APPENDIX B

ELEMENTS OF MATHEMATICAL STATISTICS

B.1 STATISTICAL INFERENCE

Statistics deals with the gathering, summarization, analysis, and interpretation of data. The two main branches of statistics are:

1. *Classical statistics*: Here the data object $\mathbf{x}$ is viewed as the outcome of a random object $\mathbf{X}$ described by a probabilistic model — usually the model is specified up to a (multidimensional) parameter; that is, $\mathbf{X} \sim f(\cdot; \boldsymbol{\theta})$ for some $\boldsymbol{\theta}$. The statistical inference is then purely concerned with the model and in particular with the parameter $\boldsymbol{\theta}$. For example, on the basis of the data one may wish to
   (a) estimate the parameter,
   (b) perform statistical tests on the parameter, or
   (c) validate the model.
2. *Bayesian statistics*: In this approach the model parameter $\boldsymbol{\theta}$ is itself random: $\boldsymbol{\theta} \sim f(\boldsymbol{\theta})$. Bayes' formula $f(\boldsymbol{\theta} \mid \mathbf{x}) \propto f(\mathbf{x} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})$ is used to update the distribution of the parameter based on the observed data $\mathbf{x}$.

Mathematical statistics uses probability theory and other branches of mathematics to study data from a purely mathematical standpoint.

Handbook of Monte Carlo Methods. By D. P. Kroese, T. Taimre, Z. I. Botev

B.1.1 Classical Models

Let $\mathbf{x}$ represent the observed data, viewed as the outcome of the random data $\mathbf{X}$. For example, $\mathbf{X}$ could be a random vector $(X_1, \ldots, X_n)^\top$. A real- or vector-valued function of the data is called a **statistic**. For example, if $\mathbf{X} = (X_1, \ldots, X_n)^\top$, then the sample mean $T = T(\mathbf{X}) = (X_1 + \cdots + X_n)/n$ is one such statistic. It is customary to use the *same* letter for both the function T and the random variable $T(\mathbf{X})$. We write T for statistics taking values in $\mathbb{R}$ and $\mathbf{T}$ for statistics taking values in $\mathbb{R}^d$ for some $d \geqslant 2$. It is important that a statistic be *computable*; that is, it cannot depend on any unknown parameters.

We summarize some classical models for data.

B.1.1.1 Iid Sample The data $X_1, \ldots, X_n$ are independent and identically distributed:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Dist}\,,$$

according to some known or unknown distribution Dist. Often the sampling distribution is specified up to an unknown parameter $\boldsymbol{\theta}$, with $\boldsymbol{\theta} \in \Theta$. An iid sample is often called a **random sample** in the statistics literature. Note that the word "sample" can refer to both a collection of random variables and to a single random variable. It should be clear from the context which meaning is being used.

A standard model for data is:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)\,,$$

in which case $\boldsymbol{\theta} = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times \mathbb{R}_+$.

B.1.1.2 Analysis of Variance In a one-way analysis of variance the objective is to compare the means $\mu_1, \ldots, \mu_k$ of k independent groups (or **levels**) of normal **responses**, all responses having the same variance σ^2. Specifically, denoting the i-th response at level j by X_{ij}, $i = 1, \ldots, n_j$, $j = 1, \ldots, k$, where n_j is the sample size of the j-th group, the model is

$$X_{ij} = \mu_j + \varepsilon_{ij}, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, k\,, \quad \{\varepsilon_{ij}\} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)\,,$$

or, equivalently,

$$X_{ij} \sim \mathsf{N}(\mu_j, \sigma^2), \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, k\,, \text{ independently}\,.$$

B.1.1.3 Regression Regression models are used to describe functional relationships between **explanatory** variables and **response** variables. In a **linear regression** model, the relationship is linear. Defining Y_i as the i-th response variable and x_i as the fixed (that is, deterministic) i-th explanatory variable, a standard model is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \quad \{\varepsilon_i\} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2) \tag{B.1}$$

for certain *unknown* parameters β_0, β_1, and σ^2. The line

$$y = \beta_0 + \beta_1\, x \tag{B.2}$$

is called the **regression line**. By replacing it with a general curve $y = g(x; \boldsymbol{\theta})$ one obtains a general regression model. For example, $y = \beta_0 + \beta_1 x + \beta_2 x^2$ gives a **quadratic regression** model, and $y = \mathbf{x}^\top \boldsymbol{\beta}$, where $\mathbf{x}$ is a multidimensional explanatory variable and $\boldsymbol{\beta}$ a parameter vector, gives a **multiple linear regression** model.

B.1.1.4 Linear Model A data vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is said to satisfy a **linear model** if

$$\mathbf{Y} = A\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \quad \boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2 I) \tag{B.3}$$

for some $n \times k$ matrix A (the **design matrix**), a k-dimensional vector of **parameters** $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$, and a vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ of iid $\mathsf{N}(0, \sigma^2)$-distributed **error** terms. The analysis of variance and regression models are special cases.

For an outcome $\mathbf{y}$, the **least squares method** can be used to fit the model to the data. In particular, the optimal $\widehat{\boldsymbol{\beta}}$ is chosen such that the Euclidean distance between $\mathbf{y}$ and $A\widehat{\boldsymbol{\beta}}$ is minimal. Equivalently, $\widehat{\boldsymbol{\beta}}$ is the solution to

$$\nabla_{\boldsymbol{\beta}} \|\mathbf{y} - A\boldsymbol{\beta}\|^2 = A^\top(\mathbf{y} - A\boldsymbol{\beta}) = \mathbf{0}\,.$$

These linear set of equations are called the **normal equations**. Therefore, if $A^\top A$ is *invertible* (A can always be chosen such that this is the case), then

$$\widehat{\boldsymbol{\beta}} = (A^\top A)^{-1} A^\top \mathbf{y}\,.$$

In practice, we never compute the inverse $(A^\top A)^{-1}$, but compute $\widehat{\boldsymbol{\beta}}$ from the normal equations using, for example, Gaussian elimination. Geometrically, $\widehat{\boldsymbol{\beta}}$ is the projection of $\mathbf{y}$ onto the subspace spanned by the columns of A. Moreover, it is not difficult to show that $\widehat{\boldsymbol{\beta}}$ is precisely the maximum likelihood estimate of $\boldsymbol{\beta}$ (see Section B.2.1).

B.1.2 Sufficient Statistics

A **sufficient** statistic for a parameter (vector) $\boldsymbol{\theta}$ is a statistic that captures all the information about $\boldsymbol{\theta}$ contained in the data. This means that we can *summarize* the data via a sufficient statistic, sometimes giving a tremendous reduction in data.

If $\mathbf{T}(\mathbf{X})$ is a sufficient statistic for $\boldsymbol{\theta}$, then any inference about $\boldsymbol{\theta}$ depends on the sample $\mathbf{X} = (X_1, \ldots, X_n)^\top$ only through the value $\mathbf{T}(\mathbf{X})$. More precisely, a statistic $\mathbf{T}(\mathbf{X})$ is called a **sufficient statistic** for $\boldsymbol{\theta}$ if the conditional distribution of $\mathbf{X}$ given $\mathbf{T}(\mathbf{X})$ does not depend on $\boldsymbol{\theta}$. The workhorse for establishing sufficiency is the following theorem. A proof can be found, for example, in [4].

Theorem B.1.1 (Factorization Theorem) *Let $f(\mathbf{x}; \boldsymbol{\theta})$ denote the joint pdf of the data $\mathbf{X} = (X_1, \ldots, X_n)^\top$. A statistic $\mathbf{T}(\mathbf{X})$ is sufficient for $\boldsymbol{\theta}$ if and only if there exist functions $g(\mathbf{t}, \boldsymbol{\theta})$ and $h(\mathbf{x})$ such that for all data points $\mathbf{x}$ and all parameter points $\boldsymbol{\theta}$,*

$$f(\mathbf{x}; \boldsymbol{\theta}) = g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta})\, h(\mathbf{x})\,. \tag{B.4}$$

■ **EXAMPLE B.1 (Sufficient Statistics for Exponential Families)**

Sufficiency is particularly easy to establish for exponential families. Suppose that ☞ 701 $X_1, \ldots, X_n$ is an iid sample from the exponential family with pdf

$$\mathring{f}(x; \boldsymbol{\theta}) = c(\boldsymbol{\theta})\, \mathrm{e}^{\sum_{i=1}^m \eta_i(\boldsymbol{\theta})\, t_i(x)}\, \mathring{h}(x)\,,$$

where $\{\eta_i\}$ are linearly independent. The pdf of $\mathbf{X} = (X_1, \ldots, X_n)^\top$ is therefore

$$f(\mathbf{x}; \boldsymbol{\theta}) = \underbrace{c(\boldsymbol{\theta})^n\, \mathrm{e}^{\sum_{i=1}^m \eta_i(\boldsymbol{\theta}) \sum_{k=1}^n t_i(x_k)}}_{g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta})} \underbrace{\prod_{k=1}^n \mathring{h}(x_k)}_{h(\mathbf{x})}\,.$$

A direct consequence of the factorization theorem is that

$$\mathbf{T}(\mathbf{X}) = \left(\sum_{k=1}^{n} t_1(X_k), \ldots, \sum_{k=1}^{n} t_m(X_k) \right)$$

is a sufficient statistic for $\boldsymbol{\theta}$.

