

**Part Two**  
**Statistical Inference**

**Charles A. Rohde**

Fall 2001

# Contents

<b>6</b>	<b>Statistical Inference: Major Approaches</b>	<b>1</b>
6.1	Introduction . . . . .	1
6.2	Illustration of the Approaches . . . . .	4
6.2.1	Estimation . . . . .	5
6.2.2	Interval Estimation . . . . .	8
6.2.3	Significance and Hypothesis Testing . . . . .	11
6.3	General Comments . . . . .	22
6.3.1	Importance of the Likelihood . . . . .	22
6.3.2	Which Approach? . . . . .	22
6.3.3	Reporting Results . . . . .	23
<b>7</b>	<b>Point and Interval Estimation</b>	<b>27</b>
7.1	Point Estimation - Introduction . . . . .	27
7.2	Properties of Estimators . . . . .	28
7.2.1	Properties of Estimators . . . . .	29
7.2.2	Unbiasedness . . . . .	30
7.2.3	Consistency . . . . .	33
7.2.4	Efficiency . . . . .	35

7.3	Estimation Methods . . . . .	36
7.3.1	Analog or Substitution Method . . . . .	37
7.3.2	Maximum Likelihood . . . . .	39
7.4	Interval Estimation . . . . .	44
7.4.1	Introduction . . . . .	44
7.4.2	Confidence Interval for the Mean-Unknown Variance . . . . .	47
7.4.3	Confidence Interval for the Binomial . . . . .	49
7.4.4	Confidence Interval for the Poisson . . . . .	49
7.5	Point and Interval Estimation - Several Parameters . . . . .	51
7.5.1	Introduction . . . . .	51
7.5.2	Maximum Likelihood . . . . .	52
7.5.3	Properties of Maximum Likelihood Estimators . . . . .	54
7.5.4	Two Sample Normal . . . . .	56
7.5.5	Simple Linear Regression Model . . . . .	59
7.5.6	Matrix Formulation of Simple Linear Regression . . . . .	62
7.5.7	Two Sample Problem as Simple Linear Regression . . . . .	66
7.5.8	Paired Data . . . . .	69
7.5.9	Two Sample Binomial . . . . .	70
7.5.10	Logistic Regression Formulation of the Two sample Binomial . . . . .	75
<b>8</b>	<b>Hypothesis and Significance Testing</b>	<b>77</b>
8.1	Neyman Pearson Approach . . . . .	78
8.1.1	Basic Concepts . . . . .	78
8.1.2	Summary of Neyman-Pearson Approach . . . . .	80
8.1.3	The Neyman Pearson Lemma . . . . .	82

CONTENTS

iii

8.1.4	Sample Size and Power . . . . .	88
8.2	Generalized Likelihood Ratio Tests . . . . .	93
8.2.1	One Way Analysis of Variance . . . . .	98
8.3	Significance Testing and P-Values . . . . .	103
8.3.1	P Values . . . . .	103
8.3.2	Interpretation of P-values . . . . .	104
8.3.3	Two Sample Tests . . . . .	108
8.4	Relationship Between Tests and Confidence Intervals . . . . .	111
8.5	General Case . . . . .	112
8.5.1	One Sample Binomial . . . . .	113
8.6	Comments on Hypothesis Testing and Significance Testing . . . . .	117
8.6.1	Stopping Rules . . . . .	117
8.6.2	Tests and Evidence . . . . .	119
8.6.3	Changing Criteria . . . . .	121
8.7	Multinomial Problems and Chi-Square Tests . . . . .	122
8.7.1	Chi Square Test of Independence . . . . .	128
8.7.2	Chi Square Goodness of Fit . . . . .	131
8.8	PP-plots and QQ-plots . . . . .	133
8.9	Generalized Likelihood Ratio Tests . . . . .	135
8.9.1	Regression Models . . . . .	137
8.9.2	Logistic Regression Models . . . . .	142
8.9.3	Log Linear Models . . . . .	145



# Chapter 6

## Statistical Inference: Major Approaches

### 6.1 Introduction

The problem addressed by “statistical inference” is as follows:

Use a set of sample data to draw inferences (make statements) about some aspect of the population which generated the data. In more precise terms we have data  $\mathbf{y}$  which has probability model specified by  $f(\mathbf{y}; \boldsymbol{\theta})$ , a probability density function, and we want to make statements about the parameters  $\boldsymbol{\theta}$ .

The three major types of inferences are:

- **Estimation:** what single value of the parameter is most appropriate.?
- **Interval Estimation:** what region of parameter values is most consistent with the data?
- **Hypothesis Testing:** which of two values of the parameter is most consistent with the data?

Obviously inferences must be judged by criteria as to their usefulness and there must be methods for selecting inferences.

There are three major approaches to statistical inference:

- **Frequentist:** which judges inferences based on their performance in repeated sampling i.e. based on the sampling distribution of the statistic used for making the inference. A variety of ad hoc methods are used to select the statistics used for inference.
- **Bayesian:** which assumes that the inference problem is subjective and proceeds by
  - Elicit a prior distribution for the parameter.
  - Combine the prior with the density of the data (now assumed to be the conditional density of the data given the parameter) to obtain the joint distribution of the parameter and the data.
  - Use Bayes Theorem to obtain the posterior distribution of the parameter given the data.

No notion of repeated sampling is needed, all inferences are obtained by examining properties of the posterior distribution of the parameter.

- **Likelihood:** which defines the likelihood of the parameter as a function proportional to the probability density function and states that all information about the parameter can be obtained by examination of the likelihood function. Neither the notion of repeated sampling or prior distribution is needed.

## 6.2 Illustration of the Approaches

In this section we consider a simple inference problem to illustrate the three major methods of statistical inference.

Assume that we have data  $y_1, y_2, \dots, y_n$  which are a random sample from a normal distribution with parameters  $\mu$  and  $\sigma^2$ , where we assume, for simplicity, that the parameter  $\sigma^2$  is known. The probability density function of the data is thus

$$(2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right\}$$

### 6.2.1 Estimation

The problem is to use the data to determine an estimate of  $\mu$ .

**Frequentist Approach:** The frequentist approach uses as estimate  $\bar{y}$ , the sample mean of the data. The sample mean is justified on the basis of the facts that its sampling distribution is centered at  $\mu$  and has sampling variance  $\sigma^2/n$ . (Recall that the sampling distribution of the sample mean  $\bar{Y}$  of a random sample from a  $N(\mu, \sigma^2)$  distribution is  $N(\mu, \sigma^2/n)$ ). Moreover no other estimate has a sampling distribution which is centered at  $\mu$  with smaller variance. Thus in terms of repeated sampling properties the use of  $\bar{y}$  ensures that, on average, the estimate is closer to  $\mu$  than any other estimate. The results of the estimation procedure are reported as:

“The estimate of  $\mu$  is  $\bar{y}$  with standard error (standard deviation of the sampling distribution)  $\sigma/\sqrt{n}$ ”

**Bayesian:** In the Bayesian approach we first select a prior distribution for  $\mu$ ,  $p(\mu)$ . For this problem it can be argued that a normal distribution with parameters  $\mu_0$  and  $\sigma_\mu$  is appropriate.  $\mu_0$  is called the prior mean and  $\sigma_\mu^2$  is called the prior variance. By Bayes theorem the posterior distribution of  $\mu$  is given by

$$p(\mu|\mathbf{y}) = \frac{p(\mu)f(\mathbf{y};\mu)}{f(\mathbf{y})}$$

where

$$\begin{aligned} p(\mu) &= (2\pi\sigma_\mu^2)^{-1/2} \exp\left\{\frac{-1}{2\sigma_\mu^2}(\mu - \mu_0)^2\right\} \\ f(\mathbf{y};\mu) &= (2\pi\sigma^2)^{-n/2} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^n(y_i - \mu)^2\right\} \\ f(\mathbf{y}) &= \int_{-\infty}^{+\infty} f(\mathbf{y};\mu)p(\mu)d\mu \end{aligned}$$

It can be shown, with considerable algebra, that the posterior distribution of  $\mu$  is given by

$$p(\mu|\mathbf{y}) = (2\pi v^2)^{-1/2} \exp\left\{-\frac{1}{2v^2}(\mu - \eta)^2\right\}$$

i.e. a normal distribution with mean  $\eta$  and variance  $v$ .  $\eta$ , is called the **posterior mean** and  $v$  is called the **posterior variance** where

$$\begin{aligned} \eta &= \left(\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma^2}\right)^{-1} \left[\frac{1}{\sigma_\mu^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}\right] \\ v^2 &= \left(\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma^2}\right)^{-1} \end{aligned}$$

Note that the posterior mean is simply a weighted average of the prior mean and the sample mean with weights proportional to their variances. Also note that if the prior distribution is “vague” i.e.  $\sigma_\mu^2$  is large relative to  $\sigma^2$  then the posterior mean is nearly equal to the sample mean. In the Bayes approach the estimate reported is the posterior mean or the posterior mode which in this case coincide and are equal to  $\eta$ .

**Likelihood Approach:** The likelihood for  $\mu$  on data  $\mathbf{y}$  is defined to be proportional to the density function of  $\mathbf{y}$  at  $\mu$ . To eliminate the proportionality constant the likelihood is usually standardized to have maximum value 1 by dividing by the density function of  $\mathbf{y}$  evaluated at the value of  $\mu$ ,  $\hat{\mu}$  which maximizes the density function. The result is called the likelihood function.

In this example,  $\hat{\mu}$ , called the maximum likelihood estimate can be shown to be  $\hat{\mu} = \bar{y}$  the sample mean. Thus the likelihood function is

$$\text{lik}(\mu; \mathbf{y}) = \frac{f(\mathbf{y}; \mu)}{f(\mathbf{y}; \bar{y})}$$

Fairly routine algebra can be used to show that the likelihood in this case is given by

$$\text{lik}(\mu; \mathbf{y}) = \exp \left\{ -\frac{n(\mu - \bar{y})^2}{2\sigma^2} \right\}$$

The likelihood approach uses as estimate  $\bar{y}$  which is said to be the value of  $\mu$  which is most consistent with the observed data. A graph of the likelihood function shows the extent to which the likelihood concentrates around the best supported value.

### 6.2.2 Interval Estimation

Here the problem is to determine a set (interval) of parameter values which are consistent with the data or which are supported by the data.

**Frequentist:** In the frequentist approach we determine a confidence interval for the parameter. That is, a random interval,  $[\theta_l, \theta_u]$  is determined such that the probability that this interval includes the value of the parameter is  $1 - \alpha$  where  $1 - \alpha$  is the confidence coefficient. (Usually  $\alpha = .05$ ). Finding the interval uses the sampling distribution of a statistic (exact or approximate) or the bootstrap.

For the example under consideration here we have that the sampling distribution of  $\bar{Y}$  is normal with mean  $\mu$  and variance  $\sigma^2/n$  so that the following is a valid probability statement

$$P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

and hence

$$P\left(\bar{Y} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Thus the random interval defined by

$$\bar{Y} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

has the property that it will contain  $\mu$  with probability  $1 - \alpha$ .

**Bayesian:** In the Bayesian approach we select an interval of parameter values  $\theta_l, \theta_u$  such that the posterior probability of the interval is  $1 - \alpha$ . The interval is said to be a  $1 - \alpha$  credible interval for  $\theta$ .

In the example here the posterior distribution of  $\mu$  is normal with mean  $\eta$  and variance  $v^2$  so that the interval is obtained from the probability statement

$$P\left(-z_{1-\alpha/2} \leq \frac{\mu - \eta}{v} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

Hence the interval is

$$\eta \pm z_{1-\alpha/2}v$$

or

$$\left(\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma^2}\right)^{-1} \left[\frac{1}{\sigma_\mu^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}\right] \pm \left(\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma^2}\right)^{-1}$$

We note that if the prior variance  $\sigma_\mu^2$  is large relative to the variance  $\sigma^2$  then the interval is approximately given by

$$\bar{y} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Here, however, the statement is a subjective probability statement about the parameter being in the interval not a repeated sampling statement about the interval containing the parameter.

**Likelihood:** In the likelihood approach one determines the interval of parameter values for which the likelihood exceeds some value, say  $1/k$  where  $k$  is either 8 (strong evidence) or 32 (very strong evidence). The statement made is that we have evidence that this interval of parameter values is consistent with the data (constitutes a  $1/k$  likelihood interval for the parameter).

For this example the parameter values in the interval must satisfy

$$\text{lik}(\mu; \mathbf{y}) = \exp\left\{-\frac{n(\mu - \bar{y})^2}{2\sigma^2}\right\} \geq \frac{1}{k}$$

or

$$-n(\mu - \bar{y})^2/2\sigma^2 \geq -\ln(k)$$

which leads to

$$|\mu - \bar{y}| \leq \sqrt{2\ln(k)} \frac{\sigma}{\sqrt{n}}$$

so that the  $1/k$  likelihood interval is given by

$$\bar{y} \pm \sqrt{2\ln(k)} \frac{\sigma}{\sqrt{n}}$$

### 6.2.3 Significance and Hypothesis Testing

The general area of testing is a mess. Two distinct theories dominated the 20th century but due to common usage they became mixed up into a set of procedures that can best be described as a muddle. The basic problem is to decide whether a particular set of parameter values (called the null hypothesis) is more consistent with the data than another set of parameter values (called the alternative hypothesis).

**Frequentist:** The frequentist approach has been dominated by two overlapping procedures developed and advocated by two giants of the field of statistics in the 20th century; Fisher and Neyman.

**Significance Testing (Fisher):** In this approach we have a well defined null hypothesis  $H_0$  and a statistic which is chosen so that “extreme values” of the statistic cast doubt upon the null hypothesis in the frequency sense of probability.

**example:** If  $y_1, y_2, \dots, y_n$  are observed values of  $Y_1, Y_2, \dots, Y_n$  assumed independent each normally distributed with mean value  $\mu$  and known variance  $\sigma^2$  suppose that the null hypothesis is that  $\mu = \mu_0$ . Suppose also that values of  $\mu$  smaller than  $\mu_0$  are not tenable under the scientific theory being investigated.

It is clear that values of the observed sample mean  $\bar{y}$  larger than  $\mu_0$  suggest that  $H_0$  is not true. Fisher proposed that the calculation of the  $p$ -value be used as a **test of significance** for  $H_0 : \mu = \mu_0$ . If the  $p$ -value is small we have evidence that the null hypothesis is not true. The  $p$ -value is defined as

$$\begin{aligned} p - \text{value} &= P_{H_0}(\text{sample statistic as or more extreme than actually observed}) \\ &= P_{H_0}(\bar{Y} \geq \bar{y}_{obs}) \\ &= P\left(Z \geq \frac{\sqrt{n}(\bar{y}_{obs} - \mu_0)}{\sigma}\right) \end{aligned}$$

Fisher defined three levels of “smallness”, .05, .01 and .001 which lead to a variety of silly conventions such as

- \* – statistically significant
- \* – strongly statistically significant
- \*\* – very strongly statistically significant

**Hypothesis Testing (Neyman and Pearson):** In this approach a null hypothesis is selected and an alternative is selected. Neyman and Pearson developed a theory which fixed the probability of rejecting the null hypothesis when it is true and maximized the probability of rejecting the null hypothesis when it is false. Such tests were designed as rules of “inductive behavior” and were not intended to measure the strength of evidence for or against a particular hypothesis.

**Definition:** A rule for choosing between two hypotheses  $H_0$  and  $H_1$  (based on observed values of random variables) is called a **statistical test** of  $H_0$  vs  $H_1$ .

If we represent the test as a function,  $\delta$ , on the sample space then a test is a statistic of the form

$$\delta(\mathbf{y}) = \begin{cases} 1 & H_1 \text{ chosen} \\ 0 & H_0 \text{ chosen} \end{cases}$$

The set of observations which lead to the rejection of  $H_0$  is called the **critical region** of the test i.e.

$$C_\delta = \{\mathbf{y} : \delta(\mathbf{y}) = 1\}$$

Typical terminology used in hypothesis testing is:

$$\begin{aligned}\text{choose } H_1 \text{ when } H_0 \text{ is true} &= \text{Type I Error} \\ \text{choose } H_0 \text{ when } H_1 \text{ is true} &= \text{Type II Error}\end{aligned}$$

The probability of a Type I Error is called  $\alpha$  and the probability of a Type II Error is called  $\beta$ .  $1 - \beta$ , the probability of rejecting the null hypothesis when it is false is called the power of the test. The Neyman Pearson theory of inductive behavior says to fix the probability of a Type I Error at some value  $\alpha$ , called the **significance level**, and choose the test which maximizes the power.

In terms of the test statistic we have

$$\alpha = E_0 [\delta(\mathbf{Y})] \ ; \ \text{power} = E_1 [\delta(\mathbf{Y})]$$

Thus the inference problem has been reduced to a purely mathematical optimization problem: Choose  $\delta(\mathbf{Y})$  so that  $E_1 [\delta(\mathbf{Y})]$  is maximized subject to  $E_0 [\delta(\mathbf{Y})] = \alpha$ .

**example:** If the  $Y_i$ s are i.i.d.  $N(\mu, \sigma^2)$  and

$$H_0 : \mu = \mu_0 \text{ and } H_1 : \mu = \mu_1 > \mu_0$$

consider the test which chooses  $H_1$  if  $\bar{y} > c$  i.e. the test statistic  $\delta$  is given by

$$\delta(\mathbf{y}) = \begin{cases} 1 & \bar{y} > c \\ 0 & \text{otherwise} \end{cases}$$

The critical region is

$$C_\delta = \{\mathbf{y} : \bar{y} > c\}$$

In this case

$$\begin{aligned} \alpha &= P_0(\{\mathbf{y} : \bar{y} > c\}) \\ &= P_0\left(\frac{\sqrt{n}(\bar{Y} - \mu_0)}{\sigma} > \frac{\sqrt{n}(c - \mu_0)}{\sigma}\right) \\ &= P\left(Z \geq \frac{\sqrt{n}(c - \mu_0)}{\sigma}\right) \\ \text{power} &= P_1(\{\mathbf{y} : \bar{y} \geq c\}) \\ &= P_1\left(\frac{\sqrt{n}(\bar{Y} - \mu_1)}{\sigma} \geq \frac{\sqrt{n}(c - \mu_1)}{\sigma}\right) \\ &= P\left(Z \geq \frac{\sqrt{n}(c - \mu_1)}{\sigma}\right) \end{aligned}$$

where  $Z$  is  $N(0, 1)$ .

