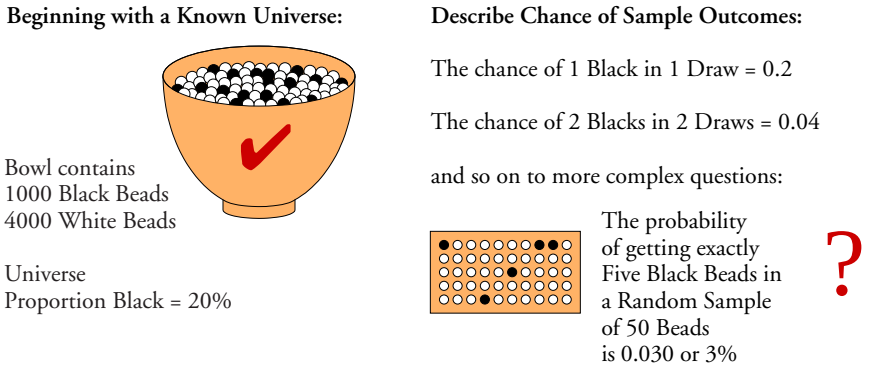

Figure 2: The Question of Probability

When we reason from a general situation, which is known, to descriptions of specific outcomes, which are presently unknown, we have an argument that is said to be deductive in nature. Deductive logic proceeds from generalities to specifics and always has a correct answer. It is a process of reasoning in which a conclusion follows necessarily from the premises presented.

When we begin with simple universes, such as beads in a bowl, we can often list all of the possible outcomes. From these enumerations it is then possible to characterize the likelihoods of different events.

Since the enumeration of outcomes quickly becomes tedious, shortcuts are sought. By developing mathematical models we can skip the enumeration step and jump directly from the known universe to the likelihoods of different possible outcomes.

As the mathematical models became increasingly sophisticated, and as the methods of computing and approximating the probabilities progressed, the models could be used to characterize more complex problems—problems that could never be handled by the enumeration approach. Thus, in probability theory we are, in effect, playing a game. We play this game to learn how things behave so that we can use this knowledge later. In introductory classes we restrict ourselves to playing this game with homogeneous and fixed universes.

Obviously, before students can make much headway in probability theory, they will need to be comfortable with deductive logic and mathematical models—two more elements of the foreign language of statistics. Fortunately, while probability theory is a necessary step in the *development* of modern statistical techniques, it is not a step that has to be mastered in order to analyze data effectively.

THE INFERENCE QUESTION

Given an unknown universe, and given a sample
that is known to have been drawn from that unknown universe,
and given that we know everything about the sample,
what can we say about the unknown universe?

This is usually thought of as the inverse of the problem addressed by the probability question. Here, it is the sample that is known and the universe that is unknown. Now the argument proceeds from the specific to the general, which makes it inductive in nature. Unfortunately, all inductive inference is fraught with uncertainty.

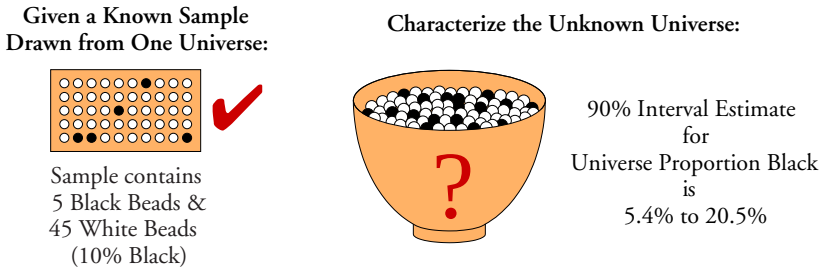

Figure 3: The Question of Inference

A sample result of 5 black beads and 45 white beads corresponds to a 90% Interval Estimate of 5.4% to 20.5% for the proportion of black beads in the bowl. The middle ninety percent of the plausible proportions fall in this interval. Thus, with inductive logic there is not a single right answer, but many plausible answers. Given this sample result, any percentage from 5.4% to 20.5% is plausible.

Statistical inference is the realm of tests of hypotheses, confidence intervals, and regression. These techniques allow us to estimate and evaluate the parameters of the unknown universe—proportions, means, and standard deviations. Of course such estimates make sense only when our outcomes are all obtained from a single universe. This assumption of a single universe is equivalent to the assumption that the behavior of these outcomes is described by one probability model. Once we have made this assumption, it is possible to use the probability model in reverse—given this outcome, these are the parameter values that are most consistent with the outcome.

While the mathematics of using the probability model in reverse makes everything seem to be rigorous and scientific, you should note that the whole argument begins with an assumption and ends with an indefinite statement. The *assumption* is that all of the outcomes came from the same universe, and the *indefinite statement* is couched in terms of interval estimates. Again, with inductive inference there is not one right answer, but many plausible answers.

THE HOMOGENEITY QUESTION

Given a collection of observations,
is it reasonable to assume that they came from one universe,
or do they show evidence of having come from multiple universes?

To understand the fundamental nature of the homogeneity question, consider what happens if the collection of values does not come from one universe.

Descriptive statistics are built on the assumption that we can use a single value to characterize a single property for a single universe. If the data come from different sources, how can any single value be used to describe what is, in effect, not one property but many? In Figure 4 the sample has 10 percent black. But if the 50 beads are the result of three separate draws from the three bowls at the bottom of Figure 4, each of which has a different number of black beads, which bowl is characterized by the sample result?

