Contents

16

Wrap-up

The final lecture of the class had several of the students presenting their final projects. These were about understanding or tuning household tasks like cuisine or entertaining young children or morning wakeup rituals or hobbies such as horticulture. It was very nice to see a range of design ideas. From a survey of the class it seemed that fractional factorials and analysis of covariance ideas turned out to be most widely used.

Prior to those examples was a short note summarizing the topics of the course. That was preceded by a brief overview of the statistics problem in general.

16.1 What statistics is about

At a very high level view, our primary challenge in statistics is to say something about numbers we don't have using numbers we do have. In prediction settings we want to know about future values of Y for some $\boldsymbol{x}$, using past $(\boldsymbol{x}_i, Y_i)$ data. Other settings manifest differently but we are still using known values to say something about unknowns. Phrased this way, our task seems at first like it might be impossible.

We connect our knowns to unknowns by choosing a model that we can think of as generating both kinds of data as depicted in the left panel of Figure 16.1. The issue of ***external validity*** that we frequently raised involves the model not changing between those two settings. ***Internal validity*** is then about the model holding for the data we do have. In statistical inference, we reverse the direction of one of the arrows, letting us learn something about the model from the known data. This is a problem of ***inductive inference*** describing the general model from the specific known data. Using what we know about the

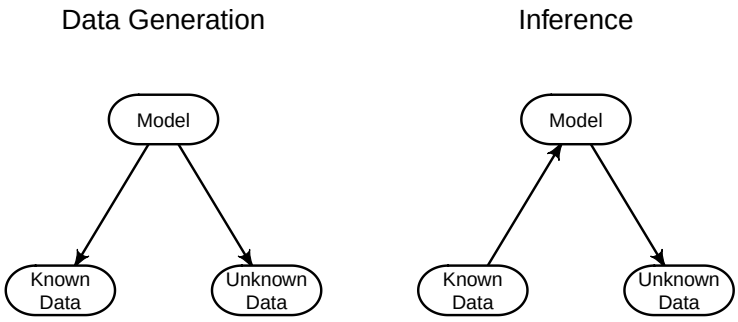

Figure 16.1: The left figure shows how we envision a one and the same statistical model produces both our known data and some unknown data of interest. In inference we reverse the arrow from the model to the known data. Then the known data tell us something about the model (with some quantified uncertainty). From that we can derive consequences about the unknown data.

model, ***deductive inference*** lets us derive consequences for the unknown data. Induction leaves us with some uncertainty about the model. When we derive something about the unknown data we can propagate that uncertainty.

One could argue that it is logically impossible to learn the general case from specific observations. For a survey of the problem of induction in philosophy, see Henderson (2018). We do it anyway, accepting that the certainty possible in mathematics may not be available in other settings.

A famous observation from George Box is that all models are wrong, but some are useful. Nobody, not even Box, could give us a list of which one is useful when. As applied statisticians, it is our responsibility to do that in any case that we handle. There are settings where we believe that we can get a usable answer from an incorrect answer. Sometimes we know that small errors in the model will yield only small errors in our inferences. This is a notion of ***robustness***. In other settings we can get consequences from our model that can be tested later in new settings. Then, even if the model had errors we can get a measure of how well it performs.

There are approaches to inference that de-emphasize or perhaps remove the model. We can imagine the path being like the right hand side of Figure 16.2 that avoids the model entirely which unlocks more comutational possibilities. There may well be an implicit model there, such as unknown $(\boldsymbol{x}, Y)$ pairs being IID from the same distribution that produced the known ones. IID sampling would provide a justification for using cross-validation or the bootstrap. For a discussion of the role of models in statistics and whether they are really necessary, see Breiman (2001).

The setting we considered in this course relates to the usual inference prob-

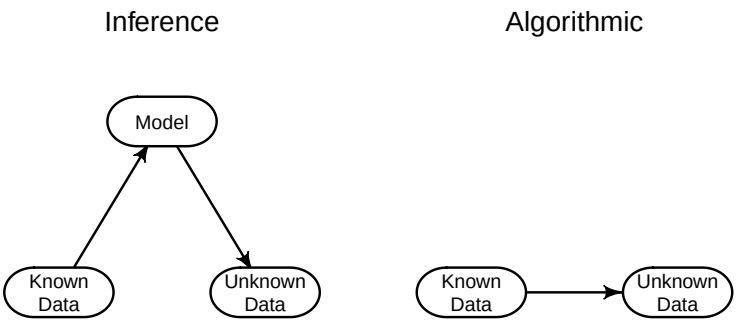

Figure 16.2: Sketch of an algorithmic approach to learning about unknown data from known data. There is only a remnant of a model.

lem as shown in Figure 16.3. We were down in the lower left hand corner looking at how to make the data that would then be fed into the inferential machinery.