Thus if we want a significance level of .05 we pick  $c$  such that

$$1.645 = \frac{\sqrt{n}(c - \mu_0)}{\sigma} \quad \text{i.e.} \quad c = \mu_0 + 1.645 \frac{\sigma}{\sqrt{n}}$$

The power is then

$$P\left(Z \geq \frac{\sqrt{n}(c - \mu_1)}{\sigma}\right) = P\left(Z \geq \frac{\mu_0 - \mu_1}{\sigma} + 1.645 \frac{\sigma}{\sqrt{n}}\right)$$

Note that  $\alpha$  and the power are functions of  $n$  and  $\sigma$  and that as  $\alpha$  decreases the power decreases. Similarly as  $n$  increases the power increases and as  $\sigma$  decreases the power increases.

In general, of two tests with the same  $\alpha$ , the Neyman Pearson theory chooses the one with the greater power.

The Neyman Pearson Fundamental Lemma states that if  $C$  is a critical region satisfying, for some  $k > 0$

$$(1) f_{\theta_1}(\mathbf{y}) \geq k f_{\theta_0}(\mathbf{y}) \text{ for all } \mathbf{y} \in C$$

$$(2) f_{\theta_1}(\mathbf{y}) \leq k f_{\theta_0}(\mathbf{y}) \text{ for all } \mathbf{y} \notin C$$

$$(3) P_{\theta_0}(\mathbf{Y} \in C) = \alpha$$

then  $C$  is the best critical region for testing the simple hypothesis  $H_0 \theta = \theta_0$  vs the simple alternative  $H_1 \theta = \theta_1$ . i.e. the test is most powerful.

The ratio

$$\frac{f_{\theta_1}(\mathbf{y})}{f_{\theta_0}(\mathbf{y})}$$

is called the **likelihood ratio**.

The test for the mean of a normal distribution with known variance obeys the Neyman-Pearson Fundamental Lemma and hence is a most powerful (best) test.

In current practice the Neyman Pearson theory is used to define the critical region and then a p-value is calculated based on the critical region's determination of extreme values of the sample. This approach thoroughly confuses the two approaches to testing.

**Note:** If instead of minimizing the probability of a Type II error (maximizing the power) for a fixed probability of a Type I error we choose to minimize a linear combination of  $\alpha$  and  $\beta$  we get an entirely different critical region.

Note that

$$\begin{aligned}\alpha + \lambda\beta &= E_0[\delta(\mathbf{Y})] + \lambda \{1 - E_1[\delta(Y)]\} \\ &= \int_C f_{\theta_0}(\mathbf{y}) \mathbf{d}\mathbf{y} + \lambda - \lambda \int_C f_{\theta_1}(\mathbf{y}) \mathbf{d}\mathbf{y} \\ &= \lambda + \int_C [f_{\theta_0}(\mathbf{y}) - \lambda f_{\theta_1}(\mathbf{y})] \mathbf{d}\mathbf{y}\end{aligned}$$

which is minimized when

$$\begin{aligned}C &= \{\mathbf{y} : f_{\theta_0}(\mathbf{y}) - \lambda f_{\theta_1}(\mathbf{y}) < 0\} \\ &= \left\{ \mathbf{y} : \frac{f_{\theta_1}(\mathbf{y})}{f_{\theta_0}(\mathbf{y})} > \frac{1}{\lambda} \right\}\end{aligned}$$

which depends only on the relative importance of the Type II Error to the Type I Error.

**Bayesian:** In the Bayesian approach to hypothesis testing we assume that  $H_0$  has a prior probability of  $p_0$  and that  $H_1$  has a prior probability of  $p_1$ . Then the posterior probability of  $H_0$  is given by

$$\frac{f_{\theta_0}(\mathbf{y})p_0}{f_{\theta_0}(\mathbf{y})p_0 + f_{\theta_1}(\mathbf{y})p_1}$$

Similarly the posterior probability of  $H_1$  is given by

$$\frac{f_{\theta_1}(\mathbf{y})p_1}{f_{\theta_0}(\mathbf{y})p_0 + f_{\theta_1}(\mathbf{y})p_1}$$

It follows that the ratio of the posterior probability of  $H_1$  to  $H_0$  is given by

$$\left[ \frac{f_{\theta_1}(\mathbf{y})}{f_{\theta_0}(\mathbf{y})} \right] \frac{p_1}{p_0}$$

We choose  $H_1$  over  $H_0$  if this ratio exceeds 1, otherwise we choose  $H_0$ . Note that the likelihood ratio again appears, this time as supplying the factor which changes the prior odds into the posterior odds. The likelihood ratio in this situation is an example of a **Bayes factor**.

For the mean of the normal distribution with known variance the likelihood ratio can be shown to be

$$\exp \left\{ \left( \bar{y} - \frac{\mu_0 + \mu_1}{2} \right) \frac{n(\mu_1 - \mu_0)}{\sigma^2} \right\}$$

so that data increase the posterior odds when the observed sample mean exceeds the value  $(\mu_0 + \mu_1)/2$ .

**Likelihood:** The likelihood approach focuses on the Law of Likelihood.

**Law of Likelihood:** If

- Hypothesis A specifies that the probability that the random variable  $X$  takes on the value  $x$  is  $p_A(x)$
- Hypothesis B specifies that the probability that the random variable  $X$  takes on the value  $x$  is  $p_B(x)$

then

- The observation  $x$  is evidence supporting A over B if and only if

$$p_A(x) > p_B(x)$$

- The likelihood ratio

$$\frac{p_A(x)}{p_B(x)}$$

measures the strength of that evidence.

The Law of Likelihood measures only the support for one hypothesis relative to another. It does not sanction support for a single hypothesis, nor support for composite hypotheses.

**example:** Assume that we have a sample  $y_1, y_2, \dots, y_n$  which are realized values of  $Y_1, Y_2, \dots, Y_n$  where the  $Y_i$  are iid  $N(\mu, \sigma^2)$  where  $\sigma^2$  is known. Of interest is  $H_0 : \mu = \mu_0$  and  $H_1 : \mu = \mu_1 = \mu_0 + \delta$  where  $\delta > 0$ .

The likelihood for  $\mu$  is given by

$$L(\theta; \mathbf{y}) = \prod_{i=1}^n (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\}$$

After some algebraic simplification the likelihood ratio for  $\mu_1$  vs  $\mu_0$  is given by

$$\frac{L_1}{L_0} = \exp \left\{ \left( \bar{y} - \mu_0 - \frac{\delta}{2} \right) \frac{n\delta}{\sigma^2} \right\}$$

It follows that

$$\frac{L_1}{L_0} \geq k$$

if and only if

$$\left( \bar{y} - \mu_0 - \frac{\delta}{2} \right) \frac{n\delta}{\sigma^2} \geq \ln(k)$$

i.e.

$$\bar{y} \geq \mu_0 + \frac{\delta}{2} + \frac{\sigma^2 \ln(k)}{n\delta}$$

or

$$\bar{y} \geq \frac{\mu_0 + \mu_1}{2} + \frac{\sigma^2 \ln(k)}{n\delta}$$

Choice of  $k$  is usually 8 or 32 (discussed later).

## 6.3 General Comments

### 6.3.1 Importance of the Likelihood

Note that each of the approaches involve the likelihood. For this reason we will spend considerable time using the likelihood to determine estimates (point and interval), test hypotheses and also to check the compatability of results with the Law of Likelihood.

### 6.3.2 Which Approach?

Each approach has its advocates, some fanatic, some less so. The important idea is to use an approach which faithfully conveys the science under investigation.

### 6.3.3 Reporting Results

Results of inferential procedures are reported in a variety of ways depending on the statistician and the subject matter area. There seems to be no fixed set of rules for reporting the results of estimation, interval estimation and testing procedures. The following is suggestion by this author on how to report results.

- **Estimation**

- **Frequentist** The estimated value of the parameter  $\theta$  is  $\hat{\theta}$  with standard error s.e. ( $\hat{\theta}$ ). The specific method of estimation might be given also.
- **Bayesian** The estimated value of the parameter is  $\hat{\theta}$  (the mean or mode) of the posterior distribution of  $\theta$ . The standard deviation of the posterior distribution is s.e. ( $\theta$ ). The prior distribution was  $g(\theta)$ . A graph of the posterior could also be provided.
- **Likelihood** The graph of the likelihood function for  $\theta$  is as follows. The maximum value (best supported value) is at  $\hat{\theta}$ . The shape of the likelihood function provides the information on “precision”.

**• Interval Estimation**

- **Frequentist** Values of  $\theta$  between  $\theta_l$  and  $\theta_u$  are consistent with the data based on a  $(1 - \alpha)$  confidence interval. The specific statistic or method used to obtain the confidence interval should be mentioned.
- **Bayesian** Values of  $\theta$  between  $\theta_l$  and  $\theta_u$  are consistent with the data based on a  $(1 - \alpha)$  credible interval. The prior distribution used in obtaining the posterior should be mentioned.
- **Likelihood** Values of  $\theta$  between  $\theta_l$  and  $\theta_u$  are consistent with the data based on a  $1/k$  likelihood interval. Presented as a graph is probably best.

- **Testing**
  - **Frequentist**
  - **Bayesian**
  - **Likelihood**



# Chapter 7

## Point and Interval Estimation

### 7.1 Point Estimation - Introduction

The statistical inference called **point estimation** provides the solution to the following problem

**Given data and a probability model find an estimate for the parameter**

There are two important features of estimation procedures:

- Desirable properties of the estimate
- Methods for obtaining the estimate

## 7.2 Properties of Estimators

Since the data in a statistical problem are subject to variability:

- Statistics calculated from the data are also subject to variability.
- The rule by which we calculate an estimate is called the **estimator** and the actual computed value is called the **estimate**.
  - An estimator is thus a random variable.
  - Its realized value is the estimate.
- In the frequentist approach to statistics the sampling distribution of the estimator:
  - determines the properties of the estimator
  - determines which of several potential estimators might be best in a given situation.

### 7.2.1 Properties of Estimators

Desirable properties of an estimator include:

- The estimator should be correct on average i.e. the sampling distribution of the estimator should be centered at the parameter being estimated. This property is called **unbiasedness**
- In large samples, the estimator should be equal to the parameter being estimated i.e.

$$P(\hat{\theta} \approx \theta) \approx 1 \quad \text{for } n \text{ large}$$

where  $\approx$  means approximately. Equivalently

$$\hat{\theta} \xrightarrow{p} \theta$$

This property is called **consistency**.

- The sampling distribution of the estimator should be concentrated closely around its center i.e. the estimator should have small variability. This property is called **efficiency**.

Of these properties most statisticians agree that consistency is the minimum criterion that an estimator should satisfy.

## 7.2.2 Unbiasedness

**Definition:** An estimator  $\hat{\theta}$  is an **unbiased** estimator of a parameter  $\theta$  if

$$E(\hat{\theta}) = \theta$$

An unbiased estimator thus has a sampling distribution centered at the value of the parameter which is being estimated.

**examples:**

- To estimate the parameter  $p$  in a binomial distribution we use the estimate  $\hat{p} = \frac{x}{n}$  where  $x$  is the number of successes in the sample. The corresponding estimator is unbiased since

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{np}{n} = p$$

- To estimate the parameter  $\lambda$  in the Poisson distribution we use the estimate  $\hat{\lambda} = \bar{x}$  where  $\bar{x}$  is the sample mean. The corresponding estimator is unbiased since

$$E(\hat{\lambda}) = E(\bar{X}) = \lambda$$

- To estimate the parameter  $\mu$  in the normal distribution we use the estimate  $\hat{\mu} = \bar{x}$  where  $\bar{x}$  is the sample mean. The corresponding estimator is unbiased since

$$E(\hat{\mu}) = E(\bar{X}) = \mu$$

- In fact the sample mean is always an unbiased estimator of the population mean, provided that the sample is a random sample from the population.

Statisticians, when possible, use unbiased estimators.

- The difficulty in finding unbiased estimators in general is that estimators for certain parameters are often complicated functions.
- The resulting expected values cannot be evaluated and hence unbiasedness cannot be checked.
- Often such estimators are, however, nearly unbiased for large sample sizes; i.e. they are **asymptotically unbiased**.

**examples:**

- The estimator for the log odds in a binomial distribution is

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$$

The expected value of this estimate is not defined since there is a positive probability that it is infinite ( $p = 0$  or  $p = 1$ )

- The estimator  $s^2$  of  $\sigma^2$  defined by

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

is an unbiased estimator of  $\sigma^2$  for a random sample from any population with variance  $\sigma^2$ .

- To see this note that

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2$$

- Since we know that

$$\text{var}(X_i) = E(X_i^2) - \mu^2 = \sigma^2 \quad \text{and} \quad \text{var}(\bar{X}) = E(\bar{X}^2) - \mu^2 = \frac{\sigma^2}{n}$$

- we have

$$E(X_i^2) = \sigma^2 + \mu^2 \quad \text{and} \quad E(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2$$

- Thus

$$E\left(\sum_{i=1}^n (X_i - \bar{X})^2\right) = n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = (n-1)\sigma^2$$

so that  $s^2$  is an unbiased estimator of  $\sigma^2$  as claimed.

### 7.2.3 Consistency

**Definition:** An estimator  $\hat{\theta}$  is **consistent** for the parameter  $\theta$  if

$$P(\hat{\theta} - \theta \approx 0) \approx 1 \quad \text{or} \quad \hat{\theta} \xrightarrow{p} \theta$$

i.e.

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| < \epsilon) \longrightarrow 1 \quad \text{as } n \rightarrow \infty$$

- For an estimator  $\hat{\theta}$  of a parameter  $\theta$  it can be shown that

$$P(|\hat{\theta} - \theta| < \delta) \geq 1 - \frac{E(\hat{\theta} - \theta)^2}{\delta^2} \quad \text{for any } \delta > 0$$

- It follows that an estimator is consistent if

$$E(\hat{\theta} - \theta)^2 \rightarrow 0$$

- The quantity  $E(\hat{\theta} - \theta)^2$  is called the **mean square error** of the estimator.
- It can be shown that the mean square error of an estimator satisfies

$$E(\hat{\theta} - \theta)^2 = \text{var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$$

- The quantity  $E(\hat{\theta}) - \theta$  is called the **bias** of the estimator.
- An estimator is thus consistent if it is asymptotically unbiased and its variance approaches zero as  $n$ , the sample size, increases.

**examples:**

- $\hat{p}$  in the binomial model is consistent since

$$E(\hat{p}) = p \quad \text{and} \quad \text{var}(\hat{p}) = \frac{p(1-p)}{n}$$

- $\hat{\lambda}$  in the Poisson model is consistent since

$$E(\hat{\lambda}) = \lambda \quad \text{and} \quad \text{var}(\hat{\lambda}) = \frac{\lambda}{n}$$

- $\hat{\mu} = \bar{X}$  in the normal model is consistent since

$$E(\hat{\mu}) = \mu \quad \text{and} \quad \text{var}(\hat{\mu}) = \frac{\sigma^2}{n}$$

- The estimators of the log odds and log odds ratio for the binomial distribution are consistent as will be shown later when we discuss maximum likelihood estimation.

### 7.2.4 Efficiency

Given two estimators  $\hat{\theta}_1$  and  $\hat{\theta}_2$  which are both unbiased estimators for a parameter  $\theta$

- We say that  $\hat{\theta}_2$  is more efficient than  $\hat{\theta}_1$  if

$$\text{var}(\hat{\theta}_2) < \text{var}(\hat{\theta}_1)$$

- Thus the sampling distribution of  $\hat{\theta}_2$  is more concentrated around  $\theta$  than is the sampling distribution of  $\hat{\theta}_1$ .
- In general we choose that estimator which has the smallest variance.

**example:**

For a random sample from a normal distribution with mean  $\mu$  and variance  $\sigma^2$  the variance of  $\bar{X}$  is  $\frac{\sigma^2}{n}$  while the variance of the sample median is  $\frac{\pi}{2} \frac{\sigma^2}{n}$ . Since

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n} < \frac{\pi}{2} \left( \frac{\sigma^2}{n} \right) = \text{var}(\text{sample median})$$

we see that the sample mean is preferred for this situation.

### 7.3 Estimation Methods

An enormous variety of methods have been proposed for obtaining estimates of parameters in statistical models.

Three methods are of general importance:

- “the analog or substitution method”
- the method of maximum likelihood.
- estimating equations.

### 7.3.1 Analog or Substitution Method

The analog or substitution method of estimation is based on selecting as the estimate the sample statistic which is the analog to the population parameter being estimated.

**examples:**

- In the binomial estimate the population proportion  $p$  by the sample proportion  $\hat{p} = \frac{x}{n}$ .
- In the case of a random sample from the normal distribution estimate the population mean  $\mu$  by the sample mean  $\bar{x}$ .
- Estimate the population median by the sample median.
- Estimate the population range by the sample range.
- Estimate the upper quartile of the population by the upper quartile of the sample.
- Estimate the population distribution using the empirical distribution.

While intuitively appealing,

- The analog method does not work in complex situations because there are not sample analogs to population parameters.
- There are also few general results regarding desirable properties of estimators obtained using the analog method.

### 7.3.2 Maximum Likelihood

The maximum likelihood method of estimation was introduced in 1921 by Sir Ronald Fisher and chooses that estimate of the parameter which “makes the observed data as likely as possible”.

