6

Models, Statistical Inference and Learning

6.1 Introduction

Statistical inference, or "learning" as it is called in computer science, is the process of using data to infer the distribution that generated the data. A typical statistical inference question is:

> Given a sample $X_1, \ldots, X_n \sim F$, how do we infer F?

In some cases, we may want to infer only some feature of F such as its mean.

6.2 Parametric and Nonparametric Models

A **statistical model** $\mathfrak{F}$ is a set of distributions (or densities or regression functions). A **parametric model** is a set $\mathfrak{F}$ that can be parameterized by a finite number of parameters. For example, if we assume that the data come from a Normal distribution, then the model is

$$\mathfrak{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \quad \mu \in \mathbb{R},\ \sigma > 0 \right\}. \tag{6.1}$$

This is a two-parameter model. We have written the density as $f(x; \mu, \sigma)$ to show that x is a value of the random variable whereas μ and σ are parameters.

In general, a parametric model takes the form

$$\mathfrak{F} = \left\{ f(x;\, \theta) :\ \theta \in \Theta \right\} \tag{6.2}$$

where θ is an unknown parameter (or vector of parameters) that can take values in the **parameter space** Θ. If θ is a vector but we are only interested in one component of θ, we call the remaining parameters **nuisance parameters**. A **nonparametric model** is a set $\mathfrak{F}$ that cannot be parameterized by a finite number of parameters. For example, $\mathfrak{F}_{\mathsf{ALL}} = \{\text{all CDF}'s\}$ is nonparametric. [1]

6.1 Example (One-dimensional Parametric Estimation)**.** Let $X_1, \ldots, X_n$ be independent Bernoulli(p) observations. The problem is to estimate the parameter p. ■

6.2 Example (Two-dimensional Parametric Estimation)**.** Suppose that $X_1, \ldots, X_n \sim F$ and we assume that the PDF $f \in \mathfrak{F}$ where $\mathfrak{F}$ is given in (6.1). In this case there are two parameters, μ and σ. The goal is to estimate the parameters from the data. If we are only interested in estimating μ, then μ is the parameter of interest and σ is a nuisance parameter. ■

6.3 Example (Nonparametric estimation of the CDF)**.** Let $X_1, \ldots, X_n$ be independent observations from a CDF F. The problem is to estimate F assuming only that $F \in \mathfrak{F}_{\mathsf{ALL}} = \{\text{all CDF}'s\}$. ■

6.4 Example (Nonparametric density estimation)**.** Let $X_1, \ldots, X_n$ be independent observations from a CDF F and let $f = F'$ be the PDF. Suppose we want to estimate the PDF f. It is not possible to estimate f assuming only that $F \in \mathfrak{F}_{\mathsf{ALL}}$. We need to assume some smoothness on f. For example, we might assume that $f \in \mathfrak{F} = \mathfrak{F}_{\mathsf{DENS}} \bigcap \mathfrak{F}_{\mathsf{SOB}}$ where $\mathfrak{F}_{\mathsf{DENS}}$ is the set of all probability density functions and

$$\mathfrak{F}_{\mathsf{SOB}} = \left\{ f :\ \int (f''(x))^2 dx < \infty \right\}.$$

The class $\mathfrak{F}_{\mathsf{SOB}}$ is called a **Sobolev space**; it is the set of functions that are not "too wiggly." ■

6.5 Example (Nonparametric estimation of functionals)**.** Let $X_1, \ldots, X_n \sim F$. Suppose we want to estimate $\mu = \mathbb{E}(X_1) = \int x\, dF(x)$ assuming only that

[1] The distinction between parametric and nonparametric is more subtle than this but we don't need a rigorous definition for our purposes.