■ **EXAMPLE B.2 (Sufficient Statistics for the Normal Distribution)**

As a particular instance of Example B.1, consider the $\mathsf{N}(\mu, \sigma^2)$ case. Thus, $\boldsymbol{\theta} =$ ☞ 702 (μ, σ^2), and from Table D.1 it follows that a sufficient statistic for $\boldsymbol{\theta}$ is $\mathbf{T} = (T_1, T_2)$, with $T_1 = \sum_{k=1}^{n} X_k$ and $T_2 = \sum_{k=1}^{n} X_k^2$. This means that for the standard data model, the data can be summarized via only T_1 and T_2.

Moreover, it is not difficult to see that any 1-to-1 function of a sufficient statistic again yields a sufficient statistic. Hence, the sample mean $\widetilde{T}_1 = \bar{X}$ and the sample variance

$$\widetilde{T}_2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{k=1}^{n} X_k^2 - n\bar{X}^2 \right)$$

also form a pair of sufficient statistics, because the mapping

$$\widetilde{T}_1 = \frac{T_1}{n} \quad \text{and} \quad \widetilde{T}_2 = \frac{1}{n-1} \left(T_2 - T_1^2/n \right)$$

is invertible.

B.1.3 Estimation

Suppose the distribution of the data $\mathbf{X}$ is completely specified up to an unknown parameter vector $\boldsymbol{\theta}$. The aim is to estimate $\boldsymbol{\theta}$ on the basis of the observed data $\mathbf{x}$ only. (An alternative could be to estimate $\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\theta})$ for some vector-valued function $\mathbf{g}$.) Specifically, the goal is to find an **estimator** $\mathbf{T} = \mathbf{T}(\mathbf{X})$ that is close to the unknown $\boldsymbol{\theta}$. The corresponding outcome $\mathbf{t} = \mathbf{T}(\mathbf{x})$ is the **estimate** of $\boldsymbol{\theta}$. The **bias** of an estimator $\mathbf{T}$ of $\boldsymbol{\theta}$ is defined as $\mathbf{T} - \boldsymbol{\theta}$. An estimator $\mathbf{T}$ of $\boldsymbol{\theta}$ is said to be **unbiased** if $\mathbb{E}_{\boldsymbol{\theta}} \mathbf{T} = \boldsymbol{\theta}$. We often write $\widehat{\boldsymbol{\theta}}$ for both an estimator and estimate of $\boldsymbol{\theta}$. The **mean square error** (MSE) of a real-valued estimator T is defined as

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

An estimator T_1 is said to be more **efficient** than an estimator T_2 if the MSE of T_1 is smaller than the MSE of T_2. The MSE can be written as the sum

$$\mathrm{MSE} = (\mathbb{E}_\theta T - \theta)^2 + \mathrm{Var}_\theta(T) .$$

The first term measures the unbiasedness and the second is the variance of the estimator. In particular, for an *unbiased* estimator the MSE of an estimator is simply equal to its variance.

For simulation purposes it is often important to include the *running time* of the estimator in efficiency comparisons. One way to compare two unbiased estimators ☞ 383 T_1 and T_2 is to compare their **relative time variance products**,

$$\frac{\tau_i \operatorname{Var}(T_i)}{(\mathbb{E}T_i)^2}\,, \quad i = 1, 2\,, \tag{B.5}$$

where τ_1 and τ_2 are the times required to calculate the estimators T_1 and T_2, respectively. In this scheme, T_1 is considered more efficient than T_2 if its relative time variance product is smaller.

Two systematic approaches for constructing sound estimators are:

- the maximum likelihood method; see Section B.2.1,
- the method of moments, discussed next.

B.1.3.1 Method of Moments Suppose $x_1, \ldots, x_n$ are outcomes from an iid sample $X_1, \ldots, X_n \sim_{\text{iid}} f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is unknown. The moments of the sampling distribution can be easily estimated. Namely, if $X \sim f(x; \boldsymbol{\theta})$, then the r-th moment of X, that is, $\mu_r(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} X^r$ (assuming it exists), can be estimated through the **sample r-th moment**

$$m_r = \frac{1}{n} \sum_{i=1}^{n} x_i^r\,.$$

The **method of moments** procedure involves choosing the estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ such that each of the first k sample and true moments are matched:

$$m_r = \mu_r(\widehat{\boldsymbol{\theta}}), \quad r = 1, 2, \ldots, k\,.$$

In general, this set of equations is nonlinear, and so its solution often has to be found numerically.

■ **EXAMPLE B.3 (Sample Mean and Sample Variance)**

Suppose the data is given by $\mathbf{X} = (X_1, \ldots, X_n)^\top$, where the $\{X_i\}$ form an iid sample from a general distribution with mean μ and variance $\sigma^2 < \infty$. Matching the first two moments gives the set of equations

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} x_i &= \mu\,, \\ \frac{1}{n} \sum_{i=1}^{n} x_i^2 &= \mu^2 + \sigma^2\,. \end{aligned}$$

The method of moments estimates for μ and σ^2 are therefore the **sample mean**

$$\widehat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i\,, \tag{B.6}$$

and

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2\,. \tag{B.7}$$

The corresponding estimator for μ, $\bar{X}$, is unbiased. However, the estimator for σ^2 is biased: $\mathbb{E}\widehat{\sigma^2} = \sigma^2(n-1)/n$. An unbiased estimator is the **sample variance**

$$S^2 = \widehat{\sigma^2}\frac{n}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 .$$

The square root of the sample variance $S = \sqrt{S^2}$ is called the **sample standard deviation.**

B.1.3.2 Confidence Interval An essential part in any estimation procedure is to provide an assessment of the *accuracy* of the estimate. Indeed, without information on its accuracy the estimate itself would be meaningless. Confidence intervals (sometimes called **interval estimates**) provide a precise way of describing the uncertainty in the estimate.

Let $X_1, \ldots, X_n$ be random variables with a joint distribution depending on a parameter $\theta \in \Theta$. Let $T_1 < T_2$ be statistics (thus, $T_i = T_i(X_1, \ldots, X_n)$, $i = 1, 2$ are functions of the data, but not of θ).

1. The random interval (T_1, T_2) is called a **stochastic confidence interval** for θ with confidence $1 - \alpha$ if

$$\mathbb{P}_\theta(T_1 < \theta < T_2) \geqslant 1 - \alpha \quad \text{for all } \theta \in \Theta . \tag{B.8}$$

2. If t_1 and t_2 are the observed values of T_1 and T_2, then the interval (t_1, t_2) is called the **(numerical) confidence interval** for θ with confidence $1 - \alpha$ for every $\theta \in \Theta$.

3. If (B.8) only holds approximately, the interval is called an **approximate confidence interval.**

4. The probability $\mathbb{P}_\theta(T_1 < \theta < T_2)$ is called the **coverage probability**. For a $1 - \alpha$ confidence interval, it must be at least $1 - \alpha$.

For multidimensional parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ the stochastic confidence interval is replaced with a stochastic **confidence region** $\mathscr{C} \subset \mathbb{R}^d$ such that $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in \mathscr{C}) \geqslant 1-\alpha$ for all $\boldsymbol{\theta}$.

The systematic construction of (approximate) confidence intervals often involves *likelihood methods*, see Section B.2. Another approach is to use the *bootstrap* ☞ 331 *method*, see Section 8.6. The analogue of a confidence interval in Bayesian analysis is called a **credible interval**; see Section B.3.

■ **EXAMPLE B.4 (Approximate Confidence Interval for the Mean)**

Let $X_1, X_2, \ldots, X_n$ be an iid sample from a distribution with mean μ and variance $\sigma^2 < \infty$ (both assumed to be unknown). By the central limit theorem and the law of large numbers,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \overset{\text{approx.}}{\sim} \mathsf{N}(0, 1) ,$$

for large n, where S is the sample standard deviation. Rearranging the approximate equality $\mathbb{P}(|T| \leqslant z_{1-\alpha/2}) \approx 1-\alpha$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution, yields

$$\mathbb{P}\left(\bar{X} - z_{1-\alpha/2}\frac{S}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right) \approx 1-\alpha\,,$$

so that

$$\left(\bar{X} - z_{1-\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right),\ \text{abbreviated as } \bar{X} \pm z_{1-\alpha/2}\frac{S}{\sqrt{n}}\,, \qquad \text{(B.9)}$$

is an approximate stochastic $1-\alpha$ confidence interval for μ.

Since (B.9) is an asymptotic result only, care should be taken when applying it to cases where the sample size is small or moderate and the sampling distribution is heavily skewed. For one- and two-sample normal (Gaussian) data Table B.1 provides *exact* confidence intervals for various parameters. The model for the two-sample data is $X_1, \ldots, X_m \sim_{\text{iid}} \mathsf{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_n \sim_{\text{iid}} \mathsf{N}(\mu_Y, \sigma_Y^2)$, where $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are independent. All parameters are assumed to be unknown.

Table B.1 Exact confidence intervals for normal data with unknown mean and variance.

Parameter	Exact $1-\alpha$ confidence interval	Condition
μ_X	$\bar{X} \pm t_{m-1;1-\alpha/2}\dfrac{S_X}{\sqrt{m}}$	
σ_X^2	$\left(\dfrac{(m-1)S_X^2}{\chi^2_{m-1;1-\alpha/2}},\ \dfrac{(m-1)S_X^2}{\chi^2_{m-1;\alpha/2}}\right)$	
$\mu_X - \mu_Y$	$\bar{X} - \bar{Y} \pm t_{m+n-2;1-\alpha/2}\, S_p\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}$	$\sigma_X^2 = \sigma_Y^2$
σ_X^2/σ_Y^2	$\left(F_{n-1,m-1;\alpha/2}\dfrac{S_X^2}{S_Y^2},\ F_{n-1,m-1;1-\alpha/2}\dfrac{S_X^2}{S_Y^2}\right)$	

Here $S_p^2 = \frac{\sum_{i=1}^m (X_i-\bar{X})^2 + \sum_{j=1}^n (Y_j-\bar{Y})^2}{m+n-2}$ is the **pooled sample variance**, $t_{n;\gamma}$ is the γ quantile of the t_n distribution, and $F_{m,n;\gamma}$ is the γ quantile of the $\mathsf{F}(m,n)$ distribution.