**Definition:** If the sample data is denoted by  $\mathbf{y}$ , the parameter by  $\theta$  and the probability density function by  $f(\mathbf{y}; \theta)$  then the **maximum likelihood** estimate of  $\theta$  is that value of  $\theta$ ,  $\hat{\theta}$  which maximizes  $f(\mathbf{y}; \theta)$

- Recall that the likelihood of  $\theta$  is defined as

$$\text{lik}(\theta; \mathbf{y}) = \frac{f(\mathbf{y}; \theta)}{f(\mathbf{y}; \hat{\theta})}$$

- The likelihood of  $\theta$  may be used to evaluate the relative importance of different values of  $\theta$  in explaining the observed data i.e. if

$$\text{lik}(\theta_2; \mathbf{y}) > \text{lik}(\theta_1; \mathbf{y})$$

then  $\theta_2$  explains the observed data better than  $\theta_1$ .

- As we have seen likelihood is the most important component of the alternative theories of statistical inference.

Maximum likelihood estimates are obtained by:

- Maximizing the likelihood using calculus. Most often we have a random sample of size  $n$  from a population with density function  $f(y; \theta)$ . In this case we have that

$$f(\mathbf{y}; \theta) = \prod_{i=1}^n f(y_i; \theta)$$

Since the maximum of a function occurs at the same value as the maximum of the natural logarithm of the function it is easier to maximize

$$\sum_{i=1}^n \ln[f(y_i; \theta)]$$

with respect to  $\theta$ . Thus we solve the equations

$$\sum_{i=1}^n \frac{d \ln[f(y_i; \theta)]}{d\theta} = 0$$

which is called the **maximum likelihood or score equation**.

- Maximizing the likelihood numerically. Most statistical software programs do this.
- Graphing the likelihood and observing the point at which the maximum value of the likelihood occurs.

**examples:**

- In the binomial,  $\hat{p} = \frac{x}{n}$  is the maximum likelihood estimate of  $p$ .
- In the Poisson,  $\hat{\lambda} = \bar{x}$  is the maximum likelihood estimate of  $\lambda$ .
- In the normal,
  - $\hat{\mu} = \bar{x}$  is the maximum likelihood estimate of  $\mu$ .
  - $s^2$  is the maximum likelihood estimate of  $\sigma^2$
  - $\sqrt{s^2} = s$  is the maximum likelihood estimate of  $\sigma$

In addition to their intuitive appeal and the fact that they are easy to calculate using appropriate software, maximum likelihood estimates have several important properties.

- Invariance. The maximum likelihood estimate of a function  $g(\theta)$  is  $g(\hat{\theta})$  where  $\hat{\theta}$  is the maximum likelihood estimate of  $\theta$ .

Assuming that we have a random sample from a distribution with probability density function  $f(y; \theta)$ :

- Maximum likelihood estimates are usually consistent i.e.

$$\hat{\theta} \xrightarrow{p} \theta_0$$

where  $\theta_0$  is the true value of  $\theta$ .

- The distribution of the maximum likelihood estimate in large samples is usually normal, centered at  $\theta$ , with a variance that can be explicitly calculated. Thus

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx N(0, v(\theta_0))$$

where  $\theta_0$  is the true value of  $\theta$  and

$$v(\theta_0) = \frac{1}{i(\theta_0)} \quad \text{where} \quad i(\theta_0) = -E_{\theta_0} \left[ \frac{d^{(2)} \ln(f(Y); \theta_0)}{d\theta_0^{(2)}} \right]$$

Thus we may obtain probabilities for  $\hat{\theta}$  as if it were normal with expected value  $\theta_0$  and variance  $v(\theta_0)$ . We may also approximate  $v(\theta_0)$  by  $v(\hat{\theta})$ .

- If  $g(\theta)$  is a differentiable function then the approximate distribution of  $g(\hat{\theta})$  satisfies

$$\sqrt{n}[g(\hat{\theta}) - g(\theta_0)] \approx N(0, v_g(\theta_0))$$

where

$$v_g(\theta_0) = [g^{(1)}(\theta_0)]^2 v(\theta_0)$$

$v_g(\theta_0)$  may be approximated by  $v_g(\hat{\theta})$

- Maximum likelihood estimators can be calculated for complex statistical models using appropriate software.

A major drawback to maximum likelihood estimates is the fact that the estimate, and more importantly, its variance, depend on the model  $f(\mathbf{y}; \theta)$ , and the assumption of large samples. Using the bootstrap allows us to obtain variance estimates which are robust (do not depend strongly on the validity of the model) and do not depend on large sample sizes.

## 7.4 Interval Estimation

### 7.4.1 Introduction

For estimating  $\mu$  when we have  $Y_1, Y_2, \dots, Y_n$  which are i.i.d.  $N(\mu, \sigma^2)$  where  $\sigma^2$  is known we know that the maximum likelihood estimate of  $\mu$  is  $\bar{Y}$ . For a given set of observations we obtain a point estimate of  $\mu$ ,  $\bar{y}$ . However, this does not give us all the information about  $\mu$  that we would like to have.

In interval estimation we find a set of parameter values which are consistent with the data.

One approach would be to sketch the likelihood function of  $\mu$  which is given by

$$L(\mu, \mathbf{y}) = \exp \left\{ -\frac{n(\mu - \bar{y})^2}{2\sigma^2} \right\}$$

which shows that the likelihood has the shape of a normal density, centered at  $\bar{y}$  and gets narrower as  $n$  increases.

Another approach is to construct a confidence interval. We use the fact that

$$\hat{\mu} = \bar{Y} \sim N \left( \mu, \frac{\sigma^2}{n} \right)$$

i.e. the sampling distribution of  $\bar{Y}$  is normal with mean  $\mu$  and variance  $\sigma^2/n$ . Thus we find that

$$P \left( \frac{|\bar{Y} - \mu|}{\sigma/\sqrt{n}} \leq 1.96 \right) = .95$$

It follows that

$$P\left(\bar{Y} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = .95$$

This last statement says that the probability is .95 that the **random interval**

$$\left[\bar{Y} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{Y} + 1.96 \frac{\sigma}{\sqrt{n}}\right]$$

will contain  $\mu$ .

Notice that for a given realization of  $\bar{Y}$ , say  $\bar{y}$ , the probability that the interval contains the parameter  $\mu$  is either 0 or 1 since there is no random variable present at this point. Thus we cannot say that there is a 95% chance that the parameter  $\mu$  is in a given observed interval.

**Definition:** An interval  $I(\mathbf{Y}) \subset \Theta$ , the parameter space, is a  $100(1 - \alpha)\%$  confidence interval for  $\theta$  if

$$P(I(\mathbf{Y}) \supset \theta) = 1 - \alpha$$

for all  $\theta \in \Theta$ .  $1 - \alpha$  is called the confidence level.

Note that we cannot say

$$P(I(\mathbf{y}) \supset \theta) = 1 - \alpha$$

but we can say

$$P(I(\mathbf{Y}) \supset \theta) = 1 - \alpha$$

What we can say with regard to the first statement is that we used a procedure which has a probability of  $1 - \alpha$  of producing an interval which contains  $\theta$ . Since the interval we observed was constructed according to this procedure we say that we have **a set of parameter values which are consistent with the data at confidence level  $1 - \alpha$** .

### 7.4.2 Confidence Interval for the Mean-Unknown Variance

In the introduction we obtained the confidence interval for  $\mu$  when the observed data was a sample from a normal distribution with mean  $\mu$  and known variance  $\sigma^2$ . If the variance is not known we use the fact that the distribution of

$$T = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

is Student's  $t$  with  $n - 1$  degrees of freedom where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

is the bias corrected maximum likelihood estimator of  $\sigma^2$ .

It follows that

$$\begin{aligned} 1 - \alpha &= P\left(\frac{|\bar{Y} - \mu|}{s/\sqrt{n}} \leq t_{1-\alpha/2}(n-1)\right) \\ &= P\left(|\bar{Y} - \mu| \leq t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right) \\ &= P\left(\bar{Y} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right) \end{aligned}$$

Thus the random interval

$$\bar{Y} \pm t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

is a  $1 - \alpha$  confidence interval for  $\mu$ . The observed interval

$$\bar{y} \pm t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

has the same interpretation as the interval for  $\mu$  with  $\sigma^2$  known.

### 7.4.3 Confidence Interval for the Binomial

Since  $\hat{p}$  is a maximum likelihood estimator for  $p$  we have that the approximate distribution of  $\hat{p}$  may be taken to be normal with mean  $p$  and variance  $p(1-p)/n$  which leads to an approximate confidence interval for  $p$  given by

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Exact confidence limits for  $p$  may be obtained by solving the equation

$$\sum_{i=y}^n \binom{n}{i} p_L^i (1-p_L)^{n-i} = \frac{\alpha}{2} = \sum_{j=0}^y \binom{n}{j} p_U^j (1-p_U)^{n-j}$$

where  $y$  is the observed number of successes. This is the procedure STATA uses to obtain the exact confidence intervals. The solutions can be shown to be

$$p_L = \frac{n_1 F_{n_1, n_2, \alpha/2}}{n_2 + n_1 F_{n_1, n_2, \alpha/2}}$$

$$p_U = \frac{m_1 F_{m_1, m_2, 1-\alpha/2}}{m_2 + m_1 F_{m_1, m_2, 1-\alpha/2}}$$

where

$$n_1 = 2y, \quad n_2 = 2(n-y+1), \quad m_1 = 2(y+1), \quad m_2 = 2(n-y)$$

and  $F_{r_1, r_2, \gamma}$  is the  $\gamma$  percentile of the  $F$  distribution with  $r_1$  and  $r_2$  degrees of freedom.

We can also use the bootstrap to obtain confidence intervals for  $p$ .

### 7.4.4 Confidence Interval for the Poisson

If we observe  $Y$  equal to  $y$  the maximum likelihood estimate of  $\lambda$  is  $y$ . If  $\lambda$  is large we have that  $\hat{\lambda}$  is approximately normal with mean  $\lambda$  and variance  $\lambda$ . Thus an approximate confidence interval for  $\lambda$  is given by

$$\hat{\lambda} \pm z_{1-\alpha/2} \sqrt{\hat{\lambda}}$$

Exact confidence interval can be obtained by solving the equations

$$e^{-\lambda_L} \sum_{i=y}^{\infty} \frac{\lambda_L^i}{i!} = \frac{\alpha}{2} = e^{-\lambda_U} \sum_{j=0}^y \frac{\lambda_U^j}{j!}$$

This is the procedure STATA uses to obtain the exact confidence interval. The solutions can be shown to be

$$\begin{aligned}\lambda_L &= \frac{1}{2}\chi_{2y, \alpha/2}^2 \\ \lambda_U &= \frac{1}{2}\chi_{2(y+1), 1-\alpha/2}^2\end{aligned}$$

where  $\chi_{r, \gamma}^2$  is the  $\gamma$  percentile of the chi-square distribution with  $r$  degrees of freedom.

The bootstrap can also be used to obtain confidence intervals for  $\lambda$ .

## 7.5 Point and Interval Estimation - Several Parameters

### 7.5.1 Introduction

We now consider the situation where we have a probability model which has several parameters.

- Often we are interested in only one of the parameters and the other is considered a **nuisance parameter**. Nevertheless we still need to estimate all of the parameters to specify the probability model.
- We may be interested in a function of all of the parameters e.g. the odds ratio when we have two binomial distributions.
- The properties of unbiasedness, consistency and efficiency are still used to evaluate the estimators.
- A variety of methods are used to obtain estimators, the most important of which is maximum likelihood.

### 7.5.2 Maximum Likelihood

Suppose that we have data  $\mathbf{y}$  which are a realization of  $\mathbf{Y}$  which has density function  $f(\mathbf{y}; \boldsymbol{\theta})$  where the parameter  $\boldsymbol{\theta}$  is now  $k$ -dimensional i.e.

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$$

As in the case of one parameter the maximum likelihood estimate of  $\boldsymbol{\theta}$  is defined as that value  $\hat{\boldsymbol{\theta}}$  which maximizes  $f(\mathbf{y}; \boldsymbol{\theta})$ .

For a  $k$  dimensional problem we find the maximum likelihood estimate of  $\boldsymbol{\theta}$  by solving the system of equations:

$$\frac{\partial \ln[f(\mathbf{y}; \boldsymbol{\theta})]}{\partial \theta_j} = 0 \quad \text{for } j = 1, 2, \dots, k$$

which are called the **maximum likelihood equations** or the **score equations**.

**example:** If  $Y_1, Y_2, \dots, Y_n$  are i.i.d.  $N(\mu, \sigma^2)$  then

$$f(\mathbf{y}; \mu, \sigma^2) = \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right\}$$

and hence

$$\ln f(\mathbf{y}; \mu, \sigma) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}$$

It follows that

$$\begin{aligned} \frac{\partial \ln[f(\mathbf{y}; \mu, \sigma)]}{\partial \mu} &= \frac{2 \sum_{i=1}^n (y_i - \mu)}{2\sigma^2} \\ \frac{\partial \ln[f(\mathbf{y}; \mu, \sigma)]}{\partial \sigma^2} &= -\frac{n}{\sigma^2} + \frac{\sum_{i=1}^n (y_i - \mu)^2}{2(\sigma^2)^2} \end{aligned}$$

Equating to 0 and solving yields

$$\hat{\mu} = \bar{y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

Note that the maximum likelihood estimator for  $\sigma^2$  is not the usual estimate of  $\sigma^2$  which is

$$s^2 = \frac{\sum_i (y_i - \bar{y})^2}{n - 1}$$

### 7.5.3 Properties of Maximum Likelihood Estimators

Maximum likelihood estimators have the following properties:

- By definition they are the parameter values best supported by the data.
- The maximum likelihood estimator of  $\gamma(\boldsymbol{\theta})$  is  $\gamma(\hat{\boldsymbol{\theta}})$  where  $\hat{\boldsymbol{\theta}}$  is the MLE of  $\boldsymbol{\theta}$ . This is called the **invariance property**.
- Consistency is generally true for maximum likelihood estimators. That is

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$$

In particular each component of  $\hat{\boldsymbol{\theta}}$  is consistent.

- The maximum likelihood estimator in the multiparameter situation is also asymptotically (approximately) normal under fairly general conditions. Let  $f(\mathbf{y}; \boldsymbol{\theta})$  denote the density function and let and let the maximum likelihood estimate of  $\boldsymbol{\theta}$  be the solution to the score equations

$$\frac{\partial \ln[f(\mathbf{y}; \boldsymbol{\theta})]}{\partial \theta_j} = 0 \quad j = 1, 2, \dots, k$$

Then the sampling distribution of  $\hat{\boldsymbol{\theta}}$  is approximately multivariate normal with mean vector  $\boldsymbol{\theta}_0$  and variance covariance matrix  $\mathbf{V}(\boldsymbol{\theta}_0)$  where

$$\mathbf{V}(\boldsymbol{\theta}_0) = [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$$

and the i-j element of  $\mathbf{I}(\boldsymbol{\theta}_0)$  is given by

$$-E \left\{ \frac{\partial^{(2)} \ln[f(\mathbf{y}; \boldsymbol{\theta}_0)]}{\partial \theta_i \partial \theta_j} \right\}$$

- $\mathbf{I}(\boldsymbol{\theta}_0)$  is called **Fisher's information matrix**.
- As in the case of one parameter we may replace  $\boldsymbol{\theta}_0$  by its estimate to obtain an estimate of  $\mathbf{V}(\boldsymbol{\theta}_0)$
- If  $g(\boldsymbol{\theta})$  is a function of  $\boldsymbol{\theta}$  then its maximum likelihood estimator,  $g(\hat{\boldsymbol{\theta}})$ , is approximately normal with mean  $g(\boldsymbol{\theta}_0)$  and variance  $v_g(\boldsymbol{\theta}_0)$  where

$$v_g(\boldsymbol{\theta}_0) = \boldsymbol{\nabla}_g^T \mathbf{V}(\boldsymbol{\theta}_0) \boldsymbol{\nabla}_g$$

and the ith element of  $\boldsymbol{\nabla}_g$  is given by

$$\frac{\partial g(\boldsymbol{\theta}_0)}{\partial \theta_i}$$

- We replace  $\boldsymbol{\theta}_0$  by  $\hat{\boldsymbol{\theta}}$  to obtain an estimate of  $v_g(\boldsymbol{\theta}_0)$

### 7.5.4 Two Sample Normal

Suppose that  $y_{11}, y_{12}, \dots, y_{1n_1}$  is a random sample from a distribution which is  $N(\mu_1, \sigma^2)$  and  $y_{21}, y_{22}, \dots, y_{2n_2}$  is an independent random sample from a distribution which is  $N(\mu_2, \sigma^2)$ . Then the likelihood of  $\mu_1, \mu_2$  and  $\sigma^2$  is given by

$$f(\mathbf{y}; \mu_1, \mu_2, \sigma^2) = \left[ \prod_{j=1}^{n_1} (2\pi\sigma^2)^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} (y_{1j} - \mu_1)^2 \right\} \right] \left[ \prod_{j=1}^{n_2} (2\pi\sigma^2)^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} (y_{2j} - \mu_2)^2 \right\} \right]$$

which simplifies to

$$(2\pi)^{-(n_1+n_2)/2} \sigma^{-(n_1+n_2)/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_1} (y_{1j} - \mu_1)^2 - \frac{1}{2\sigma^2} \sum_{j=1}^{n_2} (y_{2j} - \mu_2)^2 \right\}$$

It follows that the log likelihood is

$$-\frac{n_1 + n_2}{2} \ln(2\pi) - \frac{n_1 + n_2}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^{n_1} (y_{1j} - \mu_1)^2 - \frac{1}{2\sigma^2} \sum_{j=1}^{n_2} (y_{2j} - \mu_2)^2$$

The partial derivatives are thus

$$\begin{aligned} \frac{\partial \ln f(\mathbf{y}; \mu_1, \mu_2, \sigma^2)}{\partial \mu_1} &= \frac{1}{2\sigma^2} \sum_{j=1}^{n_1} (y_{1j} - \mu_1) \\ \frac{\partial \ln f(\mathbf{y}; \mu_1, \mu_2, \sigma^2)}{\partial \mu_2} &= \frac{1}{2\sigma^2} \sum_{j=1}^{n_2} (y_{2j} - \mu_2) \\ \frac{\partial \ln f(\mathbf{y}; \mu_1, \mu_2, \sigma^2)}{\partial \sigma^2} &= -\frac{n_1 + n_2}{2\sigma^2} - \frac{1}{2\sigma^4} \left[ \sum_{j=1}^{n_1} (y_{1j} - \mu_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \mu_2)^2 \right] \end{aligned}$$

Equating to 0 and solving yields the maximum likelihood estimators:

$$\begin{aligned}\hat{\mu}_1 &= \bar{y}_{1+} \\ \hat{\mu}_2 &= \bar{y}_{2+} \\ \hat{\sigma}^2 &= \frac{1}{n_1+n_2} \left[ \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_{1+})^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_{2+})^2 \right]\end{aligned}$$

The estimators for  $\mu_1$  and  $\mu_2$  are unbiased while the estimator for  $\sigma^2$  is biased. An unbiased estimator for  $\sigma^2$  is

$$\hat{\sigma}^2 = s_p^2 = \frac{1}{n_1 + n_2 - 2} \left[ \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_{1+})^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_{2+})^2 \right]$$

which is easily seen to be equal to

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$s_p^2$  is called the pooled estimate of  $\sigma^2$ .