μ exists. The mean μ may be thought of as a function of F: we can write $\mu = T(F) = \int x\, dF(x)$. In general, any function of F is called a **statistical functional**. Other examples of functionals are the variance $T(F) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$ and the median $T(F) = F^{-1}(1/2)$. ■

6.6 Example (Regression, prediction, and classification)**.** Suppose we observe pairs of data $(X_1, Y_1), \ldots (X_n, Y_n)$. Perhaps X_i is the blood pressure of subject i and Y_i is how long they live. X is called a **predictor** or **regressor** or **feature** or **independent variable**. Y is called the **outcome** or the **response variable** or the **dependent variable**. We call $r(x) = \mathbb{E}(Y|X = x)$ the **regression function**. If we assume that $r \in \mathfrak{F}$ where $\mathfrak{F}$ is finite dimensional — the set of straight lines for example — then we have a **parametric regression model**. If we assume that $r \in \mathfrak{F}$ where $\mathfrak{F}$ is not finite dimensional then we have a **nonparametric regression model**. The goal of predicting Y for a new patient based on their X value is called **prediction**. If Y is discrete (for example, live or die) then prediction is instead called **classification.** If our goal is to estimate the function r, then we call this **regression** or **curve estimation**. Regression models are sometimes written as

$$Y = r(X) + \epsilon \tag{6.3}$$

where $\mathbb{E}(\epsilon) = 0$. We can always rewrite a regression model this way. To see this, define $\epsilon = Y - r(X)$ and hence $Y = Y + r(X) - r(X) = r(X) + \epsilon$. Moreover, $\mathbb{E}(\epsilon) = \mathbb{E}\mathbb{E}(\epsilon|X) = \mathbb{E}(\mathbb{E}(Y - r(X))|X) = \mathbb{E}(\mathbb{E}(Y|X) - r(X)) = \mathbb{E}(r(X) - r(X)) = 0$. ■

What's Next? It is traditional in most introductory courses to start with parametric inference. Instead, we will start with nonparametric inference and then we will cover parametric inference. In some respects, nonparametric inference is easier to understand and is more useful than parametric inference.

Frequentists and Bayesians. There are many approaches to statistical inference. The two dominant approaches are called **frequentist inference** and **Bayesian inference**. We'll cover both but we will start with frequentist inference. We'll postpone a discussion of the pros and cons of these two until later.

Some Notation. If $\mathfrak{F} = \{f(x;\, \theta) :\ \theta \in \Theta\}$ is a parametric model, we write $\mathbb{P}_\theta(X \in A) = \int_A f(x;\, \theta)dx$ and $\mathbb{E}_\theta(r(X)) = \int r(x)f(x;\, \theta)dx$. The subscript θ indicates that the probability or expectation is with respect to $f(x;\, \theta)$; it does not mean we are averaging over θ. Similarly, we write $\mathbb{V}_\theta$ for the variance.

6.3 Fundamental Concepts in Inference

Many inferential problems can be identified as being one of three types: estimation, confidence sets, or hypothesis testing. We will treat all of these problems in detail in the rest of the book. Here, we give a brief introduction to the ideas.

6.3.1 Point Estimation

Point estimation refers to providing a single "best guess" of some quantity of interest. The quantity of interest could be a parameter in a parametric model, a CDF F, a probability density function f, a regression function r, or a prediction for a future value Y of some random variable.

> By convention, we denote a point estimate of θ by $\widehat{\theta}$ or $\widehat{\theta}_n$. Remember that θ is a fixed, unknown quantity. The estimate $\widehat{\theta}$ depends on the data so $\widehat{\theta}$ is a random variable.

More formally, let $X_1, \ldots, X_n$ be n IID data points from some distribution F. A point estimator $\widehat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\widehat{\theta}_n = g(X_1, \ldots, X_n).$$

The bias of an estimator is defined by

$$\mathsf{bias}(\widehat{\theta}_n) = \mathbb{E}_\theta(\widehat{\theta}_n) - \theta. \tag{6.4}$$

We say that $\widehat{\theta}_n$ is **unbiased** if $\mathbb{E}(\widehat{\theta}_n) = \theta$. Unbiasedness used to receive much attention but these days is considered less important; many of the estimators we will use are biased. A reasonable requirement for an estimator is that it should converge to the true parameter value as we collect more and more data. This requirement is quantified by the following definition:

6.7 Definition. *A point estimator $\widehat{\theta}_n$ of a parameter θ is* **consistent** *if $\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta$.*

The distribution of $\widehat{\theta}_n$ is called the **sampling distribution**. The standard deviation of $\widehat{\theta}_n$ is called the **standard error**, denoted by se:

$$\mathsf{se} = \mathsf{se}(\widehat{\theta}_n) = \sqrt{\mathbb{V}(\widehat{\theta}_n)}. \tag{6.5}$$

Often, the standard error depends on the unknown F. In those cases, se is an unknown quantity but we usually can estimate it. The estimated standard error is denoted by $\widehat{\mathsf{se}}$.