For one- and two-sample data from the binomial distribution, described by the model $X \sim \mathsf{Bin}(m, p_X)$ and $Y \sim \mathsf{Bin}(n, p_Y)$ independently, approximate $(1-\alpha)$ confidence intervals for p_X and $p_X - p_Y$ are given in Table B.2. We use the notation $\widehat{p}_X = X/m$ and $\widehat{p}_Y = Y/n$.

Table B.2 Approximate confidence intervals for binomial data.

Parameter	Approximate $1-\alpha$ confidence interval
p_X	$\widehat{p}_X \pm z_{1-\alpha/2}\sqrt{\dfrac{\widehat{p}_X(1-\widehat{p}_X)}{m}}$
$p_X - p_Y$	$\widehat{p}_X - \widehat{p}_Y \pm z_{1-\alpha/2}\sqrt{\dfrac{\widehat{p}_X(1-\widehat{p}_X)}{m} + \dfrac{\widehat{p}_Y(1-\widehat{p}_Y)}{n}}$

Finally, Table B.3 gives exact confidence intervals for various parameters of the linear regression model (B.1).

Table B.3 Exact confidence intervals for normal regression data.

Parameter	Exact $1-\alpha$ confidence interval
β_0	$\widehat{\beta}_0 \pm t_{n-2;1-\alpha/2}\,\widetilde{S}\sqrt{\dfrac{\sum_{i=1}^n x_i^2}{nS_{xx}}}$
β_1	$\widehat{\beta}_1 \pm t_{n-2;1-\alpha/2}\,\widetilde{S}\sqrt{\dfrac{1}{S_{xx}}}$
$\beta_0 + \beta_1 x$	$\widehat{Y} \pm t_{n-2;1-\alpha/2}\,\widetilde{S}\sqrt{\dfrac{1}{n} + \dfrac{(x-\bar{x})^2}{S_{xx}}}$
σ^2	$\left(\dfrac{(n-2)\widetilde{S}^2}{\chi^2_{n-2;1-\alpha/2}}, \dfrac{(n-2)\widetilde{S}^2}{\chi^2_{n-2;\alpha/2}}\right)$

Here $\widehat{\beta}_1 = \frac{S_{xY}}{S_{xx}}$, $\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1\bar{x}$, $\widetilde{S}^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$, $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$, and $S_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})$.

B.1.4 Hypothesis Testing

Suppose the model for the data $\mathbf{X}$ is described by a family of probability distributions that depend on a parameter $\boldsymbol{\theta} \in \Theta$. The aim of **hypothesis testing** is to decide, on the basis of the observed data $\mathbf{x}$, which of two competing hypotheses, $H_0 : \boldsymbol{\theta} \in \Theta_0$ (the **null hypothesis**) and $H_1 : \boldsymbol{\theta} \in \Theta_1$ (the **alternative hypothesis**), holds true.

In classical statistics the null hypothesis and alternative hypothesis do not play equivalent roles. H_0 contains the "status quo" statement, and is only rejected if the observed data are very unlikely to have happened under H_0.

The decision whether to accept or reject H_0 is dependent on the outcome of a **test statistic** $\mathbf{T} = \mathbf{T}(\mathbf{X})$. For simplicity, we discuss only the one-dimensional case $\mathbf{T} \equiv T$. Two (related) types of decision rules are generally used:

1. **Decision rule 1**: *Reject H_0 if T falls in the critical region.*

Here the **critical region** is any appropriately chosen region in $\mathbb{R}$. In practice a critical region is one of the following:

- *left one-sided*: $(-\infty, c]$,
- *right one-sided*: $[c, \infty)$,
- *two-sided*: $(-\infty, c_1] \cup [c_2, \infty)$.

For example, for a right one-sided test, H_0 is rejected if the outcome of the test statistic is too large. The endpoints c, c_1, and c_2 of the critical regions are called **critical values**.

2. **Decision rule 2**: *Reject H_0 if the p-*value *is smaller than some p_0.*

   The ***p*-value** is the probability that under H_0 the (random) test statistic takes a value as extreme as or more extreme than the one observed. In particular, if t is the observed outcome of the test statistic T, then

   - *left one-sided test*: $p = \mathbb{P}_{H_0}(T \leqslant t)$,
   - *right one-sided*: $p = \mathbb{P}_{H_0}(T \geqslant t)$,
   - *two-sided*: $p = \min\{2\mathbb{P}_{H_0}(T \leqslant t),\ 2\mathbb{P}_{H_0}(T \geqslant t)\}$.

   The smaller the p-value, the greater the strength of the evidence against H_0 provided by the data. As a rule of thumb:

   $p < 0.10$ suggestive evidence,
   $p < 0.05$ reasonable evidence,
   $p < 0.01$ strong evidence.

Whether the first or the second decision rule is used, one can make two types of errors, as depicted in Table B.4.

Table B.4 Type I and II errors in hypothesis testing.

	True statement	
Decision	H_0 *is true*	H_1 *is true*
Accept H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

The **power** of the test at $\theta \in \Theta_1$ is defined as the probability that H_0 is rejected (correctly). That is,

$$\text{Power}(\theta) = \mathbb{P}_\theta(T \in \text{Critical Region}) = 1 - \mathbb{P}_\theta(\text{Type II Error}) .$$

The function $\theta \mapsto \text{Power}(\theta)$, with $\theta \in \Theta_1$ is called the **power curve**.

The choice of the test statistic and the corresponding critical region involves a multiobjective optimization criterion, whereby both the probabilities of a type I and type II error should, ideally, be chosen as small as possible. Unfortunately,

these probabilities compete with each other. For example, if the critical region is made larger (smaller), the probability of a type II error is reduced (increased), but at the same time the probability of a type I error is increased (reduced).

Since the type I error is considered more serious, Neyman and Pearson [8] suggested the following approach: choose the critical region such that the probability of a type II error is as small as possible, while keeping the probability of a type I error below a predetermined small **significance level** α.

Remark B.1.1 (Equivalence of Decision Rules) Note that decision rule 1 and 2 are equivalent in the following sense:

$$\begin{gathered}\text{Reject } H_0 \text{ if } T \text{ falls in the critical region, at significance level } \alpha. \\ \Leftrightarrow \\ \text{Reject } H_0 \text{ if the } p\text{-value is} \leqslant \text{significance level } \alpha.\end{gathered}$$

In other words, the p-value of the test is the smallest level of significance that would lead to the rejection of H_0.

In general, a statistical test involves the following steps:

1. Formulate an appropriate statistical model for the data.
2. Give the null and alternative hypotheses.
3. Determine the test statistic.
4. Determine the distribution of the test statistic under H_0.
5. Calculate the outcome of the test statistic.
6. Calculate the p-value *or* calculate the critical region, given a preselected significance level α.
7. Accept or reject H_0.

The actual choice of an appropriate test statistic is akin to selecting a good estimator for the unknown parameter $\boldsymbol{\theta}$. The test statistic should summarize the information about $\boldsymbol{\theta}$ and make it possible to distinguish between the alternative hypotheses. The likelihood ratio test provides a systematic approach to constructing powerful test statistics; see Section B.2.3.

We conclude with a number of standard tests involving normal and binomial data. Below, z_γ denotes the γ quantile of the $\mathsf{N}(0,1)$ distribution. The γ quantiles of the χ^2_n, t_n, and $\mathsf{F}(m,n)$ distributions are denoted by $\chi^2_{n;\gamma}$, $t_{n;\gamma}$, and $F_{m,n;\gamma}$, respectively. Details may be found in [1], for example.

Table B.5 Normal distribution, one sample: testing μ.

Model:	$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$	
H_0:	$\mu = \mu_0$	
Test statistic:	$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	
Null distribution:	$T \sim \mathsf{t}_{n-1}$	
Reject H_0 if:	$T \geqslant t_{n-1;1-\alpha}$	$H_1 : \mu > \mu_0$
	$T \leqslant -t_{n-1;1-\alpha}$	$H_1 : \mu < \mu_0$
	$T \leqslant -t_{n-1;1-\alpha/2}$ or $T \geqslant t_{n-1;1-\alpha/2}$	$H_1 : \mu \neq \mu_0$

Table B.6 Normal distribution, one sample: testing σ^2.

Model:	$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$		
H_0:	$\sigma^2 = \sigma_0^2$		
Test statistic:	$T = S^2(n-1)/\sigma_0^2$		
Null distribution:	$T \sim \chi^2_{n-1}$		
Reject H_0 if:	$T \geqslant \chi^2_{n-1;1-\alpha}$		$H_1 : \sigma^2 > \sigma_0^2$
	$T \leqslant \chi^2_{n-1;\alpha}$		$H_1 : \sigma^2 < \sigma_0^2$
	$T \geqslant \chi^2_{n-1;1-\alpha/2}$	or $T \leqslant \chi^2_{n-1;\alpha/2}$	$H_1 : \sigma^2 \neq \sigma_0^2$

Table B.7 Normal distribution, two samples: testing $\mu_X - \mu_Y$.

