CHAPTER 2

STATISTICAL INFERENCE

We typically presume, at least, that a random sample of data $x_1, ..., x_n$ are available, that is that n independent observations of random variables with probability mass/density function f_X have been recorded. Much of "non-inferential" statistics is concerned with *summary* and *presentation* of the raw data. Here, we will concentrate on *statistical inference* and *testing*.

2.1 ESTIMATION

The objective is to learn about parameters in a probability model from data. We assume that the mass/density function f_X depends on (vector) parameter $\theta = (\theta_1, ..., \theta_P)$. Two methods of estimation of θ are used:

- **Method of moments:** match sample moments to theoretical theoretical moments implied by the model, that is, solve the system of P equations

$$E_{f_X}\left[X^p\right] = \frac{1}{n}\sum_{i=1}^{n} x_i^p \qquad\qquad p = 1, 2, ..., P$$

- **Method of Maximum Likelihood:** choose estimate $\widehat{\theta}$ as

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} f_X(x_1, ..., x_n; \theta)$$

For an independent sample,

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} f_X\left(x;\theta\right) = \arg\max_{\theta\in\Theta} \log\left\{\prod_{i=1}^{n} f_X\left(x;\theta\right)\right\} = \arg\max_{\theta\in\Theta}\left\{\sum_{i=1}^{n} \log f_X\left(x_i;\theta\right)\right\}$$

where

$$\log L(\theta) = \sum_{i=1}^{n} \log f_X\left(x_i;\theta\right)$$

is the log-likelihood. If $\log L(\theta)$ is differentiable on parameter space Θ, then $\widehat{\theta}$ satisfies

$$\frac{\partial}{\partial\theta_p}\left\{\log L(\theta)\right\} = \sum_{i=1}^{n} \frac{\partial}{\partial\theta_p}\left\{\log f_X\left(x_i;\theta\right)\right\} = 0 \tag{2.1.1}$$

for all $p = 1, 2, ...P$.

Definition 2.1.1 *The **standard error** of an estimator T of parameter θ is*

$$s.e.\left(T;\theta\right) = \sqrt{Var_{f_{T|\theta}}\left[T\right]} = s_e\left(\theta\right)$$

*for some function s_e. The **estimated standard error** is*

$$e.s.e\left(T\right) = s_e\left(\widehat{\theta}\right)$$

where $\widehat{\theta}$ is the estimate of θ (that is, the observed value of T).

2.1.1 LIKELIHOOD INFERENCE AND SCORE EQUATIONS

We consider likelihood inference for a (possibly vector) parameter θ taking values in parameter space Θ, with data $x = (x_1, ..., x_n)$ and likelihood function

$$L(\theta) = f_{X|\theta}(x;\theta) = \prod_{i=1}^{n} f_{X_i}(x_i;\theta).$$

In all the results below, $f_{X|\theta}(x;\theta)$ will denote the likelihood for all of $x_1, ..., x_n$.

Definition 2.1.2 ***Score Function***

For univariate θ the ***observed score function*** *is denoted $U_\theta(x)$*

$$U_\theta(x) = \frac{\partial}{\partial\theta} \log f_{X|\theta}(x;\theta).$$

The score function can be shown to exhibit certain key properties. In the univariate case, we have the following;

(1) If $\phi = g(\theta)$ is some reparameterization, then by the chain rule for differentiation

$$\frac{\partial h(g(\theta))}{\partial\theta} = \frac{\partial h(g(\theta))}{\partial g(\theta)} \frac{\partial (g(\theta))}{\partial\theta} \Longleftrightarrow \frac{\partial h(g(\theta))}{\partial\theta} = \frac{\partial h(\phi)}{\partial\phi} \frac{\partial\phi}{\partial\theta}$$

we have that

$$U_\phi(x) = \left(\frac{\partial g(\theta)}{\partial\theta}\right)^{-1} U_\theta(x)$$

(2) The function $U_\theta(X)$ is a random variable.

(3) The expectation of $U_\theta(X)$ with respect to the joint distribution of $X_1, ..., X_n$ is zero. First note that

$$\begin{aligned} \frac{\partial}{\partial\theta}\left\{\log f_{X|\theta}(x;\theta)\right\} \times f_{X|\theta}(x;\theta) &= \frac{1}{f_{X|\theta}(x;\theta)} \times \frac{\partial}{\partial\theta}\left\{f_{X|\theta}(x;\theta)\right\} \times f_{X|\theta}(x;\theta) \\ &= \frac{\partial}{\partial\theta}\left\{f_{X|\theta}(x;\theta)\right\} \end{aligned}$$

Hence

$$\begin{aligned} E_{f_{X|\theta}}[U_\theta(X)] &= \int \left[\frac{\partial}{\partial\theta}\left\{\log f_{X|\theta}(x;\theta)\right\}\right] f_{X|\theta}(x;\theta)\, dx \\ &= \int \left\{\frac{\partial}{\partial\theta} f_{X|\theta}(x;\theta)\right\} dx \\ &= \frac{\partial}{\partial\theta}\left\{\int f_{X|\theta}(x;\theta)\, dx\right\} = \frac{\partial}{\partial\theta}\{1\} = 0 \end{aligned}$$

From the result in equation (2.1.1), maximum likelihood estimation can be seen to be an operation requiring the solution of the equations

$$\sum_{i=1}^{n} U_\theta(x_i) = 0$$

that is, an equation of the form

$$\sum_{i=1}^{n} G_i(\theta) = 0 \tag{2.1.2}$$

where

$$G_i(\theta) = \frac{\partial}{\partial\theta}\left\{\log f_{X|\theta}(x_i;\theta)\right\}$$

is a function derived from the ith data point. The form in (2.1.2) is known as an **estimating equation.** Writing $G_i(\theta) \equiv G_i(\theta; x_i)$, the function $G_i(\theta; X_i)$ is a random variable; the vector function $G(\theta) = (G_1(\theta), ..., G_n(\theta))$ is an estimating function.

Definition 2.1.3 ***Fisher Information***

The ***Fisher Information*** *is denoted $I(\theta)$, and is defined (equivalently) by*

$$\begin{aligned} I(\theta) &= Var_{f_{X|\theta}}[U_\theta(X)] = E_{f_{X|\theta}}\left[\{U_\theta(X)\}^2\right] \\ &= E_{f_{X|\theta}}\left[\left\{\frac{\partial}{\partial\theta}\log f_{X|\theta}(X;\theta)\right\}^2\right] = E_{f_{X|\theta}}\left[-\frac{\partial^2}{\partial\theta^2}\log f_{X|\theta}(X;\theta)\right] \\ &= E_{f_{X_1|\theta}}\left[-n\frac{\partial^2}{\partial\theta^2}\log f_{X_1|\theta}(X_1;\theta)\right]. \end{aligned}$$

