Chapter 5

Statistics Primer

Probability primarily concerns random variables and studying what happens to those random variables under various transformations or when repeated measurements are made. Implicit in the study is the fact that the underlying probability distribution is fixed and known (or described implicitly) and we want to infer properties of that distribution.

Statistics in large part turns this picture around. We have a collection of data, which are assumed to be random samples from some underlying, but unknown, probability distribution. We would like to develop methods to infer properties of that distribution from the data. Furthermore, we would like to perform hypothesis tests, that is, determine whether or not the data is likely to have been generated from a certain family of distributions, and measure how confident we are in that assessment.

Statistics is a rich subject and we only briefly touch on many topics in that area. Our focus for much of the book will be on frequentist methods of statistics, which is concerned with the development of statistical procedures based on the long-term behavior of those methods under repeated trials or large sample sizes. We also discuss Bayesian methods in Section 5.5, where statistical inference is concerned with how data can be used to update our prior beliefs about an unknown parameter. Algebraic methods for Bayesian statistics will appear in Chapters 17 and 18.

5.1. Statistical Models

Much of statistics is carried out with respect to statistical models. A statistical model is just a family of probability distributions. Often these models

are parametric. In this section, we give the basic definitions of statistical models and give a number of important examples.

Definition 5.1.1. A *statistical model* $\mathcal{M}$ is a collection of probability distributions or density functions. A *parametric statistical model* $\mathcal{M}_\Theta$ is a mapping from a finite-dimensional parameter space $\Theta \subseteq \mathbb{R}^d$ to a space of probability distributions or density functions, i.e.,

$$p_\bullet : \Theta \to \mathcal{M}_\Theta, \quad \theta \mapsto p_\theta.$$

The model is the image of the map $p_\bullet$, $\mathcal{M}_\Theta = \{p_\theta : \theta \in \Theta\}$.

A statistical model is called *identifiable* if the mapping $p_\bullet$ is one-to-one, that is, $p_{\theta_1} = p_{\theta_2}$ implies that $\theta_1 = \theta_2$. We will return to discussions of identifiability in Chapter 16.

We have already seen examples of statistical models in the previous chapters, which we remind the reader of and put into context here.

Example 5.1.2 (Binomial random variable)**.** Let X be a discrete random variable with $r+1$ states, labeled $0, 1, \ldots, r$. Let $\Theta = [0,1]$ and for $\theta \in \Theta$ consider the probability distribution

$$P_\theta(X = i) = \binom{r}{i} \theta^i (1-\theta)^{r-i},$$

which is the distribution of a binomial random variable with r samples and parameter θ. The binomial random variable model $\mathcal{M}_\Theta$ consists of all the probability distributions that arise this way. It sits as a subset of the probability simplex Δ_r.

Example 5.1.3 (Multivariable normal random vector)**.** Let $X \in \mathbb{R}^m$ be an m-dimensional real random vector. Let $\Theta = \mathbb{R}^m \times PD_m$, where PD_m is the cone of symmetric $m \times m$ positive definite matrices. For $\theta = (\mu, \Sigma) \in \Theta$ let

$$p_\theta(x) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

be the density of a jointly normal random variable with mean vector μ and covariance matrix Σ. The model $\mathcal{M}_\Theta$ consists of all such density functions for an m-dimensional real random vector.

Statistical models with infinite-dimensional parameter spaces are called *nonparametric* or *semiparametric* depending on the context. Note that this does not necessarily mean that we are considering a statistical model which does not have a parametrization, as so-called nonparametric models might still have nice parametrizations. On the other hand, models which are defined via constraints on probability distributions or densities we will refer to as *implicit statistical models*. Note, however, that in the statistics literature

such models defined by implicit constraints on a finite-dimensional parameter space are still called parametric models. We have already seen examples of implicit statistical models when looking at independence constraints.

Example 5.1.4 (Independent random variables)**.** Let X_1 and X_2 be two discrete random variables with state spaces $[r_1]$ and $[r_2]$, respectively, with $\mathcal{R} = [r_1] \times [r_2]$. The model of independence $\mathcal{M}_{X_1 \perp\!\!\!\perp X_2}$ consists of all distributions $p \in \Delta_{\mathcal{R}}$ such that

$$P(X_1 = i_1, X_2 = i_2) = P(X_1 = i_1)P(X_2 = i_2)$$

for all $i_1 \in [r_1]$ and $i_2 \in [r_2]$. Thus, the model of independence is an implicit statistical model. As discussed in Section 4.1, this model consists of all joint distribution matrices which are rank one matrices. From that description we can also realize this as a parametric statistical model. Indeed, let $\Theta = \Delta_{r_1-1} \times \Delta_{r_2-1}$, and given $\theta = (\alpha, \beta) \in \Theta$, define the probability distribution

$$P_\theta(X_1 = i_1, X_2 = i_2) = \alpha_{i_1}\beta_{i_2}.$$

Our main example of an implicit statistical model in this book comes from looking at conditional independence constraints. If

$$\mathcal{C} = \{A_1 \perp\!\!\!\perp B_1 | C_1, A_2 \perp\!\!\!\perp B_2 | C_2, \ldots\}$$

is a collection of conditional independence statements, for each fixed set of states $r_1, r_2, \ldots, r_m$ we get a discrete conditional independence model

$$\mathcal{M}_{\mathcal{C}} = V(I_{A_1 \perp\!\!\!\perp B_1|C_1} + I_{A_2 \perp\!\!\!\perp B_2|C_2} + \cdots) \cap \Delta_{\mathcal{R}},$$

which is an implicit statistical model. Similarly, we get a Gaussian conditional independence model

$$\mathcal{M}_{\mathcal{C}} = V(J_{A_1 \perp\!\!\!\perp B_1|C_1} + J_{A_2 \perp\!\!\!\perp B_2|C_2} + \cdots) \cap PD_m.$$

An important concept in the study of statistical models is the notion of sufficiency of a statistic. Sufficient statistics explain the way that the state of a variable enters into the calculation of the probability of that state in a model, and the interaction of the state of the random variable with the parameters.