6.8 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\widehat{p}_n = n^{-1}\sum_i X_i$. Then $\mathbb{E}(\widehat{p}_n) = n^{-1}\sum_i \mathbb{E}(X_i) = p$ so $\widehat{p}_n$ is unbiased. The standard error is $\mathsf{se} = \sqrt{\mathbb{V}(\widehat{p}_n)} = \sqrt{p(1-p)/n}$. The estimated standard error is $\widehat{\mathsf{se}} = \sqrt{\widehat{p}(1-\widehat{p})/n}$. ■

The quality of a point estimate is sometimes assessed by the **mean squared error**, or MSE defined by

$$\text{MSE} = \mathbb{E}_\theta(\widehat{\theta}_n - \theta)^2. \tag{6.6}$$

Keep in mind that $\mathbb{E}_\theta(\cdot)$ refers to expectation with respect to the distribution

$$f(x_1, \ldots, x_n;\, \theta) = \prod_{i=1}^{n} f(x_i;\, \theta)$$

that generated the data. It does not mean we are averaging over a distribution for θ.

6.9 Theorem. *The* MSE *can be written as*

$$\text{MSE} = \mathsf{bias}^2(\widehat{\theta}_n) + \mathbb{V}_\theta(\widehat{\theta}_n). \tag{6.7}$$

PROOF. Let $\overline{\theta}_n = E_\theta(\widehat{\theta}_n)$. Then

$$\begin{aligned}
\mathbb{E}_\theta(\widehat{\theta}_n - \theta)^2 &= \mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n + \overline{\theta}_n - \theta)^2 \\
&= \mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n)^2 + 2(\overline{\theta}_n - \theta)\mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n) + \mathbb{E}_\theta(\overline{\theta}_n - \theta)^2 \\
&= (\overline{\theta}_n - \theta)^2 + \mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n)^2 \\
&= \mathsf{bias}^2(\widehat{\theta}_n) + \mathbb{V}(\widehat{\theta}_n)
\end{aligned}$$

where we have used the fact that $\mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n) = \overline{\theta}_n - \overline{\theta}_n = 0$. ■

6.10 Theorem. *If* $\mathsf{bias} \to 0$ *and* $\mathsf{se} \to 0$ *as* $n \to \infty$ *then* $\widehat{\theta}_n$ *is consistent, that is,* $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$.

PROOF. If $\mathsf{bias} \to 0$ and $\mathsf{se} \to 0$ then, by Theorem 6.9, $\mathsf{MSE} \to 0$. It follows that $\widehat{\theta}_n \xrightarrow{\text{qm}} \theta$. (Recall Definition 5.2.) The result follows from part (b) of Theorem 5.4. ■

6.11 Example. Returning to the coin flipping example, we have that $\mathbb{E}_p(\widehat{p}_n) = p$ so the $\mathsf{bias} = p - p = 0$ and $\mathsf{se} = \sqrt{p(1-p)/n} \to 0$. Hence, $\widehat{p}_n \xrightarrow{\text{P}} p$, that is, $\widehat{p}_n$ is a consistent estimator. ■

Many of the estimators we will encounter will turn out to have, approximately, a Normal distribution.

6.12 Definition. *An estimator is* **asymptotically Normal** *if*

$$\frac{\widehat{\theta}_n - \theta}{\mathsf{se}} \rightsquigarrow N(0,1). \tag{6.8}$$

6.3.2 Confidence Sets

A $1-\alpha$ **confidence interval** for a parameter θ is an interval $C_n = (a,b)$ where $a = a(X_1,\ldots,X_n)$ and $b = b(X_1,\ldots,X_n)$ are functions of the data such that

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha, \quad \text{for all } \theta \in \Theta. \tag{6.9}$$

In words, (a,b) traps θ with probability $1-\alpha$. We call $1-\alpha$ the **coverage** of the confidence interval.

Warning! C_n is random and θ is fixed.