For vector parameter $\theta = (\theta_1, ..., \theta_K)$ the definitions for score function and Fisher information can be extended naturally; the vector score function is

$$U_\theta(x) = (U_{\theta_1}(x), ..., U_{\theta_K}(x))^T$$

with

$$U_{\theta_k}(x) = \frac{\partial}{\partial\theta_k}\left\{\log f_{X|\theta}(x;\theta)\right\}$$

and the Fisher information becomes a $K \times K$ variance-covariance matrix with (i,j)th entry

$$\begin{aligned} [I(\theta)]_{ij} &= E_{f_{X|\theta}}\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\left\{\log f_{X|\theta}(X;\theta)\right\}\right] \\ &= -nE_{f_{X|\theta}}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\left\{\log f_{X_1|\theta}(X_1;\theta)\right\}\right] \end{aligned}$$

2.1.2 DESIRABLE PROPERTIES OF ESTIMATORS

Suppose that T (a function of $(X_1, ..., X_n)$) is an estimator of parameter θ that has sampling density $f_{T|\theta}$. It may be desirable that the estimator has the following properties

- **Unbiasedness:** T is **unbiased** for θ if

$$E_{f_{T|\theta}}[T] = \theta.$$

The **bias** of T is

$$b(\theta) = E_{f_{T|\theta}}[T] - \theta.$$

The **mean square error (MSE)** of T is

$$MSE(\theta) = E_{f_{T|\theta}}\left[(T-\theta)^2\right] = Var_{f_{T|\theta}}[T] + \{b(\theta)\}^2$$

- **Consistency**: T is a **consistent** estimator of θ if

$$\lim_{n\to\infty} P[|T-\theta| < \varepsilon] = 1$$

that is T converges in probability to θ as $n \to \infty$.

- **Efficiency:** An unbiased estimator T is more **efficient** than another unbiased estimator T^* if, for all θ

$$Var_{f_{T|\theta}}[T] \leq Var_{f_{T^*|\theta}}[T^*].$$

In fact, there is a lower bound on the variance of unbiased estimators in many problems; the **Cramer Rao Lower Bound** indicates that if T is an unbiased estimator of θ, then for each θ

$$Var_{f_{T|\theta}}[T] \geq \frac{1}{I(\theta)}$$

with equality if an only if the score function satisfies

$$U_\theta(\theta) = g(\theta)(T-\theta)$$

for some function g. If equality is found, then T is the minimum variance unbiased estimator (MVUE).

- **Asymptotic properties**: two of the properties above relate to finite sample results. We can extend these ideas by considering **asymptotic unbiasedness** and **asymptotic efficiency**, that is, as $n \to \infty$.

It can be shown that, under mild regularity conditions, **maximum likelihood estimators** exist

- **consistent,**
- **asymptotically unbiased**,
- **efficient** estimators with variance equal to

$$[I(\theta)]^{-1}$$

in the limit as $n \to \infty$. In fact, the maximum likelihood estimator is **asymptotically normally distributed**

$$\widehat{\theta}_{ML} \xrightarrow{d} N\left(\theta, [I(\theta)]^{-1}\right)$$

2.2 HYPOTHESIS TESTING : NORMAL SAMPLES

Given a sample $x_1, ..., x_n$ from a probability model $f(x;\theta)$ depending on parameter θ, we can produce an estimate $\hat{\theta}$ of θ, and in some circumstances understand how $\hat{\theta}$ varies for repeated samples. Now we might want to test, say, whether or not there is evidence from the sample that true (but unobserved) value of θ is not equal to a specified value. To do this, we use an estimate of θ, and the corresponding estimator and its sampling distribution, to quantify this evidence.. First, we concentrate on data samples that we can presume to have a normal distribution, and look at two situations, namely **one sample** and **two sample** experiments.

- **ONE SAMPLE:** Random variables: $X_1, ..., X_n \sim N(\mu, \sigma^2)$, sample observations $x_1, ...x_n$ - possible comparisons of interest: $\mu = \mu_0$, $\sigma = \sigma_0$
- **TWO SAMPLE:** Random variables $X_1, ..., X_n \sim N(\mu_X, \sigma_X^2)$, $Y_1, ..., Y_n \sim N(\mu_Y, \sigma_Y^2)$, data $x_1, ...x_{n_X}$ and $y_1, ...y_{n_Y}$- possible comparisons of interest $\mu_X = \mu_Y, \sigma_X = \sigma_Y$

2.2.1 HYPOTHESIS TESTS FOR NORMAL DATA I - THE Z-TEST (σ KNOWN)

If $X_1, ..., X_n \sim N(\mu, \sigma^2)$ are the i.i.d. outcome random variables of n experimental trials, then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{and} \qquad \frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$$

with $\bar{X}$ and S^2 statistically independent. Suppose we want to test the **hypothesis** that $\mu = \mu_0$, for some specified constant μ_0, (where, for example, $\mu_0 = 20.0$) is a plausible model; more specifically, we want to test the hypothesis $H_0 : \mu = \mu_0$ against the hypothesis $H_1 : \mu \neq \mu_0$, that is, we want to test whether H_0 is true, or whether H_1 is true. Now, we know that, in the case of a Normal sample, the distribution of the estimator $\bar{X}$ is Normal, and

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Longrightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

where Z is a **random variable**. Now, when we have observed the data sample, we can calculate $\bar{x}$, and therefore we have a way of testing whether $\mu = \mu_0$ is a plausible model; we calculate $\bar{x}$ from $x_1, ..., x_n$, and then calculate

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

If H_0 is true, and $\mu = \mu_0$, then the **observed** z should be an observation from an $N(0, 1)$ distribution (as $Z \sim N(0, 1)$), that is, it should be near zero with high probability. In fact, z should lie between -1.96 and 1.96 with probability $1 - \alpha = 0.95$, say, as

$$\mathrm{P}[-1.96 \leq Z < 1.96] = \Phi(1.96) - \Phi(-1.96) = 0.975 - 0.025 = 0.95$$

.If we observe z to be outside of this range, then there is evidence that H_0 is **not true**. Alternatively, we could calculate the probability p of observing a z value that is **more extreme** than the z we did observe; this probability is given by

$$p = \begin{cases} 2\Phi(z) & z < 0 \\ 2(1 - \Phi(z)) & z \geq 0 \end{cases}$$

If p is very small, say $p \leq \alpha = 0.05$, then again. there is evidence that H_0 is **not true**. In summary, we need to assess whether z is a **surprising** observation from an $N(0, 1)$ distribution - if it is, then we can **reject** H_0.

2.2.2 HYPOTHESIS TESTING TERMINOLOGY

There are five crucial components to a hypothesis test, namely

- **TEST STATISTIC**
- **NULL DISTRIBUTION**
- **SIGNIFICANCE LEVEL**, denoted α
- **CRITICAL VALUE(S)** (C_{R_1}, C_{R_2})
- **P-VALUE**, denoted p.

In the Normal example given above, we have that

z is the **test statistic**
The distribution of random variable Z if H_0 is true is the **null distribution**
$\alpha = 0.05$ is the **significance level** of the test (we could use $\alpha = 0.01$ if we require a "stronger" test).
The solution C_R of $\Phi(C_R) = 1 - \alpha/2$ ($C_R = 1.96$ above) gives the **critical values** of the test $\pm C_R$.
p is the **p-value** of the test statistic under the null distribution

EXAMPLE : A sample of size 10 has sample mean $\bar{x} = 19.7$. Suppose we want to test the hypothesis

$$H_0 : \mu = 20.0$$
$$H_1 : \mu \neq 20.0$$

under the assumption that the data follow a Normal distribution with $\sigma = 1.0$.

