Philosophy of Statistics

André-Ignace Ghonda Lukoki

July 15, 2024

1 Introduction

The American Philosophical Association describes philosophy as a field pursuing questions in every dimension of human life. It is a resoned pursuit of fundamental truths, a quest for understanding. It seeks to establish standards of evidence, to provide rational methods of resolving conflicts, and to create techniques for evaluating ideas and arguments. The discipline can be futher divided into subdisplines such as *logic*, *metaphysics*, *epistemology*, and *ethics*. However, there is overlap between the subdisciplines.

1.1 Logic

The branch of philospohy that focuses on the analysis of arguments is called *logic*. There are deductive arguments where the conclusion is logically entailled by the premises. In such cases, it is impossible for the conclusion to be false while the premises are true. Inductive arguments are arguments where the premises do not logically entail the conclusion, but the premises provide good reasons to believe it.

1.2 Metaphysics

Defining the causal relationships that exist can be difficult. The philosopher David Hume believed that causality could only be found through experimentation. Experience can only reveal temporal relations and cannot establish the necessary connection between cause and effect. The discussion of causality should concern those interested in scientific knowledge because much of modern science relies on statistical methods to estimate causal relationships. However, one may question whether those statistical methods are well-equipped to account for more than correlations between variables. Those questions can be thought of metaphysical questions. *Metaphysics* is the study of the fundamental nature of reality.

1.3 Epistemology

Epistemologist are interested in questions about the definition, the sources, the limits of knowledge but also the meaning of justification. They are especially interested in scientific discoveries, and their methodologies. Some epistemologist make use of statistics as means of reliable knowledge generation while others question the reliability of certain statistical methods for generating knowledge.

1.4 Ethics

Ethical questions are questions that attempt to define what we ought to do as individuals or as a society. These questions have a moral aspects and reasonable people might disagree on the answers.

2 Definitions

In the empirical sciences that allow for the collection of data, inferential statistics can be thought of as a set of methods for drawing conclusions about the world from limited information. The conclusions go beyond the data at hand meaning that the arguments presented by statistics are inductive. In his book, *The Seven Pillars of Statistical Wisdom*, Stigler presents seven principles that form a conceptual foundation for statistics as a discipline.

2.1 Aggregation

Aggregation is the combining of observations for the purposes of information gain. It might be done by taking the mean or median of a variable for a sample. In some context, the sample mean describes well the typical observation across the sample. The mean has limits as it is not particularly resistant to outliers thus, in the presence of outliers, the median is more appropriate. There are other forms of aggregation such as the variance or the range that measure the level of variablilty within the sample. Note that aggregation does not only occur as summary statistics, Least squares or Maximum Likelihood estimates can be thought of as "weighted aggregates of data that submerge the identity of individuals".

2.2 Information

The previous section established that we can gain information by combining observations Expanding on this observation, consider the following example: There is a jar containing an unknown amount of candy beans c. To estimate c, we ask a group of n people to give an estimate of c denoted by $X_i,\ i = 1, \ldots, n$ and summarize it by the sample mean. The observation of the mean leads to the following questions: How precise is this estimate? How much information do we gain by doubling the number of guessses? Those can be answered by looking at the variance and the standard deviation.

$$Var(\bar{X}) = \sigma^2/n$$
$$sd(\bar{X}) = \sigma/\sqrt{n}$$

Using the inverse of the standard deviation to measure precision, we see to increase the precision of our estimator by a factor of k, we need to increase the sample by k^2 since $\frac{k}{sd(\bar{X})} = \frac{k\sqrt{n}}{\sigma} = \frac{\sqrt{k^2 n}}{\sigma}$.

2.3 Likelihood

Consider the following example: A woman claims that she is able to distinguish the cases where the tea was poured first then the milk or the other way around. To test the woman's ability, we collected data on the experimentand see how likely the ddata is under the assumption that she does not have this ability. This assumption is called the *null hypothesis*, denoted H_0. For the test, she is required to drink 8 cups, with 4 for each kind. In this case, we would expect her to correctly identify all four cups approximately 1.4% of the time. The result $X = 4$ is rare under the null hypothesis, then if we observe $X = 4$, we have evidence against H_0. It is important to note that the data can never strictly contradict a hypothesis, it might only provide evidence against the null when the data is improbable under under that hypothesis.

2.4 Intercomparison

The variance σ^2 is a measure that can be used to quantity the amount of information that is present in a given sample mean however, this variable is not known. There exist an estimate of σ^2 that only uses internal information to assess the variability of $\bar{X}$, the sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

The effects of the approximation are only felt in small sample analysis. In large samples, the distribution of the sample means approaches the normal distribution. The link between the sample estimates and population estimates is important in statistics since much of statistics relies on inductive reasoning.

2.5 Regression

Regression analysis attempts to estimate the relationship between at least two variables. Regressions can be used for prediction where we look for information about a response variable based on known measurements of known predictor variables. Regressions can also be used to explain a change in one variable based on the changes in the independent variables. Those models often raise the issue of causation.

2.6 Design

In the medical sector, one may be interested in estimating the effect of a new medication on a given condition. Each treatment can be thought of as a categorical variable, called a factor, with two levels: either the treatment has been given to a patient, or it hasn't. The first experiment would be to give the treatment to one group and a placebo to another. Such procedure is called a *one factor at a time* (OFAT) design. The alternative to OFAT designs are *factorial* designs where we allow more than one variable to vary. This allows the designer to estimate the interactions between the factors as well. There are many important principles in experimental design to help us assess the effectiveness of the experimental treatment.

Randomization The use of randomization helps block the negative effect of confounding variables.

Blocking Blocking is a technique for including a factor (or factors) in an experiment that lead to undesirable variation in the outcome. We control for those factors by randomly assigning treatment levels within each factor.

Replication Replication is the repetition of an experiment on many dierent units.

2.7 Residual

The quality of a model can be assessed by analyzing its residuals. In theory, the residuals ϵ are assumed to be a noisy, random, and normally distributed elements $\epsilon \overset{iid}{\backsim} N(0, \sigma^2)$. If the model is well-specified the error term should be normally distributed. The true population parameter are unknown and are estimated, resulting in sample parameters $\hat{\beta}$ in vector notation.

$$Y - f(x|\hat{\beta}_0, \hat{\beta}_1) = \bar{\epsilon}$$

The residuals of the model can be considered an estimate of the population error term.

3 Philosophy & Statistics

There are a number of philosophical issues underpinning many of the commonly-used statistical methods. The reliability of any analysis presented in a scientific paper depends on a set of conceptual issues. There exist a misconception that meta-analysis, *the combination of the results of several studies*, can create a more precise estimate of an effect. However, few meta-analyses meet all criteria for correctness. Other conceptual issues in statitics are the issues of response bias in random surveys. The set of participants that respond to a survey is most likely not a random sample and might be correlated with other variables resulting in a weaker statistical analysis. As mentionned earlier, the estimated

relations lead to the question of causality but its existence is, in many cases, not guaranteed.