

# MT427 Notebook 2

prepared by Professor Jenny Baglivo

---

© Copyright 2009 by Jenny A. Baglivo. All Rights Reserved.

<b>2</b>	<b>MT427 Notebook 2</b>	<b>3</b>
2.1	Definitions and Examples . . . . .	3
2.2	Performance Measures for Estimators . . . . .	4
2.2.1	Measuring Accuracy: Bias . . . . .	5
2.2.2	Measuring Precision: Mean Squared Error (MSE) . . . . .	7
2.2.3	Comparing Estimators: Efficiency . . . . .	8
2.2.4	Another Performance Measure: Consistency . . . . .	9
2.3	Interval Estimation . . . . .	11
2.3.1	Error Probability, Confidence Coefficient, Confidence Interval . . . . .	11
2.3.2	Confidence Interval Procedures for Normal Distributions . . . . .	11
2.4	Method of Moments (MOM) Estimation . . . . .	17
2.4.1	$K$ th Moments and $K$ th Sample Moments . . . . .	17
2.4.2	MOM Estimation Method . . . . .	17
2.4.3	MOM Estimation Method for Multiple Parameters . . . . .	20
2.5	Maximum Likelihood (ML) Estimation . . . . .	22
2.5.1	Likelihood and Log-Likelihood Functions . . . . .	22
2.5.2	ML Estimation Method . . . . .	22
2.5.3	Cramer-Rao Lower Bound . . . . .	27
2.5.4	Efficient Estimators; Minimum Variance Unbiased Estimators . . . . .	27
2.5.5	Large Sample Theory: Fisher's Theorem . . . . .	30
2.5.6	Approximate Confidence Interval Procedures; Special Cases . . . . .	31
2.5.7	Multinomial Experiments . . . . .	34
2.5.8	ML Estimation Method for Multiple Parameters . . . . .	38



## 2 MT427 Notebook 2

There are two core areas of mathematical statistics: estimation theory and hypothesis testing theory. This notebook is concerned with the first core area, estimation theory. The notes include material from Chapter 8 of the Rice textbook.

### 2.1 Definitions and Examples

1. *Random Sample*: A random sample of size  $n$  from the  $X$  distribution is a list,

$$X_1, X_2, \dots, X_n,$$

of  $n$  mutually independent random variables, each with the same distribution as  $X$ .

2. *Statistic*: A *statistic* is a function of a random sample (or random samples).
3. *Sampling Distribution*: The probability distribution of a statistic is known as its *sampling distribution*. Sampling distributions are often difficult to find.
4. *Point Estimator*: A *point estimator* (or *estimator*) is a statistic used to estimate an unknown parameter of a probability distribution.
5. *Estimate*: An *estimate* is the value of an estimator for a given set of data.
6. *Test Statistic*: A *test statistic* is a statistic used to test an assertion.

**Example: Normal distribution.** Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .

1. *Sample Mean*: The sample mean,  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ , is an estimator of  $\mu$ . This statistic has a normal distribution with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ .
2. *Sample Variance*: The sample variance,

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

is an estimator of  $\sigma^2$ . The distribution of this statistic is related to the chi-square. Specifically, the statistic  $V = \frac{(n-1)}{\sigma^2} S^2$  has a chi-square distribution with  $(n-1)$  *df*.

If the numbers 90.8, 98.0, 113.0, 134.7, 80.5, 97.6, 117.6, 119.9 are observed, for example, then

1.  $\bar{x} = 106.513$  is an estimate of  $\mu$ , and
2.  $s^2 = 316.316$  is an estimate of  $\sigma^2$

based on these observations.

$\bar{X}$  and  $S^2$  can also be used to test assertions about the parameters of a normal distribution. For example,  $\bar{X}$  can be used to test the assertion that

“The mean of the distribution is 100.”

**Example: Multinomial experiments.** Consider a multinomial experiment with  $k$  outcomes whose probabilities are  $p_i$ , for  $i = 1, 2, \dots, k$ .

The results of  $n$  independent trials of a multinomial experiment are usually presented to us in summary form: Let  $X_i$  be the number of occurrences of the  $i^{\text{th}}$  outcome in  $n$  independent trials of the experiment, for  $i = 1, 2, \dots, k$ . Then

1. *Sample Proportions:* The  $i^{\text{th}}$  *sample proportion*,

$$\hat{p}_i = \frac{X_i}{n},$$

(read “ $p_i$ -hat”) is an estimator of  $p_i$ , for  $i = 1, 2, \dots, k$ . The distribution of this statistic is related to the binomial distribution. Specifically,  $X_i = n\hat{p}_i$  has a binomial distribution with parameters  $n$  and  $p_i$ .

2. *Pearson’s statistic:* Pearson’s statistic,

$$\mathbf{X}^2 = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i},$$

can be used to test the assertion that the true probability list is  $(p_1, p_2, \dots, p_k)$ . The exact probability distribution of  $\mathbf{X}^2$  is hard to calculate. But, if  $n$  is large, the distribution of Pearson’s statistic is approximately chi-square with  $(k - 1)$  degrees of freedom.

If  $(x_1, x_2, x_3, x_4, x_5, x_6) = (5, 7, 8, 5, 11, 4)$  is observed, for example, then

$$\hat{p}_1 = 0.125, \hat{p}_2 = 0.175, \hat{p}_3 = 0.20, \hat{p}_4 = 0.125, \hat{p}_5 = 0.275, \hat{p}_6 = 0.10$$

are point estimates of the model proportions based on the observed list of counts. (It is common practice to use the same notation for estimators and estimates of proportions.)

If, in addition, we were interested in testing the assertion that all outcomes are equally likely, (that is, the assertion that  $p_i = \frac{1}{6}$  for each  $i$ ), for example, then  $\mathbf{x}^2_{\text{obs}} = 5.0$  is the value of the test statistic for the observed list of counts.

## 2.2 Performance Measures for Estimators

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with parameter  $\theta$ . The notation

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$$

(read “theta-hat”) is used to denote an estimator of  $\theta$ .

Estimators should be both *accurate* (measured using the center of the sampling distribution) and *precise* (measured using both the center and the spread of the sampling distribution).

### 2.2.1 Measuring Accuracy: Bias

1. *Bias*: The *bias* of the estimator  $\hat{\theta}$  is defined as follows:

$$BIAS(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

An accurate estimator will have little or no bias.

2. *Unbiased/Biased Estimators*: If  $E(\hat{\theta}) = \theta$ , then  $\hat{\theta}$  is said to be an *unbiased estimator* of the parameter  $\theta$ ; otherwise,  $\hat{\theta}$  is said to be a *biased estimator* of  $\theta$ .
3. *Asymptotically Unbiased Estimator*: If  $\hat{\theta}$  is a biased estimator satisfying

$$\lim_{n \rightarrow \infty} E(\hat{\theta}) = \theta, \quad \text{where } n \text{ is the sample size,}$$

then  $\hat{\theta}$  is said to be an *asymptotically unbiased estimator* of  $\theta$ .