Definition 5.1.5. A *statistic* is a function from the state space of a random variable to some other set. For a parametric statistical model $\mathcal{M}_\Theta$, a statistic T is *sufficient* for the model if

$$P(X = x \mid T(X) = t, \theta) = P(X = x \mid T(X) = t).$$

Equivalently, the joint distribution $p_\theta(x)$ should factorize as

$$p_\theta(x) = f(x)g(T(x), \theta),$$

where f is a function that does not depend on θ. A statistic T is *minimal sufficient* if every other sufficient statistic is a function of T.

Examples of sufficiency will appear in the next section as we discuss different types of data.

5.2. Types of Data

Statistical models are used to analyze data. If we believe our data follows a distribution that belongs to the model, we would like to estimate the model parameters from the data and use the resulting quantities to make some conclusions about the underlying process that generated the data. Alternately, we might not be sure whether or not a given model is a good explanation of the data. To this end, we would perform a hypothesis test, to decide whether or not the model is a good fit. Before we can talk about any of these statistical settings, however, we need to discuss what we mean by data, and, in some sense, different types of models. In particular, natural classes of data generating procedures include: independent and identically distributed data, exchangeable data, and time-series and spatial data. The techniques from algebraic statistics that can be applied depend on what was the underlying data generation assumption.

The most common type of data generation we will encounter in this book is the case of *independent and identically distributed* data. In this setting, we assume that we have a sample of n data $D = X^{(1)}, X^{(2)}, \ldots, X^{(n)}$, each distributed identically and independently like the distribution $p_\theta(X)$ for some distribution p_θ in our model. If our random variables are discrete, the probability of observing the particular sequence of data, given θ, is

$$p_\theta(D) = \prod_{i=1}^{n} p_\theta(X^{(i)}).$$

For continuous distributions, the density function is represented similarly in product form. Note that in the discrete case, where our random variable has state space $[r]$, we can compute the vector of counts $u \in \mathbb{N}^r$, defined by

$$u_j = \#\{i : X^{(i)} = j\}$$

in which case

$$p_\theta(D) = \prod_{j=1}^{r} p_\theta(j)^{u_j}$$

so that the vector of counts gives sufficient statistics for a model under the i.i.d. assumption. Note that for discrete random variables, the vector of counts u is a summary of the data from which it is possible to compute the probability of observing the particular sequence $i_1, \ldots, i_n$. When working

with such models and discussing data for such models, we usually immediately pass to the vector of counts as our data for such models.

A second type of assumption that arises in Bayesian statistics is known as *exchangeable* random variables. Discrete random variables $X^{(1)}, \ldots, X^{(n)}$ are exchangeable if and only if

$$P(X_1 = i_1, \ldots, X_n = i_n) = P(X_1 = i_{\sigma(1)}, \ldots, X_n = i_{\sigma(n)})$$

for any permutation σ of the index set $[n]$. For continuous random variables, we require the same permutation condition on the density function. Note that independent identically distributed random variables are exchangeable, but the reverse is not true. For example, sampling balls from an urn without replacement gives a distribution on sequences of balls that is exchangeable but not i.i.d. Although sampling without replacement is a common sampling mechanism, we will typically assume i.i.d. sampling in most contexts. We will discuss exchangeable random variables again in Section 14.1 where De Finetti's theorem is discussed.

The first two data types we have discussed have the property that the order that the data arrives does not matter in terms of the joint distribution of all the data. Only the vector of counts is important in both i.i.d. and exchangeable random variables. A third type of data to consider is when the order of the variables does matter. This type of data typically arises in the context of *time series* data, or *spatial* data.

A typical example of a time series model, where the resulting order of the observations matters, is a Markov chain (see Chapter 1). In a Markov chain, we have a sequence of random variables $X_1, \ldots, X_n$ with the same state space $[r]$ such that

$$P(X_j = i_j \mid X_{j-1} = i_{j-1}, \ldots, X_1 = i_1) = P(X_j = i_j \mid X_{j-1} = i_{j-1})$$

for $j = 2, \ldots, n$, $i_1, \ldots, i_n \in [r]^n$. Note that by applying the multiplication rule of conditional probability and using the first condition of being a Markov chain, we see that the joint probability factorizes as

$$\begin{aligned} &P(X_1 = i_1, \ldots, X_n = i_n) \\ &\quad = P(X_1 = i_1)P(X_2 = i_2 \mid X_1 = i_1) \cdots P(X_n = i_n \mid X_{n-1} = i_{n-1}). \end{aligned}$$

The further assumption of homogeneity requires that

$$P(X_j = i' \mid X_{j-1} = i) = P(X_2 = i' \mid X_1 = i)$$

for all $j = 3, \ldots, n$ and $i', i \in [r]$. From this we see that the sufficient statistics for a homogeneous Markov chain model are X_1 and the table of counts u such that

$$u_{i'i} = \#\{j : X_j = i', X_{j-1} = i\}.$$

Note that u counts the number of each type of transition that occurred in the sequences $X_1, X_2, \ldots, X_n$.

