MS&E 226: Fundamentals of Data Science

Lecture 8: Introduction to inference

Ramesh Johari
ramesh.johari@stanford.edu

What is inference?

Where did our data come from?

Recall our *sample* is:

- $\mathbf{Y}$, the vector of n observed outcomes
- $\mathbf{X}$, the corresponding matrix of covariates: n rows, with p covariates in each row

What process generated $\mathbf{X}$ *and* $\mathbf{Y}$*?*

Population model

Recall that the *population model* is a probabilistic representation of the process that generated $\mathbf{X}$ and $\mathbf{Y}$.

We represent this through a *joint distribution* of $\mathbf{X}$ and $\mathbf{Y}$ (or equivalently, a distribution for $\mathbf{X}$, and then a conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$).

Prediction

Our focus thus far has been on *prediction*:

Given a dataset $\mathbf{X}$ and $\mathbf{Y}$, and a new covariate vector $\vec{X}$, use the data to build the best model you can of the corresponding new observation Y.

Recall that we could do fairly well at this, even *without* having properly understood the population model. (*Examples*: k-nearest-neighbors, collinear data, etc.)

The basic reason is that correlations in the data are useful for predictions, even if they don't provide the correct explanation of a phenomenon.

Inference: "Opening the box"

We can think of prediction as *black box* modeling:

It doesn't matter if we properly describe the population model, as long as we can make good predictions.

By contrast, *inference* refers to "opening the black box", and trying to understand and explain the population model itself.

Inference: Why should we care?

Why do we care about inference?

- *Prediction*. Ostensibly, good understanding of the population model should let us make better predictions. Of course, as we just discussed, we can often make great predictions *without* understanding the population model...

Inference: Why should we care?

Why do we care about inference?

- *Prediction*. Ostensibly, good understanding of the population model should let us make better predictions. Of course, as we just discussed, we can often make great predictions *without* understanding the population model...
- *Decisions*. Often it is important to understand the population model correctly, because it guides other things we do: experiments we try, decisions we make, etc. This is why we so often try to *interpret* the population model.

Inference: Why should we care?

Why do we care about inference?

- *Prediction*. Ostensibly, good understanding of the population model should let us make better predictions. Of course, as we just discussed, we can often make great predictions *without* understanding the population model...
- *Decisions*. Often it is important to understand the population model correctly, because it guides other things we do: experiments we try, decisions we make, etc. This is why we so often try to *interpret* the population model.
- *Causality*. Ultimately, inference is the basis for extracting "if-then" relationships; these are the insights that form the foundation of data-driven decision-making.

A formal statement of the inference problem

Given: Sample ("training data") $\mathbf{X}$, $\mathbf{Y}$.

The goal of inference is to produce an estimate of the population model that generated the data. In other words, *estimate the joint distribution of $\vec{X}$'s and Y's in the population.*

Note that this can be broken into two pieces:

- Estimate the distribution of the $\vec{X}$'s.
- Estimate the conditional distribution of Y given $\vec{X}$.

Overview of inference

What kinds of questions does inference try to answer?

Inference is principally concerned with two (closely related) goals:

- *Estimation*. The first goal is to develop an estimate of the population model itself: what is our best guess for the process that generated the data?

What kinds of questions does inference try to answer?

Inference is principally concerned with two (closely related) goals:

- *Estimation*. The first goal is to develop an estimate of the population model itself: what is our best guess for the process that generated the data?
- *Quantifying uncertainty*. No matter what our guess is, we can't be sure we are right. How do we quantify the uncertainty around our guess?

An example: Biased coin flipping

Suppose I flip a coin five times.

The flips lead to H, H, T, H, T.

- *Estimation*: What is your best guess of the bias of the coin?
- *Quantifying uncertainty*: How sure are you of your answer?

Important dichotomies

We start with two important dichotomies:

- Nonparametric vs. parametric inference
- Frequentist vs. Bayesian inference

Nonparametric vs. parametric

In *nonparametric* inference, we make no assumptions about the nature of the distribution that generated our data.

In *parametric* inference, we assume we know the “shape” of the distribution that generated our data, but don’t know parameters that determine the exact distribution.

This is really a *false dichotomy*: parametric and nonparametric approaches live on a “sliding” scale of modeling complexity, and there are close relationships between them.

Nonparametric vs. parametric

Nonparametric inference is appealing, and increasingly feasible (as computational power and data availability increases).

Why make assumptions if you don't have to? Parametric inference:

- Can be simpler (more parsimonious).
- Can be easier to interpret, especially relationships between covariates and the outcome.
- Can be less prone to overfitting (lower variance).
- Requires less data to estimate, and therefore can be more robust to model misspecification.

Ultimately both are valuable approaches. We focus primarily in parametric inference in this class (with linear regression as a primary example).

Frequentist vs. Bayesian

The other important dichotomy in approaches to inference is between *frequentist* and *Bayesian* inference.

Frequentist vs. Bayesian

The other important dichotomy in approaches to inference is between *frequentist* and *Bayesian* inference.

- Frequentists treat the *parameters* as fixed (deterministic).
    - Considers the training data to be a random draw from the population model.
    - Uncertainty in estimates is quantified based on their performance if they are repeated over and over again, over many sets of training data ("parallel universes").

Frequentist vs. Bayesian

The other important dichotomy in approaches to inference is between *frequentist* and *Bayesian* inference.

- Frequentists treat the *parameters* as fixed (deterministic).
    - Considers the training data to be a random draw from the population model.
    - Uncertainty in estimates is quantified based on their performance if they are repeated over and over again, over many sets of training data ("parallel universes").
- Bayesians treat the *parameters* as random.
    - Key element is a *prior* distribution on the parameters.
    - Using Bayes' theorem, combine prior with data to obtain a *posterior* distribution on the parameters.
    - Uncertainty in estimates is quantified through the posterior distribution.

Frequentist vs. Bayesian

We will have much more to say about the comparison between frequentist and Bayesian methods in later lectures.

Despite the intense debates this dichotomy provokes, both viewpoints are quite relevant in the modern data scientist's toolbox.

We start with frequentist inference, and will compare to Bayesian methods later.

Our goal in the next few lectures

In the next few lectures we:

- Develop the maximum likelihood approach to parameter estimation
- Quantify the uncertainty in estimates derived using maximum likelihood, using standard errors, confidence intervals, and hypothesis tests
- Apply these techniques to linear regression, and to logistic regression (a method used for binary data)