Model:	$X_1, \ldots, X_m \overset{\text{iid}}{\sim} \mathsf{N}(\mu_X, \sigma^2)$, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu_Y, \sigma^2)$	
	$X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are independent	
H_0:	$\mu_X = \mu_Y$	
Test statistic:	$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}}$	
Null distribution:	$T \sim \mathsf{t}_{n+m-2}$	
Reject H_0 if:	$T \geqslant t_{n+m-2;1-\alpha}$	$H_1 : \mu_X > \mu_Y$
	$T \leqslant -t_{n+m-2;1-\alpha}$	$H_1 : \mu_X < \mu_Y$
	$T \leqslant -t_{n+m-2;1-\alpha/2}$ or $T \geqslant t_{n+m-2;1-\alpha/2}$	$H_1 : \mu_X \neq \mu_Y$

Here $S_p^2 = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2}{m+n-2}$ is the *pooled* sample variance. Note that in Table B.7 the variances of the two samples are assumed to be equal. If $\{X_i\}$ and $\{Y_i\}$ are assumed to have different variances and the sample sizes are large, then one can use the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}},$$

which under H_0 approximately has a $\mathsf{N}(0,1)$ distribution. An alternative approach is to use Welch's t-test [9].

Table B.8 Normal distribution, two samples: testing σ_X^2/σ_Y^2.

Model:	$X_1, \ldots, X_m \overset{\text{iid}}{\sim} \mathsf{N}(\mu_X, \sigma_X^2)$, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu_Y, \sigma_Y^2)$		
	$X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are independent		
H_0:	$\sigma_X^2 = \sigma_Y^2$		
Test statistic:	$T = S_X^2/S_Y^2$		
Null distribution:	$T \sim \mathsf{F}(m-1, n-1)$		
Reject H_0 if:	$T \geqslant F_{m-1,n-1;1-\alpha}$		$H_1 : \sigma_X^2 > \sigma_Y^2$
	$T \leqslant F_{m-1,n-1;\alpha}$		$H_1 : \sigma_X^2 < \sigma_Y^2$
	$T \geqslant F_{m-1,n-1;1-\alpha/2}$	or $T \leqslant F_{m-1,n-1;\alpha/2}$	$H_1 : \sigma_X^2 \neq \sigma_Y^2$

Table B.9 Binomial distribution, one sample: testing p.

Model:	$X \sim \mathsf{Bin}(n, p)$	
H_0:	$p = p_0$	
Test statistic:	$T = X$	
Null distribution:	$\mathsf{Bin}(n, p_0)$	
Reject H_0 if:	$X \geqslant c$, where c is the smallest integer such that $\mathbb{P}_{H_0}(X \geqslant c) \leqslant \alpha$	$H_1 : p > p_0$
	$X \leqslant c$, where c is the largest integer such that $\mathbb{P}_{H_0}(X \leqslant c) \leqslant \alpha$	$H_1 : p < p_0$
	$X \leqslant c_1$ or $X \geqslant c_2$, where c_1 is the largest integer such that $\mathbb{P}_{H_0}(X \leqslant c_1) \leqslant \alpha/2$ and c_2 is the smallest integer such that $\mathbb{P}_{H_0}(X \geqslant c_2) \leqslant \alpha/2$	$H_1 : p \neq p_0$

For large n an alternative is to use the test statistic

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}},$$

which under H_0 approximately has a $\mathsf{N}(0,1)$ distribution. The null hypothesis is then rejected if $Z \geqslant z_{1-\alpha}$ for $H_1 : p > p_0$, $Z \leqslant -z_{1-\alpha}$ for $H_1 : p < p_0$, and $[Z \leqslant -z_{1-\alpha/2}$ or $Z \geqslant z_{1-\alpha/2}]$ for $H_1 : p \neq p_0$.

Table B.10 Binomial distribution, two samples: testing $p_X - p_Y$.

Model:	$X \sim \mathsf{Bin}(m, p_X)$ and $Y \sim \mathsf{Bin}(n, p_Y)$ independent	
H_0:	$p_X = p_Y$	
Test statistic:	$T = \frac{\widehat{p}_X - \widehat{p}_Y}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{m}+\frac{1}{n}\right)}}$	
Null distribution:	$\mathsf{N}(0,1)$ (approx.)	
Reject H_0 if:	$Z \geqslant z_{1-\alpha}$	$H_1 : p_X > p_Y$
	$Z \leqslant -z_{1-\alpha}$	$H_1 : p_X < p_Y$
	$Z \geqslant z_{1-\alpha/2}$ or $Z \leqslant -z_{1-\alpha/2}$	$H_1 : p_X \neq p_Y$

Here $\widehat{p}_X = X/m$, $\widehat{p}_Y = Y/n$, and $\widehat{p} = (X+Y)/(m+n)$.

B.2 LIKELIHOOD

The concept of *likelihood* is central in statistics. It describes in a precise way the information about model parameters that is contained in the observed data.

Let $\mathbf{X} = (X_1, \ldots, X_n)^\top$ be a random vector that is distributed according to a pdf $f(\mathbf{x}; \boldsymbol{\theta})$ (discrete or continuous) with parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top \in \Theta$. Let $\mathbf{x}$ be an outcome of $\mathbf{X}$. The function $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, is called the **likelihood function** of $\boldsymbol{\theta}$, based on $\mathbf{x}$. The (natural) logarithm of the likelihood function is called the **log-likelihood function** and is denoted by l. The gradient

of the log-likelihood function l is called the **score function**, and is denoted by $\mathcal{S}$. Hence,

$$\mathcal{S}(\boldsymbol{\theta};\mathbf{x}) = \begin{pmatrix} \frac{\partial l(\boldsymbol{\theta};\mathbf{x})}{\partial \theta_1} \\ \frac{\partial l(\boldsymbol{\theta};\mathbf{x})}{\partial \theta_2} \\ \vdots \\ \frac{\partial l(\boldsymbol{\theta};\mathbf{x})}{\partial \theta_d} \end{pmatrix} = \nabla_{\boldsymbol{\theta}} \ln \mathcal{L}(\boldsymbol{\theta};\mathbf{x}) = \frac{\nabla_{\boldsymbol{\theta}} f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x};\boldsymbol{\theta})} \,. \tag{B.10}$$

If θ is one-dimensional, the score function is thus defined as

$$\mathcal{S}(\theta;\mathbf{x}) = \frac{\mathrm{d}}{\mathrm{d}\theta} l(\theta;\mathbf{x}) = \frac{\mathrm{d}}{\mathrm{d}\theta} \ln \mathcal{L}(\theta;\mathbf{x}) = \frac{\frac{\mathrm{d}}{\mathrm{d}\theta} f(\mathbf{x};\theta)}{f(\mathbf{x};\theta)} \,.$$

The *random vector* $\mathcal{S}(\boldsymbol{\theta}) = \mathcal{S}(\boldsymbol{\theta};\mathbf{X})$ with $\mathbf{X} \sim f(\cdot;\boldsymbol{\theta})$ is called the **efficient score** or simply **score**. The covariance matrix $\mathcal{I}(\boldsymbol{\theta})$ of the score $\mathcal{S}(\boldsymbol{\theta})$ is called the **Fisher information matrix**. Note that $\mathcal{L}$ is a function of $\boldsymbol{\theta}$ for fixed $\mathbf{x}$, whereas $f(\mathbf{x};\boldsymbol{\theta})$ is viewed as a function of $\mathbf{x}$ for fixed $\boldsymbol{\theta}$. Similarly, l and $\mathcal{S}$ and $\mathcal{I}$ are functions of $\boldsymbol{\theta}$. The expectation of the score $\mathcal{S}(\boldsymbol{\theta})$ is equal to the zero vector:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} \mathcal{S}(\boldsymbol{\theta}) &= \int \frac{\nabla_{\boldsymbol{\theta}} f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x};\boldsymbol{\theta})} f(\mathbf{x};\boldsymbol{\theta}) \, \mathrm{d}\mathbf{x} \\ &= \int \nabla_{\boldsymbol{\theta}} f(\mathbf{x};\boldsymbol{\theta}) \, \mathrm{d}\mathbf{x} = \nabla_{\boldsymbol{\theta}} \int f(\mathbf{x};\boldsymbol{\theta}) \, \mathrm{d}\mathbf{x} = \nabla_{\boldsymbol{\theta}} 1 = \mathbf{0} \,, \end{aligned}$$

provided the interchange of differentiation and integration is justified. In particular, this is allowed for natural exponential families; see [6]. We will assume henceforth that $\mathbb{E}_{\boldsymbol{\theta}} \mathcal{S}(\boldsymbol{\theta}) = \mathbf{0}$.

Table B.11 displays the score functions $\mathcal{S}(\boldsymbol{\theta};x)$ calculated from (B.10) for some commonly used distributions. In this table ψ refers to the digamma function. ☞ 716

The concepts of likelihood and score are particularly useful in the case where $X_1,\ldots,X_n$ form an iid sample from some pdf $\mathring{f}$; that is, $X_1,\ldots,X_n \sim_{\text{iid}} \mathring{f}(\cdot;\boldsymbol{\theta})$. In that case, the likelihood of $\boldsymbol{\theta}$ given the data $\mathbf{x} = (x_1,\ldots,x_n)^\top$ is the product

$$\mathcal{L}(\boldsymbol{\theta};\mathbf{x}) = \prod_{i=1}^{n} \mathring{f}(x_i;\boldsymbol{\theta}) \,. \tag{B.11}$$

Consequently, the log-likelihood is the sum $l(\boldsymbol{\theta};\mathbf{x}) = \sum_{i=1}^{n} \ln \mathring{f}(x_i;\boldsymbol{\theta})$, and the score is

$$\mathcal{S}(\boldsymbol{\theta};\mathbf{X}) = \sum_{i=1}^{n} \mathring{\mathcal{S}}(\boldsymbol{\theta};X_i) \,, \tag{B.12}$$

where $\mathring{\mathcal{S}}(\boldsymbol{\theta};x)$ is the score function corresponding to $\mathring{f}(x;\boldsymbol{\theta})$. It follows that the information matrix satisfies

$$\mathcal{I}(\boldsymbol{\theta}) = n\,\mathring{\mathcal{I}}(\boldsymbol{\theta}) \,,$$

where $\mathring{\mathcal{I}}(\boldsymbol{\theta})$ is the information matrix corresponding to $\mathring{f}$.

