6 Introduction to Statistical Inference

6.1 Introduction

Statistical inference, or "learning" as it is sometimes called in Computer Science, is the process of using data to infer the distribution that generated the data. The basic statistical inference problem is this: we observe $X_1, \ldots, X_n \sim F$ and we want to guess (or infer or estimate) F or some feature of F, such as the mean.

Sometimes we will make very assumptions about F. For example, we might assume only that $F \in \mathcal{F}$ where $\mathcal{F}$ is the set of all distribution functions. Inferential methods that use few assumptions about F, are called *nonparametric methods.* The set $\mathcal{F}$ is an example of a *nonparametric model.*

Sometimes we will make stronger assumptions about F, such as F has a density $f \in \mathcal{F}$ where $\mathcal{F}$ is the set of all Normal density functions. In this case we can write $\mathcal{F} = \{f(x;\theta) : \quad \theta \in \Theta\}$ where θ is the unknown *parameter* and Θ is the *parameter space.* We will write a typical density in a parametric model as $f(x;\theta)$. In the Normal example, $\theta = (\mu, \sigma)$ and $\Theta = \{(\mu, \sigma) : \quad \mu \in \mathcal{R}, \sigma > 0\}$. We call $\mathcal{F}$ a *parametric model* since it is indexed by a finite-dimensional parameter θ. Inferential methods that are based on the assumption that F is in a parametric model are called *parametric methods.* Here are some examples to make these ideas clear.

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

EXAMPLE 6.2 (Two-dimensional Parametric Estimation.) *Suppose that $X_1, \ldots, X_n \sim F$ and suppose we assume that the pdf $f \in \mathcal{F}$ where*

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

We call $\mathcal{F}$ a two-dimensional parametric model. *In this case there are two parameters, μ and σ. The goal is to estimate the parameters from the data.*

EXAMPLE 6.3 (Nonparametric estimation of the cdf.) *Let $X_1, \ldots, X_n$ be independent observations from a cdf F. The problem is to estimate F without making any assumptions about F.*

EXAMPLE 6.4 (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 *while making weak assumptions about* f. *For example, if we have reason to believe that* f *is smooth, we might assume that* $f \in \mathcal{F}$ *where*

$$\mathcal{F} = \left\{ g : \ g \geq 0, \ \int g(x)dx = 1, \ \int (g''(x))^2 dx < \infty \right\}.$$

The class $\mathcal{F}$ *is the set of pdf's that are not "too wiggly."*

EXAMPLE 6.5 (Nonparametric estimation of functionals.) *Let* $X_1, \ldots, X_n \sim F$. *Suppose we want to estimate* $\mu = E(X_1) = \int x\, dF(x)$ *assuming only that* μ *exists. The mean* μ *may be thought of as a function of* F*; we can write* $\mu = T(F)$ *where* $T(F) = \int x\, dF(x)$. *In general any function of* F *is called a* statistical functional. *Other examples of functions are the variance* $T(F) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$ *and the median* $T(F) = F^{-1/2}$.

EXAMPLE 6.6 (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. Define* $f(x) = E(Y|X = x)$. *We call* $f(x)$ *the* regression function. *Let* $\epsilon = Y - f(X)$. *Note that* $E(\epsilon) = EE(\epsilon|X) = E(E(Y - f(X))|X) = E(E(Y|X) - f(X)) = E(f(X) - f(X)) = 0$. *We can write the data as*

$$Y_i = f(X_i) + \epsilon_i.$$

The problem is to estimate the function f. *If we assume that* f *has a linear form such as* $f(x) = \beta_0 + \beta_1 x$ *then this is a parametric problem and is called* linear regression. *We might instead make only a weak smoothness assumptions on* f. *In that case we refer to the problem as* nonparametric regression. *If we don't want to estimate* f *but instead we only want to predict* Y *given a new value of* X, *we call this* prediction. *When* Y *is discrete, the prediction problem is usually called* classification.

WHAT'S NEXT? It is traditional in most introductory courses to start with parametric inference. I am going to take a radical approach and do the opposite. We will start with nonparametric inference and then we will cover parametric inference. I think that nonparametric inference is easier to understand. It is also, in my opinion, 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 pro's and con's of these two approaches until we've learned a bit of both.

6.2 Point Estimates, Confidence Intervals and Hypothesis Tests

The particular goals of inference are problem dependent, but many inferential problems can be identified as being of one of three types: point estimation, confidence intervals or hypothesis testing.

POINT ESTIMATION. Consider the simple problem of coin flipping. Suppose $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. If we want a "best guess" of p, then we might use $\widehat{p} = n^{-1} \sum_i X_i$. The quantity $\widehat{p}$ is called a *point estimate*.

A point estimator $\widehat{\theta}_n$ of a parameter θ is *unbiased* if $E(\widehat{\theta}_n) = \theta$. Unbiasedness used to receive much attention. These days it is not considered very important and many of the estimators we use are biased. A point estimator $\widehat{\theta}_n$ of a parameter θ is *consistent* if $\widehat{\theta}_n \xrightarrow{p} \theta$. Consistency is important and all the estimators we will use will be consistent estimators.