Commonly, people use 95 percent confidence intervals, which corresponds to choosing $\alpha = 0.05$. If θ is a vector then we use a **confidence set** (such as a sphere or an ellipse) instead of an interval.

Warning! There is much confusion about how to interpret a confidence interval. A confidence interval is not a probability statement about θ since θ is a fixed quantity, not a random variable. Some texts interpret confidence intervals as follows: if I repeat the experiment over and over, the interval will contain the parameter 95 percent of the time. This is correct but useless since we rarely repeat the same experiment over and over. A better interpretation is this:

> On day 1, you collect data and construct a 95 percent confidence interval for a parameter θ_1. On day 2, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_2. On day 3, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_3. You continue this way constructing confidence intervals for a sequence of unrelated parameters $\theta_1, \theta_2, \ldots$ Then 95 percent of your intervals will trap the true parameter value. There is no need to introduce the idea of repeating the same experiment over and over.

6.13 Example. Every day, newspapers report opinion polls. For example, they might say that "83 percent of the population favor arming pilots with guns." Usually, you will see a statement like "this poll is accurate to within 4 points

95 percent of the time." They are saying that 83 ± 4 is a 95 percent confidence interval for the true but unknown proportion p of people who favor arming pilots with guns. If you form a confidence interval this way every day for the rest of your life, 95 percent of your intervals will contain the true parameter. This is true even though you are estimating a different quantity (a different poll question) every day. ■

6.14 Example. The fact that a confidence interval is not a probability statement about θ is confusing. Consider this example from Berger and Wolpert (1984). Let θ be a fixed, known real number and let X_1, X_2 be independent random variables such that $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$. Now define $Y_i = \theta + X_i$ and suppose that you only observe Y_1 and Y_2. Define the following "confidence interval" which actually only contains one point:

$$C = \begin{cases} \{Y_1 - 1\} & \text{if } Y_1 = Y_2 \\ \{(Y_1 + Y_2)/2\} & \text{if } Y_1 \neq Y_2. \end{cases}$$

You can check that, no matter what θ is, we have $\mathbb{P}_\theta(\theta \in C) = 3/4$ so this is a 75 percent confidence interval. Suppose we now do the experiment and we get $Y_1 = 15$ and $Y_2 = 17$. Then our 75 percent confidence interval is $\{16\}$. However, we are certain that $\theta = 16$. If you wanted to make a probability statement about θ you would probably say that $\mathbb{P}(\theta \in C | Y_1, Y_2) = 1$. There is nothing wrong with saying that $\{16\}$ is a 75 percent confidence interval. But is it not a probability statement about θ. ■

In Chapter 11 we will discuss Bayesian methods in which we treat θ as if it were a random variable and we do make probability statements about θ. In particular, we will make statements like "the probability that θ is in C_n, given the data, is 95 percent." However, these Bayesian intervals refer to degree-of-belief probabilities. These Bayesian intervals will not, in general, trap the parameter 95 percent of the time.

6.15 Example. In the coin flipping setting, let $C_n = (\widehat{p}_n - \epsilon_n,\ \widehat{p}_n + \epsilon_n)$ where $\epsilon_n^2 = \log(2/\alpha)/(2n)$. From Hoeffding's inequality (4.4) it follows that

$$\mathbb{P}(p \in C_n) \geq 1 - \alpha$$

for every p. Hence, C_n is a $1 - \alpha$ confidence interval. ■

As mentioned earlier, point estimators often have a limiting Normal distribution, meaning that equation (6.8) holds, that is, $\widehat{\theta}_n \approx N(\theta, \widehat{\mathsf{se}}^2)$. In this case we can construct (approximate) confidence intervals as follows.