Note that the random vectors $\{\mathring{\mathcal{S}}(\boldsymbol{\theta};X_i)\}$ are independent and identically distributed with mean vector $\mathbf{0}$ and covariance matrix $\mathring{\mathcal{I}}(\boldsymbol{\theta})$. The law of large numbers and the central limit theorem now lead directly to two important properties of the ☞ 625

Table B.11 Score functions for commonly used distributions.

Distribution	$\boldsymbol{\theta}$	$\mathcal{S}(\boldsymbol{\theta}; x)$
$\mathsf{Exp}(\lambda)$	λ	$\lambda^{-1} - x$
$\mathsf{Gamma}(\alpha, \lambda)$	(α, λ)	$\big(\ln(\lambda x) - \psi(\alpha),\ \alpha\lambda^{-1} - x\big)^\top$
$\mathsf{N}(\mu, \sigma^2)$	(μ, σ)	$\big(\sigma^{-2}(x-\mu),\ -\sigma^{-1} + \sigma^{-3}(x-\mu)^2\big)^\top$
$\mathsf{Weib}(\alpha, \lambda)$	(α, λ)	$\big(\alpha^{-1} + \ln(\lambda x)[1 - (\lambda x)^\alpha],\ \frac{\alpha}{\lambda}[1 - (\lambda x)^\alpha]\big)^\top$
$\mathsf{Bin}(n, p)$	p	$\dfrac{x - np}{p(1-p)}$
$\mathsf{Poi}(\lambda)$	λ	$\dfrac{x}{\lambda} - 1$
$\mathsf{Geom}(p)$	p	$\dfrac{1 - px}{p(1-p)}$

score of an iid sample.

1. *Law of large numbers*: As $n \to \infty$,

$$\frac{1}{n}\mathcal{S}(\boldsymbol{\theta}; \mathbf{X}) \to \mathbb{E}_{\boldsymbol{\theta}}\mathring{\mathcal{S}}(\boldsymbol{\theta}; X) = \mathbf{0}\,, \tag{B.13}$$

   since the expected score is the zero vector.

2. *Central limit theorem*: For large n

$$\mathcal{S}(\boldsymbol{\theta}; \mathbf{X}) \overset{\text{approx.}}{\sim} \mathsf{N}(\mathbf{0}, n\,\mathring{\mathcal{I}}(\boldsymbol{\theta}))\,. \tag{B.14}$$

■ EXAMPLE B.5 (Bernoulli Random Sample)

Let $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{Ber}(p)$. Then, for a given observation $\mathbf{x} = (x_1, \ldots, x_n)^\top$, the likelihood of p is given by

$$\mathcal{L}(p; \mathbf{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^x(1-p)^{n-x}, \quad 0 < p < 1\,, \tag{B.15}$$

where $x = x_1 + \cdots + x_n$. The log-likelihood is $l(p) = x \ln p + (n - x)\ln(1-p)$. Through differentiation with respect to p, we find the score function:

$$\mathcal{S}(p; x) = \frac{x}{p} - \frac{n-x}{1-p} = \frac{x}{p(1-p)} - \frac{n}{1-p}\,. \tag{B.16}$$

The corresponding score $\mathcal{S}(p)$ is obtained by replacing x with $X \sim \mathsf{Bin}(n, p)$. The expectation of $\mathcal{S}(p)$ is 0 and its variance (the information matrix/number) is

$$\mathcal{I}(p) = \frac{\mathrm{Var}(X)}{p^2(1-p)^2} = \frac{n}{p(1-p)}\,.$$

Hence, for large n, $\mathcal{S}(p)$ approximately has a $\mathsf{N}(0, n/(p(1-p)))$ distribution.

Other properties of the likelihood and score include (for proofs see, for example, [1]):

1. *Natural exponential family:* For an exponential family in canonical form ☞ 701

$$f(\mathbf{x};\boldsymbol{\eta}) = \mathrm{e}^{\boldsymbol{\eta}^\top \mathbf{t}(\mathbf{x}) - A(\boldsymbol{\eta})} h(\mathbf{x}) \;, \tag{B.17}$$

with A as in (D.3), the log-likelihood function is $l(\boldsymbol{\eta};\mathbf{x}) = \boldsymbol{\eta}^\top \mathbf{t}(\mathbf{x}) - A(\boldsymbol{\eta}) + \ln h(\mathbf{x})$, so that the score function becomes

$$\mathcal{S}(\boldsymbol{\eta};\mathbf{x}) = \mathbf{t}(\mathbf{x}) - \nabla A(\boldsymbol{\eta}) \;= \mathbf{t}(\mathbf{x}) - \mathbb{E}_{\boldsymbol{\eta}} \mathbf{t}(\mathbf{X}) \;. \tag{B.18}$$

It follows that the information matrix is the covariance matrix of $\mathbf{t}(\mathbf{X})$:

$$\mathcal{I}(\boldsymbol{\eta}) = \mathrm{Cov}(\mathbf{t}(\mathbf{X})) = \nabla^2 A(\boldsymbol{\eta}) \;. \tag{B.19}$$

2. *Information matrix:* An alternative expression for the information matrix is

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}} H(\boldsymbol{\theta};\mathbf{X}) \;, \tag{B.20}$$

where $H(\boldsymbol{\theta};\mathbf{X})$ is the Hessian of $l(\boldsymbol{\theta};\mathbf{X})$; that is, the (random) matrix

$$H(\boldsymbol{\theta};\mathbf{X}) = \left(\frac{\partial^2 \ln f(\mathbf{X};\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} \right) = \left(\frac{\partial^2 l(\boldsymbol{\theta};\mathbf{X})}{\partial\theta_i \partial\theta_j} \right) = \left(\frac{\partial \mathcal{S}_i(\boldsymbol{\theta};\mathbf{X})}{\partial\theta_j} \right) \;,$$

where $\mathcal{S}_i$ denotes the i-th component of the score. This alternative expression is valid under mild conditions (which are satisfied for exponential families) that allow the interchange of the order of integration and differentiation [6].

3. *Cramér–Rao:* Let $\mathbf{X} \sim f(\mathbf{x};\boldsymbol{\theta})$. The variance of any unbiased estimator $Z = Z(\mathbf{X})$ of $g(\boldsymbol{\theta})$, where g is a $\mathcal{C}^1$ function, is bounded from below by

$$\mathrm{Var}(Z) \geqslant (\nabla g(\boldsymbol{\theta}))^\top \mathcal{I}^{-1}(\boldsymbol{\theta}) \, \nabla g(\boldsymbol{\theta}) \;. \tag{B.21}$$

4. *Location–scale families:* For location–scale families $\{f(x;\mu,\sigma)\}$ the Fisher information does not depend on μ. In particular, for location families it is constant.

B.2.1 Likelihood Methods for Estimation

Let $\mathbf{x}$ be the observed data from the model $\mathbf{X} \sim f(\mathbf{x};\boldsymbol{\theta})$, yielding the likelihood function $\mathcal{L}(\boldsymbol{\theta};\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta})$. The **maximum likelihood estimate** (MLE) of $\boldsymbol{\theta}$ is a vector $\widehat{\boldsymbol{\theta}}$ such that $\mathcal{L}(\widehat{\boldsymbol{\theta}};\mathbf{x}) \geqslant \mathcal{L}(\boldsymbol{\theta};\mathbf{x})$ for all $\boldsymbol{\theta}$ in the parameter space Θ. The corresponding random variable (a function of $\mathbf{X}$), also denoted $\widehat{\boldsymbol{\theta}}$, is called the **maximum likelihood estimator** (also abbreviated as MLE).

Since the natural logarithm is an increasing function, maximization of $\mathcal{L}(\boldsymbol{\theta};\mathbf{x})$ is equivalent to maximization of the log-likelihood $l(\boldsymbol{\theta};\mathbf{x})$. This is often easier, especially when $\mathbf{X}$ is an iid sample from some sampling distribution.

If $l(\boldsymbol{\theta};\mathbf{x})$ is a differentiable function with respect to $\boldsymbol{\theta}$ and the maximum is attained in the *interior* of Θ, *and there exists a unique maximum*, then the MLE

of $\boldsymbol{\theta}$ can be found by differentiating $l(\boldsymbol{\theta}; \mathbf{x})$ with respect to $\boldsymbol{\theta}$ — more precisely, by solving

$$\nabla_{\boldsymbol{\theta}}\, l(\boldsymbol{\theta}; \mathbf{x}) = \mathbf{0}\,.$$

In other words, the MLE is obtained by finding a root of the score; that is, by solving

$$\mathcal{S}(\boldsymbol{\theta}; \mathbf{x}) = \mathbf{0}\,. \tag{B.22}$$

Properties of the maximum likelihood estimator include (see, for example, [6, Page 444]):

1. *Consistency*: The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ is **consistent**. That is, with probability tending to 1 as $n \to \infty$ the likelihood equation has a solution $\widehat{\boldsymbol{\theta}}$ such that for all $\varepsilon > 0$

$$\mathbb{P}(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| > \varepsilon) \to 0\,.$$

2. *Asymptotic normality*: Suppose that $\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2, \ldots$ is a sequence of consistent maximum likelihood estimators for $\boldsymbol{\theta}$. Then, $\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ converges in distribution to a $\mathsf{N}(\mathbf{0}, \mathring{\mathcal{I}}^{-1}(\boldsymbol{\theta}))$-distributed random vector as $n \to \infty$. In other words

$$\widehat{\boldsymbol{\theta}}_n \overset{\text{approx.}}{\sim} \mathsf{N}(\boldsymbol{\theta}, \mathring{\mathcal{I}}^{-1}(\boldsymbol{\theta})/n)\,.$$

3. *Invariance*: Let $\widehat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$. Then, for any function $\mathbf{g}$ the MLE of $\mathbf{g}(\boldsymbol{\theta})$ is $\mathbf{g}(\widehat{\boldsymbol{\theta}})$.