**Example: Normal distribution.** Let  $\bar{X}$  be the sample mean,  $S^2$  be the sample variance, and  $S$  be the sample standard deviation of a random sample of size  $n$  from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .

1. *Sample Mean*:  $\bar{X}$  is an unbiased estimator of  $\mu$  since  $E(\bar{X}) = \mu$ .
2. *Sample Variance*:  $S^2$  is an unbiased estimator of  $\sigma^2$  since  $E(S^2) = \sigma^2$ .
3. *Sample Standard Deviation*:  $S$  is a biased estimator of  $\sigma$ , with expectation

$$E(S) = \sigma \sqrt{\frac{2}{n-1}} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \neq \sigma.$$

Since  $E(S) \rightarrow \sigma$  as  $n \rightarrow \infty$ ,  $S$  is an asymptotically unbiased estimator of  $\sigma$ .

*Exercise.* Another commonly used estimator of  $\sigma^2$  is

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

Demonstrate that  $\widehat{\sigma^2}$  is a biased estimator of  $\sigma^2$ , but that it is asymptotically unbiased.

*Exercise.* Let  $\bar{X}$  be the sample mean of a random sample of size  $n$  from a uniform distribution on the interval  $(0, b)$ , and let  $\hat{b} = 2\bar{X}$  be an estimator of the upper endpoint  $b$ .

- (a) Demonstrate that  $\hat{b}$  is an unbiased estimator of  $b$ , and find the variance of the estimator.
- (b) Use  $\hat{b}$  to estimate  $b$  using the following observations: 3.64, 3.76, 2.53, 7.06, 5.71.

### 2.2.2 Measuring Precision: Mean Squared Error (MSE)

The *mean squared error* (MSE) of an estimator is the expected value of the square of the difference between the estimator and the true parameter:

$$MSE(\hat{\theta}) = E((\hat{\theta} - \theta)^2).$$

A precise estimator will have a small mean squared error.

**Theorem (Mean Squared Error).** Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with parameter  $\theta$ , and let  $\hat{\theta}$  be an estimator of  $\theta$ . Then

1.  $MSE(\hat{\theta}) = Var(\hat{\theta}) + (BIAS(\hat{\theta}))^2$ .
2. If  $\hat{\theta}$  is an unbiased estimator of  $\theta$ , then  $MSE(\hat{\theta}) = Var(\hat{\theta})$ .

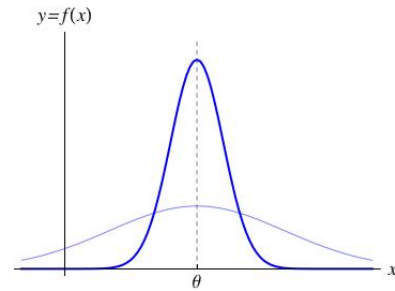
*Exercise.* Use the definitions of mean squared error and variance, and properties of expectation, to prove the theorem above.

### 2.2.3 Comparing Estimators: Efficiency

1. *Unbiased Estimators:* Let  $\hat{\theta}_1$  and  $\hat{\theta}_2$  be unbiased estimators of  $\theta$ , each based on a random sample of size  $n$  from the  $X$  distribution.

$\hat{\theta}_1$  is said to be *more efficient* than  $\hat{\theta}_2$  if

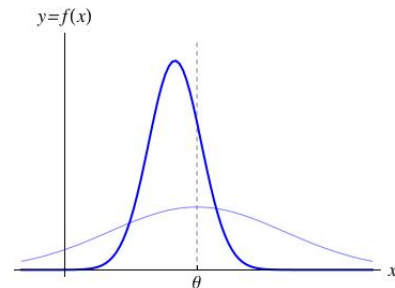
$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2).$$



2. *General Estimators:* More generally, let  $\hat{\theta}_1$  and  $\hat{\theta}_2$  be two estimators of  $\theta$ , each based on a random sample of size  $n$  from the  $X$  distribution.

$\hat{\theta}_1$  is said to be *more efficient* than  $\hat{\theta}_2$  if

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



Given two unbiased estimators, we would prefer to use the one with the smaller variance (as illustrated in the *top plot*); otherwise, we would prefer to use the one whose average squared deviation from the true parameter is smaller (as illustrated in the *bottom plot*).

*Exercise.* Let  $X_1, X_2, X_3, X_4$  be a random sample of size 4 from a distribution with mean  $\mu$  and standard deviation  $\sigma$ . The following two estimators are unbiased estimators of  $\mu$ :

$$\hat{\mu}_1 = \frac{1}{2}X_1 + \frac{1}{6}X_2 + \frac{1}{6}X_3 + \frac{1}{6}X_4 \quad \text{and} \quad \hat{\mu}_2 = \frac{1}{4}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3 + \frac{1}{4}X_4.$$

Which is the more efficient estimator (or are they equally efficient)?

**A natural question to ask** is whether a most efficient estimator exists. If we consider unbiased estimators only, we have the following definition.

**MVUE:** The unbiased estimator  $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$  is said to be a *minimum variance unbiased estimator (MVUE)* of  $\theta$  if

$$\text{Var}(\hat{\theta}) \leq \text{Var}(\hat{\theta}^*) \text{ for all unbiased estimators } \hat{\theta}^* = \hat{\theta}^*(X_1, X_2, \dots, X_n).$$

We will consider the question of finding an MVUE (if one exists) in Section 2.5.4 (page 28) of these notes.

#### 2.2.4 Another Performance Measure: Consistency

The estimator  $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$  is said to be a *consistent estimator* of  $\theta$  if, for every positive number  $\epsilon$ ,

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| \geq \epsilon) = 0, \text{ where } n \text{ is the sample size.}$$

(A consistent estimator is unlikely to be far from the true parameter when  $n$  is large enough.)

*Note* that consistent estimators are not necessarily unbiased, but they are generally asymptotically unbiased. If an estimator is unbiased, however, a quick check for consistency is given in the following theorem.

**Theorem (Consistency).** If  $\hat{\theta}$  is an unbiased estimator of  $\theta$  satisfying

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}) = 0, \text{ where } n \text{ is the sample size,}$$

then  $\hat{\theta}$  is a consistent estimator of  $\theta$ .

*Exercise.* Let  $X$  be the number of successes in  $n$  independent trials of a Bernoulli experiment with success probability  $p$ , and let  $\hat{p} = \frac{X}{n}$  be the sample proportion.

Demonstrate that  $\hat{p}$  is a consistent estimator of  $p$ .

*Exercise.* Let  $\bar{X}$  be the sample mean and  $S^2$  be the sample variance of a random sample of size  $n$  from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .

- (a) Demonstrate that  $\bar{X}$  is a consistent estimator of  $\mu$ .
- (b) Demonstrate that  $S^2$  is a consistent estimator of  $\sigma^2$ .