We have

$$z = \frac{19.7 - 20.0}{1/\sqrt{10}} = -0.95$$

which lies between the critical values ± 1.96, and therefore we have no reason to reject H_0. Also, the p-value is given by $p = 2\Phi(-0.95) = 0.342$, which is greater than $\alpha = 0.05$, which confirms that we have no reason to reject H_0.

2.2.3 HYPOTHESIS TESTS FOR NORMAL DATA II - THE T-TEST (σ UNKNOWN)

In practice, we will often want to test hypotheses about μ when σ is unknown. We cannot perform the Z-test, as this requires knowledge of σ to calculate the z statistic.

We proceed as follows; recall that we know the sampling distributions of $\bar{X}$ and s^2, and that the two estimators are statistically independent. Now, from the properties of the Normal distribution, if we have independent random variables $Z \sim N(0, 1)$ and $Y \sim \chi^2_\nu$, then we know that random variable T defined by

$$T = \frac{Z}{\sqrt{Y/\nu}}$$

has a Student-t distribution with ν degrees of freedom. Using this result, and recalling the sampling distributions of $\bar{X}$ and s^2, we see that

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2/\sigma^2}{(n-1)}}} = \frac{(\bar{X}-\mu)}{s/\sqrt{n}} \sim t_{n-1}$$

and T has a Student-t distribution with $n-1$ degrees of freedom, denoted $St(n-1)$. Thus we can repeat the procedure used in the σ known case, but use the sampling distribution of T rather than that of Z to assess whether the test statistic is "surprising" or not.

Specifically, we calculate

$$t = \frac{(\bar{x}-\mu)}{s/\sqrt{n}}$$

and find the critical values for a $\alpha = 0.05$ significance test by finding the ordinates corresponding to the 0.025 and 0.975 percentiles of a Student-t distribution, $St(n-1)$ (rather than a $N(0,1)$) distribution.

EXAMPLE : A sample of size 10 has sample mean $\bar{x} = 19.7.$ and $s^2 = 0.78^2$. Suppose we want to carry out a test of the hypotheses

$$H_0 : \mu = 20.0$$
$$H_1 : \mu \neq 20.0$$

under the assumption that the data follow a Normal distribution with σ unknown.

We have test statistic t given by

$$t = \frac{19.7 - 20.0}{0.78/\sqrt{10}} = -1.22.$$

The upper critical value C_R is obtained by solving

$$F_{t_{n-1}}(C_R) = 0.975$$

where $F_{St(n-1)}$ is the c.d.f. of a Student-t distribution with $n-1$ degrees of freedom; here $n = 10$, so we can use the statistical tables to find $C_R = 2.262$, and not that, as Student-t distributions are symmetric the lower critical value is $-C_R$. Thus t lies between the critical values, and therefore we have no reason to reject H_0. The p-value is given by

$$p = \begin{cases} 2F_{t_{n-1}}(t) & t < 0 \\ 2(1 - F_{t_{n-1}}(t)) & t \geq 0 \end{cases}$$

so here, $p = 2F_{t_{n-1}}(-1.22)$ which we can find to give $p = 0.253$; this confirms that we have no reason to reject H_0.

2.2.4 HYPOTHESIS TESTS FOR NORMAL DATA III - TESTING σ.

The Z-test and T-test are both tests for the parameter μ. Suppose that we wish to test a hypothesis about σ, for example

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

We construct a test based on the estimate of variance, s^2. In particular, we saw from the Theorem on p.32 that the random variable Q, defined by

$$Q = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

if the data have an $N(\mu, \sigma^2)$ distribution. Hence if we define test statistic q by

$$q = \frac{(n-1)s^2}{\sigma_0^2}$$

then we can compare q with the critical values derived from a χ^2_{n-1} distribution; we look for the 0.025 and 0.975 quantiles - note that the Chi-squared distribution is not symmetric, so we need two distinct critical values.

In the above example, to test

$$H_0 : \sigma^2 = 1.0$$
$$H_1 : \sigma^2 \neq 1.0$$

we compute test statistic

$$q = \frac{(n-1)s^2}{\sigma_0^2} = \frac{90.78^2}{1.0} = 5.43.75$$

and compare with

$$C_{R_1} = F_{\chi^2_{n-1}}(0.025) \implies C_{R_1} = 2.700$$
$$C_{R_2} = F_{\chi^2_{n-1}}(0.975) \implies C_{R_2} = 19.022$$

so q is not a surprising observation from a χ^2_{n-1} distribution, and hence we cannot reject H_0.

2.2.5 TWO SAMPLE TESTS

It is straightforward to extend the ideas from the previous sections to two sample situations where we wish to compare the distributions underlying two data samples. Typically, we consider sample one, $x_1, ..., x_{n_X}$, from a $N(\mu_X, \sigma_X^2)$ distribution, and sample two, $y_1, ..., y_{n_Y}$, independently from a $N(\mu_Y, \sigma_Y^2)$ distribution, and test the equality of the parameters in the two models. Suppose that the sample mean and sample variance for samples one and two are denoted $(\bar{x}, s_X^2)$ and $(\bar{y}, s_Y^2)$ respectively.

First, consider testing the hypothesis

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y$$

when $\sigma_X = \sigma_Y = \sigma$ is known. Now, we have from the sampling distributions theorem we have

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma^2}{n_X}\right) \qquad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma^2}{n_Y}\right) \implies \bar{X} - \bar{Y} \sim N\left(0, \frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}\right)$$

and hence

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim N(0,1)$$

giving us a test statistic z defined by

$$z = \frac{\bar{x} - \bar{y}}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

which we can compare with the standard normal distribution; if z is a surprising observation from $N(0,1)$, and lies outside of the critical region, then we can reject H_0. This procedure is the Two Sample Z-Test.

If $\sigma_X = \sigma_Y = \sigma$ is unknown, we parallel the one sample T-test by replacing σ by an estimate in the two sample Z-test. First, we obtain an estimate of σ by "pooling" the two samples; our estimate is the **pooled estimate**, s_P^2, defined by

$$s_P^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}$$

which we then use to form the test statistic t defined by

$$t = \frac{\bar{x} - \bar{y}}{s_P\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

It can be shown that, if H_0 is true then t should be an observation from a Student-t distribution with $n_X + n_Y - 2$ degrees of freedom. Hence we can derive the critical values from the tables of the Student-t distribution.

If $\sigma_X \neq \sigma_Y$, but both parameters are known, we can use a similar approach to the one above to derive test statistic z defined by

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

which has an $N(0,1)$ distribution if H_0 is true.