For i.i.d. samples, we typically observe many replicates from the same underlying model. When we use a time series or spatial model, the usual way data arrives is as a single sample from that model, whose length or size might not be a priori specified. For these models to be useful in practice, we need them to be specified with a not very large set of parameters, so that as the data grows (i.e., as the sequence gets longer) we have a hope of being able to estimate the parameters. Of course, it might be the situation that for a time-series or spatial model we have not just one sample, but i.i.d. data. For instance, in Chapter 1 during our discussion of the maximum likelihood estimation in a very short Markov chain, we analyzed the case where we received many data sets of the chain of length 3, where each was an i.i.d. sample from the same underlying distribution.

5.3. Parameter Estimation

Given a parametric statistical model and some data, a typical problem in statistics is to estimate some of, or all of, the parameters of the model based on the data. At this point we do not necessarily assume that the model accurately models the data. The problem of testing whether or not a model actually fits the data is the subject of the next section, on hypothesis testing.

Ideally, we would like a procedure which, as more and more data arrives, if the underlying distribution that generated the data comes from the model, the parameter estimate converges to the true underlying parameter. Such an estimator is called a consistent estimator.

Definition 5.3.1. Let $\mathcal{M}_\Theta$ be a parametric statistical model with parameter space Θ. A *parameter* of a statistical model is a function $s : \Theta \to \mathbb{R}$. An *estimator* of s is a function from the data space D to $\mathbb{R}$, $\hat{s} : D \to \mathbb{R}$. The estimator $\hat{s}$ is *consistent* if $\hat{s} \longrightarrow_p s$ as the sample size tends to infinity.

Among the simplest examples of estimators are the plug-in estimators. As the name suggests, a plug-in estimator is obtained by plugging in values obtained from the data to estimate parameters.

Example 5.3.2. As a simple example, consider the case of a binomial random variable, with $r + 1$ states, $0, 1, \ldots, r$. The model consists of all distributions of the form

$$\left\{ \left(\theta^r, \binom{r}{1}\theta^{r-1}(1-\theta), \ldots, (1-\theta)^r \right) : \theta \in [0,1] \right\}.$$

Under i.i.d. sampling, data consists of n repeated draws $X^{(1)}, \ldots, X^{(n)}$ from an underlying distribution p_θ in this model. The data is summarized by a

vector of counts $u = (u_0, \ldots, u_r)$, where $u_i = \#\{j : X^{(j)} = i\}$. We would like to estimate the parameter θ from the data of counts u. The value $p_\theta(0) = \theta^r$, hence, if we had a consistent estimator of $p_\theta(0)$, we could obtain a plug-in estimate for θ by extracting the rth root. For example, the formula

$$\sqrt[r]{\frac{1}{n}\sum_{i=1}^{n} 1_{x=0}(X^{(i)})} = \sqrt[r]{\frac{u_0}{n}}$$

gives a consistent plug-in estimator of the parameter θ.

Intuitively, the plug-in estimator from Example 5.3.2 is unlikely to be a very useful estimator, since it only uses very little information from the data to obtain an estimate of the parameter θ. When choosing a consistent plug-in estimator, we would generally like to use one whose variance rapidly tends to zero as $n \to \infty$. The estimator from Example 5.3.2 has high variance and so is an inefficient estimator of the parameter θ.

Another natural choice for an estimator is a *method of moments* estimator. The idea of the method of moments is to choose the probability distribution in the model whose moments match the empirical moments of the data.

Definition 5.3.3. Given a random vector $X \in \mathbb{R}^m$ and an integer vector $\alpha \in \mathbb{N}^m$, the *αth moment* is

$$\mu_\alpha = \mathrm{E}[X_1^{\alpha_1} \cdots X_m^{\alpha_m}].$$

Given i.i.d. data $X^{(1)}, \ldots, X^{(n)}$, their αth *empirical moment* is the estimate

$$\hat{\mu}_\alpha = \frac{1}{n}\sum_{i=1}^{m} (X_1^{(i)})^{\alpha_1} \cdots (X_m^{(i)})^{\alpha_m}.$$

So in method of moments estimation, we find formulas for some of the moments of the random vector $X \sim p_\theta \in \mathcal{M}_\Theta$ in terms of the parameter vector θ. If we calculate enough such moments, we can find a probability distribution in the model whose moments match the empirical moments. For many statistical models, formulas for the moments in terms of the parameters are given by polynomial or rational formulas in terms of the parameters. So finding the method of moments estimator will turn into the problem of solving a system of polynomial equations.