## 2.3 Interval Estimation

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with parameter  $\theta$ . The goal in interval estimation is to find two statistics

$$L = L(X_1, X_2, \dots, X_n) \quad \text{and} \quad U = U(X_1, X_2, \dots, X_n)$$

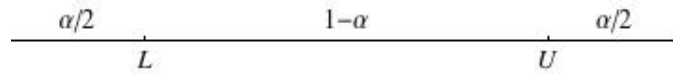
with the property that  $\theta$  lies in the interval  $[L, U]$  with high probability.

### 2.3.1 Error Probability, Confidence Coefficient, Confidence Interval

1. *Error Probability:* It is customary to let  $\alpha$  (called the *error probability*) equal the probability that  $\theta$  is not in the interval, and to find statistics  $L$  and  $U$  satisfying

(a)  $P(\theta < L) = P(\theta > U) = \frac{\alpha}{2}$  and

(b)  $P(L \leq \theta \leq U) = 1 - \alpha$ .



2. *Confidence Coefficient:* The probability  $(1 - \alpha)$  is called the *confidence coefficient*.
3. *Confidence Interval:* The interval  $[L, U]$  is called a  $100(1 - \alpha)\%$  *confidence interval* for  $\theta$ .

A  $100(1 - \alpha)\%$  confidence interval is a random quantity. In applications, we substitute sample values for the lower and upper endpoints and report the interval  $[\ell, u]$ .

The reported interval is not guaranteed to contain the true parameter, but  $100(1 - \alpha)\%$  of reported intervals will contain the true parameter.

**Example: Normal distribution with known  $\sigma$ .** Let  $X$  be a normal random variable with mean  $\mu$  and known standard deviation  $\sigma$ , and let  $\bar{X}$  be the sample mean of a random sample of size  $n$  from the  $X$  distribution.

The following steps can be used to find expressions for the lower limit  $L$  and the upper limit  $U$  of a  $100(1 - \alpha)\%$  confidence interval for the mean  $\mu$ :

1. First note that  $Z = (\bar{X} - \mu)/\sqrt{\sigma^2/n}$  is a standard normal random variable.
2. For convenience, let  $p = \alpha/2$ . Further, let  $z_p$  and  $z_{1-p}$  be the  $p^{\text{th}}$  and  $(1 - p)^{\text{th}}$  quantiles of the standard normal distribution. Then

$$P(z_p \leq Z \leq z_{1-p}) = 1 - 2p.$$

3. The following statements are equivalent:

$$\begin{aligned} z_p \leq Z \leq z_{1-p} &\iff z_p \leq \frac{(\bar{X} - \mu)}{\sqrt{\sigma^2/n}} \leq z_{1-p} \\ &\iff z_p \sqrt{\frac{\sigma^2}{n}} \leq (\bar{X} - \mu) \leq z_{1-p} \sqrt{\frac{\sigma^2}{n}} \\ &\iff \bar{X} - z_{1-p} \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} - z_p \sqrt{\frac{\sigma^2}{n}} \end{aligned}$$

Thus,  $P\left(\bar{X} - z_{1-p} \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} - z_p \sqrt{\frac{\sigma^2}{n}}\right) = 1 - 2p$ .

4. Since  $z_p = -z_{1-p}$  and  $p = \alpha/2$ , the expressions we want are

$$L = \bar{X} - z_{1-\alpha/2} \sqrt{\frac{\sigma^2}{n}} \quad \text{and} \quad U = \bar{X} - z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}} = \bar{X} + z_{1-\alpha/2} \sqrt{\frac{\sigma^2}{n}}.$$

*Example.* I used the computer to generate 50 random samples of size 16 from the normal distribution with mean 0 and standard deviation 10. For each sample, I computed 90% confidence intervals for  $\mu$  using the expressions for  $L$  and  $U$  from above.

The following plot shows the observed intervals as vertical line segments.



46 intervals contain  $\mu = 0$  (solid lines), while 4 do not (dashed lines).

### 2.3.2 Confidence Interval Procedures for Normal Distributions

The following tables give confidence interval (CI) procedures for the parameters of the normal distribution. In each case,  $\bar{X}$  is the sample mean and  $S^2$  is the sample variance of a random sample of size  $n$  from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .

1. 100(1 -  $\alpha$ )% CI for  $\mu$ , when  $\sigma$  is known:

$$\bar{X} \pm z(\alpha/2) \sqrt{\frac{\sigma^2}{n}},$$

where  $z(\alpha/2)$  is the 100(1 -  $\alpha/2$ )% point of the standard normal distribution.

2. 100(1 -  $\alpha$ )% CI for  $\mu$ , when  $\sigma$  is estimated:

$$\bar{X} \pm t_{n-1}(\alpha/2) \sqrt{\frac{S^2}{n}},$$

where  $t_{n-1}(\alpha/2)$  is the 100(1 -  $\alpha/2$ )% point of the Student t distribution with  $n - 1$  df.

3. 100(1 -  $\alpha$ )% CI for  $\sigma^2$ , when  $\mu$  is known:

$$\left[ \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_n^2(\alpha/2)}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_n^2(1 - \alpha/2)} \right],$$

where  $\chi_n^2(p)$  is the 100(1 -  $p$ )% point of the chi-square distribution with  $n$  df.

4. 100(1 -  $\alpha$ )% CI for  $\sigma^2$ , when  $\mu$  is estimated:

$$\left[ \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\chi_{n-1}^2(\alpha/2)}, \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\chi_{n-1}^2(1 - \alpha/2)} \right],$$

where  $\chi_{n-1}^2(p)$  is the 100(1 -  $p$ )% point of the chi-square distribution with  $(n - 1)$  df.

5. 100(1 -  $\alpha$ )% CI for  $\sigma$ :

If  $[L, U]$  is a 100(1 -  $\alpha$ )% CI for  $\sigma^2$ , then  $[\sqrt{L}, \sqrt{U}]$  is a 100(1 -  $\alpha$ )% CI for  $\sigma$ .

Notes:

1. *Upper Tail Notation:* It is common practice to use upper tail notation for the quantiles used in confidence interval procedures. Thus, we use  $x(p)$  to denote the quantile  $x_{1-p}$ .
2. *(Estimated) Standard Error of the Mean:* The standard deviation of  $\bar{X}$ ,

$$SD(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

is often called the *standard error* of the mean, and the approximation  $\sqrt{\frac{S^2}{n}} = \frac{S}{\sqrt{n}}$  is often called the *estimated standard error* of the mean.

*Exercise.* Demonstrate that the confidence interval procedure for estimating  $\sigma^2$  when  $\mu$  is known is correct.

*Exercise.* Assume that the following numbers are the values of a random sample from a normal distribution with known standard deviation 10:

97.864	103.689	101.945	89.416	104.230	90.190	108.890	102.993
85.746	104.759	89.685	117.986	83.084	96.150	96.324	

Sample summaries:  $n = 15$ ,  $\bar{x} = 98.1967$

Construct 80% and 90% confidence intervals for  $\mu$ .