16.2 Principals from experimental design

Statistical inference from data faces several obstacles:

1) cost of data,
2) confounding of causes,
3) correlation of predictors,
4) noise,
5) interactions,
6) missing variables, and
7) external validity.

In the face of these obstacles, experimental design offers the following techniques:

a) Randomization,
b) Blocking and balancing,
c) Factorials,
d) Fractional replication,
e) Covariate adjustments (ancova),
f) Adaptation (bandits and sequential DOE),
g) Nesting and split-plots,
h) Random effects models, and
i) Replication.

Obstacle 1, the cost of data, is often forgotten in data analysis because, once the data are available that cost is not pertinent to its analysis. It is a sunk cost. Cost is important in experimental design and many of the designs we saw were chosen to reduce that cost. For instance in fractional factorial experiments, we would purposely sacrifice statistical correctness (e.g., an unbiased variance

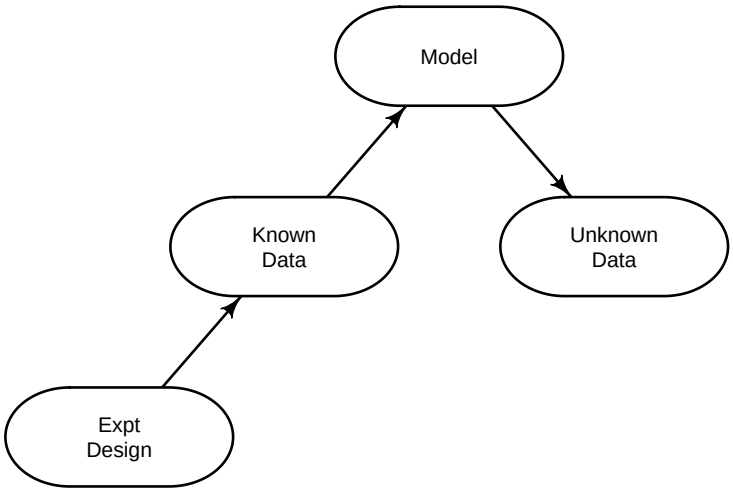

Figure 16.3: The place held by experimental design in statistical inference.

estimate) in order to study more variables at a fixed cost. Nesting and split-cost designs take advantage of the fact that some experimental factors are cheaper to change than others. Adaptive sampling via bandits is cost driven. It can reduce the cost incurred on the experimental units themselves. Sequential experiments also counter the cost of continuing to experiment once the better treatment has been clearly identified.

Confounding of potential causes was one of the primary motivations for the use of randomization. Randomization reduces the risk that some important cause other than the treatment is perfectly or even strongly associated with the treatment. Missing or unmeasured variables interfere with causal claims. We can think of them as continuously varying quantities that cause similar problems to confounding. Randomization ensures that those missing variables cannot be strongly associated with the treatment.

Correlated predictors can be problematic in regression settings because they make it harder to tell which variable is important. Many of the designs we studied produced perfectly uncorrelated predictors.

Regression methods average away the noise. In optimal designs we found ways to minimize the effect of noise on the variance of regression coefficients. Replication raises the sample size and thus helps us reduce noise. We also used blocking methods to get better comparisons of "like with like" by arranging that treatment and control would both be applied to similar experimental units, with similarity defined by one or more categorical variables. The analysis of

covariance was useful to balance out impacts of continuous variables.

Interactions severely complicate interpretation of the effect of variables. We looked at factorial designs that allowed us to estimate those interaction effects.

External validity is critical problem for causal inferences. Having an experiment contain a wide variety of settings helps to improve its external validity. Random effect models also provide greater external validity.

Bibliography

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231.

Henderson, L. (2018). The problem of induction. In *The Stanford Encyclopedia of Philosophy*. `https://plato.stanford.edu/entries/induction-problem/`.