Clearly, the choice of test depends on whether $\sigma_X = \sigma_Y$ or otherwise; we may test this hypothesis formally; to test

$$H_0 : \sigma_X = \sigma_Y$$
$$H_1 : \sigma_X \neq \sigma_Y$$

we compute the test statistic

$$q = \frac{s_X^2}{s_Y^2}$$

which has a null distribution known as the **Fisher** or F distribution with $(n_X - 1, n_Y - 1)$ degrees of freedom; this distribution can be denoted $F(n_X - 1, n_Y - 1)$, and its quantiles are tabulated. Hence we can look up the 0.025 and 0.975 quantiles of this distribution (the F distribution is not symmetric), and hence define the critical region; informally, if the test statistic q is very small or very large, then it is a surprising observation from the F distribution and hence we reject the hypothesis of equal variances.

2.2.6 ONE-SIDED AND TWO-SIDED TESTS

So far we have considered hypothesis tests of the form

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

which is referred to as a **two-sided test**, that is, the alternative hypothesis is supported by an extreme test statistic in **either** tail of the distribution. We may also consider a **one-sided test** of the form

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned} \quad \text{or} \quad \begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu < \mu_0 \end{aligned} .$$

Such a test proceeds exactly as the two-sided test, except that a significant result can only occur in the right (or left) tail of the null distribution, and there is a single critical value, placed, for example, at the 0.95 (or 0.05) probability point of the null distribution.

2.2.7 CONFIDENCE INTERVALS

The procedures above allow us to test specific hypothesis about the parameters of probability models. We may complement such tests by reporting a **confidence interval**, which is an interval in which we believe the "true" parameter lies with high probability. Essentially, we use the sampling distribution to derive such intervals. For example, in a one sample Z-test, we saw that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

that is, that, for critical values $\pm C_R$ in the test at the 5 % significance level

$$\mathrm{P}\left[-C_R \leq Z \leq C_R\right] = \mathrm{P}\left[-C_R \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq C_R\right] = 0.95$$

Now, from tables we have $C_R = 1.96$, so re-arranging this expression we obtain

$$\mathrm{P}\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

from which we deduce a **95 % Confidence Interval** for μ based on the sample mean $\bar{x}$ of

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

We can derive other confidence intervals (corresponding to different significance levels in the equivalent tests) by looking up the appropriate values of the critical values. The general approach for construction of confidence interval for generic parameter θ proceeds as follows. ¿From the modelling assumptions, we derive a **pivotal quantity**, that is, a statistic, T_{PQ}, say, (usually the test statistic random variable) that depends on θ, but whose sampling distribution is "parameter-free" (that is, does not depend on θ). We then look up the critical values C_{R_1} and C_{R_2}, such that

$$\mathrm{P}\left[C_{R_1} \leq T_{PQ} \leq C_{R_2}\right] = 1 - \alpha$$

where α is the significance level of the corresponding test. We then rearrange this expression to the form

$$\mathrm{P}\left[c_1 \leq \theta \leq c_2\right] = 1 - \alpha$$

where c_1 and c_2 are functions of C_{R_1} and C_{R_2} respectively. Then a $1 - \alpha$ % Confidence Interval for θ is $[c_1, c_2]$.

SUMMARY

For a data sample $x_1, ..., x_n$, with corresponding random variables $X_1, ..., X_n$, we

1. consider a pair of competing **hypotheses**, H_0 and H_1
2. define a suitable test statistic $T = T(X_1, ..., X_n)$ (that is, some function of the original random variables; this will define the **test statistic**), and a related pivotal random variable $T_{PQ} = T_{PQ}(X)$
3. **assume that H_0 is true**, and compute the sampling distribution of T, f_T or F_T; this is the **null distribution**
4. compute the **observed** value of T, $t = T(x_1, ..., x_n)$; this is the **test statistic**
5. assess whether t is a surprising observation from the distribution f_T. If it **is** surprising, we have evidence to **reject** H_0; if it is not surprising, we **cannot reject** H_0

This strategy can be applied to more complicated normal examples, and also non-normal and non-parametric testing situations. It is a general strategy for assessing the statistical evidence for or against a hypothesis. For the tests discussed in previous sections, the calculation of the form of the confidence intervals is straightforward: in each case, C_{R_1} and C_{R_2} are the $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of the pivotal quantity.

One Sample Tests

Test	Pivotal Quantity T_{PQ}	Null Distribution	Parameter	Confidence Interval
Z-TEST	$Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	$N(0,1)$	μ	$\bar{x} \pm C_R\ \sigma/\sqrt{n}$
T-TEST	$T = \dfrac{\bar{X} - \mu}{s/\sqrt{n}}$	$St(n-1)$	μ	$\bar{x} \pm C_R\ s/\sqrt{n}$
Q-TEST	$Q = \dfrac{(n-1)s^2}{\sigma^2}$	χ^2_{n-1}	σ^2	$\left[\dfrac{(n-1)s^2}{C_{R_2}} : \dfrac{(n-1)s^2}{C_{R_1}}\right]$

Two Sample Tests

Test	Pivotal Quantity T_{PQ}	Null Distribution	Parameter	Confidence Interval
Z-TEST(1)	$Z = \dfrac{(\bar{X} - \mu_X) - (\bar{Y} - \mu_Y)}{\sigma\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}}$	$N(0,1)$	$\mu_X - \mu_Y$	$(\bar{x} - \bar{y}) \pm C_R\ \sigma\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}$
T-TEST(2)	$T = \dfrac{(\bar{X} - \mu_X) - (\bar{Y} - \mu_Y)}{s_P\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}}$	$St(n_X + n_Y - 2)$	$\mu_X - \mu_Y$	$(\bar{x} - \bar{y}) \pm C_R\ s_P\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}$
Z-TEST(2)	$Z = \dfrac{(\bar{X} - \mu_X) - (\bar{Y} - \mu_Y)}{\sqrt{\dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}}}$	$N(0,1)$	$\mu_X - \mu_Y$	$(\bar{x} - \bar{y}) \pm C_R\sqrt{\dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}}$
Q-TEST	$Q = \dfrac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2}$	$F(n_X - 1, n_Y - 1)$	$\dfrac{\sigma_X^2}{\sigma_Y^2}$	$\left[\dfrac{s_X^2}{C_{R_2}s_Y^2} : \dfrac{s_X^2}{C_{R_1}s_Y^2}\right]$

2.3 HYPOTHESIS TESTING : NEYMAN-PEARSON AND EXTENSIONS

The general strategy for statistical hypothesis testing was outlined previously and proceeds as follows: for a data sample $x_1, ..., x_n$, with corresponding random variables $X_1, ..., X_n$, we

1. consider a pair of competing **hypotheses**, H_0 and H_1
2. define a suitable **test statistic** $T = T(X_1, ..., X_n)$ (that is, some function of the original random variables)
3. **assume that H_0 is true**, and compute the sampling distribution of T, f_T or F_T; this is the **null distribution**
4. compute the **observed** value of T, $t = T(x_1, ..., x_n)$; this is the **test statistic**
5. assess whether t is a surprising observation from the distribution f_T. If it **is** surprising, we have evidence to **reject** H_0; if it is not surprising, we **cannot reject** H_0. The highest level of acceptable "surprise" is related to the specified significance level α.