*Exercise.* Assume that the following numbers are the values of a random sample from a normal distribution:

2.432 4.618 0.730 2.128 3.347 4.651 2.545 3.393  
3.250 5.858 -1.871 2.761 2.738 1.910 5.062

Sample summaries:  $n = 15$ ,  $\bar{x} = 2.90347$ ,  $s^2 = 3.53177$

Construct 90% confidence intervals for  $\sigma^2$  and  $\sigma$ .

*Exercise.* As part of a study on normal body temperatures, 30 temperatures (in degrees Fahrenheit) were recorded:

98.2 98.2 98.4 98.4 98.6 98.7 98.7 98.7 98.8 98.8  
98.8 98.9 99.0 99.1 99.9 97.1 97.4 97.4 97.6 97.6  
97.8 97.8 98.0 98.2 98.6 98.6 98.8 99.0 99.1 99.4

Sample summaries:  $n = 30$ ,  $\bar{x} = 98.4533$ ,  $s^2 = 0.417747$

Assume these data are the values of a random sample from a normal distribution.

Construct 95% confidence intervals for  $\mu$  and  $\sigma^2$ .

## 2.4 Method of Moments (MOM) Estimation

The method of moments is a general method for estimating one or more unknown parameters. The method of moments was introduced by Karl Pearson in the 1880's. In general, MOM estimators are consistent, but they are not necessarily unbiased.

### 2.4.1 $K$ th Moments and $K$ th Sample Moments

1. The  $k^{\text{th}}$  *moment* of the  $X$  distribution is

$$\mu_k = E(X^k) \quad \text{for } k = 1, 2, 3, \dots$$

2. The  $k^{\text{th}}$  *sample moment* is the statistic

$$\widehat{\mu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k \quad \text{for } k = 1, 2, 3, \dots,$$

where  $X_1, X_2, \dots, X_n$  is a random sample of size  $n$  from the  $X$  distribution.

*Note* that the  $k^{\text{th}}$  sample moment is an unbiased and consistent estimator of  $\mu_k$  for each  $k$ .

### 2.4.2 MOM Estimation Method

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with parameter  $\theta$ , and suppose that  $\mu_k = E(X^k)$  is a function of  $\theta$  for some  $k$ .

Then, a *method of moments estimator* of  $\theta$  is obtained using the following procedure:

$$\text{Solve } \mu_k = \widehat{\mu}_k \text{ for the parameter } \theta.$$

*For example,* let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the uniform distribution on the interval  $(0, b)$ . Since  $E(X) = b/2$ , a method of moments estimator of the upper endpoint  $b$  can be obtained as follows:

$$\mu_1 = \widehat{\mu}_1 \implies E(X) = \bar{X} \implies \frac{b}{2} = \bar{X} \implies b = 2\bar{X}.$$

Thus,  $\widehat{b} = 2\bar{X}$  is a MOM estimator of  $b$ .

*Exercise.* Let  $X_1, X_2, \dots, X_n$  be a random sample from a uniform distribution on the interval  $(-b, b)$ , where  $b > 0$  is unknown.

- (a) Find a method of moments estimator of  $b$ .
- (b) Suppose that the numbers  $-6.2, -4.9, -0.4, 5.8, 6.6, 6.9$  are observed. Use the formula derived in part (a) to estimate the upper endpoint  $b$ .

*Exercise.* Let  $X$  be a continuous random variable with PDF as follows:

$$f(x) = \frac{1}{2}(1 + \alpha x) \quad \text{when } -1 < x < 1 \text{ and } 0 \text{ otherwise,}$$

where  $\alpha \in (-1, 1)$  is an unknown parameter. Find a method of moments estimator of  $\alpha$  based on a random sample of size  $n$  from the  $X$  distribution.

### 2.4.3 MOM Estimation Method for Multiple Parameters

The method of moments procedure can be generalized to any number of unknown parameters.

For example, if the  $X$  distribution has two unknown parameters (say  $\theta_1$  and  $\theta_2$ ), then MOM estimators are obtained using the procedure:

Solve  $\mu_{k_1} = \widehat{\mu}_{k_1}$  and  $\mu_{k_2} = \widehat{\mu}_{k_2}$  simultaneously for  $\theta_1$  and  $\theta_2$ .

**Example: Normal distribution.** Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , where both parameters are unknown. To find method of moments estimators of  $\mu$  and  $\sigma^2$ , we need to solve two equations in the two unknowns.

Since  $E(X) = \mu$  and  $E(X^2) = \text{Var}(X) + (E(X))^2 = \sigma^2 + \mu^2$ , we can use the equations

$$\mu_1 = \widehat{\mu}_1 \quad \text{and} \quad \mu_2 = \widehat{\mu}_2$$

to find the MOM estimators. Now (complete the derivation),

**Example: Gamma distribution.** Let  $X_1, X_2, \dots, X_n$  be a random sample from a gamma distribution with parameters  $\alpha$  and  $\beta$ , where both parameters are unknown. To find method of moments estimators of  $\alpha$  and  $\beta$  we need to solve two equations in the two unknowns.

Since  $E(X) = \alpha\beta$  and  $E(X^2) = \text{Var}(X) + (E(X))^2 = \alpha\beta^2 + (\alpha\beta)^2$ , we can use the equations

$$\mu_1 = \widehat{\mu}_1 \quad \text{and} \quad \mu_2 = \widehat{\mu}_2$$

to find the MOM estimators. Now (complete the derivation),

## 2.5 Maximum Likelihood (ML) Estimation

The method of maximum likelihood is a general method for estimating one or more unknown parameters. Maximum likelihood estimation was introduced by R.A. Fisher in the 1920's. In general, ML estimators are consistent, but they are not necessarily unbiased.

### 2.5.1 Likelihood and Log-Likelihood Functions

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with parameter  $\theta$ .

1. The *likelihood function* is the joint PDF of the random sample thought of as a function of the parameter  $\theta$ , with the random  $X_i$ 's left unevaluated:

$$Lik(\theta) = \begin{cases} \prod_{i=1}^n p(X_i) & \text{when } X \text{ is discrete} \\ \prod_{i=1}^n f(X_i) & \text{when } X \text{ is continuous} \end{cases}$$

2. The *log-likelihood function* is the natural logarithm of the likelihood function:

$$\ell(\theta) = \log(Lik(\theta))$$

*Note* that it is common practice to use  $\log()$  to denote the natural logarithm, and that this convention is used in the *Mathematica* system as well.

*For example*, if  $X$  is an exponential random variable with parameter  $\lambda$ , then

$$Lik(\lambda) = \prod_{i=1}^n f(X_i) = \prod_{i=1}^n \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_i X_i},$$

and  $\ell(\lambda) = \log(\lambda^n e^{-\lambda \sum_i X_i}) = n \log(\lambda) - \lambda \sum_i X_i$ .