Example 5.3.4 (Binomial random variable)**.** Consider the example of the model of a binomial random variable from Exercise 5.3.2. Given a binomial random variable $X \sim \mathrm{Bin}(\theta, r)$, the first moment $\mathrm{E}[X] = r\theta$. The empirical first moment of $X^{(1)}, \ldots, X^{(n)}$ is the sample mean $\bar{X}$. Hence the method of moments estimate for θ in the binomial model is

$$\hat{\theta} = \tfrac{1}{r}\bar{X}.$$

The method of moments estimators can often lead to interesting algebraic problems [**AFS16**]. One potential drawback to the method of moments estimators is that the empirical higher moments tend to have high variability, so there can be a lot of noise in the estimates.

Among the many possible estimators of a parameter, one of the most frequently used is the maximum likelihood estimator (MLE). The MLE is one of the most commonly used estimators in practice, both for its intuitive appeal and for useful theoretical properties associated with it. In particular, it is usually a consistent estimator of the parameters and, with certain smoothness assumptions on the model, it is asymptotically normally distributed. We will return to these properties in Chapter 7.

Definition 5.3.5. Let D be data from some model with parameter space Θ. The *likelihood function*

$$L(\theta \mid D) := p_\theta(D)$$

in the case of discrete data and

$$L(\theta \mid D) := f_\theta(D)$$

in the case of continuous data. Here $p_\theta(D)$ is the probability of observing the data given the parameter θ in the discrete case, and $f_\theta(D)$ is the density function evaluated at the data in the continuous case. The *maximum likelihood estimate* (MLE) $\hat{\theta}$ is the maximizer of the likelihood function:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta \mid D).$$

Note that we consider the likelihood function as a function of θ with the data D fixed. This contrasts the interpretation of the probability distribution where the parameter is considered fixed and the random variable is the unknown (random) quantity.

In the case of i.i.d. sampling, so $D = X^{(1)}, \ldots, X^{(n)}$, the likelihood function factorizes as

$$L(\theta \mid D) = L(\theta \mid X^{(1)}, \ldots, X^{(n)}) = \prod_{i=1}^{n} L(\theta \mid X^{(i)}).$$

In the case of discrete data, this likelihood function is thus only a function of the vector of counts u, so that

$$L(\theta \mid X^{(1)}, \ldots, X^{(n)}) = \prod_{j} p_\theta(j)^{u_j}.$$

In the common setting in which we treat the vector of counts itself as the data, we need to multiply this quantity by an appropriate multinomial coefficient

$$L(\theta \mid u) = \binom{n}{u} \prod_j p_\theta(j)^{u_j},$$

which does not change the maximum likelihood estimate but will change the value of the likelihood function when evaluated at the maximizer.

It is common to replace the likelihood function with the *log-likelihood function*, which is defined as

$$\ell(\theta \mid D) = \log L(\theta \mid D).$$

In the case of i.i.d. data, this has the advantage of turning a product into a sum. Since the logarithm is a monotone function both the likelihood and log-likelihood have the same maximizer, which is the maximum likelihood estimate.

Example 5.3.6 (Maximum likelihood of a binomial random variable)**.** Consider the model of a binomial random variable with r trials. The probability $p_\theta(i) = \binom{r}{i}\theta^i(1-\theta)^{r-i}$. Given a vector of counts u, the log-likelihood function is

$$\begin{aligned} \ell(\theta, u) &= C + \sum_{i=0}^{r} u_i \log(\theta^i(1-\theta)^{r-i}) \\ &= C + \sum_{i=0}^{r} \left(iu_i \log\theta + (r-i)u_i \log(1-\theta)\right), \end{aligned}$$

where C is a constant involving logarithms of binomial coefficients but does not depend on the parameter θ. To calculate the maximum likelihood estimate, we differentiate the log-likelihood function with respect to θ and set it equal to zero, arriving at:

$$\frac{\sum_{i=0}^{r} iu_i}{\theta} - \frac{\sum_{i=0}^{r}(r-i)u_i}{1-\theta} = 0.$$

Hence, the maximum likelihood estimator, $\hat{\theta}$, is given by

$$\hat{\theta} = \frac{\sum_{i=0}^{r} iu_i}{rn}.$$

Note that $\frac{1}{n}\sum_{i=0}^{r} iu_i$ is the sample mean $\bar{X}$, so the maximum likelihood estimate of θ is the same as the method of moments estimate of θ.

The gradient of the log-likelihood function is called the *score function*. Since the gradient of a function is zero at a global maximum of a differentiable function, the equations obtained by setting the score function to zero are called the *score equations* or the critical equations. In many cases,

these equations are algebraic and the algebraic nature of the equations will be explored in later chapters.

For some of the most standard statistical models, there are well-known closed formulas for the maximum likelihood estimates of parameters.

Proposition 5.3.7. *For a multivariate normal random variable, the maximum likelihood estimates for the mean and covariance matrix are*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}, \qquad \hat{\Sigma} = \frac{1}{n}(X^{(i)} - \hat{\mu})(X^{(i)} - \hat{\mu})^T.$$

Proof. The log-likelihood function has the form

$$\log(\mu, \Sigma \mid D) = -\frac{1}{2}\sum_{i=1}^{n}\left(m\log(2\pi) + \log|\Sigma| + (X^{(i)} - \mu)^T\Sigma^{-1}(X^{(i)} - \mu)\right).$$

The *trace trick* is useful for rewriting this log-likelihood as

$$\begin{aligned}&\log(\mu, \Sigma \mid D)\\ &= -\frac{1}{2}\left(nm\log(2\pi) + n\log|\Sigma| + \operatorname{tr}(\sum_{i=1}^{n}((X^{(i)} - \mu)(X^{(i)} - \mu)^T)\Sigma^{-1})\right).\end{aligned}$$

Differentiating with respect to μ and setting equal to zero yields the equation

$$-\frac{1}{2}\sum_{i=1}^{n}\Sigma^{-1}(X^{(i)} - \mu) = 0.$$

From this we deduce that $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}$, the sample mean.