2.3.1 CRITERIA FOR ASSESSING TESTS

Definition 2.3.1 ***TYPE I and TYPE II ERRORS.***

*A **Type I error** occurs when the null hypothesis H_0 is REJECTED when it is in fact TRUE.*
*A **Type II error** occurs when the null hypothesis H_0 is ACCEPTED when it is in fact FALSE.*

The concept of these types of error can be best illustrated by the following 2×2 table

		UNOBSERVED TRUTH	
		H_0 True	H_0 not True
OBSERVED RESULT OF TEST	H_0 rejected	**TYPE I ERROR**	✓
	H_0 not rejected	✓	**TYPE II ERROR**

Consider in this testing context a partition of the parameter space $\Theta \equiv \Theta_0 \cup \Theta_1$ where Θ_0 and Θ_1 are the ranges of values of θ implied by H_0 and H_1 respectively

Definition 2.3.2 *For a particular test, let $R \subset \mathbb{X}$ be the critical region. For $\theta \in \Theta_0$, define the **Type I error probability** $\alpha(\theta)$ by*

$$\alpha(\theta) = P[X \in R | \theta \in \Theta_0] \tag{2.3.3}$$

If Θ_0 comprises a single value, then $\alpha = P[X \in R | H_0$ is TRUE$]$

*For $\theta \in \Theta_1$ define the **Type II error probability** $\beta(\theta)$ by*

$$\beta(\theta) = P[X \notin R | \theta \in \Theta_1] \tag{2.3.4}$$

*and the **power function** by*

$$1 - \beta(\theta) = P[X \in R | \theta \in \Theta_1]$$

Intuitively, the power function should be **high**, as it represents the probability of correctly rejecting H_0.

Definition 2.3.3 *For constant α, if $\alpha(\theta) \leq \alpha < 1$ for all $\theta \in \Theta_0$ and if $X \in R$, then H_0 is* ***rejected at significance level*** *α.*

NOTE: The value α is any upper bound of $\alpha(\theta)$, so if H_0 is rejected at level γ then (by definition) it is rejected at level $\gamma + \delta$ for $\delta > 0$

Definition 2.3.4 *The* ***size*** *of a statistical test is*

$$\sup_{\theta \in \Theta_0} \alpha(\theta)$$

which is equal to α if Θ_0 comprises a single value.

The task remaining is to construct the critical region R; this is achieved by specifying the test statistic and the probabilities described in the previous definitions,, and identifying the subset of for which the target probabilities are met, via (2.3.3) and (2.3.4). It would be ideal to be able to specify $\alpha(\theta)$ and $\beta(\theta)$ small, uniformly in θ. These objectives cannot be achieved simultaneously, and hence a compromise is sought. Now H_0 and H_1 are treated asymmetrically; in reality the testing context relates to finding sufficient evidence to reject H_0, rather than deciding between the two hypotheses. Hence rejecting H_0 when it is true (a Type I error) is a more serious error than accepting H_0 when it is false (a Types II error), and this implies that the size of the test, α, should be fixed at some small level (0.05 or 0.01 say), and subject to this constraint, the test constructed that minimizes $\beta(\theta)$ uniformly in θ. That is, set $\alpha(\theta) \leq \gamma$, for all $\theta \in \Theta_0$, and then attempt to find a test to minimize $\beta(\theta)$ for all θ.

Definition 2.3.5 *If a test is of size α and if $\beta(\theta)$ uniformly in θ minimizes the probability of a Type II error among all tests with size $\leq \alpha$, then the test is termed the* ***uniformly most powerful (UMP)*** *test of size α.*

Definition 2.3.6 *A* ***simple hypothesis*** *is one which specifies the distribution of the data completely.*

Example 2.3.1 Consider the parameter space $\Theta = \{\theta_0, \theta_1\}$ where $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, and

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1 \end{aligned}$$

These are two simple hypotheses, and

$$\beta(\theta) = \beta(\theta_1) = \beta, \text{ say}$$

The next result constructs the UMP test for this example as based on a test statistic which is a function of the ratio

$$T = \frac{f_{X|\theta}(X; \theta_1)}{f_{X|\theta}(X; \theta_0)}$$

where X here is a (vector) random variable.

Theorem 2.3.1 ***The Neyman-Pearson Lemma***

Suppose that $\Theta = \{\theta_0, \theta_1\}$ *and a test of the hypotheses*

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1 \end{aligned}$$

is required. Suppose that, as a requirement, the size is to be less than or equal to α*. Then the most powerful test of* H_0 *against* H_1 *is defined by any the critical region*

$$R = \left\{ x : \frac{f_{X|\theta}(x;\theta_1)}{f_{X|\theta}(x;\theta_0)} \geq k \right\}$$

where $k > 0$ *is defined by*

$$\alpha(\theta_0) = P\left[T \in R | \theta = \theta_0\right] = \alpha$$

To evaluate the value of constant k that appears in the Theorem, need to compute $\mathrm{P}[T \in R|\theta = \theta_0]$ for a fixed size α. To do this, the sampling distribution of T given that $\theta = \theta_0$ (that is, the null distribution) must be used.

2.3.2 COMPOSITE HYPOTHESES.

Often the hypotheses do not specify the distribution of the data completely. For example, hypotheses

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

could be of interest. If, in general, a UMP test of size α is to be found, then its power must equal the power of the most powerful test against a point alternative hypothesis, $H_1 : \theta = \theta_1$, **for all** $\theta_1 \in \Theta \backslash \theta_0$.

Note that (i) It is possible that, for given alternative hypotheses, no UMP test exists, and (ii) for discrete data, it may not be possible to solve the equation $P\left[T \in R|\theta = \theta_0\right] = \alpha$, for every value of α, and hence only specific values of α may be attained.

UMP tests do not always exist, so it is convenient to also consider **locally most powerful** tests. Consider the hypothesis

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta > \theta_0 \end{aligned}$$

and a partial alternative $\theta_1 = \theta_0 + \delta$ for δ small. Then

$$\log \frac{f_{X|\theta}(X;\theta_1)}{f_{X|\theta}(X;\theta_0)} = \log \frac{f_{X|\theta}(X;\theta_0+\delta)}{f_{X|\theta}(X;\theta_0)} \approx \delta \frac{\partial}{\partial \theta} \log f_{X|\theta}(X;\theta)\big|_{\theta=\theta_0} = \delta U_{\theta_0}(X)$$

where $U_{\theta_0}(X)$ is the **score function** defined previously. For these alternatives, H_0 would be rejected if $U_{\theta_0}(x)$ is too large; this test is **locally most powerful** in the sense that it maximizes the **slope** of the power function at $\theta = \theta_0$. To complete the calculation, the distribution of $U_{\theta_0}(X)$ is needed. Using the Central Limit Theorem, it follows that

$$U_{\theta_0}(X) \sim N(0, I(\theta_0)) \tag{2.3.5}$$

where $I(\theta_0)$ is the Fisher information at $\theta = \theta_0$. This result forms the basis for the **Rao Score Test** described below.

2.3.3 THE LIKELIHOOD RATIO TEST

As an alternative to the search for most powerful tests, consider a general approach to test construction, and assess the qualities of the types of test constructed.

