American Journal of Epidemiology **Submitted Manuscript**

Special Collection: None

Title: Inferential Statistics and Direct versus Inverse Problems

Authors: Ashley I. Naimi, PhD 1 and Brian W. Whitcomb, PhD 2

ORCiD IDs: Brian W Whitcomb https://orcid.org/0000-0002-8646-5823 Ashley I Naimi https://orcid.org/0000-0002-1510-8175

Correspondence Address: Department of Epidemiology Rollins School of Public Health Emory University 1518 Clifton Road Atlanta, GA 30322 ashley.naimi@emory.edu

Joint Authorship: Click or tap here to enter text.

Affiliations: 1 Department of Epidemiology, Emory University. 2 Department of Biostatistics and Epidemiology, University of Massachusetts at Amherst.

Key words: Statistical Inference; Propositional Logic; Inverse Problems; Direct Problems; Epidemiologic Methods

Acknowledgments: We thank Dr Stephen R Cole for comments on an initial draft of this manuscript.

Funding: AIN was funded by R01HD102313 and R01HL174652

Conflict of Interest: None

Disclaimer: None

Data Availability Statement: No data

AJE Classroom

Inferential Statistics and Direct versus Inverse Problems

Ashley I. Naimi, PhD[1*]
Brian W. Whitcomb, PhD[2]

1 Department of Epidemiology, Emory University.
2 Department of Biostatistics and Epidemiology, University of Massachusetts at Amherst.

* Correspondence: Department of Epidemiology
Rollins School of Public Health
Emory University
1518 Clifton Road
Atlanta, GA 30322
ashley.naimi@emory.edu

Acknowledgements:

Sources of Funding:

Target Journal: AJE
Text word count: 1,500 / 1,500
Number of Figures: 1
Number of Tables: 0
Number of References: 5
Running head: Direct and Inverse Problems

Statistical methods are fundamental to science. However, scientists routinely misinterpret p values, confidence intervals, and other statistical metrics. This partly results from a lack of clarity around core concepts in statistical reasoning. These include ideas about the structure of scientific arguments, as well as the assumptions involved in constructing statistical measures [1].

For many fields, there is an important distinction between problems that can be classified as "direct" or "inverse." These problems relate to the foundations of statistical inference. Here, we explain the structure of direct and inverse problems, connect them to inductive and deductive reasoning, and comment on how understanding these issues can bring clarity to the interpretation of statistical results.

Consider the following scenarios:

Scenario 1

A research team is given exactly 500,000 red and 500,000 blue marbles, uniformly mixed in a container. The researchers know that the probability of drawing a red marble is 50%. The team is asked "in a random sample of 4 marbles from the container, what's the probability the sample will include exactly 2 red marbles?" If we let Y denote the number of red marbles, N denote the marbles sampled from the container, and θ denote the proportion of red and blue marbles in the container, we might frame this as a question about: $P(Y = 2 \mid \theta = 0.5, N = 4)$ or more generally, $P(\text{Data} \mid \theta = 0.5)$.

Based on available information, one can document all possible color combinations in samples of 4 from the container, including samples with no red marbles, with one, with two, etc.

Figure 1 shows all 16 possible samples of 4 marbles. Because the underlying distribution of red:blue marbles is 50:50, the probability of drawing a sample of four with two red marbles can be obtained by dividing the number of samples with two red marble (6 samples) by the total

number of possible samples (16 total samples), yielding 6/16 = 0.375. Alternatively, we can use binomial probability distribution as:

$$P(Y = 2 \mid \theta, N = 4) = \binom{4}{2}\theta^2(1-\theta)^{(4-2)}$$

where $\binom{4}{2} = \frac{4!}{2!(4-2)!} = 6$ and where $\theta = 0.5$.

This first scenario is an example of a direct problem. Direct problems are those whose answers are based on information that is known. Here, we started with the binomial distribution with $\theta = 0.5$, and determined what might be observed under this model.

Direct problems can be framed as deductive arguments, where conclusions can be deduced from statements that comprise the premise of the argument. For our direct problem, we can write:

$$\frac{\text{If } A \text{ then } B \qquad A}{\therefore B}$$

Which, in the context of our scenario 1, can be interpreted as:

(if A then B): if the true model is binomial with θ = 0.5, then the probability of sampling two blue and two red marbles is 0.375.

(A): because the container contains exactly 500,000 red and 500,000 blue marbles, the true probabilty model is binomial with θ = 0.5.

(∴ therefore): the probability of sampling two blue and two red marbles is 0.375.

This syllogism is referred to as “affirming the antecedent” or “modus ponens” and is a valid logical argument.

Scenario 2

In contrast to scenario 1, suppose the researchers are simply given a container with a very large number of red and blue marbles of unknown proportions. The team is asked, "If you draw a simple random sample of four marbles and observe two red and two blue marbles, what's the proportion of red and blue marbles in the population (container)?" This question is distinct from scenario 1 because the team must use sampled observations to infer properties of the unknown probability distribution. This "inverse problem" is considerably more difficult to answer, and is the type of problem upon which modern statistics is built [2].

Formalizing the problem slightly, we may ask, "if a simple random sample of two red and two blue marbles is drawn, what's the probability that the proportion of red and blue marbles in the container is 50:50?" Notationally, we can write this inverse problem as asking about $P(\theta = 0.5 \mid Y = 2, N = 4)$ or more informally as $P(\theta = 0.5 \mid \text{Data})$. Though similar to the direct problem above, the conditioning statements are inverted.