Since  $\bar{y}_{1+}$  is a linear combination of independent normal random variables it has a sampling distribution which is normal with mean  $\mu_1$  and variance  $\sigma^2/n_1$ . Similarly  $\bar{y}_{2+}$  is normal with mean  $\mu_2$  and variance  $\sigma^2/n_2$ . It follows that the sampling distribution of  $\bar{y}_{2+} - \bar{y}_{1+}$  is normal with mean  $\mu_2 - \mu_1$  and variance  $\sigma^2(1/n_1 + 1/n_2)$  and is the maximum likelihood estimator of  $\mu_2 - \mu_1$ .

It can be shown that the sampling distribution of  $(n_1 + n_2 - 2)s_p^2/\sigma^2$  is chi-square with  $n_1 + n_2 - 2$  degrees of freedom and is independent of  $\bar{y}_{2+} - \bar{y}_{1+}$ . It follows that the sampling distribution of

$$T = \frac{(\bar{Y}_{2+} - \bar{Y}_{1+}) - (\mu_2 - \mu_1)/\sqrt{\sigma^2(1/n_1 + 1/n_2)}}{\sqrt{s_p^2/\sigma^2}} = \frac{(\bar{Y}_{2+} - \bar{Y}_{1+}) - (\mu_2 - \mu_1)}{s_p\sqrt{1/n_1 + 1/n_2}}$$

is Student's  $t$  with  $n_1 + n_2 - 2$  degrees of freedom. Hence we have that

$$P\left(-t_{1-\alpha/2}(n_1 + n_2 - 2) \leq \frac{(\bar{Y}_{2+} - \bar{Y}_{1+}) - (\mu_2 - \mu_1)}{s_p\sqrt{1/n_1 + 1/n_2}} \leq t_{1-\alpha/2}(n_1 + n_2 - 2)\right) = 1 - \alpha$$

It follows that a  $1 - \alpha$  confidence interval for  $\mu_2 - \mu_1$  is given by

$$\bar{Y}_{2+} - \bar{Y}_{1+} \pm t_{1-\alpha/2}(n_1 + n_2 - 2)s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

### 7.5.5 Simple Linear Regression Model

Suppose that  $y_1, y_2, \dots, y_n$  are realized values of  $Y_1, Y_2, \dots, Y_n$  which are independent normal with common variance  $\sigma^2$  and mean

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 x_i$$

where the  $x_i$  are known. This is called a simple linear regression model or a regression model with one covariate and an intercept.

Note that the parameter  $\beta_1$  in this model represents the change in the expected response associated with a unit change in the covariate  $x$ .

The likelihood is given by

$$f(\mathbf{y}; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

Thus the log likelihood is given by

$$-\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

It follows that the partial derivatives are given by

$$\begin{aligned} \frac{\partial \ln f(\mathbf{y}; \beta_0, \beta_1, \sigma^2)}{\partial \beta_0} &= \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \\ \frac{\partial \ln f(\mathbf{y}; \beta_0, \beta_1, \sigma^2)}{\partial \beta_1} &= \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i \\ \frac{\partial \ln f(\mathbf{y}; \beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

Equating to 0 and denoting the estimates by  $b_0$ ,  $b_1$  and  $\hat{\sigma}^2$  yields the three equations

$$\begin{aligned}nb_0 + n\bar{x}b_1 &= n\bar{y} \\n\bar{x}b_0 + \sum_{i=1}^n x_i^2 b_1 &= \sum_{i=1}^n x_i y_i \\n\hat{\sigma}^2 &= \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2\end{aligned}$$

It follows that

$$b_0 = \bar{y} - b_1 \bar{x}$$

Substituting this value of  $b_0$  into the second equation yields

$$n\bar{x}(\bar{y} - b_1 \bar{x}) + \sum_{i=1}^n x_i^2 b_1 = \sum_{i=1}^n x_i y_i$$

Combining terms and using the facts that

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2 \quad \text{and} \quad \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$$

gives  $b_1$  as:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Define

$$\hat{y}_i = b_0 + b_1 x_i$$

to be the estimated or “fitted” value of  $y_i$  and

$$y_i - \hat{y}_i$$

to be the residual or error made when we estimate  $y$  at  $x_i$  by  $\hat{y}_i$ . Then the estimate of  $\sigma^2$  is equal to

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n - 2}$$

where

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

is called the residual or error sum of squares.

### 7.5.6 Matrix Formulation of Simple Linear Regression

It is useful to rewrite the simple linear regression model in matrix notation. It turns out that in this formulation we can add as many covariates as we like and obtain essentially the same results. Define

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

Then the model may be written as

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \quad \text{var}(\mathbf{Y}) = \mathbf{I}\sigma^2$$

We now note that

$$\begin{aligned} \mathbf{X}^T \mathbf{X} \mathbf{b} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \mathbf{b} \\ &= \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \\ &= \begin{bmatrix} nb_0 + n\bar{x}b_1 \\ n\bar{x}b_0 + \sum_{i=1}^n x_i^2 b_1 \end{bmatrix} \end{aligned}$$

and

$$\mathbf{X}^T \mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

Hence the maximum likelihood equations for  $b_0$  and  $b_1$  are, in matrix terms,

$$\mathbf{X}^T \mathbf{X} \mathbf{b} = \mathbf{X}^T \mathbf{y}$$

From this representation we see that

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

From our earlier work on expected values and variance-covariances of multivariate normal distributions we see that  $\mathbf{b}$  has a multivariate normal distribution with mean vector

$$E(\mathbf{b}) = E[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(\mathbf{y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{I} \boldsymbol{\beta} = \boldsymbol{\beta}$$

and variance-covariance matrix

$$\begin{aligned} \text{var}(\mathbf{b}) &= \text{var}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \text{var}(\mathbf{y} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T]^T) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T [\mathbf{I} \sigma^2] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2 \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2 \end{aligned}$$

It follows that  $b_0$  and  $b_1$  are unbiased estimators of  $\beta_0$  and  $\beta_1$ . The variances are obtained as elements of  $(\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$  e.g. the variance of  $b_1$  is the element in the second row and second column of  $(\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$ .

Since

$$\begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} = (n \sum_{i=1}^n x_i^2 - n^2 \bar{x}^2)^{-1} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}$$

we see that the variance of  $b_1$  is given by

$$\frac{n}{n \sum_{i=1}^n x_i^2 - n^2 \bar{x}^2} \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Thus  $b_1$  has a normal distribution with mean  $\beta_1$  and variance given by the above expression.

It can be shown that  $\text{SSE}/\sigma^2$  has a chi-squared distribution with  $n-2$  degrees of freedom and is independent of  $b_1$ . It follows that the sampling distribution of

$$T = \frac{(b_1 - \beta_1)/\sqrt{\sigma^2/\sum_{i=1}^n (x_i - \bar{x})^2}}{\sqrt{\text{SSE}/(n-2)\sigma^2}} = \frac{b_1 - \beta_1}{\sqrt{\hat{\sigma}^2/\sum_{i=1}^n (x_i - \bar{x})^2}}$$

is Student's  $t$  with  $n-2$  degrees of freedom.

Hence a  $1-\alpha$  confidence interval for  $\beta_1$  is given by

$$b_1 \pm t_{1-\alpha/2}(n-2) \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

which may be rewritten as:

$$b_1 \pm t_{1-\alpha/2}(n-2) \text{s.e.}(b_1)$$

### 7.5.7 Two Sample Problem as Simple Linear Regression

In simple linear regression suppose that the covariate is given by

$$x_i = \begin{cases} 0 & i = 1, 2, \dots, n_1 \\ 1 & i = n_1 + 1, n_1 + 2, \dots, n_1 + n_2 \end{cases}$$

where  $n_1 + n_2 = n$ . Such a covariate is called a **dummy** or **indicator** variable since its values describe which group the observations belong to.

The simple linear regression model

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

becomes

$$E(Y_i) = \begin{cases} \beta_0 & i = 1, 2, \dots, n_1 \\ \beta_0 + \beta_1 & i = n_1 + 1, n_1 + 2, \dots, n_1 + n_2 \end{cases}$$

We now note that

$$\begin{aligned} \sum_{i=1}^n x_i &= n_2 ; \quad \sum_{i=1}^n y_i = n\bar{y} = n_1\bar{y}_{1+} + n_2\bar{y}_{2+} \\ \sum_{i=1}^n x_i^2 &= n_2 ; \quad \sum_{i=1}^n x_i y_i = \sum_{i=n_1+1}^n y_i = n_2\bar{y}_{2+} \end{aligned}$$

where we define

Group 1	Group 2
$y_{11} = y_1$	$y_{21} = y_{n_1+1}$
$y_{12} = y_2$	$y_{22} = y_{n_1+2}$
$y_{13} = y_3$	$y_{23} = y_{n_1+3}$
$\vdots$	$\vdots$
$y_{1n_1} = y_{n_1}$	$y_{2n_2} = y_{n_1+n_2}$

Thus the maximum likelihood equations become

$$\begin{aligned} (n_1 + n_2)b_0 + n_2b_1 &= n_1\bar{y}_{1+} + n_2\bar{y}_{2+} \\ n_2b_0 + n_2b_1 &= n_2\bar{y}_{2+} \end{aligned}$$

Subtract the second equation from the first to get

$$n_1b_0 = n_1\bar{y}_{1+} \quad \text{and hence} \quad b_0 = \bar{y}_{1+}$$

It follows that

$$b_1 = \bar{y}_{2+} - \bar{y}_{1+}$$

Moreover the fitted values are given by

$$\hat{y}_i = \begin{cases} b_0 = \bar{y}_{1+} & i = 1, 2, \dots, n_1 \\ b_0 + b_1 = \bar{y}_{2+} & i = n_1 + 1, n_1 + 2, \dots, n_1 + n_2 \end{cases}$$

so that the error sum of squares is given by

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n_1} (y_i - \bar{y}_{1+})^2 + \sum_{i=n_1+1}^{n_1+n_2} (y_i - \bar{y}_{2+})^2$$

Thus the estimate of  $\sigma^2$  is just the pooled estimate  $s_p^2$ .

It follows that a two sample problem is a special case of simple linear regression using a dummy variable to indicate group membership. The result holds for more than 2 groups i.e. a  $k$  sample problem is just a special case of multiple regression on  $k - 1$  dummy variables which indicate sample or group measurement. This is called a one-way analysis of variance and will be discussed in a later section.

### 7.5.8 Paired Data

Often we have data in which a response is observed on a collection of individuals at two points in time or under two different conditions. Since individuals are most likely independent but observations on the same individual are probably not independent a two sample procedure is not appropriate. The simplest approach is to take the difference between the two responses, individual by individual and treat the differences as a one sample problem.

Thus the data are

Subject	Response 1	Response 2	Difference
1	$y_{11}$	$y_{21}$	$d_1 = y_{21} - y_{11}$
2	$y_{12}$	$y_{22}$	$d_2 = y_{22} - y_{12}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n$	$y_{1n}$	$y_{2n}$	$d_n = y_{2n} - y_{1n}$

The confidence interval for the true mean difference is then based on  $\bar{d}$  with variance  $s_d^2/n$  exactly as in the case of a one sample problem.

### 7.5.9 Two Sample Binomial

Suppose that we have two observations  $y_1$  and  $y_2$  which come from two independent binomial distributions. One with  $n_1$  Bernoulli trials having probability  $p_1$  and the other with  $n_2$  Bernoulli trials having probability  $p_2$ .

The likelihood is given by

$$f(y_1, y_2, p_1, p_2) = \left[ \binom{n_1}{y_1} p_1^{y_1} (1 - p_1)^{n_1 - y_1} \right] \left[ \binom{n_2}{y_2} p_2^{y_2} (1 - p_2)^{n_2 - y_2} \right]$$

Thus the log likelihood is given by

$$\ln \left[ \binom{n_1}{y_1} \binom{n_2}{y_2} \right] + y_1 \ln(p_1) + (n_1 - y_1) \ln(1 - p_1) + y_2 \ln(p_2) + (n_2 - y_2) \ln(1 - p_2)$$

Hence the maximum likelihood equations are

$$\begin{aligned} \frac{\partial \ln[f(y_1, y_2; p_1, p_2)]}{\partial p_1} &= \frac{y_1}{p_1} - \frac{n_1 - y_1}{1 - p_1} = 0 \\ \frac{\partial \ln[f(y_1, y_2; p_1, p_2)]}{\partial p_2} &= \frac{y_2}{p_2} - \frac{n_2 - y_2}{1 - p_2} = 0 \end{aligned}$$

It follows that

$$\widehat{p}_1 = \frac{y_1}{n_1} ; \quad \widehat{p}_2 = \frac{y_2}{n_2}$$

The second derivatives of the log likelihood are

$$\begin{aligned}\frac{\partial^2 \ln[f(y_1, y_2; p_1, p_2)]}{\partial p_1^2} &= -\frac{y_1}{p_1^2} - \frac{n_1 - y_1}{(1-p_1)^2} \\ \frac{\partial^2 \ln[f(y_1, y_2; p_1, p_2)]}{\partial p_2^2} &= -\frac{y_2}{p_2^2} - \frac{n_2 - y_2}{(1-p_2)^2} \\ \frac{\partial^2 \ln[f(y_1, y_2; p_1, p_2)]}{\partial p_1 \partial p_2} &= 0 \\ \frac{\partial^2 \ln[f(y_1, y_2; p_1, p_2)]}{\partial p_2 \partial p_1} &= 0\end{aligned}$$

The expected values are given by

$$\begin{aligned}E \left\{ \frac{\partial^2 \ln[f(y_1, y_2; p_1, p_2)]}{\partial p_1^2} \right\} &= -\frac{n_1}{p_1} - \frac{n_1}{(1-p_1)} = -\frac{n_1}{p_1(1-p_1)} \\ E \left\{ \frac{\partial^2 \ln[f(y_1, y_2; p_1, p_2)]}{\partial p_2^2} \right\} &= -\frac{n_2}{p_2} - \frac{n_2}{(1-p_2)} = -\frac{n_2}{p_2(1-p_2)} \\ E \left\{ \frac{\partial^2 \ln[f(y_1, y_2; p_1, p_2)]}{\partial p_1 \partial p_2} \right\} &= 0 \\ E \left\{ \frac{\partial^2 \ln[f(y_1, y_2; p_1, p_2)]}{\partial p_2 \partial p_1} \right\} &= 0\end{aligned}$$

It follows that Fisher's Information matrix is given by

$$\begin{bmatrix} \frac{n_1}{p_1(1-p_1)} & 0 \\ 0 & \frac{n_2}{p_2(1-p_2)} \end{bmatrix}$$

Thus we may treat  $\widehat{p}_1$  and  $\widehat{p}_2$  as if they were normal with mean vector and variance covariance matrix

$$\begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \quad \begin{bmatrix} \frac{p_1(1-p_1)}{n_1} & 0 \\ 0 & \frac{p_2(1-p_2)}{n_2} \end{bmatrix}$$

**Estimate and Confidence Interval for  $p_2 - p_1$** 

The maximum likelihood estimate of  $g(p_1, p_2) = p_2 - p_1$  is given by

$$g(\hat{p}_2, \hat{p}_1) = \hat{p}_2 - \hat{p}_1 = \frac{y_2}{n_2} - \frac{y_1}{n_1}$$

Since

$$\nabla_g = \begin{bmatrix} \frac{\partial g(p_1, p_2)}{\partial p_1} \\ \frac{\partial g(p_1, p_2)}{\partial p_2} \end{bmatrix} = \begin{bmatrix} -1 \\ +1 \end{bmatrix}$$

the approximate variance of  $\hat{p}_2 - \hat{p}_1$  is given by

$$\begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} \frac{p_1(1-p_1)}{n_1} & 0 \\ 0 & \frac{p_2(1-p_2)}{n_2} \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

which we approximate by replacing  $p_1$  and  $p_2$  by their maximum likelihood estimates.

It follows that an approximate  $1 - \alpha$  confidence interval for  $p_2 - p_1$  is given by

$$(\hat{p}_2 - \hat{p}_1) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

provided both  $n_1$  and  $n_2$  are large.