Definition 2.3.7 ***Likelihood Ratio Test***

The ***Likelihood Ratio Test*** *statistic for testing H_0 against H_1 is*

$$T = T(X) = \frac{\sup\limits_{\theta \in \Theta_1} f_{X|\theta}(X;\theta)}{\sup\limits_{\theta \in \Theta_0} f_{X|\theta}(X;\theta)}$$

where H_0 is rejected if T is too large, that is, if

$$P[T \geq k | H_0] = \alpha$$

Theorem 2.3.2 *If H_0 imposes q independent constraints on H_1, then, as $n \to \infty$*

$$2\log T = 2\log \frac{\sup\limits_{\theta \in \Theta_1} f_{X|\theta}(X;\theta)}{\sup\limits_{\theta \in \Theta_0} f_{X|\theta}(X;\theta)} \overset{A}{\sim} \chi^2_q \tag{2.3.6}$$

that is, $2\log T$ has an approximate Chi-squared distribution with q degrees of freedom.
(The notation $\overset{A}{\sim}$ means "is approximately distributed as").

Equation (2.3.6) gives one method of testing two hypotheses; there are two other related testing approaches, both of which rely on asymptotic normal approximations. These methods are the **Rao Score Test** and the **Wald** test. First, some notation is needed. For convenience, let

$$l_n(\theta) = f_{X|\theta}(x;\theta)$$

be the likelihood function for vector parameter $\theta = (\theta_1, ..., \theta_K)^T$, and define vector quantity

$$l'_n(\theta) = (U_{\theta_1}(x), ...U_{\theta_K}(x))^T$$

as the observed vector score function with k^{th} element

$$U_{\theta_k}(x) = \frac{\partial}{\partial \theta_k}\left\{\log f_{X|\theta}(x;\theta)\right\} = \sum_{i=1}^{n} \log f_{X_i|\theta}(x_i;\theta) \qquad k = 1, ..., K$$

When required, we will also use $l'_n(\theta)$ to denote the corresponding vector random variable

$$(U_{\theta_1}(X), ...U_{\theta_K}(X))^T$$

We now wish to test a null hypothesis $H_0 : \theta = \theta_0$. First, let $\widetilde{\theta}_n$ be an estimate (or the corresponding estimator) derived from solving the score equations which, in the current notation, can be written

$$l'_n(\theta) = 0$$

(in the simplest case, $\widetilde{\theta}_n$ is the MLE)

2.3.4 THE RAO(SCORE) TEST

The **Rao (Score) Test** statistic, R_n, for testing H_0 against $H_1 : \theta \neq \theta_0$ is defined by

$$R_n = Z_n^T \left[I_1(\theta_0)\right]^{-1} Z_n \tag{2.3.7}$$

where

$$Z_n = \frac{1}{\sqrt{n}} l_n' \left(\theta_0\right)$$

and where $I_1(.)$ is the Fisher information evaluated at θ_0 for a **single** X_i, that is a $K \times K$ matrix with $(i,j)^{th}$ element

$$\left[I\left(\theta\right)\right]_{ij} = -E_{f_{X|\theta}} \left[\frac{\partial^2}{\partial\theta_i \partial\theta j} \log f_{X|\theta}\left(X;\theta\right)\right]$$

and

$$\left[I_1\left(\theta\right)\right]_{ij} = -E_{f_{X|\theta}} \left[\frac{\partial^2}{\partial\theta_i \partial\theta j} \log f_{X_1|\theta}\left(X_1;\theta\right)\right]$$

as $I\left(\theta\right) = nI_1\left(\theta\right)$. For large n, if H_0 is true,

$$R_n \overset{A}{\sim} \chi_K^2$$

and H_0 is rejected if R_n is too large, that is, if $R_n \geq C$, and where $P\left[R_n \geq C | H_0\right] = \alpha$ for significance level α.

Interpretation and Explanation: we have previously calculated the expectation and variance of the score function; in particular, the expected score is zero, and the variance covariance is determined by the Fisher information. The score test uses these results; if H_0 is true, we would have that, from (2.3.5),

$$U_{\theta_0}\left(X\right) \overset{A}{\sim} N_K\left(0, I\left(\theta_0\right)\right) \equiv N_K\left(0, nI_1\left(\theta_0\right)\right)$$

so that the standardized score

$$V_n = \left\{A\left(\theta_0\right)\right\}^{-1} U_{\theta_0}\left(X\right) \overset{A}{\sim} N_K\left(0, 1_K\right)$$

where 1_K is the $K \times K$ identity matrix, and where matrix $A\left(\theta\right)$ is given by

$$\left\{A\left(\theta\right)\right\}^T \left\{A\left(\theta\right)\right\} = I\left(\theta\right).$$

Hence, by the usual normal distribution theory

$$R_n = V_n^T V_n = \left\{U_{\theta_0}\left(X\right)\right\}^T \left[I\left(\theta_0\right)\right]^{-1} U_{\theta_0}\left(X\right) = Z_n^T \left\{I_1\left(\theta_0\right)\right\}^{-1} Z_n \overset{A}{\sim} \chi_K^2$$

so that observed test statistic

$$r_n = z_n^T \left\{I_1(\theta_0)\right\}^{-1} z_n \qquad \text{where } z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_{\theta_k}(x_i)$$

should be an observation from a χ_K^2 distribution.

Extension: It is legitimate, if required, to replace the Fisher Information at $I_1(\theta_0)$ by a suitable estimate $\hat{I}_n\left(\tilde{\theta}_n\right)$. There are several ways of estimating the Fisher information, for example

$$\hat{I}_n\left(\tilde{\theta}_n\right) = \begin{cases} I_1(\tilde{\theta}_n) \\ \\ \frac{1}{n}\sum_{i=1}^{n}\left\{l'\left(\tilde{\theta}_n; x_i\right)\right\}\left\{l'\left(\tilde{\theta}_n; x_i\right)\right\}^T \\ \\ -\frac{1}{n}\sum_{i=1}^{n} l''\left(\tilde{\theta}_n; x_i\right) \end{cases} \tag{2.3.8}$$

where the j^{th} element of the $K \times 1$ vector $l'\left(\tilde{\theta}_n; x_i\right)$ is

$$\frac{\partial}{\partial \theta_j}\left\{\log f_{X_i|\theta}\left(x_i;\theta\right)\right\}\Big|_{\theta=\tilde{\theta}_n}$$

and $l''\left(\tilde{\theta}_n; x_i\right)$ is the $K \times K$ matrix with $(j,k)^{th}$ entry

$$\frac{\partial^2}{\partial \theta_j \partial \theta_k}\left\{\log f_{X_i|\theta}\left(x_i;\theta\right)\right\}\Big|_{\theta=\tilde{\theta}_n}.$$

These two terms represent the contribution of the i^{th} data point to the observed (or **empirical**) score function and Fisher information.

