PART 3

Statistical inference

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

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

3.1

PARAMETRIC AND NONPARAMETRIC MODELS

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

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

This is a two-parameter model. Writing the density as $f\left(x;\mu,\sigma\right)$ shows that x is a value of the RV while μ and σ are parameters. In general, a parametric model takes the form

$$\mathscr{F} = \{f\left(x;\theta\right) : \theta\in\Theta\}$$

where θ is an unknown parameter (or vector of parameters) that can take values in the **parameter space** Θ. A **nonparametric model** is a set $\mathscr{F}$ that cannot be parameterized by a finite number of parameters. For example, $\mathscr{F}_{ALL} = \{\text{all CDF's}\}$ is nonparametric[1].

Example 3.1.1 (Regression, prediction, and classification)**.** Suppose we observe pairs of data $\left(X_1,Y_1\right),...,\left(X_n,Y_n\right)$. 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 $r\left(x\right)=\mathbb{E}\left[Y|X=x\right]$ the **regression function**. If we assume that $r\in\mathscr{F}$ where $\mathscr{F}$ is finite dimensional, then we have a **parametric regression model**. If $\mathscr{F}$ 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, then prediction is called **classification**. If we wish to estimate the function r, then we call this **regression** or **curve estimation**. Regression models are sometimes written as

$$Y = r\left(X\right)+\epsilon \tag{7}$$

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

Notation: If $\mathscr{F}=\{f\left(x;\theta\right):\theta\in\Theta\}$ is a parametric model, we write $\mathbb{P}_\theta\left(X\in A\right)=\int_A f\left(x;\theta\right)dx$ and $\mathbb{E}_\theta\left[r\left(X\right)\right]=\int r\left(x\right)f\left(x;\theta\right)dx$. The subscript θ indicates that the probability or expectation is w.r.t $f\left(x;\theta\right)$, not that we are averaging over θ. Similarly, write $\mathbb{V}_\theta$ for the variance[2].

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

[2] There's no need to understand the logic behind the notation here, this is more for me so I don't lose track of what's happening

Most inferential problems can be divided into either estimation, confidence sets, or hypothesis testing. We are interested mainly in hypothesis testing so I will only provide a basic description of the other two.

Point Estimation

Point estimation refers to providing a single "best guess" of some quantity of interest. We denote a point estimate of θ by $\hat{\theta}$ or $\hat{\theta}_n$.

Confidence Sets

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

$$\mathbb{P}_\theta\left(\theta \in C_n\right) \geq 1-\alpha,\ \forall \theta \in \Theta. \tag{8}$$

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. If θ is a vector then we use a **confidence set** instead of an interval.

Interpretation: Don't think of it as, 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 experiments are rarely repeated that many times. A better interpretation is that if you construct 95 percent confidence intervals with many parameters (e.g. $\theta_1, \theta_2, \theta_3, ...$) then 95 percent of the intervals you construct will trap the true parameter value.

Note: Remember that a confidfence interval is not a probability statement about θ.

Hypothesis Testing

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

Example 3.1.2 (Test if a coin is fair)**.** Let

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

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. H_0 is called the **null hypothesis** and H_1 the **alternative hypothesis**. They can be written as

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

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

Another question we could ask is if exposure to asbestos is associated with lung disease. We take some rats and randomly divide them into two groups. We expose one group to asbestos and leave the second group unexposed. Then we compare the disease rates in the two groups. Consider the following two hypotheses:

The Null Hypothesis: The disease rate is the same in the two groups.
The Alternative Hypothesis: The disease rate is not the same in the two groups.

If the exposed group has a much higher rate of disease than the unexposed group then we will reject the null hypothesis. This is an example of hypothesis testing. More formally, suppose that we partititon the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1. \tag{9}$$

We call H_0 the null hypothesis and H_1 the alternative hypothesis.

Let X be a random variable and let $\mathscr{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $R \subset \mathscr{X}$ called the rejection region. If $X \in R$ we reject the null hypothesis, otherwise, we do not reject the null hypothesis:
$X \in R \Rightarrow$ reject H_0
$X \notin R \Rightarrow$ retain (do not reject) H_0.

Usually the rejection region R is of the form

$$R = \{x : T(x) > c\} \tag{10}$$

where T is a test statistic and c is a critical value. The problem in hypothesis testing is to find an appropriate test statistic T and an appropriate critical value c.

Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggests that they are guilty. Similarly, we retain H_0 unless there is strong evidence to reject H_0. Rejecting H_0 when H_0 is true is called a type I error and retaining H_0 when H_1 is true is called a type II error.