If θ is a parameter and $\widehat{\theta}_n$ is a point estimate of θ, then the standard deviation of $\widehat{\theta}_n$ is called the *standard error*, denoted by se:

$$\text{se}(\widehat{\theta}_n) = \sqrt{\text{Var}(\widehat{\theta}_n)}.$$

Often, it is not possible to compute the standard error but usually we can estimate the standard error. The estimated standard error is denoted by $\widehat{se}$. As an example, consider the estimate $\widehat{p} = n^{-1} \sum_i X_i$ for the Bernoulli parameter p. Its true standard error is $\text{se} = \sqrt{\text{Var}(\widehat{p}_n)} = \sqrt{p(1-p)/n}$. Since we do not know p, we can't compute se. But we can estimate the standard error with $\widehat{se} = \sqrt{\widehat{p}(1-\widehat{p})/n}$.

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

$$\begin{aligned} \text{MSE} &= E_\theta(\widehat{\theta}_n - \theta)^2 \\ &= \int \left(\widehat{\theta}_n(x_1, \ldots, x_n) - \theta\right)^2 f(x_1, \ldots, x_n; \theta) dx_1 \cdots dx_n \end{aligned}$$

where $E_\theta(\cdot)$ refers to expectation with respect to the distribution $f(x_1, \ldots, x_n; \theta)$ that generated the data. It does not mean we are averaging over a density for θ.

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

$$\begin{aligned} E_\theta(\widehat{\theta}_n - \theta)^2 &= E_\theta(\widehat{\theta}_n - \overline{\theta}_n + \overline{\theta}_n - \theta)^2 \\ &= E_\theta(\widehat{\theta}_n - \overline{\theta}_n)^2 + 2(\overline{\theta}_n - \theta)^2 E_\theta(\widehat{\theta}_n - \overline{\theta}_n) + E_\theta(\overline{\theta}_n - \theta)^2 \\ &= (\overline{\theta}_n - \theta)^2 + E_\theta(\widehat{\theta}_n - \overline{\theta}_n)^2 \\ &= \text{bias}^2 + \text{Variance} \\ &= \text{bias}^2 + \text{se}^2 \end{aligned}$$

where bias $= E_\theta(\widehat{\theta}_n) - \theta$. We see that if bias $\to 0$ and se $\to 0$ then MSE $\to 0$ and hence $\widehat{\theta}_n \stackrel{q.m.}{\to} 0$. Since convergence in quadratic mean implies convergence in probability, we conclude that if bias $\to 0$ and se $\to 0$ then $\widehat{\theta}_n \stackrel{p}{\to} \theta$. Returning to the coin flipping example, we have that $E_p(\widehat{p}_n) = p$ so that bias $= p - p = 0$ and $se = \sqrt{p(1-p)/n} \to 0$. Hence, $\widehat{p}_n \stackrel{p}{\to} p$, that is, $\widehat{p}_n$ is a consistent estimator.

Confidence Intervals. A 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 (a, b) traps the θ with some pre-specified probability. **Note that C_n is random and θ is fixed.** In the coin flipping setting, let $C_n = (\widehat{p}_n - \epsilon_n,\ \widehat{p}_n + \epsilon_n)$ where $\epsilon_n^2 = \log(2/\alpha)/(2n)$. Earlier, we showed, using Hoeffding's inequality, that $P(p \in C_n) \geq 1 - \alpha$, so C_n is a $1 - \alpha$ confidence interval.

In some cases, point estimators have a limiting Normal distribution, $\widehat{\theta}_n \approx N(\theta, \text{se}^2)$. In this case we can construct (approximate) confidence intervals as follows. Let $z_{\alpha/2} = \Phi^{-1}(1 - (\alpha/2))$. Hence, $P(Z > z_{\alpha/2}) = \alpha/2$ and $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$ where $Z \sim N(0, 1)$. Let

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

Then,

$$\begin{aligned} P(\theta \in C_n) &= P\left(\widehat{\theta}_n - z_{\alpha/2}\,\text{se} < \theta < \widehat{\theta}_n + z_{\alpha/2}\,\text{se}\right) \\ &= P\left(-z_{\alpha/2} < \frac{\widehat{\theta}_n - \theta}{\text{se}} < z_{\alpha/2}\right) \end{aligned}$$

$$
\begin{aligned}
&\approx P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) \\
&= 1 - \alpha.
\end{aligned}
$$

The same is true if an estimate $\widehat{se}$ is inserted for se. In summary,

$$\boxed{\widehat{\theta}_n \pm z_{\alpha/2}\,\widehat{se} \text{ is an approximate } 1-\alpha \text{ confidence interval.}}$$

Often, people use 95 per cent confidence intervals which corresponds to $\alpha = 0.05$. In this case, $z_{\alpha/2} = 1.96 \approx 2$ leading to the approximate 95 per cent confidence interval $\widehat{\theta}_n \pm 2\widehat{se}$.

HYPOTHESIS TESTING. Suppose we want to know if a coin is fair ($p = 1/2$). 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 $H_0 : p = 1/2$ versus $H_1 : p \neq 1/2$. If we toss a coin 100 times and get 100 heads, we might reasonably take this as evidence against H_0. This is an example of *hypothesis testing*.