2.3.5 THE WALD TEST

The Wald test uses similar logic and asymptotic approximations to construct another test statistic. To test a null hypothesis $H_0 : \theta = \theta_0$ for vector parameter $\theta = (\theta_1, ..., \theta_K)$, the **Wald Test** statistic, W_n, for testing H_0 against $H_1 : \theta \neq \theta_0$ is defined by

$$W_n = \sqrt{n}\left(\tilde{\theta}_n - \theta_0\right)^T \left[\hat{I}_n\left(\tilde{\theta}_n\right)\right] \sqrt{n}\left(\tilde{\theta}_n - \theta_0\right) \tag{2.3.9}$$

Then, for large n, if H_0 is true,

$$W_n \overset{A}{\sim} \chi^2_K$$

and H_0 is rejected if W_n is too large, that is, if $W_n \geq C$, and where $\mathrm{P}[W_n \geq C|H_0] = \alpha$ for significance level α.

Interpretation and Explanation: the logic of the Wald test depends on the asymptotic Normal distribution of the score equation derived estimates

$$\sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow{d} N_K\left(0, [I_1(\theta)]^{-1}\right)$$

so that

$$\tilde{\theta}_n \overset{A}{\sim} N\left(\theta, [nI_1(\theta)]^{-1}\right)$$

and standard multivariate Normal Theory. Again, estimates of the Fisher Information such as those in (2.3.8) can be substituted for $I(\theta_0)$ in (2.3.9).

2.3.6 EXTENSION TO TESTS FOR COMPONENTS OF θ.

The theory above concerns tests for the whole parameter vector θ. Often it is of interest to consider components of θ, that is, if $\theta = (\theta_1, \theta_2)$, we might wish to test

$$\begin{aligned} H_0 &: \theta_1 = \theta_{10}, \text{ with } \theta_2 \text{ unspecified} \\ H_1 &: \theta_1 \neq \theta_{10}, \text{ with } \theta_2 \text{ unspecified} \end{aligned}$$

The Rao Score and Wald tests can be developed to allow for testing in this slightly different context. Suppose that θ_1 has dimension m and θ_2 has dimension $K - d$. Let the Fisher information matrix $I_1(\theta)$ and its inverse be partitioned

$$I_1 = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \qquad [I_1(\theta)]^{-1} = \begin{bmatrix} [I_{11.2}]^{-1} & -[I_{11.2}]^{-1} I_{12} [I_{22}]^{-1} \\ -[I_{22.1}]^{-1} I_{21} [I_{11}]^{-1} & [I_{22.1}]^{-1} \end{bmatrix}$$

be a partition of the information matrix, where

$$I_{11.2} = I_{11} - I_{12} [I_{22}]^{-1} I_{21}$$

$$I_{22.1} = I_{22} - I_{21} [I_{11}]^{-1} I_{12}$$

and all quantities depend on θ.

- The **Rao score** statistic is given by

$$R_n = Z_{n0}^T \left[\hat{I}_n \left(\tilde{\theta}_n^{(0)} \right) \right]^{-1} Z_{n0} \overset{A}{\sim} \chi_m^2$$

  where $\tilde{\theta}_n^{(0)}$ is the estimate of θ under H_0

$$Z_{n0} = \frac{1}{\sqrt{n}} l_n' \left(\tilde{\theta}_n^{(0)} \right)$$

  and $\hat{I}_n \left(\tilde{\theta}_n^{(0)} \right)$ is the estimated Fisher information I_1, evaluated at $\tilde{\theta}_n^{(0)}$, obtained using any of the estimates in (2.3.8).

- The **Wald** statistic is given by

$$W_n = \sqrt{n} \left(\tilde{\theta}_{n1} - \theta_{10} \right)^T \left[\hat{I}_n^{(11.2)} \left(\tilde{\theta}_n \right) \right] \sqrt{n} \left(\tilde{\theta}_{n1} - \theta_{10} \right) \overset{A}{\sim} \chi_m^2$$

  where $\tilde{\theta}_{n1}$ is the vector component of $\tilde{\theta}_n$ corresponding to θ_1 under H_1, and $\hat{I}_n^{(11.2)} \left(\tilde{\theta}_n \right)$ is the estimated version of $I_{11.2}$ (using the sample data, under H_1) evaluated at $\tilde{\theta}_n$, obtained using any of the estimates in (2.3.8).

These tests can be carried out in any likelihood-based estimation problem: for example, in a two sample Binomial problem, a test of

$$H_0 : \theta_1 = \theta_2$$

can be re-written in terms of parameters (ϕ_1, ϕ_2) where $\phi_1 = \theta_1 - \theta_2, \phi_2 = \theta_2$, with the null hypothesis re-phrased

$$H_0 : \phi_1 = 0 \text{ with } \phi_2 \text{ unspecified}$$

and the theory of component-wise tests is important.

2.4 BAYESIAN THEORY

The classical (maximum-likelihood/Neyman Pearson) view of Statistical Inference Theory contrasts with the alternative **Bayesian** approach. In Bayesian theory, the likelihood function still plays a central role, but is combined with a **prior** probability distribution to give a **posterior** distribution for the parameters in the model. Inference, estimation, uncertainty reporting and hypothesis testing can be carried out within the Bayesian framework.

2.4.1 PRIOR AND POSTERIOR DISTRIBUTIONS

In the Bayesian framework, inference about an unknown parameter θ is carried out via the **posterior probability distribution** that combines prior opinion about the parameter with the information contained in the likelihood $f_{X|\theta}(x;\theta)$ which represents the data contribution. In terms of events, Bayes Theorem says that

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

that is, it relates the two conditional probabilities $P(A|B)$ and $P(B|A)$. Carrying this idea over to probability distributions, it follows that we can carry out inference via the conditional probability distribution for parameter θ **given** data $X = x$. Specifically for parameter θ, the **posterior probability distribution** for θ is denoted $p_{\theta|X}(\theta|x)$, and is calculated as

$$p_{\theta|X}(\theta|x) = \frac{f_{X|\theta}(x;\theta)\, p_\theta(\theta)}{\int f_{X|\theta}(x;\theta)\, p_\theta(\theta)\, d\theta} = c(x) f_{X|\theta}(x;\theta)\, p_\theta(\theta) \tag{2.4.10}$$

say, where $f_{X|\theta}(x;\theta)$ is the likelihood, and $p_\theta(\theta)$ is the **prior probability distribution** for θ. The denominator in (2.4.10) can be regarded as the **marginal distribution** (or **marginal likelihood**) for data X evaluated at the observed data x

$$f_X(x) = \int f_{X|\theta}(x;\theta)\, p_\theta(\theta)\, d\theta. \tag{2.4.11}$$

2.4.2 BAYESIAN INFERENCE: ESTIMATION AND UNCERTAINTY INTERVALS

Inference for the parameter θ via the posterior $\pi_{\theta|Y}(\theta|y)$ can be carried out once the posterior has been computed. Intuitively appealing methods rely on summaries of this probability distribution, that is, moments or quantiles. For example, one Bayes estimate, $\widehat{\theta}_B$ of θ is the **posterior expectation**

$$\widehat{\theta}_B = E_{p_{\theta|X}}[\theta|X = x] = \int \theta p_{\theta|X}(\theta|x) d\theta$$

whereas another is the **posterior mode**, $\widehat{\theta}_B$, that is, the value of θ at which $p_{\theta|X}(\theta|x)$ is maximized, and finally the **posterior median** that satisfies

$$\int_{-\infty}^{\widehat{\theta}_B} p_{\theta|X}(\theta|x) d\theta = \frac{1}{2}$$