### 2.5.2 ML Estimation Method

The *maximum likelihood estimator* (or *ML estimator*) of  $\theta$  is the value that maximizes either the likelihood function ( $Lik(\theta)$ ) or the log-likelihood function ( $\ell(\theta)$ ).

In general, ML estimators are obtained using methods from calculus.

***Example: Bernoulli distribution.*** Let  $X_1, X_2, \dots, X_n$  be a random sample from a Bernoulli distribution with success probability  $p$  and let  $Y = \sum_i X_i$  be the sample sum. Then the ML estimator of  $p$  is the sample proportion  $\hat{p} = Y/n$ .

To see this,

1. Since there are  $Y$  successes (each occurring with probability  $p$ ) and  $n - Y$  failures (each occurring with probability  $1 - p$ ), the likelihood function is

$$Lik(p) = \prod_{i=1}^n p(X_i) = p^Y (1 - p)^{n-Y}$$

and the log-likelihood function is  $\ell(p) = Y \log(p) + (n - Y) \log(1 - p)$ .

2. The derivative of the log-likelihood function is

$$\ell'(p) = \frac{Y}{p} - \frac{(n - Y)}{(1 - p)}.$$

3. There are three cases to consider:

- (a) If  $Y = 0$ , then  $\ell'(p) < 0$  for all  $p$ . This implies that the log-likelihood function is decreasing, so is maximized when  $\hat{p} = 0$ .
- (b) If  $Y = n$ , then  $\ell'(p) > 0$  for all  $p$ . This implies that the log-likelihood function is increasing, so is maximized when  $\hat{p} = 1$ .
- (c) If  $0 < Y < n$ , then

$$\ell'(p) = 0 \Rightarrow \frac{Y}{p} = \frac{(n - Y)}{(1 - p)} \Rightarrow p = \frac{Y}{n} \Rightarrow \hat{p} = \frac{Y}{n} \text{ is the ML estimator.}$$

To demonstrate that we have a maximum, we compute

$$\ell''(p) = -\frac{Y}{p^2} - \frac{(n - Y)}{(1 - p)^2}.$$

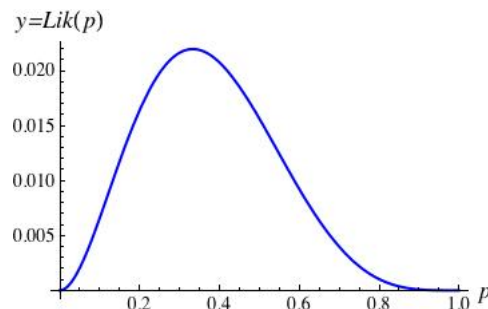
Then  $\ell''(\hat{p}) < 0$  since it is the sum of two negative numbers.

in all cases, the estimator is  $\hat{p} = \frac{Y}{n}$ .

*To illustrate* maximum likelihood estimation in the Bernoulli case, assume that 2 successes are observed in 6 trials. The graph shows the observed likelihood function,

$$y = Lik(p) = p^2(1 - p)^4,$$

for  $p \in (0, 1)$ . This function is maximized at  $1/3$ . Thus, the ML estimate is  $1/3$ .



**Example: Slope-parameter model.** Let  $X$  be a continuous random variable with PDF as follows:

$$f(x) = \frac{1}{2}(1 + \alpha x) \quad \text{when } -1 < x < 1 \text{ and } 0 \text{ otherwise,}$$

where  $\alpha \in (-1, 1)$  is an unknown parameter.

Given a random sample of size  $n$  from the  $X$  distribution, the likelihood function is

$$Lik(\alpha) = \prod_{i=1}^n f(X_i) = \prod_{i=1}^n \frac{(1 + \alpha X_i)}{2} = \frac{1}{2^n} \prod_{i=1}^n (1 + \alpha X_i).$$

Since the likelihood and log-likelihood functions are not easy to analyze by hand, we will use the computer to find the ML estimate given specific numbers.

For example, assume that the following data are the values of a random sample from the  $X$  distribution:

0.455   -0.995   -0.101   0.568   0.298   -0.485   0.081   0.906  
 -0.971   -0.337   0.278   0.714   0.175   0.414   -0.345   0.888

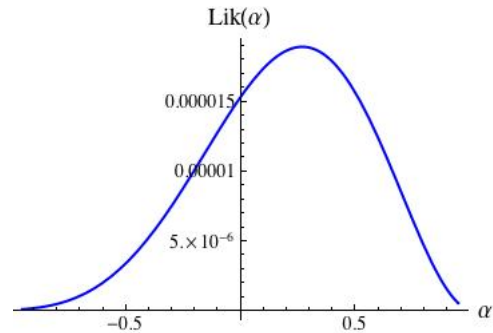
Sample summaries:  $n = 16$ ,  $\bar{x} = 0.0964375$

The graph shows the function

$$y = Lik(\alpha) = \frac{1}{2^{16}} \prod_{i=1}^{16} (1 + \alpha x_i)$$

for  $\alpha \in (-1, 1)$ , and the  $x_i$ 's from above.

This function is maximized at 0.270822. Thus, the ML estimate of alpha for these data is 0.270822.



Using the work we did on page 20,

a MOM estimate of  $\alpha$  for the data above is \_\_\_\_\_

**Exercise: Poisson distribution.** Let  $X_1, X_2, \dots, X_n$  be a random sample from a Poisson distribution with parameter  $\lambda$ , where  $\lambda > 0$  is unknown. Let  $Y = \sum_i X_i$  be the sample sum, and assume that  $Y > 0$ .

Demonstrate that the ML estimator of  $\lambda$  is  $\hat{\lambda} = Y/n$ .

**Exercise: Uniform distribution on  $(0, b)$ .** Let  $X_1, X_2, \dots, X_n$  be a random sample from a uniform distribution on the interval  $[0, b]$ , where  $b > 0$  is unknown.

Demonstrate that  $\hat{b} = \max(X_1, X_2, \dots, X_n)$  is the ML estimator of  $b$ .

### 2.5.3 Cramer-Rao Lower Bound

The following theorem, proven by both Cramer and Rao in the 1940's, gives a formula for the lower bound on the variance of an unbiased estimator of  $\theta$ . The formula is valid under certain "smoothness conditions" on the  $X$  distribution.

**Theorem (Cramer-Rao Lower Bound).** Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with parameter  $\theta$  and let  $\hat{\theta}$  be an unbiased estimator of  $\theta$ . Under smoothness conditions on the  $X$  distribution:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}, \quad \text{where } nI(\theta) = -E(\ell''(\theta)).$$

In this formula,  $\ell''(\theta)$  is the second derivative of the log-likelihood function, and the expectation is computed using the joint distribution of the  $X_i$ 's for fixed  $\theta$ .