To find the maximum likelihood estimate for the covariance matrix we substitute $K = \Sigma^{-1}$ and differentiate the log-likelihood with respect to an entry of K. One makes use of the classical adjoint formula for the inverse of a matrix to see that

$$\frac{\partial}{\partial k_{ij}}\log|K| = (1 + \delta_{ij})\sigma_{ij},$$

where δ_{ij} is the Dirac delta function. Similarly,

$$\frac{\partial}{\partial k_{ij}}\operatorname{tr}(\sum_{i=1}^{n}((X^{(i)} - \mu)(X^{(i)} - \mu)^T)K) = n(1 + \delta_{ij})s_{ij},$$

where $S = \frac{1}{n}\sum_{i=1}^{n}((X^{(i)} - \mu)(X^{(i)} - \mu)^T)$ is the sample covariance matrix. Putting these pieces together with our solution that the maximum likelihood estimate of μ is $\hat{\mu}$ gives that $\hat{\Sigma} = \frac{1}{n}(X^{(i)} - \hat{\mu})(X^{(i)} - \hat{\mu})^T$. □

Proposition 5.3.8. *Let $\mathcal{M}_{1 \perp\!\!\!\perp 2}$ be the model of independence of two discrete random variables, with r_1 and r_2 states respectively. Let $u \in \mathbb{N}^{r_1 \times r_2}$ be the table of counts for this model obtained from i.i.d. samples from the model. Let $u_{i_1+} = \sum_{i_2} u_{i_1 i_2}$ and $u_{+i_2} = \sum_{i_1} u_{i_1 i_2}$ be the table marginals, and $n = \sum_{i_1, i_2} u_{i_1 i_2}$ the sample size. Then the maximum likelihood estimate for a distribution $p \in \mathcal{M}_{1 \perp\!\!\!\perp 2}$ given the data u is*

$$\hat{p}_{i_1 i_2} = \frac{u_{i_1+} u_{+i_2}}{n^2}.$$

Proof. A distribution $p \in \Delta_{\mathcal{R}}$ belongs to the independence model if and only if we can write $p_{i_1 i_2} = \alpha_{i_1} \beta_{i_2}$ for some $\alpha \in \Delta_{r_1 - 1}$ and $\beta \in \Delta_{r_2 - 1}$. We solve the likelihood equations in terms of α and β and use them to find $\hat{p}$. Given a table of counts u, the log-likelihood function for a discrete random variable has the form

$$\begin{aligned} \ell(\alpha, \beta \mid u) &= \sum_{i_1, i_2 \in \mathcal{R}} u_{i_1 i_2} \log p_{i_1 i_2} \\ &= \sum_{i_1, i_2 \in \mathcal{R}} u_{i_1 i_2} \log \alpha_{i_1} \beta_{i_2} \\ &= \sum_{i_1 \in [r_1]} u_{i_1+} \log \alpha_{i_1} + \sum_{i_2 \in [r_2]} u_{+i_2} \log \beta_{i_2}. \end{aligned}$$

From the last line, we see that we have two separate optimization problems that are independent of each other: maximizing with respect to α and maximizing with respect to β. Remembering that $\alpha_{r_1} = 1 - \sum_{i_1 = 1}^{r_1 - 1} \alpha_{i_1}$ and computing partial derivatives to optimize shows that $\hat{\alpha}_{i_1} = \frac{u_{i_1+}}{n}$. Similarly, $\hat{\beta}_{i_2} = \frac{u_{+i_2}}{n}$. □

Unlike the three preceding examples, most statistical models do not possess closed form expressions for their maximum likelihood estimates. The algebraic geometry of solving the critical equations will be discussed in later chapters, as will some numerical hill-climbing methods for approximating solutions.

5.4. Hypothesis Testing

A hypothesis test is a procedure given data for deciding whether or not a statistical hypothesis might be true. Typically, statistical hypotheses are phrased in terms of statistical models: for example, does the unknown distribution, about which we have collected i.i.d. samples, belong to a given model, or not. Note that statisticians are conservative so we rarely say that we accept the null hypothesis after performing a hypothesis test, only that

the test has not provided evidence against a particular hypothesis. The introduction of p-values provides more information about the hypothesis test besides simply whether or not we reject the null hypothesis.

As a simple example of a hypothesis test, suppose we have two random variables X and Y on the real line, and we suspect that they might be the same distribution. Hence the *null hypothesis* H_0 is that $X = Y$ and the *alternative hypothesis* H_a is that $X \neq Y$. This is stated as saying we are testing:

$$H_0 : X = Y \text{ vs. } H_a : X \neq Y.$$

We collect samples $X^{(1)}, \ldots, X^{(n_1)}$ and $Y^{(1)}, \ldots, Y^{(n_2)}$. If $X = Y$, then these two empirical distributions should be similar, in particular, they should have nearly the same mean. Of course, the sampling means

$$\overline{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X^{(i)}, \qquad \overline{\mu}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y^{(i)}$$

will not be exactly equal, and so we need a way to measure that these are close to each other up to the error induced by sampling.

