prepared by Professor Jenny Baglivo

2 MATH4427 Notebook 2 **3**

2.1 Definitions and Examples 3

2.2 Performance Measures for Estimators 5

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 12

2.3.1 Error Probability, Confidence Coefficient, Confidence Interval 12

2.3.2 Confidence Interval Procedures for Normal Distributions 14

2.4 Method of Moments (MOM) Estimation 19

2.4.1 Kth Moments and Kth Sample Moments 19

2.4.2 MOM Estimation Method 19

2.4.3 MOM Estimation Method for Multiple Parameters 22

2.5 Maximum Likelihood (ML) Estimation 24

2.5.1 Likelihood and Log-Likelihood Functions 24

2.5.2 ML Estimation Method 24

2.5.3 Cramér-Rao Lower Bound 29

2.5.4 Efficient Estimators; Minimum Variance Unbiased Estimators 29

2.5.5 Large Sample Theory: Fisher's Theorem 32

2.5.6 Approximate Confidence Interval Procedures; Special Cases 33

2.5.7 Multinomial Experiments 37

2.5.8 ML Estimation Method for Multiple Parameters 42

2 MATH4427 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,

$$X_1, X_2, \ldots, X_n,$$

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 $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and standard deviation σ.

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

2. *Sample Variance:* The sample variance,

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

is an estimator of σ^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. $\overline{x} = 106.513$ is an estimate of μ, and
2. $s^2 = 316.316$ is an estimate of σ^2

based on these observations.

$\overline{X}$ and S^2 can also be used to test assertions about the parameters of a normal distribution. For example, $\overline{X}$ 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 $i = 1, 2, \ldots, k$.

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^{th} outcome in n independent trials of the experiment, for $i = 1, 2, \ldots, k$. Then

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

$$\widehat{p_i} = \frac{X_i}{n},$$

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

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

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

can be used to test the assertion that the true probability list is $(p_1, p_2, \ldots, p_k)$. The exact probability distribution of $\boldsymbol{X^2}$ 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.

Suppose that $n = 40$, $k = 6$ and $(x_1, x_2, x_3, x_4, x_5, x_6) = (5, 7, 8, 5, 11, 4)$ is observed.

Then

$$\widehat{p_1} = 0.125,\ \widehat{p_2} = 0.175,\ \widehat{p_3} = 0.20,\ \widehat{p_4} = 0.125,\ \widehat{p_5} = 0.275,\ \widehat{p_6} = 0.10$$

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), then

$$\boldsymbol{x^2}_{\text{obs}} = \sum_{i=1}^{6} \frac{(x_i - 40/6)^2}{40/6} = 5.0$$

is the value of the test statistic for the observed list of counts.

2.2 Performance Measures for Estimators

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

$$\widehat{\theta} \;=\; \widehat{\theta}\,(X_1, X_2, \ldots, X_n)$$

(read "theta-hat") is used to denote an estimator of θ.

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 $\widehat{\theta}$ is defined as follows:

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

   An accurate estimator will have little or no bias.

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

3. *Asymptotically Unbiased Estimator:* If $\widehat{\theta}$ is a biased estimator satisfying

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

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

Example: Normal distribution. Let $\overline{X}$ 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 μ and standard deviation σ.

Then

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

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

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

Exercise. Another commonly used estimator of σ^2 is

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

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

Exercise. Let $\overline{X}$ be the sample mean of a random sample of size n from a continuous uniform distribution on the interval $[0, b]$, and let $\widehat{b} = 2\overline{X}$ be an estimator of the upper endpoint b.

(a) Demonstrate that $\widehat{b}$ is an unbiased estimator of b, and find the variance of the estimator.

(b) Use $\widehat{b}$ 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:

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

A precise estimator will have a small mean squared error.

> ***Mean Squared Error Theorem.*** Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ, and let $\widehat{\theta}$ be an estimator of θ. Then
>
> 1. $MSE(\widehat{\theta}) = Var(\widehat{\theta}) + (BIAS(\widehat{\theta}))^2$.
> 2. If $\widehat{\theta}$ is an unbiased estimator of θ, then $MSE(\widehat{\theta}) = Var(\widehat{\theta})$.