*Notes:*

1. *Smoothness Conditions:* If the following three conditions hold:

- (a) The PDF of  $X$  has continuous second partial derivatives (except, possibly, at a finite number of points),
- (b) The parameter  $\theta$  is not at the boundary of possible parameter values, and
- (c) The range of  $X$  does not depend on  $\theta$ ,

then  $X$  satisfies the "smoothness conditions" of the theorem.

The theorem excludes, for example, the Bernoulli distribution with  $p = 1$  (condition 2 is violated) and the uniform distribution on  $[0, b]$  (condition 3 is violated).

2. *Fisher Information:* The quantity  $nI(\theta)$  is called the *information in a sample of size n*.
3. *Cramer-Rao Lower Bound:* The lower bound for the variance given in the theorem is called the *Cramer-Rao lower bound*; it is the reciprocal of the information,  $\frac{1}{nI(\theta)}$ .

### 2.5.4 Efficient Estimators; Minimum Variance Unbiased Estimators

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with parameter  $\theta$ , and let  $\hat{\theta}$  be an estimator based on this sample. Then  $\hat{\theta}$  is said to be an *efficient estimator* of  $\theta$  if

$$E(\hat{\theta}) = \theta \quad \text{and} \quad \text{Var}(\hat{\theta}) = \frac{1}{nI(\theta)}.$$

*Note* that if the  $X$  distribution satisfies the smoothness conditions listed above, and  $\hat{\theta}$  is an efficient estimator, then the Cramer-Rao theorem implies that  $\hat{\theta}$  is the minimum variance unbiased estimator (MVUE) of  $\theta$ .

**Example: Bernoulli distribution.** Let  $X_1, X_2, \dots, X_n$  be a random sample from a Bernoulli distribution with success probability  $p$  and let  $Y = \sum_i X_i$  be the sample sum.

Assume that  $0 < Y < n$ . Then  $\hat{p} = \frac{Y}{n}$  is an efficient estimator of  $p$ .

To see this,

1. From the work we did starting on page 23, we know that  $\hat{p} = \frac{Y}{n}$  is the ML estimator of the success probability  $p$ , and that the second derivative of the log-likelihood function is

$$\ell''(p) = -\frac{Y}{p^2} - \frac{(n-Y)}{(1-p)^2}.$$

2. The following computations demonstrate that the information  $nI(p) = \frac{n}{p(1-p)}$ :

$$\begin{aligned} nI(p) &= -E(\ell''(p)) \\ &= -E\left(-\frac{Y}{p^2} - \frac{(n-Y)}{(1-p)^2}\right) \\ &= \frac{1}{p^2}E(Y) + \frac{1}{(1-p)^2}E(n-Y) \\ &= \frac{1}{p^2}(np) + \frac{1}{(1-p)^2}(n-np) && \text{since } E(Y) = nE(X) = np \\ &= \frac{n}{p} + \frac{n}{(1-p)} \\ &= \frac{n}{p(1-p)} \end{aligned}$$

Thus, the Cramer-Rao lower bound is  $\frac{p(1-p)}{n}$ .

3. The following computations demonstrate that  $\hat{p}$  is an unbiased estimator of  $p$  with variance equal to the Cramer-Rao lower bound:

$$\begin{aligned} E(\hat{p}) &= \frac{1}{n} E(Y) = \frac{1}{n}(nE(X)) = \frac{1}{n}(np) = p \\ \text{Var}(\hat{p}) &= \frac{1}{n^2} \text{Var}(Y) = \frac{1}{n^2}(n\text{Var}(X)) = \frac{1}{n^2}(np(1-p)) = \frac{p(1-p)}{n}. \end{aligned}$$

Thus,  $\hat{p}$  is an efficient estimator of  $p$ .

**Exercise: Poisson Distribution.** Let  $X_1, X_2, \dots, X_n$  be a random sample from a Poisson distribution with parameter  $\lambda$  and let  $Y = \sum_i X_i$  be the sample sum. Assume that  $Y > 0$ .

Demonstrate that  $\hat{\lambda} = Y/n$  is an efficient estimator of  $\lambda$ .

### 2.5.5 Large Sample Theory: Fisher's Theorem

R.A. Fisher proved a generalization of the Central Limit Theorem for ML estimators.

**Fisher's Theorem.** Let  $\hat{\theta}_n$  be the ML estimator of  $\theta$  based on a random sample of size  $n$  from the  $X$  distribution, and

$$Z_n = \frac{\hat{\theta}_n - \theta}{\sqrt{1/(nI(\theta))}}, \text{ where } nI(\theta) \text{ is the information.}$$

Under smoothness conditions on the  $X$  distribution,

$$\lim_{n \rightarrow \infty} P(Z_n \leq x) = \Phi(x) \text{ for every real number } x,$$

where  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal random variable.

*Notes:*

1. *Smoothness Conditions:* The smoothness conditions in Fisher's theorem are the same ones mentioned earlier.
2. *Asymptotic Efficiency:* If the  $X$  distribution satisfies the smoothness conditions and the sample size is large, then the sampling distribution of the ML estimator is approximately normal with mean  $\theta$  and variance equal to the Cramer-Rao lower bound. Thus, the ML estimator is said to be *asymptotically efficient*.

*It is instructive* to give an outline of a key idea Fisher used to prove the theorem above.

1. Let  $\ell(\theta) = \log(\text{Lik}(\theta))$  be the log-likelihood function, and  $\hat{\theta}$  be the ML estimator.
2. Using the second order Taylor polynomial of  $\ell(\theta)$  expanded around  $\hat{\theta}$ , we get

$$\ell(\theta) \approx \ell(\hat{\theta}) + \ell'(\hat{\theta})(\theta - \hat{\theta}) + \frac{\ell''(\hat{\theta})}{2}(\theta - \hat{\theta})^2 = \ell(\hat{\theta}) + \frac{\ell''(\hat{\theta})}{2}(\theta - \hat{\theta})^2 \text{ since } \ell'(\hat{\theta}) = 0.$$

3. Since  $\text{Lik}(\theta) = e^{\ell(\theta)} = \exp(\ell(\theta))$ , we get

$$\text{Lik}(\theta) \approx \exp\left(\ell(\hat{\theta}) + \frac{\ell''(\hat{\theta})}{2}(\theta - \hat{\theta})^2\right) = \text{Lik}(\hat{\theta})e^{-(\theta - \hat{\theta})^2/(2(-1/\ell''(\hat{\theta})))}.$$

The rightmost expression has the approximate form of the density function of a normal distribution with mean  $\hat{\theta}$  and variance equal to the reciprocal of  $nI(\theta) = -E(\ell''(\theta))$ .