The following *permutation test* (also called an *exact test*) gives a method, free from any distributional assumptions on the variables X and Y, to test the null hypothesis versus the alternative hypothesis:

For the given data $X^{(1)}, \ldots, X^{(n_1)}$ and $Y^{(1)}, \ldots, Y^{(n_2)}$ calculate the absolute difference of means $\overline{d} = |\overline{\mu}_1 - \overline{\mu}_2|$. Now, group all of these samples together into one large set of samples $Z_1, \ldots, Z_{n_1+n_2}$. For each partition $A|B$ of $[n_1 + n_2]$ into a set A of size n_1 and a set B of size n_2, we compute the difference of sample means

$$d(A, B) = \left| \frac{1}{n_1} \sum_{a \in A} Z_a - \frac{1}{n_2} \sum_{b \in B} Z_b \right|.$$

If $d(A, B)$ is larger than $\overline{d}$ for *few* partitions $A|B$, we are likely to conclude that $X \neq Y$, since the test seems to indicate that the sample means are farther apart, more than would be expected due to sampling error.

Of course, the crucial word "few" in the last sentence needs clarification. The proportion of partitions $A|B$ such that $d(A, B) \geq \overline{d}$ is called the *p-value* of the permutation test. A large p-value indicates that the permutation test provides no evidence against the null hypothesis, whereas a small p-value provides evidence against the null hypothesis. Fixing a level α, if the p-value is less than α we reject the null hypothesis at level α. Otherwise, we are unable to reject the null hypothesis. Tradition in statistics sets the usual rejection level at $\alpha = .05$.

We will discuss a number of other hypothesis tests throughout the book, including the likelihood ratio test and Fisher's exact test. In each case, we must define the p-value in a way that is appropriate to the given null hypothesis. This is stated most generally as follows. Most generally, when testing a specific hypothesis, we have a test statistic T which we evaluate on the data, which evaluates to zero if the data exactly fits the null hypothesis H_0, and which increases moving away from the null hypothesis. The p-value is then the probability under the null hypothesis that the test statistic has a larger value than the given observed value. In the case of the permutation test, we calculate the probability that a random partition of the data has differences of means above the value calculated from the data. Under the null hypothesis $X = Y$, so the sample mean of each part of the partition converges to the common mean by the law of large numbers, so the test statistic is zero for perfect data that satisfies the null hypothesis.

Example 5.4.1 (Jackal mandible lengths)**.** The following data set gives the mandible lengths in millimeters for 10 male and 10 female golden jackals in the collection of the British museum (the data set is taken from [**Man07**]).

Males
120 107 110 116 114 111 113 117 114 112
Females
110 111 107 108 110 105 107 106 111 111

A basic question about this data is whether the distribution of mandible lengths is the same or different for males versus females of this species. A quick look at the data suggests that the males tend to have larger mandibles than females, but is there enough data that we can really conclude this? Let μ_m denote the mean of the male population and let μ_f denote the mean of the female population. According to our discussion above, we can perform the hypothesis test to compare

$$H_0 : \mu_m = \mu_f \quad \text{vs.} \quad H_a : \mu_m \neq \mu_f.$$

An exact calculation yields the p-value of 0.00333413, suggesting we can reject the null hypothesis that the means are equal. We run the following code in `R` [**R C16**] to perform the hypothesis test.

```
m <- c(120, 107, 110, 116, 114, 111, 113, 117, 114, 112)
f <- c(110, 111, 107, 108, 110, 105, 107, 106, 111, 111)
observed_tstat <- abs(mean(m) - mean(f))
resampled_tstats <- replicate(1e6, {
  ndcs <- sample(length(c(m, f)), length(m))
  abs(mean(c(m, f)[ndcs]) - mean(c(m, f)[-ndcs]))
})
mean(resampled_tstats >= observed_tstat)
```

In this case the p-value is approximated by taking random samples, rather than computing all partitions. The approximate p-value on one run with 1000000 samples was 0.003422.

An important remark about hypothesis testing is that we rarely calculate the p-value exactly. Even in the permutation test, it is not possible to give a closed form formula for the p-value nor is it possible for large sample sizes to compute the sum over all partitions $A|B$. In that case, we can randomly sample partitions to approximate the p-value. That is, let $A_1|B_1, \ldots, A_k|B_k$ be random partitions of the data of size where A_i has n_1 elements and B_i has n_2 elements. An approximation of the p-value of the permutation test is

$$\frac{1}{k}\sum_{i=1}^{k} 1_{d(A_i|B_i)>\bar{d}},$$

which converges to the exact p-value as k tends to infinity by the law of large numbers.

Example 5.4.2 (Calculator random digits)**.** A sequence of 300 pseudo-random digits was generated on a Casio fx-3600p calculator [**DDt87**]. The following table displays the number of times that each digit occurred.