Inverse problems can be framed as inductive arguments, where the flow of the argument's information occurs from the specific observations to properties of the unknown data generating mechanism. This is the opposite direction (or inverse) of direct problems, where the flow occurs from properties of the known data generating mechanism to specific observations.

Deductive and Inductive Inference

Understanding distinctions between the strengths and weakness of inductive versus deductive methods can help frame difficulties in solving inverse problems. First, deduction is subject to errors if the premises are wrong, and it's often not possible to establish the truth of the premises. Second, even if the premises are true, deduction can only transform the information contained within them, and thus cannot generate information not already contained within the

premises. However, valid deductive arguments based on true premises are sufficient to *guarantee* the truth of the conclusions. This feature is a primary strength of deduction.

In contrast, inductive reasoning uses observations (e.g., marble colors in the sample) as evidence for general statements about the world (e.g., the unknown ratio of red:blue marbles in the container). Because scientific induction is based on observations or measurements, it is possible to generate new information about phenomena under study. This is one reason inductive inference is central to the empirical sciences. However, using specific observations to reason about underlying mechanisms or characteristics famously results in the "problem of induction": though a given observation (e.g., a sample of two red and two blue marbles) may suggest a particular data generating mechanism (the ratio of red:blue marbles in the container is 50:50), inductive arguments cannot guarantee true conclusions [3, p 247].

Framing Inductive Arguments Deductively

Induction can be linked to deductive arguments with a different effect. In the context of scenario 2, one might state:

(if A then B): if the true model is binomial with $\theta = 0.5$, then we are most likely to observe a sample ratio of red:blue marbles of 50:50.

(B): the sample ratio is, in fact, 50:50.

($\therefore$ therefore A): the true model is binomial with $\theta = 0.5$.

This syllogism, referred to as "affirming the consequent", is a formal logical fallacy, and occurs frequently in the empirical sciences. Concluding that the true probabilty $\theta = 0.5$ may well be true and entirely reasonable, but the structure of this argument is not a sufficient basis to justify its truth. Reasons other than a $\theta = 0.5$ may lead to a sample ratio of 50:50.

Inverse Problems and Statistics

The consequences of direct/inverse problems lie at the heart of many statistical misinterpretations. Say we conduct a randomized trial in 180 patients to evaluate whether a new drug reduces the risk of illness. We find 18 of the 84 patients assigned to treatment and 32 of the 96 patients assigned to placebo experience illness. This yields an estimated risk difference of -0.12, but we are primarily interested in the true (not estimated) risk difference. This is an inverse problem.

We can compute a measure of how compatible our estimated risk difference is with some assumed true risk difference using the p value [4]. If we assume that the true risk difference is zero (the test hypothesis), the probability of observing an estimated risk difference greater than |-0.12| is 0.04 (two-sided p value). However, because this is an inverse problem, and our interest lies in the true risk difference, we might be inclined to interpret this value of 0.04 as a measure of $P(\theta = 0 \mid \text{Data})$. Indeed, many people do [1]. But the p value measures the probability of the data given an assumed test hypothesis, or $P(\text{Data} \mid \theta)$, not the other way around.

Alternatively, we may be inclined to reason that, if the true risk difference is, say, 0.15, then we'd expect to see a large p value when we compute it using a test hypothesis value of 0.15. In the sample from our randomized trial, the two-sided p value under a test hypothesis of 0.15 is 0.64. However, if we then conclude that the true risk difference is 0.15, we would be affirming the consequent, and thus committing a logical fallacy.

Inferential statistics are notoriously difficult to interpret. Part of the challenge lies in the structure of direct versus inverse problems. Understanding the distinctions between these types of problems can bring some clarity to the interpretation of statistical measures. Notably, by helping us avoid logical fallacies such as affirming the consequent, or naively inverting conditional probabilities. Our goal here was to highlight the difference between direct and

inverse problems, and convey why the structure of inverse problems make certain statistical summaries difficult to interpret. However, direct and inverse problems are foundational, and underlie a number of related topics we could not discuss here. These include [5]: probabilistic induction and the associated formalizations of evidence for and against hypotheses; Fiducial and Bayesian approaches to solving inverse problems; and logical strategies (e.g., falsification, modus tollens) and their relationship to Neyman-Pearson testing and (neo)-Fisherian statistics.

References

1. Greenland, S., et al., *Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations.* Eur J Epidemiol, 2016. **31**(4): p. 337-50.
2. Stark, P.B., *Inverse Problems as Statistics*, in *Surveys on Solution Methods for Inverse Problems*, D. Colton, et al., Editors. 2000, Springer Vienna: Vienna. p. 253-275.
3. Hacking, I., *An Introduction to Probability and Inductive Logic*. 2001: Cambridge University Press.
4. Greenland, S., *Divergence versus decision P-values: A distinction worth making in theory and keeping in practice: Or, how divergence P-values measure evidence even when decision P-values do not.* Scandinavian Journal of Statistics, 2023. **50**(1): p. 54-88.
5. Greenland, S., *Probability logic and probabilistic induction.* Epidemiology, 1998. **9**(3): p. 322-32.

Figure Legends

Figure 1. Tabulation of all possible scenarios in which four marbles are drawn from a container containing red and blue marbles.