MS&E 226: Fundamentals of Data Science

Lecture 8: Introduction to frequentist statistical inference

Ramesh Johari

From prediction to inference

Recall: Prediction

In the first unit of the course, we focused 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 to predict the corresponding new outcome Y.

We can think of prediction as *black box* modeling: it doesn't matter if we "understand" the population model, as long as we can make good predictions.

Example: Housing prices

Recall our housing price example. Using the Saratoga Housing dataset, suppose we fit an OLS linear regression:

$$\texttt{price} \approx 13{,}439.394 + 113.123 \times \texttt{livingArea}$$

This fitted model allows us to predict the price of a new house given its living area.

But this model also raises questions that go beyond prediction...

Questions beyond prediction

The coefficient on `livingArea` reveals a positive association with `price`. How should we interpret this?

Given this fitted model, we might ask:

- Could the true association between `livingArea` and `price` be larger or smaller, or even nonexistent?
- How confident are we in the quantitative relationship between `livingArea` and `price`?

Questions like these are the domain of *statistical inference*.

Statistical inference

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

In other words: Inference tries to understand and quantify *which* relationships between covariates and outcome are actually present in the population model.

Formally: Statistical inference focuses on *learning the structure of the population model itself*, rather than only making predictions.

Why should we care?

Why do we care about inference, beyond prediction?

- *Interpretation*. We often want to understand what the fitted model tells us about the population (e.g., the association of `livingArea` with `price`).

Why should we care?

Why do we care about inference, beyond prediction?

- *Interpretation*. We often want to understand what the fitted model tells us about the population (e.g., the association of `livingArea` with `price`).
- *Causality*. Ultimately, inference is the basis for extracting "if-then" relationships. E.g.: If I increase living area, how much will the house price increase?

Why should we care?

Why do we care about inference, beyond prediction?

- *Interpretation*. We often want to understand what the fitted model tells us about the population (e.g., the association of `livingArea` with `price`).
- *Causality*. Ultimately, inference is the basis for extracting "if-then" relationships. E.g.: If I increase living area, how much will the house price increase?
- *Decisions*. Understanding the population model correctly guides actions we take, experiments we try, etc.

The two goals of inference

Inference is principally concerned with two closely related goals:

- *Estimation*. What is our best guess for the process that generated the data? I.e., what is our best guess for the population model?

The two goals of inference

Inference is principally concerned with two closely related goals:

- *Estimation*. What is our best guess for the process that generated the data? I.e., what is our best guess for the population model?
- *Quantifying uncertainty*. How do we quantify our uncertainty in the guess we made?

Parametric inference

Nonparametric vs. parametric inference

There are two broad approaches to inference:

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.

Why parametric inference?

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 on parametric inference in this class (with linear regression as a primary example).

Parametric inference: An example

We've actually already seen an example of parametric inference:

Suppose our *population* is all valid flights in 2024, i.e., flights with a recorded arrival delay; recall we denoted the population mean by μ, and the population variance by σ^2.

Suppose we get a *sample* of $n = 10$ data points $Y_1, \ldots, Y_n$ independently drawn from this data.

Questions: How would we estimate μ? How confident are we in our estimate?

Parametric estimation

A formal description of estimation:

- The *population* model is characterized by certain *parameters* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ (e.g., μ and σ^2 in the flights example).
- We obtain a data *sample* $\mathscr{D}$ (e.g., $\mathscr{D} = \mathbf{Y}$ in the flights example, or $\mathscr{D} = (\mathbf{X}, \mathbf{Y})$ when there are also features available), consisting of independent random samples from the population.
- We use $\mathscr{D}$ to form an *estimator* $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathscr{D})$ of $\boldsymbol{\theta}$.

Since we have a particular data sample, we obtain a specific numerical estimate of each parameter using this procedure.

Parametric estimation: An example

Suppose we get a *sample* of $n = 10$ data points $Y_1, \ldots, Y_n$ independently drawn from this data.

A natural estimator of the population mean μ is the *sample mean* $\overline{Y} = (1/n) \sum_{i=1}^{n} Y_i$.

(Clearly the estimator depends on the sample.)

Parametric inference: A frequentist view

How do we quantify our uncertainty in $\hat{\boldsymbol{\theta}}$? In other words: *How sure are we of our estimate?*

The frequentist view is that our uncertainty in the estimate is inherently due to *uncertainty in the data we used to form the estimate*.

In other words, the frequentist:

- Believes the true parameters $\boldsymbol{\theta}$ are *fixed* (not random!)
- The *only* randomness is in the data sample from the population ("parallel universes").
- We get only one such data sample (one "universe"), and have to use it to reason about the true parameters $\boldsymbol{\theta}$.