### 2.5.6 Approximate Confidence Interval Procedures; Special Cases

Under the conditions of Fisher's theorem, an approximate  $100(1 - \alpha)\%$  confidence interval for  $\theta$  has the following form:

$$\hat{\theta} \pm z(\alpha/2) \sqrt{\frac{1}{nI(\hat{\theta})}}$$

where  $z(\alpha/2)$  is the  $100(1 - \alpha/2)\%$  point of the standard normal distribution.

Several special cases are of interest:

1. *Bernoulli Distribution.* Let  $Y = \sum_i X_i$  be the sample sum of a random sample of size  $n$  from a Bernoulli distribution with parameter  $p$ , where  $p \in (0, 1)$ . If  $n$  is large and  $0 < Y < n$ , then an approximate  $100(1 - \alpha)\%$  confidence interval for  $p$  has the following form:

$$\hat{p} \pm z(\alpha/2) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad \text{where } \hat{p} = \frac{Y}{n}$$

and  $z(\alpha/2)$  is the  $100(1 - \alpha/2)\%$  point of the standard normal distribution.

2. *Poisson Distribution.* Let  $Y = \sum_i X_i$  be the sample sum of a random sample of size  $n$  from a Poisson distribution with parameter  $\lambda$ . If  $n$  is large and  $Y > 0$ , then an approximate  $100(1 - \alpha)\%$  confidence interval for  $\lambda$  has the following form:

$$\hat{\lambda} \pm z(\alpha/2) \sqrt{\frac{\hat{\lambda}}{n}} \quad \text{where } \hat{\lambda} = \frac{Y}{n}$$

and  $z(\alpha/2)$  is the  $100(1 - \alpha/2)\%$  point of the standard normal distribution.

3. *Exponential Distribution.* Let  $Y = \sum_i X_i$  be the sample sum of a random sample of size  $n$  from an exponential distribution with parameter  $\lambda$ . If  $n$  is large, then an approximate  $100(1 - \alpha)\%$  confidence interval for  $\lambda$  has the following form:

$$\hat{\lambda} \pm z(\alpha/2) \sqrt{\frac{(\hat{\lambda})^2}{n}} \quad \text{where } \hat{\lambda} = \frac{n}{Y}$$

and  $z(\alpha/2)$  is the  $100(1 - \alpha/2)\%$  point of the standard normal distribution.

**Margin of error in survey analysis.** Researchers are often interested in determining the proportion,  $p$ , of individuals in a population with a certain property. (For example, the researchers may be interested in determining the proportion of individuals who favor a certain tax plan). To estimate  $p$ , they collect information from a simple random sample of individuals from the population, and report the sample proportion,  $\hat{p}$ , and the margin of error,

$$z(\alpha/2) \sqrt{\hat{p}(1 - \hat{p})/n}.$$

The *margin of error* is the half-width of the large sample confidence interval for Bernoulli models; unless otherwise specified,  $\alpha = 0.05$  is used to calculate the margin of error.

*Exercise.* In a recent Gallup Poll of 1570 adults in the United States, 30% (471/1570) said they considered the most serious problem facing the nation's public schools to be drugs.

Assume this information is a summary of the values of a random sample from a Bernoulli distribution. Use Fisher's theorem to construct an approximate 95% confidence interval for the proportion  $p$  of all adults in the United States who consider the most serious problem facing public schools to be drugs.

*Exercise.* Assume the following data are the values of a random sample from an exponential distribution:

0.62	0.64	0.70	0.85	0.91	1.03	1.10	2.27	2.37	3.61	3.87	4.94
5.28	5.87	6.35	6.91	7.08	8.22	8.41	8.74	9.38	9.44	9.64	11.83
12.09	12.60	13.40	13.49	14.14	14.41	15.45	16.42	16.46	16.91	17.76	17.88
18.27	18.68	18.68	19.37	19.67	20.68	21.14	21.48	21.49	22.69	22.95	23.32
24.73	25.40	26.40	26.40	26.47	26.60	26.61	29.25	30.31	30.64	33.32	34.40
35.29	35.54	36.34	37.82	40.97	44.00	51.36	51.47	51.73	51.95	52.77	52.79
55.40	55.93	56.26	59.74	64.19	67.23	67.73	69.69	69.99	70.76	73.77	73.87
76.80	77.85	79.11	79.55	82.01	82.38	83.46	87.28	100.34	100.41	108.48	108.93
135.24	150.86	201.54	214.53								

Sample summaries:  $n = 100$ ,  $\sum_i x_i = 3985.38$

Find the ML estimate of  $\lambda$ , and use Fisher's theorem to construct an approximate 90% confidence interval for  $\lambda$ .

**Exercise: Exponential distribution.** Let  $Y = \sum_i X_i$  be the sample sum of a random sample of size  $n$  from an exponential distribution with parameter  $\lambda$ .

- (a) Demonstrate that the ML estimator of  $\lambda$  is  $\hat{\lambda} = n/Y$ .
- (b) Demonstrate that  $nI(\lambda) = n/\lambda^2$ .

### 2.5.7 Multinomial Experiments

Consider a multinomial experiment with  $k$  outcomes, each of whose probabilities is a function of the parameter  $\theta$ :

$$p_i = p_i(\theta), \quad \text{for } i = 1, 2, \dots, k.$$

The results of  $n$  independent trials of a multinomial experiment are usually presented to us in summary form: Let  $X_i$  be the number of occurrences of the  $i^{\text{th}}$  outcome in  $n$  independent trials of the experiment, for  $i = 1, 2, \dots, k$ .

Since the random  $k$ -tuple  $(X_1, X_2, \dots, X_k)$  has a multinomial distribution, the likelihood function is the joint PDF thought of as a function of  $\theta$  with the  $X_i$ 's left unevaluated:

$$Lik(\theta) = \binom{n}{X_1, X_2, \dots, X_k} p_1(\theta)^{X_1} p_2(\theta)^{X_2} \dots p_k(\theta)^{X_k}.$$

We work with this likelihood when finding the ML estimator, the information, and so forth.

*Exercise.* Consider a multinomial experiment with 3 outcomes, whose probabilities are

$$p_1 = p_1(\theta) = (1 - \theta)^2, \quad p_2 = p_2(\theta) = 2\theta(1 - \theta), \quad p_3 = p_3(\theta) = \theta^2,$$

where  $\theta \in (0, 1)$  is unknown. Let  $X_i$  be the number of occurrences of the  $i^{\text{th}}$  outcome in  $n$  independent trials of the multinomial experiment, for  $i = 1, 2, 3$ .

- (a) Set up the likelihood and log-likelihood functions for this model. Simplify each function as much as possible.

- (b) Find the ML estimator,  $\hat{\theta}$ .
- (c) Find the information,  $nI(\theta)$ .

(d) Find the form of an approximate  $100(1 - \alpha)\%$  confidence interval for  $\theta$ .

