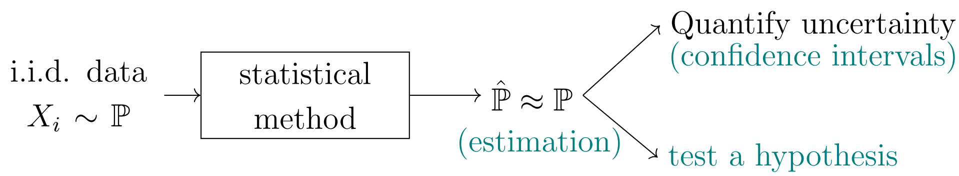
18.650: Fundamentals of Statistics	February 17, 2026

Lecture 7 — Models and point estimation

Mohammad Reza Karimi

Overview

Let's add to the statistics pipeline from Lecture 1:



"Estimation" refers to estimating the unknown distribution $\mathbb{P}$ from the data, and is the focus of this lecture. In later lectures, we will learn about two other important downstream tasks: quantifying uncertainty by constructing a confidence interval, and testing hypotheses.

Without knowing anything about $\mathbb{P}$ except that it's a probability distribution, it's too difficult to estimate it. We must therefore first specify a *model* for the data, based on our apriori knowledge.

1 Models

Definition 1.1: Statistical model

A model is a set of probability distributions, which is typically a strict subset of the set of *all* probability distributions.

There are many ways to specify a model.

Example.

- using a common family of distributions, e.g.,
  - $\{\mathrm{Ber}(p) : p \in (0,1)\}$.
  - $\{\mathrm{Exp}(\lambda) : \lambda \in [0, 22]\}$
- in terms of pdfs/pmfs, e.g., $\left\{ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : \mu \in \mathbb{R},\, \sigma^2 > 0 \right\}$.

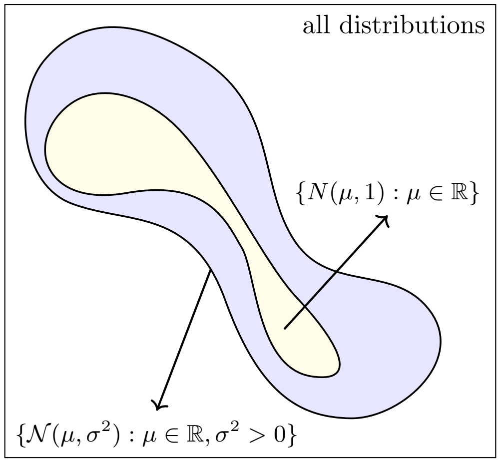


Figure 1: By imposing a model, we restrict our study to a certain subset of the set of all possible probability distributions. For example, we might consider all normal distributions, or all normal distributions with variance 1.

- in terms of CDFs, e.g., $\{F(x) : F \text{ is a continuous CDF}\}$, which rules out CDFs that have jumps.

To make statements about general statistical models, we will refer to the model as

$$\{\mathbb{P}_\theta : \theta \in \Theta\}.$$

Here, θ is the **parameter** and Θ is the **parameter space** that θ lives in.

Definition 1.2: Parametric vs nonparametric model

If Θ has finite dimension, we call it a parametric model. If Θ has infinite dimension, we call it a nonparametric model.

Remark.

Note that in the nonparametric case, there is still a "parameter", it's just infinite-dimensional. Usually, the parameter space is some sort of function class.

Example.

1. $\{\text{Exp}(\lambda) : \lambda \in (0, \infty)\}$ is parametric
2. $\{\text{pdf } f \text{ is a polynomial}\}$ is nonparametric.
3. $\{\text{pdf } f \text{ is a polynomial with degree at most } d\}$ is parametric.

We'll focus mostly on parametric statistics in this class.

2 Point estimation

Before discussing point estimation, a bit of **notation**: for a model $\{\mathbb{P}_\theta : \theta \in \Theta\}$, we indicate that some statistic is evaluated under the distribution $\mathbb{P}_\theta$ with a subscript θ, e.g.,

$$\mathbb{P}_\theta(X \geq 1), \qquad \mathbb{E}_\theta[X], \qquad \mathbb{V}_\theta[X].$$

So for a normal family, we would write $\mathbb{P}_{\mu,\sigma^2}(X \geq 1)$ or $\mathbb{E}_{\mu,\sigma^2}[X]$ (which equals μ).

A **point estimate** $\hat{\theta}$ or $\hat{\theta}_n$ (to emphasize its dependence on n data points) is a single guess for a parameter θ.

We'll use the notation $\theta \rightsquigarrow \hat{\theta}$ to say "θ is estimated by $\hat{\theta}$. " For example, $\mu \rightsquigarrow \bar{X}_n$. (Recall that $\rightsquigarrow$ also stands for weak convergence, but the difference between the two meanings will be clear in context.)

Definition 2.1: Estimator

An estimator $\hat{\theta}$ is a function of the data: $\hat{\theta} = g(X_1, \ldots, X_n)$.

Remark.

> An estimator is a random variable. Indeed, a function of random variables $X_1, \ldots, X_n$ is itself a random variable.