**Estimate and Confidence Interval for the log odds ratio and the odds ratio**

The maximum likelihood estimate of the odds ratio is

$$\frac{\hat{p}_2/(1-\hat{p}_2)}{\hat{p}_1/(1-\hat{p}_1)}$$

while the maximum likelihood estimate of the log odds ratio is

$$\ln\left(\frac{\hat{p}_2}{1-\hat{p}_2}\right) - \ln\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right)$$

If we define

$$g(p_1, p_2) = \ln\left(\frac{p_2}{1-p_2}\right) - \ln\left(\frac{p_1}{1-p_1}\right) = \ln(p_2) - \ln(1-p_2) - \ln(p_1) + \ln(1-p_1)$$

we have that

$$\nabla_g = \begin{bmatrix} \frac{\partial g(p_1, p_2)}{\partial p_1} \\ \frac{\partial g(p_1, p_2)}{\partial p_2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{p_1} - \frac{1}{1-p_1} \\ \frac{1}{p_2} + \frac{1}{1-p_2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{p_1(1-p_1)} \\ \frac{1}{p_2(1-p_2)} \end{bmatrix}$$

Thus the variance of the approximate distribution of the log odds ratio is

$$\begin{bmatrix} -\frac{1}{p_1(1-p_1)} & \frac{1}{p_2(1-p_2)} \end{bmatrix} \begin{bmatrix} \frac{p_1(1-p_1)}{n_1} & 0 \\ 0 & \frac{p_2(1-p_2)}{n_2} \end{bmatrix} \begin{bmatrix} -\frac{1}{p_1(1-p_1)} \\ \frac{1}{p_2(1-p_2)} \end{bmatrix} = \frac{1}{n_1 p_1 (1-p_1)} + \frac{1}{n_2 p_2 (1-p_2)}$$

We approximate this by

$$\frac{1}{n_1 \hat{p}_1 (1 - \hat{p}_1)} + \frac{1}{n_2 \hat{p}_2 (1 - \hat{p}_2)} = \frac{1}{n_1 \hat{p}_1} + \frac{1}{n_1 (1 - \hat{p}_1)} + \frac{1}{n_2 \hat{p}_2} + \frac{1}{n_2 (1 - \hat{p}_2)}$$

It follows that a  $1 - \alpha$  confidence interval for the log odds ratio is given by

$$\ln \left( \frac{\hat{p}_2 / (1 - \hat{p}_2)}{\hat{p}_1 / (1 - \hat{p}_1)} \right) \pm z_{1-\alpha/2} \sqrt{\frac{1}{n_1 \hat{p}_1} + \frac{1}{n_1 (1 - \hat{p}_1)} + \frac{1}{n_2 \hat{p}_2} + \frac{1}{n_2 (1 - \hat{p}_2)}}$$

To obtain a confidence interval for the odds ratio simply exponentiate the endpoints of the confidence interval for the log odds ratio.

### 7.5.10 Logistic Regression Formulation of the Two sample Binomial

As in the case of the two sample normal there is a regression type formulation of the two sample binomial problem. Instead of  $p_1$  and  $p_2$  we use the equivalent parameters  $\beta_0$  and  $\beta_1$  defined by

$$\ln\left(\frac{p_1}{1-p_1}\right) = \beta_0 \quad \ln\left(\frac{p_2}{1-p_2}\right) = \beta_0 + \beta_1$$

That is we model the log odds of  $p_1$  and  $p_2$ . If we define a covariate  $x$  by

$$x_i = \begin{cases} 1 & i = 2 \\ 0 & i = 1 \end{cases}$$

then the logistic regression model states that

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

Note that  $\beta_1$  is the log odds ratio (sample 2 to sample 1).

STATA and other statistical software packages allow one to specify models of the above form in an easy fashion. STATA has three methods: logistic (used when the responses given are 0/1), blogit (used when the data are grouped as above) and glm (which handles both and other models as well).



## Chapter 8

# Hypothesis and Significance Testing

The statistical inference called **hypothesis or significance testing** provides an answer to the following problem:

**Given data and a probability model can we conclude that a parameter  $\theta$  has value  $\theta_0$ ?**

- $\theta_0$  is a specified value of the parameter  $\theta$  of particular interest and is called a **null hypothesis**.
- In the Neyman Pearson formulation of the hypothesis testing problem the choice is between the null hypothesis  $H_0 : \theta = \theta_0$  and an alternative hypothesis  $H_1 : \theta = \theta_1$ . Neyman and Pearson stressed that their approach was based on **inductive behavior**.
- In the significance testing formulation due mainly to Fisher an alternative hypothesis is not explicitly stated. Fisher stressed that his was an approach to **inductive reasoning**.
- In current practice the two approaches have been combined, the distinctions stressed by their developers has all but disappeared, and we are left with a mess of terms and concepts which seem to have little to do with advancing science.

## 8.1 Neyman Pearson Approach

### 8.1.1 Basic Concepts

A formal approach to the hypothesis testing problem is based on a **test of the null hypothesis that  $\theta=\theta_0$  versus an alternative hypothesis about  $\theta$**  e.g.

- $\theta = \theta_1$  ( simple alternative hypothesis).
- $\theta > \theta_0$  or  $\theta < \theta_0$  (one sided alternative hypotheses)
- $\theta \neq \theta_0$  (two sided alternative hypothesis).

In a problem in which we have a null hypothesis  $H_0$  and an alternative  $H_A$  there are two types of errors that can be made:

- $H_0$  is rejected when it is true.
- $H_0$  is not rejected when it is false.

The two types of errors can be summarized in the following table:

Conclusion	"Truth"	
	$H_0$ True	$H_0$ False
Reject $H_0$	Type I error	no error
Do not Reject $H_0$	no error	Type II Error

Thus

- Type I Error = reject  $H_0$  when  $H_0$  is true.
- Type II Error = do not reject  $H_0$  when  $H_0$  is false.
- Obviously we would prefer not to make either type of error.
- However, in the face of data which is subject to uncertainty we may make errors of either type.
- The Neyman-Pearson theory of hypothesis testing is the conventional approach to testing hypotheses.

### 8.1.2 Summary of Neyman-Pearson Approach

- Given the data and a probability model, choose a region of possible data values called the **critical region**.
  - If the observed data falls into the critical region reject the null hypothesis.
  - The critical region is selected so that it is consistent with departures from  $H_0$  in favor of  $H_A$ .
- The critical region is defined by the values of a **test statistic** chosen so that:
  - The probability of obtaining a value of the test statistic in the critical region is  $\leq \alpha$  if the null hypothesis is true. i.e. the probability of a Type I error (called the **size**) of the test is required to be  $\leq \alpha$ .
  - $\alpha$  is called the **significance level** of the test procedure. Typically  $\alpha$  is chosen to be .05 or .01.
  - The probability of obtaining a value of the test statistic in the critical region is as large as possible if the alternative hypothesis is true. (Equivalently the probability of a Type II error is as small as possible).
  - This probability is called the **power** of the test.

The Neyman-Pearson theory thus tests  $H_0$  vs  $H_A$  so that the probability of a Type I error is fixed at level  $\alpha$  while the power (ability to detect the alternative) is as large as possible.

Neyman and Pearson justified their approach to the problem from what they called the “inductive behavior” point of view:

“Without hoping to know whether each separate hypothesis is true or false we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.”

Thus a test is viewed as a rule of behavior.

### 8.1.3 The Neyman Pearson Lemma

In the case of a simple hypothesis  $H_0$  vs a simple alternative hypothesis  $H_1$  the Neyman Pearson Lemma establishes that there is a test which fixes the significance level and maximizes the power.

**Neyman Pearson Lemma:** Define  $C$  to be a critical region satisfying, for some  $k > 0$

- (1)  $f_1(\mathbf{x}) \geq kf_0(\mathbf{x})$  for all  $\mathbf{x} \in C$
- (2)  $f_1(\mathbf{x}) \leq kf_0(\mathbf{x})$  for all  $\mathbf{x} \notin C$
- (3)  $P_0(\mathbf{X} \in C) = \alpha$

then  $C$  is best critical region of size  $\leq \alpha$  for testing the simple hypothesis  $H_0 : f \sim f_0$  vs the simple alternative  $H_1 : f \sim f_1$ .

- All points  $\mathbf{x}$  for which

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} > k$$

are in the critical region  $C$

- Points for which the ratio is equal to  $k$  can be either in  $C$  or in  $\bar{C}$ .
- The ratio

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$$

is called the **likelihood ratio**.

- Points are in the critical region according to how strongly they support the alternative hypothesis vis a vis the null hypothesis i.e. according to the magnitude of the likelihood ratio.
  - That is, points in the critical region have the most value for discriminating between the two hypotheses subject to the restriction that their probability under the null hypothesis be less than or equal to  $\alpha$ .

**example:** Consider two densities for a random variable  $X$  defined by

$x$ value	Probability Under $\theta_0$	Probability Under $\theta_1$	Likelihood Ratio
1	.50	.01	1/50=.02
2	.30	.04	4/30=.13
3	.15	.45	45/15=3.0
4	.04	.30	30/4=7.5
5	.01	.20	20/1=20

To test  $H_0 : \theta = \theta_0$  vs  $H_1 : \theta = \theta_1$  with significance level .05 the Neyman Pearson Lemma says that the best test is

$$\text{Reject } H_0 \text{ if } x = 4 \text{ or } x = 5$$

The size is then

$$\text{size} = P_{\theta_0}(X = 4, 5) = .04 + .01 = .05$$

and the power is

$$\text{power} = P_{\theta_1}(X = 4, 5) = .30 + .20 = .50$$

Note, however, that if  $x = 3$  (which occurs 15% of the time under  $H_0$  and 45% of the time under  $H_1$ ) we would not reject  $H_0$  even though  $H_1$  is 3 times better supported than  $H_0$ .

Thus the formal theory of hypothesis testing is incompatible with the Law of Likelihood. If a prior distribution for  $\theta$  assigned equal probabilities to  $\theta_0$  and  $\theta_1$  then the posterior probability of  $\theta_1$  would be 3 times that of  $\theta_0$ . Thus the formal theory of hypothesis testing is incompatible with the Bayesian approach also.

**example:** Let the  $Y_i$ s be i.i.d.  $N(\mu, \sigma^2)$  where  $\sigma^2$  is known. For the hypothesis

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu = \mu_1 > \mu_0$$

we have that

$$\begin{aligned} k < \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} &= \exp \left\{ \frac{1}{2\sigma^2} \left[ \sum_{i=1}^n (y_i - \mu_0)^2 - \sum_{i=1}^n (y_i - \mu_1)^2 \right] \right\} \\ &= \exp \left\{ \frac{1}{2\sigma^2} \left[ -2n\bar{y}\mu_0 + n\mu_0^2 + 2n\bar{y}\mu_1 - n\mu_1^2 \right] \right\} \\ &= \exp \left\{ \frac{n(\mu_1 - \mu_0)}{\sigma^2} \left[ \bar{y} - \frac{\mu_0 + \mu_1}{2} \right] \right\} \end{aligned}$$

It follows that

$$\frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} > k \iff \bar{y} > \frac{\sigma^2 \log(k)}{n(\mu_1 - \mu_0)} + \frac{\mu_1 + \mu_0}{2} = k_1$$

It follows that  $\{\mathbf{y} : \bar{y} > k_1\}$  is the critical region for the most powerful test.

If we want the critical region to have size  $\alpha$  then we choose  $k_1$  so that

$$P_0(\bar{Y} > k_1) = \alpha$$

i.e.

$$P_0\left(\frac{\sqrt{n}(\bar{Y} - \mu_0)}{\sigma} > \frac{\sqrt{n}(k_1 - \mu_0)}{\sigma}\right) = \alpha$$

Thus

$$k_1 = \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$$

The test procedure is thus to reject when the observed value of  $\bar{Y}$  exceeds

$$k_1 = \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$$

For this test we have that the power is given by

$$\begin{aligned} P_1(\bar{Y} \geq k_1) &= P_1\left(\bar{Y} \geq \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right) \\ &= P_1\left(\frac{\sqrt{n}(\bar{Y} - \mu_1)}{\sigma} \geq -\frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma} + z_{1-\alpha}\right) \\ &= P\left(Z \geq -\frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma} + z_{1-\alpha}\right) \end{aligned}$$

If the alternative hypothesis was that  $\mu = \mu_1 < \mu_0$  the test would be to reject if

$$\bar{y} \leq \mu_0 - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$$

and the power of this test would be given by

$$P\left(Z \leq \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma} - z_{1-\alpha}\right)$$

There are several important features of the power of this test:

- As the difference between  $\mu_1$  and  $\mu_0$  increases the power increases.
- As  $n$  increases the power increases.
- As  $\sigma^2$  decreases the power increases.

### 8.1.4 Sample Size and Power

**application:**

In a study on the effects of chronic exposure to lead, a group of 34 children living near a lead smelter in El Paso, Texas were found to have elevated blood lead levels.

- A variety of tests were performed to measure neurological and psychological function.
- For IQ measurements the following data were recorded:

$$\text{sample mean} = \bar{y} = 96.44 \quad \text{and standard error} = 2.36$$

where  $y$  is the response variable and is the IQ of a subject.

- Assuming the data are normally distributed (IQs often are), the 95% confidence interval for  $\mu$ , defined as the population mean IQ for children with elevated blood lead values, is given by

$$96.44 \pm (2.035)(2.36) \quad \text{or} \quad 91.6 \text{ to } 101.2$$

where 2.035 is the .975 Student's  $t$  value with 33 degrees of freedom.

- Thus values of  $\mu$  between 91.6 and 101.3 are consistent with the data at a 95% confidence level.

Assuming a population average IQ of 100 we see that these exposed children appear to have reduced IQs. This example, when viewed in a slightly different way, has implications for public health policy.

- A difference of say, 5 points in IQ, is probably not that important for an individual.
- However, if the average IQ of a population is reduced by 5 points the proportion of individuals classified as retarded (IQ below 60) can be significantly increased.

To see this, suppose that IQs are normally distributed with mean 100 and standard deviation 20.

- In this situation the proportion of individuals having IQ below 60 is

$$P(IQ \leq 60) = P\left(Z \leq \frac{60 - 100}{20}\right) = P(Z \leq -2) = .0228$$

or about 2 per hundred.

- If the average IQ is reduced by 5 points to 95 the proportion having IQ below 60 is given by

$$P(IQ \leq 60) = P\left(Z \leq \frac{60 - 95}{20}\right) = P(Z \leq -1.75) = .0401$$

which is nearly double the previous proportion.

Given this result we may ask the question: How large a study should be performed to detect a difference of  $\Delta = 5$  points in IQ?

From the general equations given previously we would reject  $H_0 : \Delta = 0$  when

$$\bar{y} \leq \mu_0 - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$$

and the power of the test is

$$P \left( Z \leq \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma} - z_{1-\alpha} \right)$$

For the power to exceed  $1 - \beta$  where  $\beta$  is the Type II error probability we must have

$$P \left( Z \leq \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma} - z_{1-\alpha} \right) \geq 1 - \beta$$

It follows that

$$\frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma} - z_{1-\alpha} \geq z_{1-\beta}$$

or

$$\sqrt{n} \geq \frac{(z_{1-\alpha} + z_{1-\beta})\sigma}{\Delta}$$

Thus the sample size must satisfy

$$n \geq \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{\Delta^2}$$

For the example with IQ's we have that

$$\Delta = 5 \quad z_{1-\alpha} = 1.645 \quad z_{1-\beta} = .84 \quad \sigma = 20$$

for a test with size .05 and power .80. Thus we need a sample size of at least

$$n \geq \frac{(1.645 + .84)^2 \times 20^2}{5^2} = 98.8$$

i.e. we need a sample size of at least 99 to detect a difference of 5 IQ points.

Note that the formula for sample size can be “turned around” to determine what value of  $\mu$  could be detected for a given sample size and values of  $\mu_0$ ,  $\beta$ ,  $\sigma$  and  $\alpha$  as follows:

- 

$$\Delta = |\mu_1 - \mu_0| = \sigma \left( \frac{z_{1-\alpha} + z_{1-\beta}}{\sqrt{n}} \right)$$

- Thus in the example we have

$$|\mu_1 - 100| = 20 \left( \frac{1.645 + .84}{\sqrt{34}} \right) = 8.52$$

so that we can detect values of  $\mu \leq 91.5$  with a sample size of 34,  $\sigma = 20$ ,  $\alpha = .05$  and power .80.

- This kind of analysis is called **power analysis** in the social science literature.
- Power and sample size determination can be done for any test procedure although the formulas frequently become quite complicated.
- The quantity

$$\frac{|\mu_1 - \mu_0|}{\sigma}$$

is called the **effect size** and is usually denoted by ES.

**Reference:** Landigran et al Neuropsychological Dysfunction in Children with Chronic Low-Level Lead Absorption (1975). *Lancet*; March 29; 708-715.

## 8.2 Generalized Likelihood Ratio Tests

In the typical situation where the alternative and/or the null hypothesis is composite the Neyman Pearson Lemma is not applicable but can still be used to motivate development of test statistics.

Consider the problem of testing the null hypothesis that  $\boldsymbol{\theta}$  is in  $\Theta_0$  versus the alternative that  $\boldsymbol{\theta}$  is not in  $\Theta_0$ . We assume that the full parameter space is  $\Theta$  and that this set is a subset of  $R^n$ .

The test statistic is given by

$$\lambda(\mathbf{y}) = \frac{\max_{\theta \in \Theta_0} f(\mathbf{y}; \theta)}{\max_{\theta \in \Theta} f(\mathbf{y}; \theta)}$$

and we reject  $H_0$  if  $\lambda(\mathbf{y})$  is small.

The rationale for the test is clear:

- If the null hypothesis is true the maximum value of the likelihood in the numerator will be close to the maximum value of the likelihood in the denominator i.e. the test statistic will be close to one.
- If the null hypothesis is not true then  $\boldsymbol{\theta}$  which maximizes the numerator will be different from the  $\boldsymbol{\theta}$  which maximizes the denominator and the ratio will be small.

Such tests are called **generalized likelihood ratio tests** and they have some desirable properties:

- They reduce to the Neyman Pearson Lemma when the null and the alternative are simple.
- They usually have desirable large sample properties.
- They usually give tests with useful interpretations.

The procedure for developing generalized likelihood ratio tests is simple:

- (1) Find the maximum likelihood estimate of  $\theta$  under the null hypothesis and calculate  $f(\mathbf{y}; \theta)$  at this value of  $\theta$ .
- (2) Find the maximum likelihood estimate of  $\theta$  under the full parameter space and calculate  $f(\mathbf{y}; \theta)$  at this value of  $\theta$ .
- (3) Form the ratio and simplify to a statistic whose sampling distribution can be found either exactly or approximately.
- (4) Determine critical values for this statistic, compute the observed value and thus test the hypothesis.