**Application: Hardy-Weinberg equilibrium model.** In the Hardy-Weinberg equilibrium model there are three outcomes, with probabilities:

$$p_1 = (1 - \theta)^2, \quad p_2 = 2\theta(1 - \theta), \quad p_3 = \theta^2 .$$

The model is used in genetics. The usual setup is as follows:

1. Genetics is the science of inheritance. Hereditary characteristics are carried by *genes*, which occur in a linear order along chromosomes. A gene's position along the chromosome is also called its *locus*. Except for the sex chromosomes, there are two genes at every locus (one inherited from mom and the other from dad).
2. An *allele* is an alternative form of a genetic locus.
3. Consider the simple case of a gene not located on one of the sex chromosomes, and with just two alternative forms (i.e., two alleles), say  $A$  and  $a$ . This gives three *genotypes*:

$AA$ ,  $Aa$ , and  $aa$ .

4. Let  $\theta$  equal the probability that a gene contains allele  $a$ . After many generations, the proportions of the three genotypes in a population are (approximately):

$$\begin{aligned} p_1 &= P(AA \text{ occurs}) = (1 - \theta)^2 \\ p_2 &= P(Aa \text{ occurs}) = 2\theta(1 - \theta) \\ p_3 &= P(aa \text{ occurs}) = \theta^2 \end{aligned}$$

*Exercise (Source: Rice textbook, Chapter 8).* In a sample from the Chinese population in Hong Kong in 1937, blood types occurred with the following frequencies,

Blood type	$MM$	$MN$	$NN$
Frequency	342	500	187

where  $M$  and  $N$  are two types of antigens in the blood.

Assume these data summarize the values of a random sample from a Hardy-Weinberg equilibrium model, where:

$$\begin{aligned}p_1 &= P(MM \text{ occurs}) = (1 - \theta)^2 \\p_2 &= P(MN \text{ occurs}) = 2\theta(1 - \theta) \\p_3 &= P(NN \text{ occurs}) = \theta^2\end{aligned}$$

Find the ML estimate of  $\theta$ , the estimated proportions of each blood type ( $MM$ ,  $MN$ ,  $NN$ ) in the population, and an approximate 90% confidence interval for  $\theta$ .

### 2.5.8 ML Estimation Method for Multiple Parameters

If the  $X$  distribution has two or more unknown parameters, then ML estimators are computed using the techniques of multivariable calculus.

Consider the two-parameter case. Let  $\theta_1$  and  $\theta_2$  be the two parameters, and

$$\ell(\theta_1, \theta_2) = \log(Lik(\theta_1, \theta_2)) \text{ be the log-likelihood function.}$$

1. To find the ML estimators, we need to solve the following system of partial derivatives for the unknown parameters  $\theta_1$  and  $\theta_2$ :

$$\ell_1(\theta_1, \theta_2) = 0 \text{ and } \ell_2(\theta_1, \theta_2) = 0.$$

2. To check that  $(\hat{\theta}_1, \hat{\theta}_2)$  is a maximum using the second derivative test, we need to check that the following two conditions hold:

$$d_1 = \ell_{11}(\hat{\theta}_1, \hat{\theta}_2) < 0 \text{ and } d_2 = \ell_{11}(\hat{\theta}_1, \hat{\theta}_2)\ell_{22}(\hat{\theta}_1, \hat{\theta}_2) - (\ell_{12}(\hat{\theta}_1, \hat{\theta}_2))^2 > 0,$$

where the  $\ell_{ij}()$  functions are the second partial derivatives.

**Exercise: Normal distribution.** Let  $X$  be a normal random variable with mean  $\mu$  and variance  $\sigma^2$ . Given a random sample of size  $n$  from the  $X$  distribution, the likelihood function and log-likelihood functions can be written as follows:

$$\begin{aligned} Lik(\mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2}\right) \end{aligned}$$

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$$

- (a) Find the first and second partial derivatives of  $\ell(\mu, \sigma^2)$ .

(b) Demonstrate that the ML estimators of  $\mu$  and  $\sigma^2$  are

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

and check that you have a maximum.

**Example: Gamma distribution.** Let  $X$  be a gamma random variable with shape parameter  $\alpha$  and scale parameter  $\beta$ . Given a random sample of size  $n$  from the  $X$  distribution, the likelihood and log-likelihood functions are as follows:

$$\begin{aligned} Lik(\alpha, \beta) &= \prod_{i=1}^n \frac{1}{\Gamma(\alpha)\beta^\alpha} X_i^{\alpha-1} e^{-X_i/\beta} \\ &= \left( \frac{1}{\Gamma(\alpha)\beta^\alpha} \right)^n (\prod_i X_i)^{\alpha-1} e^{-\sum_i X_i/\beta} \\ \ell(\alpha, \beta) &= -n \log(\Gamma(\alpha)) - n\alpha \log(\beta) + (\alpha - 1) \log(\prod_i X_i) - \frac{\sum_i X_i}{\beta} \end{aligned}$$

The partial derivatives of the log-likelihood with respect to  $\alpha$  and  $\beta$  are as follows:

$$\begin{aligned} \ell_1(\alpha, \beta) &= \frac{-n\Gamma'(\alpha)}{\Gamma(\alpha)} - n \log(\beta) + \log(\prod_i X_i) \\ \ell_2(\alpha, \beta) &= \frac{-n\alpha}{\beta} + \frac{\sum_i X_i}{\beta^2} \end{aligned}$$

Since the system of equations  $\ell_1(\alpha, \beta) = 0$  and  $\ell_2(\alpha, \beta) = 0$  cannot be solved exactly, the computer is used to analyze specific samples.

For example, assume the following data are the values of a random sample from the  $X$  distribution:

8.68	8.91	11.42	12.04	12.47	14.61
14.82	15.77	17.85	23.44	29.60	32.26

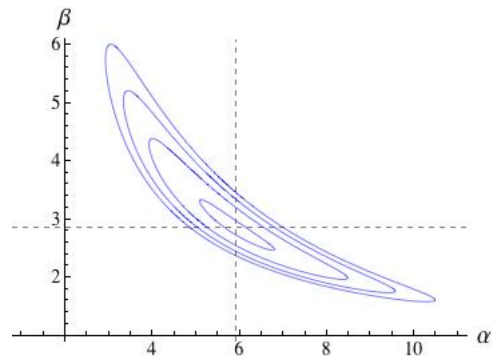
The graph below is a contour diagram of the observed likelihood function:

$$z = \ell(\alpha, \beta) = -12 \log(\Gamma(\alpha)) - 12\alpha \log(\beta) + (\alpha - 1)(32.829) - \frac{201.87}{\beta}.$$

- The contours shown in the graph correspond to

$$z = -39.6, -40.0, -40.4, -40.8.$$

- The function is maximized at  $(5.91, 2.85)$  (the point of intersection of the dashed lines).
- Thus, the ML estimate of  $(\alpha, \beta)$  is  $(5.91, 2.85)$ .



Note that, for these data, the MOM estimate of  $(\alpha, \beta)$  is  $(5.17, 3.25)$ .