Definition 2.4.1 *A* $\mathbf{100(1-\alpha)\%}$ ***Bayesian Credible Interval*** *for* θ *is a subset* C *of* Θ *such that*

$$P[\theta \in C] \geq 1 - \alpha$$

The $\mathbf{100(1-\alpha)\%}$ ***Highest Posterior Density Bayesian Credible Interval*** *for* θ*, subject to* $P[\theta \in C] \geq 1-\alpha$ *is a subset* C *of* Θ *such that* $C = \{\theta \in \Theta : p_{\theta|X}(\theta|x) \geq k\}$ *where* k *is the largest constant such that*

$$P[\theta \in C] \geq 1 - \alpha.$$

2.4.3 BAYESIAN INFERENCE AND DECISION MAKING

Suppose that, in an inference setting, a decision is to be made, and the decision is selected from some set $\mathcal{D}$ of alternatives. Regarding the parameter space Θ as a set of potential "states of nature", within which the "true" state θ lies.

Definition 2.4.2 *Define the **loss function** for decision d and state θ as the loss (or penalty) incurred when the true state of nature is θ and the decision made is d. Denote this loss as*

$$L(d, \theta)$$

Definition 2.4.3 *With prior $\pi(\theta)$ and no data, the **expected loss** (or the **Bayes loss**) is defined as*

$$E_\theta \left[L(d, \theta) \right] = \int L(d, \theta) p_\theta \left(\theta \right) d\theta$$

The optimal Bayesian decision is

$$d_B = \arg\min_{d \in \mathcal{D}} E_{p_\theta} \left[L(d, \theta) \right]$$

that is, the decision that minimizes the Bayes loss.
If data are available, the optimal decision will intuitively become a function of the data. Suppose now that the decision in light of the data is denoted $\delta(x)$ (a function from $\mathbb{X}$ to $\mathcal{D}$, and the associated loss is $L(\delta(x), \theta)$

Definition 2.4.4 *The **risk** associated with decision $\delta(X)$ is the expected loss associated with $\delta(X)$, with the expectation taken over the distribution of X given θ*

$$R_\theta(\delta) = E_{X|\theta} \left[L(\delta(X), \theta) \right] = \int L(\delta(X), \theta) f_{X|\theta}(x; \theta) dx$$

Definition 2.4.5 *The **Bayes risk** expected risk $R_\theta(\delta)$ associated with $\delta(X)$, with the expectation taken over the prior distribution of θ*

$$\begin{aligned} R(\delta) &= E_\theta \left[R_\theta(\delta) \right] = E_\theta \left[E_{X|\theta} \left[L(\delta(X), \theta) \right] \right] \\ &= \int \left\{ \int L(\delta(x), \theta) f_{X|\theta}(x; \theta) dx \right\} p_\theta \left(\theta \right) d\theta \\ &= \int \int L(\delta(x), \theta) f_X(x) p_{\theta|X}(\theta|x) dx d\theta \\ &= \int \left\{ \int L(\delta(x), \theta) p_{\theta|X}(\theta|x) d\theta \right\} f_X(x) dx \end{aligned}$$

where

$$f_X(x) = \int f_{X|\theta}(x; \theta) p_\theta \left(\theta \right) d\theta$$

Definition 2.4.6 *With prior $p_\theta \left(\theta \right)$ and fixed data x, the optimal Bayesian decision, termed the **Bayes rule** is*

$$d_B = \arg\min_{\delta \in \mathcal{D}} R(\delta) = \arg\min_{\delta \in \mathcal{D}} \int \left\{ \int L(\delta(x), \theta) p_{\theta|X}(\theta|x) d\theta \right\} f_X(x) dx = \arg\min_{\delta \in \mathcal{D}} \int L(\delta(x), \theta) p_{\theta|X}(\theta|x) d\theta$$

*that is, the decision that minimizes the Bayes risk, or equivalently **(posterior) expected loss** in making decision δ, with expectation taken with respect to the posterior distribution $p_{\theta|X}(\theta|x)$.*

2.4.4 APPLICATIONS OF DECISION THEORY TO ESTIMATION

Theorem 2.4.1 *Under squared error loss*

$$L(\delta(x), \theta) = (\delta(x) - \theta)^2$$

the Bayes rule for estimating θ is

$$\delta(x) = \widehat{\theta}_B = E_{p_{\theta|X}}[\theta|x] = \int \theta p_{\theta|X}(\theta|x) d\theta$$

that is, the posterior expectation.

Theorem 2.4.2 *Under absolute error loss*

$$L(\delta(x), \theta) = |\delta(x) - \theta|$$

the Bayes rule for estimating θ is the solution of

$$\int_{-\infty}^{\delta(x)} p_{\theta|X}(\theta|x) d\theta = \frac{1}{2}$$

that is, the posterior median.

2.4.5 BAYESIAN HYPOTHESIS TESTING

To mimic the Likelihood Ratio testing procedure outlined in previous sections. For two hypotheses H_0 and H_1 define

$$\alpha_0 = \mathrm{P}[H_0|X = x] \qquad \alpha_1 = \mathrm{P}[H_1|X = x]$$

For example,

$$\mathrm{P}[H_0|X = x] = \int_R \pi_{\theta|X}(\theta|x) d\theta$$

where R is some region of Θ. Typically, the quantity

$$\frac{\mathrm{P}[H_0|X = x]}{\mathrm{P}[H_1|X = x]}$$

(the **posterior odds** on H_0) is examined.

Example 2.4.1 To test two simple hypothesis

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1 \end{aligned}$$

define the prior probabilities of H_0 and H_1as p_0 and p_1 respectively. Then, by Bayes Theorem

$$\frac{\mathrm{P}[H_1|X = x]}{\mathrm{P}[H_0|X = x]} = \frac{\dfrac{f_{X|\theta}(x;\theta_1)p_1}{f_{X|\theta}(x;\theta_0)p_0 + f_{X|\theta}(x;\theta_1)p_1}}{\dfrac{f_{X|\theta}(x;\theta_0)p_0}{f_{X|\theta}(x;\theta_0)p_0 + f_{X|\theta}(x;\theta_1)p_1}} = \frac{f_{X|\theta}(x;\theta_1)p_1}{f_{X|\theta}(x;\theta_0)p_0}$$

More generally, two hypotheses or models can be compared via the observed marginal likelihood that appears in (2.4.11), that is if

$$\frac{f_X(x; \text{Model 1})}{f_X(x; \text{Model 0})} = \frac{\int f^{(1)}_{X|\theta}(x; \theta_1)\, p_{\theta_1}(\theta_1)\, d\theta_1}{\int f^{(0)}_{X|\theta}(x; \theta_0)\, p_{\theta_0}(\theta_0)\, d\theta_0}$$

is greater than one we would favour Model 1 (with likelihood $f^{(1)}_{X|\theta}$ and prior p_{θ_1}) over Model 0 (with likelihood $f^{(0)}_{X|\theta}$ and prior p_{θ_0}).