**example:** Let the  $Y_i$ s be i.i.d.  $N(\mu, \sigma^2)$  where  $\sigma^2$  is unknown. For the hypothesis

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$

we have that

$$\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0, 0 < \sigma^2\} \quad \Theta = \{(\mu, \sigma^2) : -\infty < \mu < +\infty, 0 < \sigma^2\}$$

The likelihood under the null hypothesis is

$$f(\mathbf{y}; \boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_0)^2 \right\}$$

which is maximized when

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mu_0)^2$$

and the maximized likelihood is given by

$$(2\pi\tilde{\sigma}^2)^{-n/2} \exp\{-n/2\}$$

Under the full parameter space the likelihood is

$$f(\mathbf{y}; \boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right\}$$

which is maximized when

$$\hat{\mu} = \bar{y} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

The resulting maximized likelihood is given by

$$(2\pi\hat{\sigma}^2)^{-n/2} \exp\{-n/2\}$$

Hence the generalized likelihood ratio test statistic is given by

$$\left[ \frac{\hat{\sigma}^2}{\tilde{\sigma}^2} \right]^{n/2} = \left\{ \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \mu_0)^2} \right\}^{n/2}$$

Since

$$\sum_{i=1}^n (y_i - \mu_0)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu_0)^2$$

the test statistic may be written as

$$\left\{ 1 + \frac{n(\bar{y} - \mu_0)^2}{(n-1)s^2} \right\}^{-n/2}$$

Thus we reject  $H_0$  when

$$\frac{|\bar{y} - \mu_0|}{\sqrt{s^2/n}}$$

is large i.e. when the statistic

$$\frac{\bar{y} - \mu_0}{\sqrt{s^2/n}} \leq -t_{1-\alpha/2} \text{ or } \geq t_{1-\alpha/2}$$

where  $t_{1-\alpha/2}$  comes from the Student's  $t$  distribution with  $n - 1$  degrees of freedom.

This test is called the one sample Student's  $t$  test.

### 8.2.1 One Way Analysis of Variance

Consider a situation in which there are  $p$  groups and  $n_i$  observations on a response variable  $y$  in each group. The data thus has the form:

Group 1	Group 2	...	Group $p$
$y_{11}$	$y_{21}$	...	$y_{p1}$
$y_{12}$	$y_{22}$	...	$y_{p2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_{1n_1}$	$y_{2n_2}$	...	$y_{pn_p}$

Thus  $y_{ij}$  is the  $j$ th observation in the  $i$ th group and define  $n$  to be the sum of the  $n_i$ s.

We assume that the  $y_{ij}$  are observed values of random variables,  $Y_{ij}$ , assumed to be independent, normal, with constant variance and

$$E(Y_{ij}) = \mu_i \text{ for } j = 1, 2, \dots, n_i$$

This set up is called a one way analysis of variance. The null hypothesis of interest is

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_p$$

and the alternative hypothesis is not  $H_0$  i.e. the null hypothesis is that there are no differences between the means of the groups while the alternative is that some of the group means are different.

Under the full model the likelihood is given by

$$f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^p \prod_{j=1}^{n_i} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \mu_i)^2\right\}$$

which reduces to

$$(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2\right\}$$

Hence the log likelihood is given by

$$-\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$$

The partial derivative with respect to  $\mu_i$  is clearly

$$\frac{1}{\sigma^2} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)$$

The partial derivative with respect to  $\sigma^2$  is

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$$

Equating to 0 yields the maximum likelihood estimates to be

$$\hat{\mu}_i = \bar{y}_{i+} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i+})^2$$

and hence the maximized likelihood is

$$(2\pi\hat{\sigma}^2)^{-n/2} \exp\{-n/2\}$$

Under the null hypothesis the likelihood is

$$f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^p \prod_{j=1}^{n_i} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \mu)^2\right\}$$

which reduces to

$$(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \mu)^2\right\}$$

Hence the log likelihood is given by

$$-\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \mu)^2$$

The partial derivative with respect to  $\mu$  is given by

$$\frac{1}{\sigma^2} \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \mu)$$

The partial derivative with respect to  $\sigma^2$  is

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \mu)^2$$

Equating to 0 and solving yields

$$\tilde{\mu} = \bar{y}_{++} \quad \tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{++})^2$$

and hence the maximized likelihood under  $H_0$  is

$$(2\pi\tilde{\sigma}^2)^{-n/2} \exp\{-n/2\}$$

The generalized likelihood ratio statistic is thus

$$\left[ \frac{\hat{\sigma}^2}{\tilde{\sigma}^2} \right]^{n/2} = \left[ \frac{\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i+})^2}{\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{++})^2} \right]^{-n/2}$$

Now we note that

$$\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{++})^2 = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i+})^2 + \sum_{i=1}^p n_i (\bar{y}_{i+} - \bar{y}_{++})^2$$

so that the generalized likelihood ratio test statistic may be written as

$$\left\{ 1 + \frac{\sum_{i=1}^p n_i (y_{i+} - \bar{y}_{++})^2}{\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i+})^2} \right\}^{-n/2}$$

so that we reject  $H_0$  when

$$\frac{\sum_{i=1}^p n_i (y_{i+} - \bar{y}_{++})^2}{\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i+})^2}$$

is large or when

$$\frac{\sum_{i=1}^p n_i (y_{i+} - \bar{y}_{++})^2 / (p-1)}{\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i+})^2 / (n-p)}$$

is large.

The sampling distribution of this later statistic is  $F$  with  $p - 1$  and  $n - p$  degrees of freedom.

Thus the generalized likelihood ratio test is to reject  $H_0$ , all group means equal when the statistic

$$F_{obs} = \frac{\sum_{i=1}^p n_i (y_{i+} - \bar{y}_{++})^2 / (p - 1)}{\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i+})^2 / (n - p)}$$

exceeds the critical value of the  $F$  distribution with  $p - 1$  and  $n - p$  degrees of freedom.

Preliminary exploration of the data should include calculation of the sample means and a boxplot for each group. These provide rough conclusions about equality of the groups and a quick check on the equality of variability between groups.

## 8.3 Significance Testing and P-Values

Long before the development of the Neyman-Pearson theory significance tests were used to investigate hypotheses. These tests were developed on the basis of intuition and were used to determine whether or a not a given hypothesis was consistent with the observed data. An alternative hypothesis was not explicitly mentioned. Fisher's thoughts about significance tests were that they are part of a process of "inductive reasoning" from the data to scientific conclusions. After Neyman and Pearson the tests developed by their theories began to be used as significance tests. Thus the two approaches merged and are today considered as branches of the same theory.

### 8.3.1 P Values

**Definition:** The **P-value** associated with a statistical test is **the probability of obtaining a result as or more extreme than that observed.**

- Note that the probability is calculated under the assumption that the null hypothesis is true.

### 8.3.2 Interpretation of P-values

In this section we discuss the conventional interpretation of P-values. By definition the P-value gives the chance of observing a result as or more extreme when the null hypothesis is true under the assumed model.

Thus finding a small P-value in an analysis means either:

- the model is wrong or
- a rare event has occurred or
- the null hypothesis is not true

Given that we assume the model to be true and that it is unlikely that a rare event has occurred, a small P-value leads to the conclusion that  $H_0$  is not true.

By convention, the P-value for a two-sided test is taken to be the twice the one-sided P-value.

By convention statisticians have chosen the following guidelines for assessing the magnitude of P values:

- P value greater than .10, not statistically significant.
- P value between .10 and .05, marginally statistically significant (R)
- P value between .05 and .01, statistically significant, (\*)
- P value between .01 and .001, statistically significant, (\*\*)
- P value less than .001, statistically significant, (\*\*\*)

**example:**

For the data set used in a previous section we had a random sample of 34 children with elevated blood lead values. For this sample the observed sample mean IQ was  $\bar{y} = 96.44$ .

- If we assume that the value of  $\sigma$  is known to be 20 consider the hypothesis that  $\mu = \mu_0 = 100$
- The P-value is given by

$$\begin{aligned} P(\bar{Y} \leq \bar{y}_{obs}) &= P\left(\frac{\sqrt{34}(\bar{Y} - 100)}{20} \leq \frac{\sqrt{34}(96.44 - 100)}{20}\right) \\ &= P(Z \leq z_{obs} = -1.04) \\ &= .1492 \end{aligned}$$

- The P-value is interpreted as “if the null hypothesis were true ( $\mu = 100$ ) we would expect to see a sample mean IQ as small as observed (96.44) 15% of the time”, not a particularly rare event.
- This leads to the conclusion that  $\mu = 100$  is consistent with the observed data.

**example:** For the same data set as in the previous example the observed sample mean IQ was  $\bar{y} = 96.44$  and the sample standard error was 2.36.

- To test the hypothesis that  $\mu = \mu_0 = 100$  vs the alternative that  $\mu < 100$ , we calculate the  $t_{obs}$  statistic as follows:

$$t_{obs} = \frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{96.44 - 100}{2.36} = -1.51$$

- Since this value is not less than the critical value of  $t_{.05} = -1.70$  with 30 degrees of freedom, we would not reject the hypothesis that  $\mu = 100$ .
- The P-value is given by

$$P(T \leq t_{obs}) = P(T \leq -1.51)$$

which is between .05 and .10

- The p-value is interpreted as “if the null hypothesis ( $\mu = 100$ ) were true we would expect to see a sample mean IQ as small as observed (96.44) between 5% and 10% of the time”, not a particularly rare event.
- This leads to the conclusion that  $\mu = 100$  is consistent with the observed data. Note, however, that the P-value is marginally significant.

### 8.3.3 Two Sample Tests

Suppose we are given two random samples

$$x_1, x_2, \dots, x_{n_1} \text{ and } y_1, y_2, \dots, y_{n_2}$$

with the  $x$  sample coming from a  $N(\mu_1, \sigma^2)$  population and the  $y$  sample coming from a  $N(\mu_2, \sigma^2)$  population. Of interest is the null hypothesis that  $\mu_1 = \mu_2$ .

- The test statistic in this case is

$$t_{obs} = \frac{\bar{y} - \bar{x}}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

is the pooled estimate of  $\sigma^2$ . We reject in this case if

- $t_{obs} \geq t_{1-\alpha}$  if the alternative is  $\mu_2 > \mu_1$
  - $t_{obs} \leq -t_{1-\alpha}$  if the alternative is  $\mu_2 < \mu_1$
  - $|t_{obs}| \geq t_{1-\alpha/2}$  if the alternative is  $\mu_2 \neq \mu_1$
- The P-values for each of the one sided hypotheses is given by

$$\text{P-value} = P(T \geq |t_{obs}|)$$

and is twice the above P-value for the two sided hypothesis.

**example:**

The following data set gives the birth weights in kilograms of 15 children born to non-smoking mothers and 14 children born to mothers who are heavy smokers. The source of the data is Kirkwood, B.R. (1988) *Essentials of Medical Statistics* Blackwell Scientific Publications page 44, Table 7.1

Of interest is whether the birthweights of children whose mothers are smokers are less than the birthweights of non-smoking mothers.

Non-Smoker	Smoker
3.99	3.52
3.79	3.75
3.60	2.76
3.73	3.63
3.21	3.23
3.60	3.59
4.08	3.60
3.61	2.38
3.83	2.34
3.31	2.84
4.13	3.18
3.26	2.90
3.54	3.27
3.51	3.85
2.71	

- For these data we find that
  - mean for non-smoking mothers = 3.593, sample variance = 0.1375
  - mean for smoking mothers = 3.203, sample variance = 0.2427
- The pooled estimate of  $\sigma^2$  is thus

$$s_p^2 = \frac{(14 \times .1375) + (13 \times .2427)}{15 + 14 - 2} = .1882$$

- The Student's  $t$  statistic is given by

$$t_{obs} = \frac{3.203 - 3.593}{\sqrt{.1882(\frac{1}{15} + \frac{1}{14})}} = -2.42$$

- From the table of the Student's  $t$  distribution with 27 degrees of freedom we find that the P-value (one-sided) is .011 so that we reject the hypothesis of equal birthweights for smoking and non-smoking mothers and conclude that smoking mothers give birth to children with lower birthweights.
- The 95% confidence interval for the difference in birth weights is given by

$$(3.203 - 3.593) \pm 2.05 \sqrt{.1882(\frac{1}{15} + \frac{1}{14})} \quad \text{or} \quad -.390 \pm .330$$

- Thus birthweight differences between  $-.72$  and  $-.06$  kilograms are consistent with the observed data.
- Whether or not such differences are of clinical importance is a matter for determination by clinicians.

## 8.4 Relationship Between Tests and Confidence Intervals

There is a close connection between confidence intervals and two-sided tests:

**If a  $100(1 - \alpha)\%$  confidence interval is constructed and a hypothesized parameter value is not in the interval, we reject that value of the parameter at significance level  $\alpha$  using a two-sided test**

- Thus values of a parameter in a confidence interval are consistent with the data in the sense that they would not be rejected if used as a value for the null hypothesis.
- Equivalently, values of the parameter not in the confidence interval are inconsistent with the data since they would be rejected if used as a value for the null hypothesis.

## 8.5 General Case

If the estimate of a parameter  $\theta$ ,  $\hat{\theta}$ , has a sampling distribution which is approximately normal, centered at  $\theta$  with estimated standard error  $s.e.(\hat{\theta})$  then an approximate test of  $H_0 : \theta = \theta_0$  may be made using the results for the normal distribution.

- Calculate the test statistic

$$z_{obs} = \frac{\hat{\theta} - \theta_0}{s.e.(\hat{\theta})}$$

and treat it exactly as for the normal distribution.

- In particular, **if the ratio of the estimate to its estimated standard error is larger than 2, then the hypothesis that the parameter value is zero is inconsistent with the data.**
- This fact allows one to assess the significance of results in a variety of complicated statistical models.

### 8.5.1 One Sample Binomial

The observed data consists of the number of successes,  $x$ , in  $n$  trials, resulting from a binomial distribution with parameter  $p$  representing the probability of success.

The null hypothesis is that  $p = p_0$  with alternative hypothesis  $p > p_0$  or  $p < p_0$  or  $p \neq p_0$ . It is intuitively clear that:

- If the alternative is that  $p > p_0$ , large values of  $x$  suggest that the alternative hypothesis is true.
- If the alternative is that  $p < p_0$ , small values of  $x$  suggest that the alternative hypothesis is true.
- If the alternative is that  $p \neq p_0$ , both large and small values of  $x$  suggest that the alternative hypothesis is true.

The principal difference between testing hypothesis for discrete distributions, such as the binomial, is

- the significance level can not be made exactly equal to  $\alpha$  as it can be for the normal distribution.
- We thus choose the critical region so that the probability of a Type I error is as close to  $\alpha$  as possible without exceeding  $\alpha$ .

If the sample size,  $n$ , in the binomial is large we use the fact that  $\hat{p}$  is approximately normal to calculate a  $z_{obs}$  statistic as

$$z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

and use the results for the normal distribution.

**example:**

It is known that the success probability for a standard surgical procedure is .6. A pilot study of a new surgical procedure results in 10 successes out of 12 patients. Is there evidence that the new procedure is an improvement over the standard procedure?

We find the P-value using STATA to be .083 indicating that there is not enough evidence that the new procedure is superior to the standard.

If we calculate the approximate confidence interval for  $p$  we find that the upper confidence limit is given by

$$.8333 + 1.96\sqrt{\frac{.8333 \times .1667}{12}} = 1.044$$

while the lower confidence limit is given by

$$.8333 - 1.96\sqrt{\frac{.8333 \times .1667}{12}} = .6224$$

or [.62, 1.0).

Since the sample size is too small for the large sample result to be valid we calculate the exact upper and lower confidence using STATA. We find that the exact confidence interval is .515 to .979.

Conclusion: There is insufficient evidence to conclude that the new treatment is superior to the standard, but, because the study is small there was little power to detect alternatives of importance.

## 8.6 Comments on Hypothesis Testing and Significance Testing

### 8.6.1 Stopping Rules

**example:** A scientist presents the results of 6 Bernoulli trials as  $(0, 0, 0, 0, 0, 1)$  and wishes to test

$$H_0 : p = \frac{1}{2} \text{ vs } H_1 : p = \frac{1}{3}$$

Under the assumed model the MP test rejects when  $\sum_i Y_i = 0$  and has  $\alpha = \left(\frac{1}{2}\right)^6 < .05$ . Thus with the observed data we do not reject  $H_0$  since  $\sum_i Y_i = 1$ .

Suppose, however, that he informs you that he ran trials until he obtained the first success. Now we note that

$$P(\text{first success on trial } r) = (1 - p)^{r-1}p$$

and to test  $H_0 : p = p_0$  vs  $H_1 : p = p_1 < p_0$  the likelihood ratio is

$$\frac{(1 - p_1)^{r-1}p_1}{(1 - p_0)^{r-1}p_0} = \frac{p_1}{p_0} \left( \frac{1 - p_1}{1 - p_0} \right)^{r-1}$$

which is large when  $r$  is large since  $1 - p_1 > 1 - p_0$  if  $p_1 < p_0$

Now note that

$$\begin{aligned} P(R \geq r) &= \sum_{y=r}^{\infty} (1 - p)^{y-1}p \\ &= p(1 - p)^{r-1} \sum_{y=r}^{\infty} (1 - p)^{y-r} \\ &= p(1 - p)^{r-1} \frac{1}{1 - (1 - p)} \\ &= (1 - p)^{r-1} \end{aligned}$$

Thus if  $p = \frac{1}{2}$  we have that

$$P(R \geq 6) = \left(\frac{1}{2}\right)^5 < .05$$

and we reject  $H_0$  since the first success occurred on trial number 6.