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	25	28	29	35	35	31	27	33	32	35

Our null hypothesis in this case is that the digits are uniformly random, and we would like to perform a hypothesis test to determine if we have evidence against this hypothesis. There are a number of ways to test this hypothesis. All involve comparing a test statistic evaluated on the data to the same test statistic on random samples of data generated under the null hypothesis. Under the null hypothesis, the expected frequency of each bit is $\hat{u}_i = 30$. A natural choice is to use the "distance" between the expected frequency vector $\hat{u}$ and a data vector u as a natural test statistic. The distance could be a standard Euclidean distance or some other measure of the dissimilarity of the observation. For example, a distance frequently used in hypothesis tests is the Pearson's X^2 statistic, which in this case is

$$X^2(u) = \sum_{i=0}^{9} \frac{(\hat{u}_i - u_i)^2}{\hat{u}_i}.$$

Note that because $\hat{u}_i = 30$ is constant in this example, both hypothesis tests will produce the same result. We generated a random sample of 100000 data sets v from the multinomial distribution with $n = 300$ and $p_i = 1/10$, and compared $X^2(v)$ to $X^2(u)$. The following `R` [**R C16**] code performs this hypothesis test.

```
data <- c (25, 28, 29, 35, 35, 31, 27, 33, 32, 25 )
compute_pearson_x2 <- function(o, e) sum((o - e)^2 / e)
(observed_x2 <- compute_pearson_x2(data, 300 * rep(1/10, 10)))
samples <- rmultinom(1e6, 300, prob = rep(1/10, 10))
sampled_x2s <- apply(samples, 2, compute_pearson_x2,
    e = 300 * rep(1/10, 10))
mean(sampled_x2s >=  observed_x2)
```

We found a p-value of approximately 0.896, which suggests that we cannot reject the null hypothesis that the samples are uniformly random.

In other contexts (e.g., the likelihood ratio test), a standard strategy is to assume that the sample size tends to infinity. In this case, it is possible to derive the asymptotic distribution of the test statistic. For smooth models, those asymptotic distributions are familiar distributions of probability theory like the chi-square distribution, and this allows us to compute asymptotic p-values. However, for models that are not smooth, nonstandard limiting distributions for the asymptotic p-values can arise. This is discussed in more detail in Section 7.4, but we illustrate the basic idea first.

Example 5.4.3 (Calculator random digits)**.** Suppose that we have i.i.d. samples from some unknown distribution $q \in \Delta_{r-1}$ and we want to test the hypothesis that the distribution is equal to some fixed distribution p. Assuming that all coordinates of p are nonzero, it can be shown that if the null hypothesis is true, as $n \to \infty$, the distribution of the Pearson X^2 statistic converges to a chi-square distribution with $r-1$ degrees of freedom. This yields the basis of an asymptotic test, since we can calculate the empirical X^2 value of our data vector u and determine the asymptotic p-value

$$P(Q > X^2(u)),$$

where $Q \sim \chi^2_{r-1}$. For the calculator random digit data set from Example 5.4.2 this is straightforward to check in `R` [**R C16**] with the command:

```
chisq.test(data,NULL, rep(1/10, 10))
```

This returns the asymptotic p-value of 0.893 which is very close to the exact test p-value calculated in Example 5.4.2.

5.5. Bayesian Statistics

Likelihood inference, hypothesis tests, and p-values are all standard tools from frequentist statistics. Frequentist statistics is typically concerned with the long-term behavior of estimators, in particular, how they behave as more and more data comes in for analyzing a problem. In frequentist statistics approaches to parametric statistics models, it is assumed that there is a

true underlying parameter vector, and we seek to find information about that parameter by collecting data.

A complementary approach in statistics is called Bayesian statistics. In Bayesian statistics, the point of collecting data is to update our beliefs about unknown parameters. Typically we have prior beliefs about the parameters, represented by a *prior distribution* π on the parameter space Θ. Data D allows us to compute an update on the distribution on Θ which gives us the *posterior distribution*. The posterior distribution is computed via the formula

$$p(\theta \mid D) = \frac{p(D \mid \theta)\pi(\theta)}{\int_\theta p(D \mid \theta)\pi(\theta)\, d\theta}.$$

Note that the quantity in the denominator $\int_\theta p(D \mid \theta)\pi(\theta)\, d\theta$ is the marginal probability of the data, and is called the *marginal likelihood integral*. This application of Bayes' theorem in the calculation of the posterior distribution is the reason for referring to these methods as Bayesian.

As with many statistical methods, the main difficulties in using Bayesian methods in practice are computational. There are usually no closed form formulas for the posterior distribution and they must be approximated, often using Monte Carlo algorithms. Similarly, the marginal likelihood is usually not easy to calculate and must itself be approximated.

Bayesian methods produce a distribution over the parameter space, which can be thought of as our posterior belief about the parameter. When the parameter space is high dimensional, it can be useful or necessary to compute a point estimate of the parameter. A natural choice in Bayesian statistics is the *maximum a posteriori* estimate or MAP estimate. This is simply the parameter value that maximizes the posterior distribution.

Definition 5.5.1. Let $\mathcal{M}_\Theta$ be a parametric statistical model, π a prior distribution over Θ, D a collection of data, and $p(\theta \mid D)$ the posterior distribution of θ given the data. The *maximum a posteriori estimator* of θ (MAP) is

$$\hat{\theta} = \arg\max_{\theta \in \Theta} p(\theta \mid D).$$

Note the similarity of the definition of the MAP estimate to the MLE. These estimates can be vastly different depending on the underlying prior distribution, although increasing amounts of data usually (as one would hope) eliminate the dependence on the prior in the limit. One advantage of the MAP estimate is that it is not necessary to compute the marginal likelihood integral $\int_\theta p(D \mid \theta)\pi(\theta)\, d\theta$ to obtain it.