6.16 Theorem (Normal-based Confidence Interval). *Suppose that* $\widehat{\theta}_n \approx N(\theta, \widehat{\mathsf{se}}^2)$. *Let* Φ *be the* CDF *of a standard Normal and let* $z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$, *that is,* $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$ *and* $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$ *where* $Z \sim N(0,1)$. *Let*

$$C_n = (\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}},\ \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}). \tag{6.10}$$

Then

$$\mathbb{P}_\theta(\theta \in C_n) \to 1-\alpha. \tag{6.11}$$

PROOF. Let $Z_n = (\widehat{\theta}_n - \theta)/\widehat{\mathsf{se}}$. By assumption $Z_n \rightsquigarrow Z$ where $Z \sim N(0,1)$. Hence,

$$\begin{aligned}
\mathbb{P}_\theta(\theta \in C_n) &= \mathbb{P}_\theta\left(\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}} < \theta < \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}\right) \\
&= \mathbb{P}_\theta\left(-z_{\alpha/2} < \frac{\widehat{\theta}_n - \theta}{\widehat{\mathsf{se}}} < z_{\alpha/2}\right) \\
&\to \mathbb{P}\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) \\
&= 1-\alpha. \quad \blacksquare
\end{aligned}$$

For 95 percent confidence intervals, $\alpha = 0.05$ and $z_{\alpha/2} = 1.96 \approx 2$ leading to the approximate 95 percent confidence interval $\widehat{\theta}_n \pm 2\,\widehat{\mathsf{se}}$.

6.17 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\widehat{p}_n = n^{-1}\sum_{i=1}^n X_i$. Then $\mathbb{V}(\widehat{p}_n) = n^{-2}\sum_{i=1}^n \mathbb{V}(X_i) = n^{-2}\sum_{i=1}^n p(1-p) = n^{-2}np(1-p) = p(1-p)/n$. Hence, $\mathsf{se} = \sqrt{p(1-p)/n}$ and $\widehat{\mathsf{se}} = \sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}$. By the Central Limit Theorem, $\widehat{p}_n \approx N(p, \widehat{\mathsf{se}}^2)$. Therefore, an approximate $1-\alpha$ confidence interval is

$$\widehat{p}_n \pm z_{\alpha/2}\widehat{\mathsf{se}} = \widehat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$

Compare this with the confidence interval in example 6.15. The Normal-based interval is shorter but it only has approximately (large sample) correct coverage. ■

6.3.3 Hypothesis Testing

In **hypothesis testing**, we start with some default theory — called a **null hypothesis** — and we ask if the data provide sufficient evidence to reject the theory. If not we retain the null hypothesis. [2]

[2] The term "retaining the null hypothesis" is due to Chris Genovese. Other terminology is "accepting the null" or "failing to reject the null."

6.18 Example (Testing if a Coin is Fair)**.** Let

$$X_1, \ldots, X_n \sim \text{Bernoulli}(p)$$

be n independent coin flips. Suppose we want to test if the coin is fair. Let H_0 denote the hypothesis that the coin is fair and let H_1 denote the hypothesis that the coin is not fair. H_0 is called the **null hypothesis** and H_1 is called the **alternative hypothesis**. We can write the hypotheses as

$$H_0 : p = 1/2 \quad \text{versus} \quad H_1 : p \neq 1/2.$$

It seems reasonable to reject H_0 if $T = |\widehat{p}_n - (1/2)|$ is large. When we discuss hypothesis testing in detail, we will be more precise about how large T should be to reject H_0. ■

6.4 Bibliographic Remarks

Statistical inference is covered in many texts. Elementary texts include DeGroot and Schervish (2002) and Larsen and Marx (1986). At the intermediate level I recommend Casella and Berger (2002), Bickel and Doksum (2000), and Rice (1995). At the advanced level, Cox and Hinkley (2000), Lehmann and Casella (1998), Lehmann (1986), and van der Vaart (1998).

6.5 Appendix

Our definition of confidence interval requires that $\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$ for all $\theta \in \Theta$. A **pointwise asymptotic** confidence interval requires that $\liminf_{n\to\infty} \mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$ for all $\theta \in \Theta$. A **uniform asymptotic** confidence interval requires that $\liminf_{n\to\infty} \inf_{\theta\in\Theta} \mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$. The approximate Normal-based interval is a pointwise asymptotic confidence interval.

6.6 Exercises

1. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ and let $\widehat{\lambda} = n^{-1} \sum_{i=1}^n X_i$. Find the bias, se, and MSE of this estimator.

2. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$ and let $\widehat{\theta} = \max\{X_1, \ldots, X_n\}$. Find the bias, se, and MSE of this estimator.

3. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$ and let $\widehat{\theta} = 2\overline{X}_n$. Find the bias, se, and MSE of this estimator.