Note however that the likelihood ratio for the first test is

$$\frac{\left(\frac{2}{3}\right)^5 \left(\frac{1}{3}\right)}{\left(\frac{1}{2}\right)^6} = \frac{2^{11}}{3^6} = 2.81$$

while the likelihood ratio for the second test is

$$\frac{\left(\frac{1}{3}\right) \left(\frac{2}{3}\right)^5}{\left(\frac{1}{2}\right)^6} = \frac{2^{11}}{3^6} = 2.81$$

Note that the two likelihood ratios are exactly the same. However, the two tests resulted in opposite conclusions. The fact that the LR provides evidence in favor of  $H_1$  with strength 2.81 does not appear in the Neyman Pearson approach.

Thus stopping rules make a difference in the classical theory but not in using the Law of Likelihood.

### 8.6.2 Tests and Evidence

**example:** Does rejection of  $H_0$  imply evidence against  $H_0$ ? **No!**

To see this let  $Y_i$  be i.i.d.  $N(\theta, 1)$  and let

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta = \theta_1 > 0$$

The MP test of size  $\alpha$  is to reject if  $\sqrt{n}\bar{Y} \geq 1.645$ . The likelihood ratio is given by

$$\frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^n(y_i - \theta_1)^2\right\}}{\exp\left\{-\frac{1}{2}\sum_{i=1}^n y_i^2\right\}} = \exp\left\{+n\bar{y}\theta_1 - \frac{n\theta_1^2}{2}\right\}$$

so that the likelihood ratio is

$$\exp\left\{n\theta_1\left(\bar{y} - \frac{\theta_1}{2}\right)\right\}$$

at the critical value i.e.  $\bar{y} = \frac{1.645}{\sqrt{n}}$  the likelihood ratio is

$$\exp\left\{n\theta_1\left(\frac{1.645}{\sqrt{n}} - \frac{\theta_1}{2}\right)\right\}$$

Suppose now that the power is large i.e. .99. Then we have

$$\begin{aligned}.99 &= P_{\theta_1}(\sqrt{n}\bar{Y} \geq 1.645) \\ &= P_{\theta_1}(\sqrt{n}(\bar{Y} - \theta_1) \geq 1.645 - \sqrt{n}\theta_1)\end{aligned}$$

so that  $1.645 - \sqrt{n}\theta_1 = -2.33$  i.e.  $\sqrt{n}\theta_1 = 3.97$  Thus if  $\theta_1 = \frac{3.97}{\sqrt{n}}$  the likelihood ratio at the critical value of  $\bar{Y}$  is

$$\exp \left\{ 3.97(1.645) - \frac{3.97^2}{2} \right\} = \exp \{-1.37\} = .254$$

Thus the MP test says to reject whenever the likelihood ratio exceeds .254. However the likelihood is higher under  $H_0$  than under  $H_1$  by a factor of

$$(.254)^{-1} = 3.9$$

### 8.6.3 Changing Criteria

**example:** If instead of minimizing the probability of a Type II error (maximizing the power) for a fixed probability of a Type I error we choose to minimize a linear combination of  $\alpha$  and  $\beta$  we get an entirely different critical region.

Note that

$$\begin{aligned}\alpha + \lambda\beta &= E_0[\delta(\mathbf{Y})] + \lambda\{1 - E_1[\delta(Y)]\} \\ &= \int_C f_0(\mathbf{y})d\mathbf{y} + \lambda - \lambda \int_C f_1(\mathbf{y})d\mathbf{y} \\ &= \lambda + \int_C [f_0(\mathbf{y}) - \lambda f_1(\mathbf{y})]d\mathbf{y}\end{aligned}$$

which is minimized when

$$\begin{aligned}C &= \{\mathbf{x} : f_0(\mathbf{y}) - \lambda f_1(\mathbf{y}) < 0\} \\ &= \left\{ \mathbf{y} : \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} > \lambda \right\}\end{aligned}$$

Thus the test statistic which minimizes  $\alpha + \lambda\beta$  is given by

$$\delta(y) = \begin{cases} 1 & \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} > \lambda \\ \text{arbitrary} & \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} = \lambda \\ 0 & \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} < \lambda \end{cases}$$

Notice that this test is essentially the Law of Likelihood.

## 8.7 Multinomial Problems and Chi-Square Tests

Recall that  $Y_1, Y_2, \dots, Y_{k-1}$  has a multinomial distribution with parameters  $\theta_1, \theta_2, \dots, \theta_{k-1}$  if its density function is of the form

$$f(\mathbf{y}; \boldsymbol{\theta}) = n! \prod_{i=1}^k \frac{\theta_i^{y_i}}{y_i!}$$

where

$$\begin{aligned} y_i &= 0, 1, 2, \dots, n ; i = 1, 2, \dots, k-1 ; y_k = n - y_1 - y_2 - \dots - y_{k-1} \\ 0 &\leq \theta_i \leq 1 ; i = 1, 2, \dots, k-1 ; \theta_k = 1 - \theta_1 - \theta_2 - \dots - \theta_{k-1} \end{aligned}$$

The log likelihood is thus

$$\ln[f(\mathbf{y}; \boldsymbol{\theta})] = \ln(n!) - \sum_{i=1}^k \ln(y_i!) + \sum_{i=1}^k y_i \ln(\theta_i)$$

The partial derivative with respect to  $\theta_j$  is thus

$$\frac{\partial \ln[f(\mathbf{y}; \boldsymbol{\theta})]}{\partial \theta_j} = \frac{y_j}{\theta_j} - \frac{y_k}{\theta_k}$$

Equating to 0 yields

$$\frac{y_j}{\hat{\theta}_j} = \frac{y_k}{\hat{\theta}_k} \quad \text{or} \quad y_j \hat{\theta}_k = y_k \hat{\theta}_j$$

Summing from  $j = 1$  to  $j = k - 1$  yields

$$(n - y_k) \hat{\theta}_k = y_k (1 - \hat{\theta}_k) \quad \text{or} \quad \hat{\theta}_k = \frac{y_k}{n}$$

and hence

$$\hat{\theta}_j = \frac{y_j \hat{\theta}_k}{y_k} = \frac{y_j}{n} \quad ; \quad j = 1, 2, \dots, k - 1$$

The second partial derivatives are given by

$$\frac{\partial^{(2)} \ln[f(\mathbf{y}; \boldsymbol{\theta})]}{\partial \theta_j^{(2)}} = -\frac{y_j}{\theta_j^2} - \frac{y_k}{\theta_k^2}$$

$$\frac{\partial^{(2)} \ln[f(\mathbf{y}; \boldsymbol{\theta})]}{\partial \theta_j \partial \theta_{j'}} = -\frac{y_k}{\theta_k^2}$$

and it follows that Fisher's Information matrix is given by

$$\mathbf{I}(\boldsymbol{\theta}) = \{i(\boldsymbol{\theta})\}_{j,j'} = \begin{cases} \frac{n}{\theta_j} + \frac{n}{\theta_k} & j = j' \\ \frac{n}{\theta_k} & j' \neq j \end{cases}$$

If we define

$$\mathbf{D}(\boldsymbol{\theta}) = \text{diag}(\theta_1, \theta_2, \dots, \theta_{k-1})$$

then we may write Fisher's information matrix as

$$\mathbf{I}(\boldsymbol{\theta}) = n \left\{ [\mathbf{D}(\boldsymbol{\theta})]^{-1} + \frac{1}{\theta_k} \mathbf{1}\mathbf{1}^T \right\}$$

Letting

$$\boldsymbol{\theta}^T = (\theta_1, \theta_2, \dots, \theta_{k-1})$$

it is easy to verify that the inverse of Fisher's information matrix, the approximate variance covariance matrix of  $\hat{\boldsymbol{\theta}}$ , is given by

$$\mathbf{V}(\boldsymbol{\theta}) = [\mathbf{I}(\boldsymbol{\theta})]^{-1} = \frac{1}{n} [\mathbf{D}(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}^T]$$

or

$$\mathbf{V}(\boldsymbol{\theta}) = \{v(\boldsymbol{\theta})\}_{j,j'} = \begin{cases} \frac{1}{n}\theta_j(1-\theta_j) & j = j' \\ -\frac{1}{n}\theta_j\theta_{j'} & j' \neq j \end{cases}$$

We now note the following result about the multivariate normal distribution: If  $\mathbf{Y}$  is MVN  $(\boldsymbol{\mu}, \mathbf{V})$  in  $p$  dimensions then the distribution of

$$(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{Y} - \boldsymbol{\mu})$$

is chi-square with  $p$  degrees of freedom.

To prove this we note that there is a matrix  $\mathbf{P}$  such that

$$\mathbf{P}^T \mathbf{P} = \mathbf{P} \mathbf{P}^T = \mathbf{I} \quad \text{and} \quad \mathbf{P} \mathbf{V} \mathbf{P}^T = \mathbf{D}$$

where  $\mathbf{D}$  is a diagonal matrix with positive diagonal elements. Define

$$\mathbf{W} = \mathbf{P}(\mathbf{Y} - \boldsymbol{\mu})$$

. Then the distribution of  $\mathbf{W}$  is multivariate normal with mean  $\mathbf{0}$  and variance covariance matrix

$$\text{var}(\mathbf{W}) = \text{var}[\mathbf{P}(\mathbf{Y} - \boldsymbol{\mu})] = \mathbf{P} \text{var}(\mathbf{Y} - \boldsymbol{\mu}) \mathbf{P}^T = \mathbf{P} \mathbf{V} \mathbf{P}^T = \mathbf{D}$$

It follows that the  $W_i$  are independent  $N(0, d_i)$  and hence

$$\sum_{i=1}^p \frac{W_i^2}{d_i} = \mathbf{W}^T \mathbf{D}^{-1} \mathbf{W}$$

is chi square with  $p$  degrees of freedom. But

$$\begin{aligned} \mathbf{W}^T \mathbf{D}^{-1} \mathbf{W} &= [\mathbf{P}(\mathbf{Y} - \boldsymbol{\mu})]^T \mathbf{D}^{-1} [\mathbf{P}(\mathbf{Y} - \boldsymbol{\mu})] \\ &= (\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{P} \mathbf{D}^{-1} \mathbf{P}^T (\mathbf{Y} - \boldsymbol{\mu}) \\ &= (\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{Y} - \boldsymbol{\mu}) \end{aligned}$$

**example:** In a multinomial problem consider testing

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ vs } [H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

Since, under  $H_0$ ,  $\hat{\boldsymbol{\theta}}$  is approximately MVN  $(\boldsymbol{\theta}_0, \mathbf{V}(\boldsymbol{\theta}_0))$  we have that

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T [\mathbf{V}(\boldsymbol{\theta}_0)]^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

is approximately chi-square with  $k - 1$  degrees of freedom.

Now note that

$$\begin{aligned} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T [\mathbf{V}(\boldsymbol{\theta}_0)]^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \left\{ [\mathbf{D}(\boldsymbol{\theta}_0)]^{-1} - \frac{1}{\theta_{k0}} \mathbf{1}\mathbf{1}^T \right\} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &= (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T [\mathbf{D}(\boldsymbol{\theta}_0)]^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &\quad + \frac{n}{\theta_{k0}} [(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)]^2 \\ &= n \sum_{i=1}^{k-1} \frac{(\hat{\theta}_i - \theta_{i0})^2}{\theta_{i0}} + n \frac{(\hat{\theta}_k - \theta_{k0})^2}{\theta_{k0}} \\ &= \sum_{i=1}^{k-1} \frac{(y_i - n\theta_{i0})^2}{n\theta_{i0}} \frac{(y_k - n\theta_{k0})^2}{n\theta_{k0}} \\ &= \sum_{i=1}^k \frac{(y_i - n\theta_{i0})^2}{n\theta_{i0}} \end{aligned}$$

Note that this expression is

$$\sum_{i=1}^k \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

If the null hypothesis is not completely specified with, say  $s$  unspecified parameters, then we simply estimate the unknown parameters by maximum likelihood, use these estimates to obtain estimated expected values and form the statistic

$$\sum_{i=1}^k \frac{(\text{observed} - \overline{\text{expected}})^2}{\overline{\text{expected}}}$$

and treat it as chi-square with  $k - s - 1$  degrees of freedom. A myriad of tests are of this form (including tests of association and goodness of fit tests). They dominated statistics for the first half of the last century. many have been replaced by likelihood ratio tests to be discussed in the next section.

### 8.7.1 Chi Square Test of Independence

Suppose that we have a random sample of  $n$  individuals who are classified according to two categories  $R$  (rows) having  $r$  values (levels) and  $C$  having  $c$  values (levels). We observe  $n_{ij}$  individuals who are classified into the cell corresponding to row  $i$  and column  $j$ . Thus the observed data are of the form:

		Column Category				
		1	2	$\cdots$	$c$	Total
Row Category	1	$n_{11}$	$n_{12}$	$\cdots$	$n_{1c}$	$n_{1+}$
	2	$n_{21}$	$n_{22}$	$\cdots$	$n_{2c}$	$n_{2+}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	$r$	$n_{r1}$	$n_{r2}$	$\cdots$	$n_{rc}$	$n_{r+}$
Total		$n_{+1}$	$n_{+2}$	$\cdots$	$n_{+c}$	$n_{++} = n$

The probabilities are given by

		Column Category				
		1	2	$\cdots$	$c$	Total
Row Category	1	$p_{11}$	$p_{12}$	$\cdots$	$p_{1c}$	$p_{1+}$
	2	$p_{21}$	$p_{22}$	$\cdots$	$p_{2c}$	$p_{2+}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	$r$	$p_{r1}$	$p_{r2}$	$\cdots$	$p_{rc}$	$p_{r+}$
Total		$p_{+1}$	$p_{+2}$	$\cdots$	$p_{+c}$	1

Thus  $p_{ij}$  is the probability that an individual is classified into row  $i$  and column  $j$ .

By the results on the multinomial these probabilities are estimated by

$$\hat{p}_{ij} = \frac{n_{ij}}{n}$$

If the classification into rows and columns is independent we have

$$p_{ij} = p_{i+}p_{+j}$$

Thus, under independence, the multinomial model is

$$f(\mathbf{n}; \mathbf{p}) = n! \prod_{i=1}^r \prod_{j=1}^c \frac{[p_{i+}p_{+j}]^{n_{ij}}}{n_{ij}!}$$

It follows that the log likelihood is

$$\ln(n!) - \sum_{i=1}^r \sum_{j=1}^c \ln(n_{ij}!) + \sum_{i=1}^r n_{i+} \ln(p_{i+}) + \sum_{j=1}^c n_{+j} \ln(p_{+j})$$

Remembering that

$$p_{r+} = 1 - \sum_{i=1}^{r-1} p_{i+} \quad \text{and} \quad p_{+c} = \sum_{j=1}^{c-1} p_{+j}$$

we see that the partial derivative with respect to  $p_{i+}$  is

$$\frac{n_{i+}}{p_{i+}} - \frac{n_{r+}}{p_{r+}}$$

Equating to 0 and solving yields

$$\hat{p}_{i+} = \frac{n_{i+}}{n}$$

Similarly

$$\hat{p}_{+j} = \frac{n_{+j}}{n}$$

Thus the estimated expected values under the hypothesis of independence are given by

$$\hat{n}_{ij} = n\hat{p}_{i+}\hat{p}_{+j} = \frac{n_{i+}n_{+j}}{n}$$

It follows that the chi square statistic is given by

$$\sum_{i=1}^r \sum_{j=1}^c \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}$$

with

$$(rc - 1) - [(r - 1) + (c - 1)] = rc - r - c + 1 = (r - 1)(c - 1)$$

degrees of freedom.

### 8.7.2 Chi Square Goodness of Fit

Suppose that we have  $y_1, y_2, \dots, y_n$  which are realized values of  $Y_1, Y_2, \dots, Y_n$  assumed to be independent each having the density function  $f(y; \theta)$  where the values of  $y$  are assumed to lie in an interval  $I$ . Usually  $I$  is  $(0, \infty)$  or  $(-\infty, +\infty)$ .

Divide  $I$  into  $k$  sub-intervals  $I_1, I_2, \dots, I_k$  defined by

$$\begin{aligned} I_1 &= \{y : y \leq c_1\} \\ I_2 &= \{y : c_1 < y \leq c_2\} \\ I_3 &= \{y : c_2 < y \leq c_3\} \\ &\vdots \\ I_{k-1} &= \{y : c_{k-2} < y \leq c_{k-1}\} \\ I_k &= \{y : y > c_{k-1}\} \end{aligned}$$

where the  $c_i$  are cut points and satisfy

$$c_1 < c_2 < \dots < c_{k-1}$$

Now define random variables  $Z_{ij}$  as follows:

$$Z_{ij} = \begin{cases} 1 & \text{if } y_i \in I_j \\ 0 & \text{otherwise} \end{cases}$$

and let  $Z_j = \sum_{i=1}^n Z_{ij}$ . Note that  $Z_j$  is the number of  $Y_i$ s that have values in  $I_j$ . It follows that the  $Z_j$  are multinomial with probabilities given by

$$p_j(\boldsymbol{\theta}) = P(Y \in I_j) = \begin{cases} \int_{I_j} f(y; \boldsymbol{\theta}) dy & \text{if } Y \text{ is continuous} \\ \sum_{I_j} f(y; \boldsymbol{\theta}) & \text{if } Y \text{ is discrete} \end{cases}$$

We estimate  $\boldsymbol{\theta}$  by maximum likelihood and then the estimated expected number in  $I_j$  is given by

$$np_j(\hat{\boldsymbol{\theta}}) \quad j = 1, 2, \dots, k$$

and the chi square statistic is given by

$$\sum_{j=1}^k \frac{[Z_j - np_j(\hat{\boldsymbol{\theta}})]^2}{np_j(\hat{\boldsymbol{\theta}})}$$

with  $k - 1 - s$  degrees of freedom, where  $s$  is the number of estimated parameters.

This test is known as the chi-square goodness of fit test and it can be used for testing the fit of any density function. It is a portmanteau test and has been replaced in the last decade by graphical tests and specialized tests (e.g. the Shapiro-Wilk test for normality).

## 8.8 PP-plots and QQ-plots

To assess whether a given distribution is consistent with an observed sample or whether two samples can be assumed to have the same distribution there are a variety of graphical methods available. The two most important are the plots known as Q-Q plots and P-P plots. Both are based on the empirical distribution function defined in the section on exploratory data analysis.