Example 5.5.2 (Prior for a discrete random variable)**.** Consider a discrete random variable with r states. We will consider a natural prior distribution,

the *Dirichlet distribution*, which is commonly used as a prior distribution on the parameter space in models with discrete random variables.

Let $\alpha \in \mathbb{R}^r_{>0}$ be a nonnegative real parameter vector. A random variable $X \in \Delta_{r-1}$ has Dirichlet distribution with parameter α (denoted $X \sim \mathrm{Dir}(\alpha)$) if the density function $f_\alpha(x) \propto \prod_{i=1}^r x_i^{\alpha_i - 1}$. The proportionality constant $B(\alpha)$ is the multivariate beta function

$$B(\alpha) = \frac{\prod_{i=1}^r \Gamma(\alpha_i)}{\Gamma(\alpha_1 + \cdots + \alpha_r)}.$$

If $\alpha_i > 1$ for all i, then Dirichlet distribution has peak (both mean and mode) at $\alpha/\|\alpha\|_1$, and the peak becomes narrower for larger values of $\|\alpha\|_1$. As a prior distribution for the multinomial distribution, the Dirichlet distribution gives concentration around the mode $\alpha/\|\alpha\|_1$ with $\|\alpha\|_1$ giving a measure of certainty about the mode.

Now suppose we receive n i.i.d. samples from a discrete random variable Y with state space $[r]$. Since we have i.i.d. samples, we can use the vector of counts u to record the data since it is a sufficient statistic. Given a sample size n, the conditional distribution of u given X is multinomial, so the joint density of u and X is

$$f(x, u) = B(\alpha) \prod_{i=1}^r x_i^{\alpha_i - 1} \times \binom{n}{u} \prod_{i=1}^r x_i^{u_i}.$$

Whatever the marginal distribution of u is, we see that the conditional density of X given u is

$$f(x \mid u) \propto \prod_{i=1}^r x_i^{u_i + \alpha_i - 1}.$$

That is, the posterior distribution of the parameter given the data u is a Dirichlet distribution with parameter $\alpha + u$. So assuming that all $u_i > 0$, the maximum a posterior estimate of the parameter given the data u is the vector $(u + \alpha)/\|u + \alpha\|_1$.

Now suppose that there is some true underlying distribution p which generates the samples. As the sample size n gets large, the law of large numbers shows that u/n will converge to p. Since α was some a priori fixed value, this will imply that $(u + \alpha)/\|u + \alpha\|_1$ also converges to p as n goes to infinity. Hence, regardless of what particular α we chose at the start of our analysis, eventually we end up with a good estimate of the true underlying parameter p.

An interesting fact about this example is that the posterior distribution has the same functional form as the prior distribution. This does not always happen, but happens here because the Dirichlet distribution is the *conjugate prior* to the multinomial distribution.

Bayesian and frequentist statistics are sometimes seen as being at odds with each other. We do not propose to "take sides" in the present work, rather seeing them both as useful methods which make sense for different situations, and, from the standpoint of algebraic statistics, both are rich with connections to algebraic geometry. Bayesian approaches make sense when we wish to incorporate our prior beliefs about parameters into our analysis of data. Frequentist perspectives make sense when we have lots of data arriving and we can analyze asymptotic properties of estimators. We will see specific discussion of Bayesian methods in the chapters on model selection and parametric inference.

5.6. Exercises

Exercise 5.1. An exponential random variable X with parameter $\lambda > 0$ has density function

$$f_\lambda(x) = \begin{cases} \lambda \exp(-\lambda x), & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let $X^{(1)}, \ldots, X^{(n)}$ be i.i.d. exponential random variables. What is the maximum likelihood estimate of λ? What is the method of moments estimate of λ?

Exercise 5.2. Consider the two lists of, respectively, 15 and 12 real numbers:

$$\begin{aligned} A &= \{1.04, -0.12, 0.22, 0.08, -0.06, -0.11, 0.06, 1.06, \\ &\quad -1.73, -0.58, 0.67, 0.06, 1.10, -0.93, 0.34\}, \end{aligned}$$

$$\begin{aligned} B &= \{3.36, 1.00, 1.15, -0.57, 0.84, 0.18, 0.28, 0.52, \\ &\quad 2.44, -4.22, 2.95, -1.83\}. \end{aligned}$$

The elements of A are i.i.d. samples from a probability distribution p and the elements of B are i.i.d. samples from a probability distribution q. Do you think $p = q$? Formulate and apply a hypothesis test, and calculate p-values.

Exercise 5.3. A gamma random variable with parameters $\alpha > 0, \beta > 0$ has density function

$$f(x \mid \alpha, \beta) = \begin{cases} \Gamma(\alpha)^{-1} \beta^\alpha x^{\alpha-1} \exp(-\beta x), & x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx$ denotes the gamma function. Suppose that we use the gamma distribution with parameters α, β as a prior on λ in an exponential random variable. We observe i.i.d. samples $X^{(1)}, \ldots, X^{(n)}$. What is the posterior distribution of λ given the data?