Note that Property 1 only says that there *exists* a sequence of MLEs $\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2, \ldots$ that converge (in probability) to the true $\boldsymbol{\theta}$. When there are multiple local maxima, a particular sequence $\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2, \ldots$ may in fact converge to a local maximum.

Theorem B.2.1 (Exponential Families) *For natural exponential families of the form* (B.17) *the MLE is found by solving*

$$\mathbf{t}(\mathbf{x}) - \nabla A(\boldsymbol{\eta}) = \mathbf{t}(\mathbf{x}) - \mathbb{E}_{\boldsymbol{\eta}}\, \mathbf{t}(\mathbf{X}) = \mathbf{0}\,. \tag{B.23}$$

That is, $\boldsymbol{\eta}$ is chosen such that the observed and expected values of $\mathbf{t}(\mathbf{X})$ are matched.

■ EXAMPLE B.6 (MLE for the Gamma Distribution)

We wish to estimate both parameters of a $\mathsf{Gamma}(\alpha, \lambda)$ distribution, based on an iid sample $\mathbf{x} = (x_1, \ldots, x_n)$. The corresponding pdf is

$$\mathring{f}(x; \alpha, \lambda) = \frac{\lambda^\alpha x^{\alpha-1} \mathrm{e}^{-\lambda x}}{\Gamma(\alpha)}, \quad x \geqslant 0\,,$$

which is of the form (B.17) with $\mathbf{t}(x) = (x, \ln x)^\top$, $\boldsymbol{\eta} = (-\lambda,\ \alpha - 1)^\top$ and $A(\boldsymbol{\eta}) = -(\eta_2+1)\ln(-\eta_1) + \ln\Gamma(\eta_2+1)$; see also Table D.1. Consequently, the score function is given by

☞ 702

$$\mathring{\mathcal{S}}(\boldsymbol{\eta}; x) = \mathbf{t}(x) - \nabla A(\boldsymbol{\eta}) = \begin{pmatrix} x + (\eta_2 + 1)/\eta_1 \\ \ln x + \ln(-\eta_1) - \psi(\eta_2 + 1) \end{pmatrix},$$

where ψ is the digamma function. The information matrix is ☞ 716

$$\mathring{\mathcal{I}}(\boldsymbol{\eta}) = \nabla^2 A(\boldsymbol{\eta}) = \begin{pmatrix} (\eta_2+1)/\eta_1^2 & -1/\eta_1 \\ -1/\eta_1 & \psi'(\eta_2+1) \end{pmatrix} .$$

The score function corresponding to the iid sample is therefore

$$\mathcal{S}(\boldsymbol{\eta}; \mathbf{x}) = \sum_{i=1}^{n} \mathring{\mathcal{S}}(\boldsymbol{\eta}; x_i) = \begin{pmatrix} \sum_{i=1}^n x_i + n(\eta_2+1)/\eta_1 \\ \sum_{i=1}^n \ln x_i + n(\ln(-\eta_1) - \psi(\eta_2+1)) \end{pmatrix}$$

and the information matrix is $\mathcal{I}(\boldsymbol{\eta}) = n\,\mathring{\mathcal{I}}(\boldsymbol{\eta})$. Estimates of the parameters are found by numerically solving $\mathcal{S}(\boldsymbol{\eta}; \mathbf{x}) = \mathbf{0}$ (continued in Example B.8).

B.2.1.1 Score Intervals The score function can also be used to construct confidence intervals. We consider only the one-dimensional case; that is, $\theta \in \mathbb{R}$. Let $\mathbf{X} = (X_1, \ldots, X_n)^\top$ be an iid sample from some sampling distribution $\mathring{f}$. Because of the normal approximation (B.14), the statistic $\mathcal{S}(\theta; \mathbf{X})/\sqrt{\mathcal{I}(\theta)}$ is approximately standard normal, and hence

$$\left\{ \theta : -z_{1-\alpha/2} \leqslant \frac{\mathcal{S}(\theta; \mathbf{X})}{\sqrt{n\,\mathring{\mathcal{I}}(\theta)}} \leqslant z_{1-\alpha/2} \right\}$$

is an approximate $1-\alpha$ **confidence set** (not necessarily an interval).

■ **EXAMPLE B.7 (Score Interval for a Bernoulli Random Sample)**

Let $\mathbf{X}$ be an iid sample from $\mathsf{Ber}(p)$. The information matrix is $\mathcal{I}(p) = n/(p(1-p))$ and the score is $\mathcal{S}(p; \mathbf{X}) = n(\bar{X} - p)/(p(1-p))$; see (B.16). So the confidence set becomes

$$\left\{ p \;:\; -z_{1-\alpha/2} \leqslant \frac{n(\bar{X}-p)}{p(1-p)} \times \sqrt{\frac{p(1-p)}{n}} \leqslant z_{1-\alpha/2} \right\}$$
$$= \left\{ p \;:\; -z_{1-\alpha/2} \leqslant \frac{\bar{X}-p}{\sqrt{p(1-p)/n}} \leqslant z_{1-\alpha/2} \right\}$$

By solving the quadratic equation $(\bar{X}-p)^2 = a^2\,p(1-p)/n$ with respect to p, this confidence set can be written as the interval $\{T_1 \leqslant p \leqslant T_2\}$ with

$$T_{1,2} = \frac{a^2 + 2n\bar{X} \mp a\sqrt{a^2 - 4n(\bar{X}-1)\bar{X}}}{2\,(a^2+n)} ,$$

where $a = z_{1-\alpha/2}$. This **score interval** has much better coverage behavior than the confidence interval in Table B.2, over the complete range of p.

B.2.2 Numerical Methods for Likelihood Maximization

It is frequently not possible to find the MLE $\widehat{\boldsymbol{\theta}}$ in an explicit form. In that case one needs to solve the equation $\mathcal{S}(\boldsymbol{\theta}) = \mathcal{S}(\boldsymbol{\theta}; \mathbf{x}) = \mathbf{0}$ numerically via a root-finding

procedure. A well-known method is the *Newton–Raphson* procedure (see also Section C.2.2.1). Starting from a guess $\boldsymbol{\theta}$, a "better" guess is obtained by approximating the score via a linear function. More precisely, suppose that $\boldsymbol{\theta}$ is the initial guess for the root. If the latter is reasonably close to $\widehat{\boldsymbol{\theta}}$, a first-order Taylor approximation around $\boldsymbol{\theta}$ gives

☞ 688

$$\mathcal{S}(\widehat{\boldsymbol{\theta}}) \approx \mathcal{S}(\boldsymbol{\theta}) + \nabla\mathcal{S}(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathcal{S}(\boldsymbol{\theta}) + H(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) ,$$

where $H(\boldsymbol{\theta}) = H(\boldsymbol{\theta}; \mathbf{x})$ is the Hessian of the log-likelihood, that is, the matrix of second-order partial derivatives of l. Since $\mathcal{S}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$, we have $\widehat{\boldsymbol{\theta}} \approx \boldsymbol{\theta} - H^{-1}(\boldsymbol{\theta})\,\mathcal{S}(\boldsymbol{\theta})$. This suggests the following iterative scheme.

Algorithm B.1 (Newton–Raphson Scheme for the MLE)

1. *Start with an initial guess* $\boldsymbol{\theta}_0$. *Set* $t = 0$.
2. *Set*
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - H^{-1}(\boldsymbol{\theta}_t)\,\mathcal{S}(\boldsymbol{\theta}_t) . \tag{B.24}$$
3. *If* $\mathcal{S}(\boldsymbol{\theta}_{t+1}) < \varepsilon$ *for some small* $\varepsilon > 0$, *then return* $\boldsymbol{\theta}_{t+1}$ *as the MLE; otherwise, set* $t = t + 1$ *and go to Step 2.*

To implement the Newton–Raphson scheme, it is often crucial to come up with a good starting value for the parameter vector. One natural way to obtain a good guess is to match the sample and theoretical moments via the method of moments; see Section B.1.3.1.

Notice that $H(\boldsymbol{\theta}) = H(\boldsymbol{\theta}; \mathbf{x})$ depends on the parameter $\boldsymbol{\theta}$ and data $\mathbf{x}$, and may be quite complicated. However, the expectation of $H(\boldsymbol{\theta}; \mathbf{X})$ under $\boldsymbol{\theta}$ is simply the negative of the information matrix $\mathcal{I}(\boldsymbol{\theta})$, which does not depend on the data. This suggests the alternative to (B.24):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \mathcal{I}^{-1}(\boldsymbol{\theta}_t)\,\mathcal{S}(\boldsymbol{\theta}_t) , \tag{B.25}$$

which may be easier to implement if the information matrix is readily available.