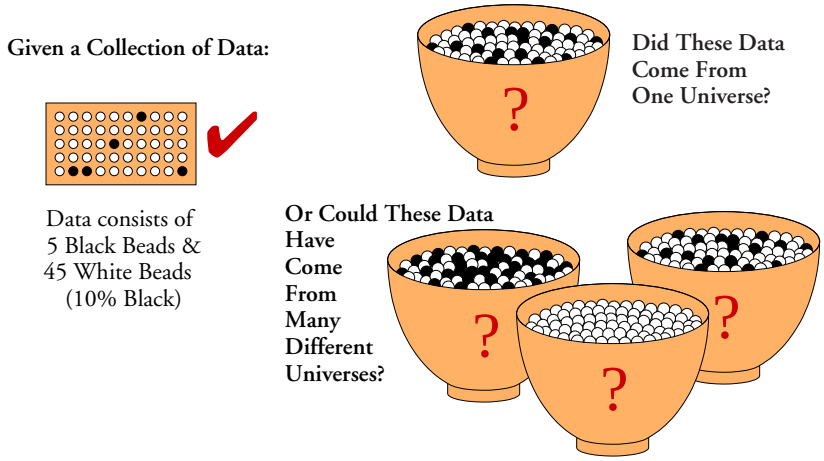

Figure 4: The Question of Homogeneity

Probability theory is focused on what happens to samples drawn from a known universe. If the data happen to come from different sources, then there are multiple universes with different probability models. If you cannot answer the homogeneity question, then you will not know if you have one probability model or many.

Statistical inference assumes that you have a sample that is known to have come from one universe. If the data come from different sources, what does your interval estimate represent? Which of the multiple universes does it characterize?

Therefore, before you can use the structure and techniques developed to answer the first three problems, you will need to examine your data for evidence of that homogeneity which is implicitly assumed by the use of descriptive statistics, the concepts of probability theory, and the techniques of statistical inference. *This implicit assumption of homogeneity, that is part of everything we do in traditional statistics classes, becomes a real obstacle whenever we try to analyze data.*

HOW TO ANALYZE DATA

When we find evidence of a changing universe in a situation where there should be only one universe we will be unable to learn anything from descriptive statistics. When the universe is changing we cannot gain from statistical inference, nor can we make predictions using probability theory. Any nonhomogeneity in our collection of values completely undermines the techniques developed to answer each of the first three questions. The lack of homogeneity is a signal that unknown things are happening, and until we discover what is happening and remove its causes, we will continue to suffer the consequences. Computations cannot remedy the problem of a lack of homogeneity; action is required.

How can we answer the homogeneity question? We can either *assume* that our data possess the appropriate homogeneity, or we can *examine* them for signs of nonhomogeneity. Since anomalous things happen in even the most carefully controlled experiments, prudence demands that we choose the second course. And the primary tool for examining a collection of values for homogeneity is the process behavior chart.

To examine our data for signs of nonhomogeneity we begin with the tentative assumption that the data are homogeneous and then look for evidence that is inconsistent with this assumption. When we reject the assumption of homogeneity we will have strong evidence which will justify taking action to remedy the situation. When we fail to reject the assumption of homogeneity we will know that any nonhomogeneity present is below the level of detection. While this is a weak result, we will at least have a reasonable basis for proceeding with estimation and prediction.

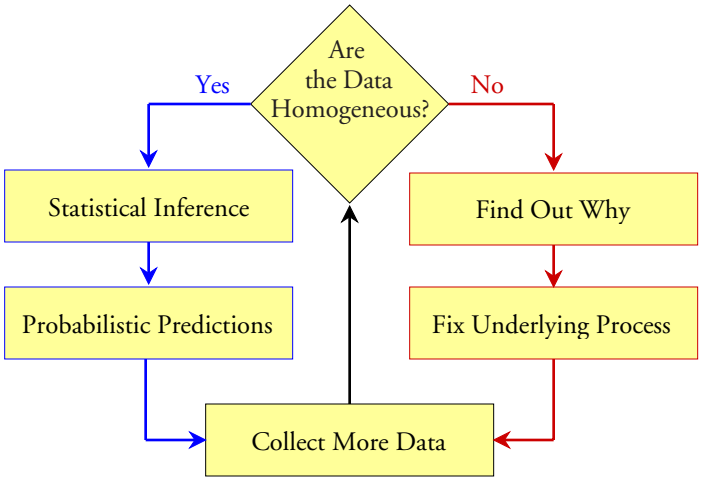

Figure 5: Homogeneity is the Primary Question of Analysis

Thus, for practitioners the first question must always be the question of homogeneity. Given a collection of data, did these data come from one universe? In fact, is it reasonable to assume that there is a universe? Only after this fundamental question has been addressed does the practitioner know how to proceed. If the assumption of a universe is reasonable, then the techniques of statistical inference may be used to characterize that universe, and then, with reasonable estimates of the parameters, probability models may be used to make predictions. But if the

assumption of a universe is not justified, the practitioner needs to find out why.

This is not the way classes in statistics are taught, but it is the way you have to do data analysis. Look at your data on a process behavior chart. If there are surprises in your data, and there often will be, then learn from these surprises. If there are no surprises, then you may proceed to analyze your data as if they came from a single universe. Any attempt to analyze data that does not begin by addressing the question of homogeneity is flawed.