Parametric inference: The sampling distribution

Thus the frequentist quantifies estimation uncertainty using sampling distributions:

Rewind time and imagine resampling data, again and again; in each "parallel universe":

- Random sampling from the population yields a dataset $\mathscr{D}$; and
- Using $\mathscr{D}$, we form an estimate $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathscr{D})$.

The resulting distribution of $\hat{\boldsymbol{\theta}}$ is the *sampling distribution of the estimator*.

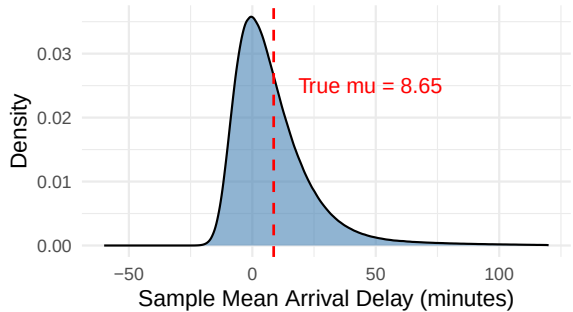
Connecting to prediction

Previously, we used sampling distributions to understand the bias and variance of *predictions*.

We are now using sampling distributions to understand the uncertainty in *parameter estimates*.

Sampling distribution of an estimator: An example

In fact, we have already previously constructed an approximate sampling distribution of the sample mean for the flights data, over 10,000,000 parallel universes:



This distribution tells us how much our sample mean varies across "parallel" universes.

Bias and variance of an estimator

Analogous to our discussion of learning algorithms, we can define the bias and variance of an estimator.

For now let's assume a single real-valued parameter θ (e.g., population mean μ in the flights example). Then for an estimator $\hat{\theta}$ of θ:

- *Bias* = $\mathbb{E}_{\mathscr{D}}[\hat{\theta}] - \theta$, i.e., the difference between the mean of the sampling distribution and the true parameter.
- *Variance* = $\mathrm{Var}_{\mathscr{D}}(\hat{\theta})$, i.e., the variance of the sampling distribution.
- *Standard error* = SE = $\sqrt{\mathrm{Var}_{\mathscr{D}}(\hat{\theta})}$, i.e., the standard deviation of the sampling distribution.

(Here the expectation and variance are over the randomness in the data sample.)

Mean squared error (MSE)

Just as for learning algorithms, a natural measure of the "quality" of an estimator is its *mean squared error* (MSE) against the true parameter:

$$\text{MSE} = \mathbb{E}_{\mathscr{D}}[(\hat{\theta} - \theta)^2].$$

This measures the average squared distance between the estimator and the true parameter, across parallel universes.

A bias-variance decomposition also holds for MSE of estimators:

$$\text{MSE} = \text{Bias}^2 + \text{Variance}.$$

Comparison to prediction

Similarities and differences between prediction and estimation:

For a *prediction* $\hat{f}(\vec{X})$:

- We care about the Average MSE at $\vec{X} = \mathbb{E}_{Y,\mathbf{X},\mathbf{Y}}[(\hat{f}(\vec{X}) - Y)^2 | \vec{X}]$.
- This decomposes as: Irreducible error $\sigma^2(\vec{X})$ + Bias^2 + Variance.

For an *estimator* $\hat{\theta}$:

- We care about the MSE: $\mathbb{E}_{\mathscr{D}}[(\hat{\theta} - \theta)^2]$.
- This decomposes as: Bias^2 + Variance.

The key difference: There is *no irreducible error* for estimators, because the parameter θ is fixed (not random)!

Two approaches to quantifying uncertainty

We will discuss two broad ways of quantifying uncertainty, both based on the sampling distribution:

1. Using the sampling distribution (e.g., its standard error) to construct *confidence intervals* for the true parameter; wider intervals represent greater uncertainty.
2. Hypothesize a particular value of the parameter of interest (e.g., "μ is zero"), then test this hypothesis using the data you observed: If the truth had been $\mu = 0$, how likely is the data that you observed? (The answer to this question is a *p-value*.)

The challenge

But first: as we've discussed, there's a problem with using the sampling distribution:

We can't actually compute the sampling distribution!

Why not?

- To compute the sampling distribution, we would need to resample from the population model.
- But we don't know the true parameters $\boldsymbol{\theta}$ that define the population model.
- We only have one dataset $\mathscr{D}$.

So how is the sampling distribution still useful?

Overcoming the challenge

Despite the challenge, there are two paths forward to working with the sampling distribution:

1. *Theory*: Under certain assumptions, we can theoretically characterize the sampling distribution of an estimator. We can use these characterizations to quantify uncertainty.

2. *Simulation*: We can use our one data sample itself to approximate the sampling distribution, by resampling from the data (the *bootstrap* procedure).

We will explore both approaches in this unit.