Example.

> $\bar{X}_n$, $\max(X_1, \ldots, X_n)$, $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ can all be used as estimators.
>
> X_1 and 4 are also valid estimators.

It's clear that X_1 and 4 are "bad" estimators (the former throws out $n-1$ data points, while the latter doesn't look at the data at all). How do we formalize this?

2.1 Bias, standard error, and MSE

We'll consider two properties of an estimator: its **bias** and its **standard error**.

Definition 2.2: Bias

The bias of an estimator $\hat{\theta}$ of θ is defined as

$$\text{bias}(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta.$$

We say $\hat{\theta}$ is *unbiased* if $\text{bias}(\hat{\theta}) = 0$. We say $\hat{\theta}_n$ is *asymptotically unbiased* if $\text{bias}(\hat{\theta}_n) \to 0$ as $n \to \infty$.

Example.

> Consider the model $\{\mathcal{N}(\mu, 1) \,:\, \mu \in \mathbb{R}\}$. In other words, $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ for an unknown μ. Consider the following three estimators of μ.
>
> 1. $\hat{\mu}_1 = \bar{X}_n$. $\text{bias}(\hat{\mu}_1) = \mathbb{E}_\mu[\bar{X}_n] - \mu = 0$. Unbiased.
> 2. $\hat{\mu}_2 = X_1$. $\text{bias}(\hat{\mu}_2) = \mathbb{E}_\mu[X_1] - \mu = 0$. Unbiased.
> 3. $\hat{\mu}_3 = 0$. $\text{bias}(\hat{\mu}_3) = \mathbb{E}_\mu[0] - \mu = -\mu$. Biased unless $\mu = 0$.

The second property of an estimator is how much it fluctuates, measured by its variance.

Definition 2.3: Standard error (se)

The standard error of an estimator $\hat{\theta}$ is

$$\text{se}(\hat{\theta}) = \sqrt{\mathbb{V}[\hat{\theta}]}.$$

In other words, the standard error of $\hat{\theta}$ equals its standard deviation.

Example.

> Consider the model $\{\mathcal{N}(\mu, 1) \,:\, \mu \in \mathbb{R}\}$, i.e., $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$.
>
> 1. $\hat{\mu}_1 = \bar{X}_n$. $\text{se}(\hat{\mu}_1) = 1/\sqrt{n}$.
> 2. $\hat{\mu}_2 = X_1$. $\text{se}(\hat{\mu}_2) = 1$.
> 3. $\hat{\mu}_3 = 0$. $\text{se}(\hat{\mu}_3) = 0$

It turns out that there is a third quantity which simultaneously captures both the bias and the standard error:

Definition 2.4: Mean squared error (MSE)

The mean squared error of an estimator $\hat{\theta}$ of θ is

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

We now show

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{bias}^2(\hat{\theta}) + \mathrm{se}^2(\hat{\theta}).$$

Indeed,

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathrm{bias})^2] = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right] + 2\mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right) \cdot \mathrm{bias}\right] + \mathrm{bias}^2 \\ &= \mathbb{V}[\hat{\theta}] + 2\mathbb{E}\left[\hat{\theta} - \mathbb{E}[\hat{\theta}]\right] \cdot \mathrm{bias} + \mathrm{bias}^2 = \mathbb{V}[\hat{\theta}] + \mathrm{bias}^2.\end{aligned}$$

Example.

Consider the model $\{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R}\}$, i.e., $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$.

1. $\hat{\mu}_1 = \bar{X}_n$. $\mathrm{MSE}(\hat{\mu}_1) = 0^2 + \frac{1}{n} = \frac{1}{n}$.
2. $\hat{\mu}_2 = X_1$. $\mathrm{MSE}(\hat{\mu}_2) = 0^2 + 1 = 1$.
3. $\hat{\mu}_3 = 0$. $\mathrm{MSE}(\hat{\mu}_3) = \mu^2 + 0 = \mu^2$. (Zero variance, but possibly large bias!)

2.2 Consistency

Definition 2.5: Consistency

An estimator $\hat{\theta}_n$ of θ is consistent if $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ as $n \to \infty$.

Example.

Consider the model $\{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R}\}$, i.e., $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$.

1. $\hat{\mu}_1 = \bar{X}_n \xrightarrow{\mathbb{P}} \mu$ by the LLN.
2. $\hat{\mu}_2 = X_1$ does *not* converge to μ
3. $\hat{\mu}_3 = 0$ does *not* converge to μ.

The LLN is one way to show consistency. Sometimes, we also need to use the *Continuous Mapping Theorem* (recall from Lecture 3).

Example.

> $\hat{\theta} = (\bar{X}_n)^2$ is a consistent estimator of $\theta = \mu^2$ because (a) $\bar{X}_n \xrightarrow{\mathbb{P}} \mu$ by the LLN, and (b) the function $\mu \mapsto \mu^2$ is continuous, so the Continuous Mapping Theorem applies.

Finally, it turns out that we can also get consistency by showing the MSE goes to zero.

Theorem 2.6

If $\text{MSE}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is consistent.