■ **EXAMPLE B.8 (MLE for the Gamma Distribution)**

We continue Example B.6 to find MLEs for the parameters of the Gamma(α, λ) distribution. The initial guess is obtained by matching the expectation and variance to the sample mean and sample variance, $\bar{x} = \sum_{i=1}^n x_i/n$ and $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$, respectively. Since for $X \sim$ Gamma(α, λ), $\mathbb{E}X = \alpha/\lambda$ and $\mathrm{Var}(X) = \alpha/\lambda^2$, this leads to the initial guess $\boldsymbol{\eta}_0 = (-\lambda_0, \alpha_0 - 1)^\top$, where $\alpha_0 = \frac{\bar{x}^2}{s^2}$ and $\lambda_0 = \frac{\bar{x}}{s^2}$. The following MATLAB program implements the Newton–Raphson scheme to find the MLE for $\alpha = 3$ and $\lambda = 0.05$.

```
%gammMLE.m
n = 100;
alpha = 3; lambda = 0.05;
x = gamrnd(alpha,1/lambda,1,n);
sumlogx = sum(log(x)); sumx = sum(x);
alp = mean(x)^2/var(x); lam = mean(x)/var(x); % initial guess
```

```
eta = [-lam;alp - 1]; S = Inf;
while  sum(abs(S) > 10^(-5)) > 0
    S = [sumx + n*(eta(2) + 1)/eta(1); ...
         sumlogx + n*(log(-eta(1)) - psi(eta(2) + 1))];
    I = n * [ (eta(2)+1)/eta(1)^2, -1/eta(1); ...
        -1/eta(1) , psi(1,eta(2)+1)];
    eta = eta + I\S;
end
fprintf('lam_hat = %g , alpha_hat = %g \n',-eta(1),1+eta(2))
```

B.2.3 Likelihood Methods for Hypothesis Testing

Let $X_1, \ldots, X_n$ be an iid sample from a distribution with unknown parameter $\boldsymbol{\theta} \in \Theta$. Write $\mathbf{X}$ for the corresponding random vector, and denote the likelihood by $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$. Suppose Θ_0 and Θ_1 are two nonoverlapping subsets of Θ, such that $\Theta_0 \cup \Theta_1 = \Theta$.

The **likelihood ratio statistic** is defined as

$$\Lambda = \frac{\max_{\boldsymbol{\theta} \in \Theta_0} \mathcal{L}(\boldsymbol{\theta}; \mathbf{X})}{\max_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta}; \mathbf{X})} = \frac{\mathcal{L}(\widehat{\boldsymbol{\theta}}_0; \mathbf{X})}{\mathcal{L}(\widehat{\boldsymbol{\theta}}; \mathbf{X})} ,$$

where $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}_0$ the maximum likelihood estimator of $\boldsymbol{\theta}$ over Θ_0 only.

The likelihood ratio statistic Λ can be used as a test statistic for testing the hypotheses

$$\begin{aligned} H_0 &: \ \boldsymbol{\theta} \in \Theta_0 , \\ H_1 &: \ \boldsymbol{\theta} \in \Theta_1 . \end{aligned}$$

The critical region is $(0, \lambda^*]$; that is, reject H_0 if Λ is smaller than some critical value λ^*. To determine λ^* one needs to know the distribution of Λ under H_0. In general this is a difficult task, but it is sometimes possible to derive the distribution of a function of Λ under H_0. This is then taken as the test statistic. The critical region follows by inspection.

■ **EXAMPLE B.9 (Likelihood Ratio Method for Gaussian Data)**

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$, with μ and σ^2 unknown. We wish to test

$$\begin{aligned} H_0 &: \ \mu = \mu_0 , \\ H_1 &: \ \mu \neq \mu_0 . \end{aligned}$$

The likelihood function is given by

$$\mathcal{L}(\mu, \sigma^2; \mathbf{X}) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}\right) .$$

Maximizing $\mathcal{L}$ over Θ gives the maximum likelihood estimate $(\widehat{\mu}, \widehat{\sigma^2})$ given in (B.6) and (B.7). Maximizing $\mathcal{L}$ over $\Theta_0 = \{(\mu_0, \sigma^2), \sigma^2 > 0\}$ gives the estimate $(\mu_0, \widetilde{\sigma^2})$,

with

$$\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2 .$$

Hence,

$$\Lambda = \frac{\mathcal{L}(\mu_0, \widetilde{\sigma^2}; \mathbf{X})}{\mathcal{L}(\widehat{\mu}, \widehat{\sigma^2}; \mathbf{X})} = \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \mu_0)^2}\right)^{n/2} = \left(1 + \frac{1}{n-1}T^2\right)^{-n/2} ,$$

where $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ and S is the sample standard deviation. Rejecting H_0 for small values of Λ is equivalent to rejecting H_0 for large values of $|T|$. Moreover, under H_0, T has a t_{n-1} distribution. Thus, the likelihood ratio method yields the test in Table B.5.

The asymptotic distribution of the likelihood ratio statistic under H_0 can be derived in several cases, in particular when Θ_0 consists of only one point $\boldsymbol{\theta}_0$. Under H_0 the log-likelihood function satisfies

$$-2\ln\Lambda = -2(l(\boldsymbol{\theta}_0) - l(\widehat{\boldsymbol{\theta}})) .$$

☞ 711

A second-order Taylor expansion at $\boldsymbol{\theta}_0$ around $\widehat{\boldsymbol{\theta}}$ gives

$$l(\boldsymbol{\theta}_0) = l(\widehat{\boldsymbol{\theta}}) + (\nabla l(\widehat{\boldsymbol{\theta}}))^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \nabla^2 l(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathcal{O}(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^3) .$$

Because $\nabla l(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$ and $\nabla^2 l(\widehat{\boldsymbol{\theta}}) \approx -\mathcal{I}(\boldsymbol{\theta}_0)$, where $\mathcal{I}$ is the information matrix, we have

$$-2\ln\Lambda \approx (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \mathcal{I}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) .$$

By the central limit theorem $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ has approximately a $\mathsf{N}(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}_0))$ distribution under H_0. Thus, for a large sample size, we have that $-2\ln\Lambda$ is approximately distributed as $\mathbf{X}^\top \mathcal{I}(\boldsymbol{\theta}_0)\mathbf{X}$ with $\mathbf{X} \sim \mathsf{N}(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}_0))$, which has a χ^2_k distribution, where k is the dimension of $\boldsymbol{\theta}$. This gives the following theorem; see also [1].

Theorem B.2.2 (Asymptotic Distribution of Likelihood Ratio Statistic) *For a k-dimensional parameter space, if the null hypothesis has only one value $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and the alternative hypothesis is $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, then under some mild regularity conditions (which are satisfied for exponential families):*

$$-2\ln\Lambda \overset{\text{approx.}}{\sim} \chi^2_k \quad \textit{for large } n .$$

B.3 BAYESIAN STATISTICS

Bayesian statistics is a branch of statistics that is centered around Bayes' formula (A.23). The type of statistical reasoning here is somewhat different from that in classical statistics. In particular, model parameters are usually treated as random rather than fixed quantities and Bayesian statistics uses different notational conventions from those in classical statistics. The two main differences in notation are:

1. Pdfs and conditional pdfs always use the *same letter* f (sometimes p is used instead of f). That is, instead of writing $f_X(x)$ and $f_{X\,|\,Y}(x\,|\,y)$ for the pdf of X and the conditional pdf of X given Y, one simply writes $f(x)$ and $f(x\,|\,y)$. If Y is a different random variable, its pdf (at y) is thus denoted by $f(y)$. This particular style of notation is typical in Bayesian analysis and can be of great descriptive value, despite its apparent ambiguity. We will use this notation whenever we work in a Bayesian setting.

2. One does not usually indicate random variables by capital letters and their outcomes by lower case letters. It is assumed that it is clear from the context whether a variable x or θ should be interpreted as a number or a random variable.

In Bayesian statistics the data $\mathbf{x}$ is modeled via a conditional pdf $f(\mathbf{x}\,|\,\boldsymbol{\theta})$, called the **likelihood**, that depends on a random parameter $\boldsymbol{\theta}$ taking values in some set Θ. The **a priori** information about $\boldsymbol{\theta}$ (that is, the knowledge about $\boldsymbol{\theta}$ without using any information from the data) is summarized by the pdf of $\boldsymbol{\theta}$, which is called the **prior** pdf. Additional knowledge about $\boldsymbol{\theta}$ obtained from the observed data $\mathbf{x}$ is given by the conditional pdf $f(\boldsymbol{\theta}\,|\,\mathbf{x})$, called the **posterior** pdf. The posterior and prior pdfs are related via Bayes' formula (replace the integral with a sum in the discrete case):

$$f(\boldsymbol{\theta}\,|\,\mathbf{x}) = \frac{f(\mathbf{x}\,|\,\boldsymbol{\theta})\,f(\boldsymbol{\theta})}{\int f(\mathbf{x}\,|\,\boldsymbol{\theta})\,f(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \propto f(\mathbf{x}\,|\,\boldsymbol{\theta})\,f(\boldsymbol{\theta})\,. \tag{B.26}$$

The denominator in (B.26),

$$f(\mathbf{x}) = \int f(\mathbf{x}\,|\,\boldsymbol{\theta})\,f(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\,,$$

is often called the **marginal likelihood** and is usually difficult to compute. A **Bayesian model** specifies the prior pdf and likelihood. Once the model is given, all inference is based on the posterior pdf in (B.26). For example, a vector for which the posterior pdf is maximal yields a point estimate for $\boldsymbol{\theta}$, called the **maximum a posteriori** estimate. Another estimate is obtained by taking the expected value of $\boldsymbol{\theta}$ under the posterior pdf. A Bayesian $1-\alpha$ confidence region, or **credible region**, is any subset $\mathscr{C} \subset \Theta$, such that

$$\mathbb{P}(\boldsymbol{\theta} \in \mathscr{C}\,|\,\mathbf{x}) = \int_{\boldsymbol{\theta}\in\mathscr{C}} f(\boldsymbol{\theta}\,|\,\mathbf{x})\,\mathrm{d}\boldsymbol{\theta} \geqslant 1-\alpha\,. \tag{B.27}$$