Suppose that we have data  $y_1, y_2, \dots, y_n$  assumed to be independent with the same distribution  $F$ , where

$$F(y) = P(Y \leq y)$$

Recall that the **sample distribution function** or **empirical distribution function** is a plot of the proportion of values in the data set less than or equal to  $y$  versus  $y$ . More precisely, let

$$z_i(y) = \begin{cases} 1 & \text{if } y_i \leq y \\ 0 & \text{otherwise} \end{cases}$$

Then the empirical distribution function at  $y$  is

$$F_n(y) = \frac{1}{n} \sum_{i=1}^n z_i(y)$$

Note that the  $z_i(y)$  are realized values of random variables which are Bernoulli with probability

$$p = E[Z_i(y)] = P(Y_i \leq y) = F(y)$$

so that the empirical distribution function at  $y$  is an unbiased estimator of the true distribution function i.e.

$$E[F_n(y)] = F(y)$$

Moreover

$$\text{var} [F_n(y)] = \frac{p(1-p)}{n} = \frac{F(y)[1-F(y)]}{n}$$

so that  $F_n(y)$  is a consistent estimator of  $F(y)$ . It can also be shown that it is the maximum likelihood estimator of  $F(y)$ . (If some of the values of  $Y$  are censored i.e. we can only observe that  $Y_i \leq c_i$  then a modification of  $F_n(y)$  is called the Kaplan-Meier estimate of the distribution and forms the basis of survival analysis.)

It follows that a plot of  $F_n(y)$  vs  $F(y)$  should be a straight line through the origin with slope equal to one. Such a plot is called a probability plot or **PP-plot** since both axes are probabilities.

It also follows that a plot of  $F_n^{-1}(p)$ , the sample quantiles vs  $F^{-1}(p)$  the quantiles of  $F$  should be a straight line through the origin with slope equal to one. Such a plot is called a quantile-quantile or **QQ-plot**.

Of the two plots QQ-plots are the most widely used. These plots can be conveniently made using current software but usually involve too much computation to be done by hand. They represent a very valuable technique for comparing observed data sets to theoretical models. STATA and other packages have a variety of programs based on the above simple ideas.

## 8.9 Generalized Likelihood Ratio Tests

The generalized likelihood ratio tests which reject when  $\lambda(\mathbf{y})$  is small have some useful properties which are fundamental in the analyses used in regression, logistic regression and Poisson regression.

Suppose we have data  $y_1, y_2, \dots, y_n$  realized values of  $Y_1, Y_2, \dots, Y_n$  which have joint pdf

$$f(\mathbf{y}; \boldsymbol{\theta})$$

The generalized likelihood ratio test of

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \text{ vs } H_1 : \boldsymbol{\theta} \notin \Theta_0$$

rejects if

$$\lambda(\mathbf{y}) = \frac{\max_{\boldsymbol{\theta} \in \Theta_0} f(\mathbf{y}; \boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \Theta} f(\mathbf{y}; \boldsymbol{\theta})}$$

is too small (small being determined by the requirement that the probability of a Type I error is less than or equal to the desired significance level).

In particular it can be shown that

$$-2 \log(\text{LR}) \xrightarrow{d} \chi^2(df)$$

where

$$df = \text{dimension}(\Theta) - \text{dimension}(\Theta_0)$$

That is, we can determine P-values for the hypothesis that  $\boldsymbol{\theta} \in \Theta_0$  using the chi-square distribution.

In a broad class of models, called generalized linear models, the parameter space  $\theta$  is specified by linear predictor for the  $i$ th observation,  $\eta_i$  defined by

$$\eta = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j$$

where the  $x_{ij}$  are known and  $\beta_0, \beta_1, \dots, \beta_p$  are unknown parameters to be estimated from the data. The  $\beta$ s are called regression coefficients and the  $x$ s are called covariates. We note that if a particular  $\beta$  is 0 then the corresponding covariate is not needed in the linear predictor.

The linear predictor is related to the expected value  $\mu_i$  of the  $i$ th response variable by a link function  $g$  defined so that

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j$$

Thus examining the  $\beta$ s allows us to determine which of the covariates explain the observed values of the response variable and which do not.

### 8.9.1 Regression Models

Suppose that the  $Y_i$ s are independent and normally distributed with the same variance  $\sigma^2$ , assumed known and that

$$E(Y_i) = \mu_i = \sum_{j=0}^p x_{ij}\beta_j = M_i$$

i.e. the linear predictor is exactly equal to the expected response. The covariate corresponding to  $\beta_0$  has each component equal to 1 and is called the intercept term. It is almost always included in any linear predictor.

The likelihood is given by

$$f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - M_i)^2 \right\}$$

From earlier work the estimates are given by

$$\hat{\boldsymbol{\beta}} = \mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

where

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \sum_{j=0}^p x_{ij} b_j)^2 = \frac{1}{n} \text{SSE}$$

If we write

$$\hat{y}_i = \sum_{j=0}^p x_{ij} b_j$$

then

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

The likelihood evaluated at  $\mathbf{b}$  and  $\hat{\sigma}^2$  is

$$(2\pi\hat{\sigma}^2)^{-n/2} \exp\{-n/2\} = (2\pi[\text{SSE}/n])^{-n/2} \exp\{-n/2\}$$

Suppose now that we are interested in the hypothesis that  $q$  of the regression coefficients are 0 i.e. that their corresponding covariates are not needed in the model. Without loss of generality we may write the full model as

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 = M_{if}$$

where  $\mathbf{X}_2$  contains all of the covariates of interest. Under the condition that  $\boldsymbol{\beta}_2$  is  $\mathbf{0}$  the model is

$$E(Y_i) = \sum_{j=0}^{p-q} x_{ij}\beta_j = M_{ic}$$

The likelihood under this conditional model is

$$f(\mathbf{y}; \boldsymbol{\beta}_1, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - M_{ic})^2\right\}$$

The estimates are given by

$$\hat{\boldsymbol{\beta}}_1 = \mathbf{b}_c = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{y}$$

where

$$\mathbf{b}_c = \begin{bmatrix} b_{0c} \\ b_{1c} \\ \vdots \\ b_{p-q,c} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{0c} \\ \hat{\beta}_{1c} \\ \vdots \\ \hat{\beta}_{p-q,c} \end{bmatrix}$$

and

$$\hat{\sigma}_c^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \sum_{j=0}^{p-q} x_{ij} b_{jc})^2 = \frac{1}{n} \text{SSCE}$$

The likelihood evaluated at  $\mathbf{b}_c$  and  $\hat{\sigma}_c^2$  is given by

$$(2\pi\hat{\sigma}_c^2)^{-n/2} \exp\{-n/2\} = (2\pi[\text{SSCE}/n])^{-n/2} \exp\{-n/2\}$$

It follows that the likelihood ratio statistic is

$$\lambda(\mathbf{y}) = \frac{(2\pi[\text{SSCE}/n])^{-n/2} \exp\{-n/2\}}{(2\pi[\text{SSE}/n])^{-n/2} \exp\{-n/2\}} = \left[ \frac{\text{SSE}}{\text{SSCE}} \right]^{n/2}$$

If we denote the estimates from this model as  $\mathbf{b}_c$  and the estimates from the full model as  $\mathbf{b}_f$  then the two sets of fitted values are

$$\hat{y}_i(f) = \mathbf{X}\mathbf{b}_f \quad \text{and} \quad \hat{y}_i(c) = \mathbf{X}_1\mathbf{b}_c$$

It can be shown that

$$\text{SSCE} = \text{SSE} + \sum_{i=1}^n [\hat{y}_i(c) - \hat{y}_i(f)]^2$$

so that the likelihood ratio is

$$\left\{ 1 + \frac{\sum_{i=1}^n [\hat{y}_i(c) - \hat{y}_i(f)]^2}{\text{SSE}} \right\}^{-n/2}$$

Thus we reject the hypothesis that the covariates defined by  $\mathbf{X}_2$  are not needed in the model if the ratio

$$\frac{\sum_{i=1}^n [\hat{y}_i(c) - \hat{y}_i(f)]^2}{\text{SSE}}$$

is large. It can be shown that

$$\frac{\sum_{i=1}^n [\hat{y}_i(c) - \hat{y}_i(f)]^2 / q}{\text{SSE} / [n - (p + 1)]}$$

has an  $F$  distribution with  $q$  and  $n - (p + 1)$  degrees of freedom. Thus we calculate the observed value of the  $F$  statistic and the P-value using the  $F$  distribution with  $q$  and  $n - (p + 1)$  degrees of freedom.

Note that the maximum likelihood equations for the regression model may be rewritten as

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}$$

or as

$$\sum_{i=1}^n (y_i - \hat{y}_i) x_{ij} = 0 \quad \text{for } j = 0, 1, 2, \dots, p$$

### 8.9.2 Logistic Regression Models

Let  $Y_1, Y_2, \dots, Y_n$  be independent binomial with parameters  $n_i$  and  $p_i$ . Then the joint density is given by

$$f(\mathbf{y}; \mathbf{p}) = \prod_{i=1}^n \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} = \prod_{i=1}^n \binom{n_i}{y_i} \left( \frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i)^{n_i}$$

Logistic regression models model the log odds using a linear model i.e.

$$\ln \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} = M_i$$

Then we have that

$$p_i = \frac{e^{M_i}}{1 + e^{M_i}} \quad ; \quad 1 - p_i = \frac{1}{1 + e^{M_i}}$$

Then the likelihood of  $\boldsymbol{\beta}$  is given by

$$\text{lik}(\boldsymbol{\beta}; \mathbf{y}) = \prod_{i=1}^n \binom{n_i}{y_i} e^{M_i y_i} (1 + e^{M_i})^{-n_i}$$

and hence the log likelihood is given by

$$\ln[\text{lik}(\boldsymbol{\beta}; \mathbf{y})] = \sum_{i=1}^n \binom{n_i}{y_i} + \sum_{i=1}^n M_i y_i - n_i \ln(1 + e^{M_i})$$

It follows that the derivative with respect to  $\beta_j$  is given by

$$\frac{\partial \ln[\text{lik}(\boldsymbol{\beta}; \mathbf{y})]}{\partial \beta_j} = \sum_{i=1}^n y_i x_{ij} - n_i \frac{e^{M_i}}{(1 + e^{M_i})} = \sum_{i=1}^n (y_i - n_i p_i) x_{ij}$$

for  $j = 0, 1, 2, \dots, p$  where  $x_{0j} \equiv 1$ .

It follows that the maximum likelihood equations are given by

$$\sum_{i=1}^n (y_i - n_i \hat{p}_i) x_{ij} \sum_{i=1}^n (y_i - \hat{y}_i) x_{ij} = 0$$

for  $j = 0, 1, 2, \dots, p$  where  $x_{0j} \equiv 1$ . Note that these equations are of the same general form as the equations for the linear regression model except that the  $\hat{y}_i$  terms are now non-linear and hence the equations must be solved iteratively.

Since

$$\frac{\partial p_i}{\partial \beta'_j} = \frac{e^{M_i} x_{ij'}}{(1 + e^{M_i})} - \frac{e^{2M_i} x_{ij'}}{(1 + e^{M_i})} = \left[ \frac{e^{M_i}}{(1 + e^{M_i})} \right] \left[ \frac{1}{(1 + e^{M_i})} \right] x_{ij'} = p_i(1 - p_i)x_{ij'}$$

we see that the Fisher information matrix is given by

$$\mathbf{I}(\boldsymbol{\beta}) = \{i(\boldsymbol{\beta})_{jj'}\} = - \sum_{i=1}^n x_{ij} p_i (1 - p_i) x_{ij'}$$

which we can write in matrix terms as

$$\mathbf{I}(\boldsymbol{\beta}) = -\mathbf{X}^T \mathbf{W} \mathbf{X}$$

where  $\mathbf{W}$  is a diagonal matrix with  $i$ th diagonal element equal to  $p_i(1 - p_i)$ .

Since this matrix is negative definite we have a maximum when we solve the equations

$$\sum_{i=1}^n (y_i - n_i p_i) x_{ij} = \sum_{i=1}^n (y_i - \hat{y}_i) x_{ij} = 0$$

for  $\beta$ . These equations are non linear and must be solved by iteration. Contrast this with equations for regression models where the equations are linear and can be solved exactly.

The approximate covariance matrix of  $\hat{\beta}$  is thus

$$\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X}$$

where  $\widehat{\mathbf{W}}$  is obtained by replacing  $p_i$  by  $\hat{p}_i$  defined by

$$\hat{p}_i = p_i(\hat{\beta})$$

### 8.9.3 Log Linear Models

Consider a classification of  $n$  individuals into  $k$  categories. If  $p_i$  is the probability that an individual is classified into category  $i$  then the probability of the observed data is

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n)$$

where  $y_i$  is the number (count) of individuals in category  $i$ . This probability is given by

$$\frac{n!}{y_1!y_2!\cdots y_k!} p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}$$

where  $y_1 + y_2 + \cdots + y_k = n$  and  $p_1 + p_2 + \cdots + p_k = 1$ .

This probability model is called the **multinomial** distribution with parameters  $n$  and  $p_1, p_2, \dots, p_k$ . The binomial is a special case when  $k = 2, p_1 = p, p_2 = 1 - p, y_1 = y$  and  $y_2 = n - y$ . We may write the multinomial distribution compactly as

$$n! \prod_{i=1}^k \frac{p_i^{y_i}}{y_i!}$$

where  $\prod_{i=1}^k$  stands for the product of the terms from  $i = 1$  to  $k$ . The type of model described by the multinomial model is called **multinomial sampling**. It can be shown that the expected value of  $Y_i$  is  $np_i$ .

Log linear models specify  $\lambda_i = np_i$  as  $\log(\lambda_i) = \mathbf{M}_i$  where  $\mathbf{M}_i$  is a linear combination of covariates. We may rewrite the multinomial distribution in terms of the  $\lambda_i$  as follows

$$n! \prod_{i=1}^k \frac{p_i^{y_i}}{y_i!} = \frac{n!}{\prod_{i=1}^k y_i!} \prod_{i=1}^k \frac{(np_i)^{y_i}}{n^{y_i}} = \frac{n!}{n^n \prod_{i=1}^k y_i!} \prod_{i=1}^k \lambda_i^{y_i}$$

Thus the likelihood of the model  $\mathbf{M}$  is

$$\text{lik}(\mathbf{M}; \mathbf{y}) = \frac{n!}{n^n \prod_{i=1}^k y_i!} \prod_{i=1}^k [\exp(y_i \mathbf{M}_i)] = \left\{ \frac{n!}{n^n \prod_{i=1}^k y_i!} \right\} \left\{ \exp \left( \sum_{i=1}^k y_i \mathbf{M}_i \right) \right\}$$

Using maximum likelihood to estimate the parameters in  $\mathbf{M}$  requires maximization of the second term in the above expression since the first term does not depend on  $\mathbf{M}$ . The resulting equations are non linear and must be solved by an iterative process.

If we now consider  $k$  independent Poisson random variables  $Y_1, Y_2, \dots, Y_k$  then

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_k = y_k) = \prod_{i=1}^k \lambda_i^{y_i} \exp(-\lambda_i) / y_i!$$

and we have a **Poisson sampling** setup. Recall that  $E(Y_i) = \lambda_i$  for the Poisson distribution. If we use a log linear model for  $\lambda_i$ , that is we model

$$\log(\lambda_i) = \log(E(Y_i)) = \mathbf{M}_i$$

where  $\mathbf{M}_i$  is a linear combination of covariates then the likelihood for Poisson sampling is given by

$$\text{lik}(\mathbf{M}; \mathbf{y}) = \frac{1}{\prod_{i=1}^k y_i!} \prod_{i=1}^k \exp(y_i \mathbf{M}_i) \exp(-\lambda_i) = \left\{ \frac{\exp(\sum_{i=1}^k \lambda_i)}{\prod_{i=1}^k y_i!} \right\} \left\{ \exp\left(\sum_{i=1}^k y_i \mathbf{M}_i\right) \right\}$$

Maximum likelihood applied to this model chooses estimates of the parameters to maximize the second term in the above expression since the first term does not involve the parameters of the model **provided** that  $\sum_{i=1}^k \lambda_i = n$ .

**Conclusion:**

If we use the Poisson sampling model and maximize the likelihood under the condition that  $\sum_{i=1}^k \lambda_i = n$  we will obtain the same estimates, standard errors, etc. as if we had used the multinomial sampling model. The technical reason for this equivalence is that estimates and standard errors depend only on the expected value of the derivatives of the log of the likelihood function with respect to the parameters. Since these expected values are the same for the two likelihoods the assertion follows.

It follows that any program which maximizes Poisson likelihoods can be used for multinomial problems. This fact was recognized in the early 1960's but was not of much use until appropriate software was developed in the 1970's and 1980's.

The same results hold when we have **product multinomial sampling** i.e. when

group 1 is multinomial  $(n_1, p_{11}, p_{12}, \dots, p_{1k})$

group 2 is multinomial  $(n_2, p_{21}, p_{22}, \dots, p_{2k})$

etc. provided the log linear model using Poisson sampling fixes the group totals i.e.  $\sum_{j=1}^k \lambda_{1j} = n_1, \sum_{j=1}^k \lambda_{2j} = n_2$ , etc. In fitting these models a group term treated as a factor must be included in the model.

**Summary:** Any cross-classified data set involving counts may be modelled by log linear models and the Poisson distribution using a log link, provided that any restrictions implied by the experimental setup are included as terms in the fitting process. This implies that any logistic regression problem can be considered as a log linear model provided we include in the fitting process a term for (success, failure), (exposed, non-exposed), etc.

The resulting equations can be shown to be of the form

$$\sum_{i=1}^n (y_i - \hat{y}_i) x_{ij} \text{ for } j = 0, 1, 2, \dots, p$$

where the fitted values, as in logistic regression are non linear functions of the estimated regression coefficients so that the equations are non linear and must be solved by iteration.