Bayesian models are often constructed in a *hierarchical* way. For example, a three-parameter **hierarchical model** could be specified as follows:

$$\begin{aligned} a &\sim f(a)\,,\\ (b\,|\,a) &\sim f(b\,|\,a)\,,\\ (c\,|\,a,b) &\sim f(c\,|\,a,b)\,,\\ (\mathbf{x}\,|\,a,b,c) &\sim f(\mathbf{x}\,|\,a,b,c)\,. \end{aligned} \tag{B.28}$$

In other words, first specify the prior pdf of a, then given a specify the pdf of b, etc., until finally the likelihood as a function of all the parameters is given. This

procedure allows for a straightforward evaluation of the joint pdf as the product of the conditional pdfs:

$$f(\mathbf{x}, a, b, c) = f(\mathbf{x} \,|\, a, b, c)\, f(c \,|\, a, b)\, f(b \,|\, a) f(a) \,.$$

To find the posterior $f(a, b, c \,|\, \mathbf{x})$, simply view $f(\mathbf{x}, a, b, c)$ as a function of a, b, and c for fixed $\mathbf{x}$. To find the marginal posterior pdfs, $f(a \,|\, \mathbf{x})$, $f(b \,|\, \mathbf{x})$, and $f(a \,|\, \mathbf{x})$, one needs to *integrate out* the other parameters. For example,

$$f(c \,|\, \mathbf{x}) = \iint f(a, b, c \,|\, \mathbf{x}) \,\mathrm{d}a \,\mathrm{d}b \,.$$

■ EXAMPLE B.10 (Coin Flipping and Bayesian Learning)

Consider the random experiment where we toss a biased coin n times. Suppose that the outcomes are $x_1, \ldots, x_n$, with $x_i = 1$ if the i-th toss is heads and $x_i = 0$ otherwise, for $i = 1, \ldots, n$. A possible Bayesian model for the data is

$$\begin{aligned} p &\sim \mathsf{U}(0,1) \\ (x_1, \ldots, x_n \,|\, p) &\overset{\text{iid}}{\sim} \mathsf{Ber}(p) \,. \end{aligned}$$

The likelihood is therefore

$$f(\mathbf{x} \,|\, p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^s \,(1-p)^{n-s} \,,$$

where $s = x_1 + \cdots + x_n$ is the total number of heads. Since $f(p) = 1$, the posterior pdf is

$$f(p \,|\, \mathbf{x}) = c\, p^s \,(1-p)^{n-s} \,, \quad p \in [0,1] \,,$$

which is the pdf of the $\mathsf{Beta}(s+1, n-s+1)$ distribution. The normalization constant is $c = (n+1)\binom{n}{s}$. The maximum a posteriori estimate of p is s/n, which coincides with the classical maximum likelihood estimate. The expectation of the posterior pdf is $(s+1)/(n+2)$. The graph of the pdf for $n = 100$ and $s = 1$ is given in Figure B.1. For this case a left one-sided 95% credible interval for p is $[0, 0.0461]$, where 0.0461 is the 0.95 quantile of the $\mathsf{Beta}(2, 100)$ distribution.

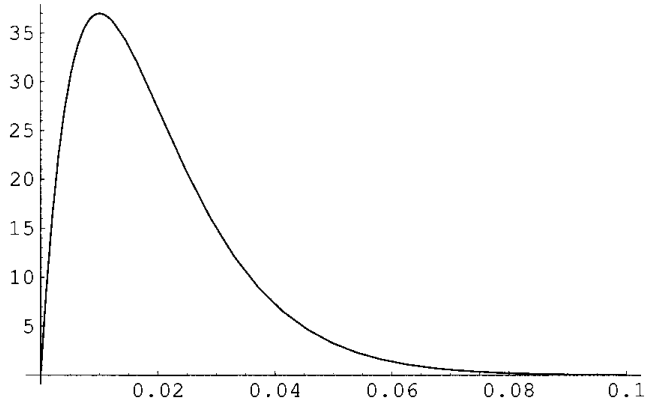

Figure B.1 Posterior pdf for p, with $n = 100$ and $s = 1$.

Evaluating or drawing from the marginal posterior distributions may not always be easy or feasible. When this is the case, one often turns to Markov chain Monte Carlo techniques; see Chapter 6. For example, in the three-parameter model (B.28) one could use the *Gibbs sampler* to sample from the posterior pdf:

☞ 225

1. Initialize a, b, c. Then iterate the following steps:
2. Draw a from $f(a \mid b, c, \mathbf{x})$.
3. Draw b from $f(b \mid a, c, \mathbf{x})$.
4. Draw c from $f(c \mid a, b, \mathbf{x})$.

After we obtain a (dependent) sample $\{(a_t, b_t, c_t)\}$ from $f(a, b, c \mid \mathbf{x})$, process only the variables of interest, for example, only the $\{c_t\}$ to obtain a dependent sample from $f(c \mid \mathbf{x})$.

B.3.1 Conjugacy

In Bayesian analysis it is convenient to have the posterior and prior pdfs in the *same* family of distributions. This property is called **conjugacy**. The advantage of conjugacy is that only the parameters of the distribution need to be updated.

Exponential families provide natural conjugate families. In particular, consider the m-dimensional exponential family

☞ 701

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = c(\boldsymbol{\theta})^n \mathrm{e}^{\sum_{i=1}^m \eta_i(\boldsymbol{\theta}) \sum_{k=1}^n t_i(x_k)} \prod_{i=1}^n h(x_k) \,, \tag{B.29}$$

which is the joint pdf of an iid sample from an exponential family — see Example B.1. Suppose the prior pdf is chosen of the form

$$f(\boldsymbol{\theta}) \propto c(\boldsymbol{\theta})^b \mathrm{e}^{\sum_{i=1}^m \eta_i(\boldsymbol{\theta}) a_i} ,$$

where the proportionality constant only depends on $\mathbf{a} = (a_1, \ldots, a_m, b)$. Then, the posterior pdf becomes

$$f(\boldsymbol{\theta} \mid \mathbf{x}) \propto f(\boldsymbol{\theta}) f(\mathbf{x} \mid \boldsymbol{\theta}) \propto c(\boldsymbol{\theta})^{n+b} \mathrm{e}^{\sum_{i=1}^m \eta_i(\boldsymbol{\theta}) \left(a_i + \sum_{k=1}^n t_i(x_k)\right)} ,$$

where the proportionality constant does not depend on $\boldsymbol{\theta}$. Thus, $f(\boldsymbol{\theta})$ and $f(\boldsymbol{\theta} \mid \mathbf{x})$ are in the same $(m+1)$-dimensional exponential family.

■ **EXAMPLE B.11 (Conjugate Prior for the Poisson Distribution)**

Let $x_1, \ldots, x_n \sim_{\text{iid}} \mathsf{Poi}(\lambda)$, with sample mean $\bar{x} = (x_1 + \cdots + x_n)/n$. The joint pdf can be written in the form (B.29) as

$$f(\mathbf{x} \mid \lambda) = \mathrm{e}^{-n\lambda} \mathrm{e}^{n\bar{x} \ln \lambda} \prod_{i=1}^n \frac{1}{x_i!} ,$$

which suggests a conjugate prior of the form $f(\lambda) \propto \mathrm{e}^{-b\lambda} \mathrm{e}^{a \ln \lambda} = \mathrm{e}^{-b\lambda} \lambda^a$, corresponding to the Gamma distribution. In particular, if we take a $\mathsf{Gamma}(\alpha, \beta)$ prior for λ, that is,

$$f(\lambda) \propto \mathrm{e}^{-\beta\lambda} \lambda^{\alpha - 1} ,$$

then the posterior pdf is

$$f(\lambda \mid \mathbf{x}) \propto \mathrm{e}^{-(n+\beta)\lambda} \lambda^{\alpha - 1 + n\bar{x}} ,$$

which corresponds to the $\mathsf{Gamma}(\alpha + n\bar{x}, \beta + n)$ distribution.

Further Reading

For an accessible introduction to mathematical statistics with simple applications see [5]. For more detailed overview of statistical inference, see Casella and Berger [2]. A standard reference for classical or frequentist statistical inference is [6]. An applied reference for Bayesian inference is [3]. For a survey of numerical techniques relevant to computational statistics see [7].

REFERENCES

1. P. J. Bickel and K. A. Doksum. *Mathematical Statistics*, volume I. Pearson Prentice Hall, Upper Saddle River, NJ, second edition, 2007.
2. G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, Pacific Grove, CA, second edition, 2001.
3. A. Gelman. *Bayesian Data Analysis*. Chapman & Hall, New York, second edition, 2004.
4. R. V. Hogg and T. A. Craig. *Introduction to Mathematical Statistics*. Prentice Hall, New York, fifth edition, 1995.
5. R. J. Larsen and M. L. Marx. *An Introduction to Mathematical Statistics and Its Applications*. Prentice Hall, New York, third edition, 2001.
6. E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, New York, second edition, 1998.
7. J. F. Monahan. *Numerical Methods of Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, London, 2010.
8. J. Neyman and E. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, 231:289–337, 1933. DOI:10.1098/rsta.1933.0009.
9. B. L. Welch. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.