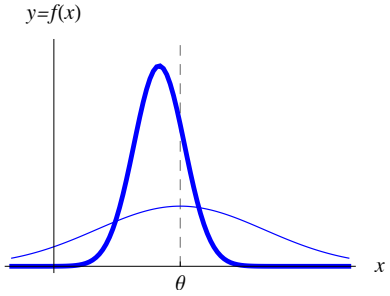
Exercise. Use the definitions of mean squared error and variance, and properties of expectation, to prove the theorem above.

2.2.3 Comparing Estimators: Efficiency

Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two estimators of θ, each based on a random sample of size n from the X distribution.

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

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



Note that the more efficient estimator could be a biased estimator of θ, as suggested above.

When both estimators are unbiased, then the work we did on the previous page implies that we just need to compare variances. That is,

If both $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are unbiased estimators of θ, then

$\widehat{\theta}_1$ is said to be *more efficient* than $\widehat{\theta}_2$ if $Var(\widehat{\theta}_1) < Var(\widehat{\theta}_2)$.

Exercise. Let X_1, X_2, X_3, X_4 be a random sample of size 4 from a distribution with mean μ and standard deviation σ. The following two statistics are unbiased estimators of μ:

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

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 $\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$ is said to be a *minimum variance unbiased estimator (MVUE)* of θ if
>
> $$Var(\widehat{\theta}) \leq Var(\widehat{\theta}^*) \text{ for all unbiased estimators } \widehat{\theta}^* = \widehat{\theta}^*(X_1, X_2, \ldots, X_n).$$

We will consider the question of finding an MVUE in Section 2.5.4 (page 29) of these notes.

2.2.4 Another Performance Measure: Consistency

The estimator $\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$ is said to be a *consistent estimator* of θ if, for every positive number ϵ,

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

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

Note that sample means are, in general, consistent estimators of distribution means. This follows from the Law of Large Numbers from probability theory:

> ***Law of Large Numbers.*** Let X be a random variable with mean $\mu = E(X)$ and standard deviation $\sigma = SD(X)$. Further, let
>
> $$\overline{X} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i$$
>
> be the sample mean of a random sample of size n from the X distribution. Then, for every positive real number ϵ,
>
> $$\lim_{n\to\infty} P(|\overline{X} - \mu| \geq \epsilon) = 0.$$

<u>*Importantly,*</u> consistent estimators are not necessarily unbiased, although they are often asymptotically unbiased. If an estimator is unbiased, however, a quick check for consistency is given in the following theorem.

> ***Consistency Theorem.*** If $\widehat{\theta}$ is an unbiased estimator of θ satisfying
>
> $$\lim_{n\to\infty} Var(\widehat{\theta}) = 0, \quad \text{where } n \text{ is the sample size,}$$
>
> then $\widehat{\theta}$ is a consistent estimator of θ.

Exercise. One form of the Chebyshev inequality from probability theory says that

$$P(|Y - \mu_y| \geq k\sigma_y) \leq \frac{1}{k^2},$$

where Y is a random variable with mean $\mu_y = E(Y)$ and standard deviation $\sigma_y = SD(Y)$, and k is a positive constant. Use this form of the Chebyshev inequality to prove the Consistency Theorem for unbiased estimators.

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

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

Exercise. Let S^2 be the sample variance of a random sample of size n from a normal distribution with mean μ and standard deviation σ.

Demonstrate that S^2 is a consistent estimator of σ^2.

2.3 Interval Estimation

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

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

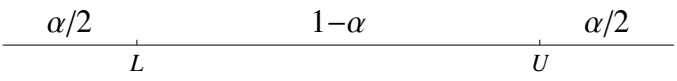
with the property that θ lies in the interval $[L, U]$ with high probability.

2.3.1 Error Probability, Confidence Coefficient, Confidence Interval

1. *Error Probability:* It is customary to let α (called the *error probability*) equal the probability that θ 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 θ.

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 σ. Let X be a normal random variable with mean μ and known standard deviation σ, and let $\overline{X}$ 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 μ:

1. First note that $Z = (\overline{X} - \mu)/\sqrt{\sigma^2/n}$ is a standard normal random variable.

2. For convenience, let $p = \alpha/2$. Further, let z_p and z_{1-p} be the p^{th} and $(1-p)^{\text{th}}$ quantiles of the standard normal distribution. Then

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

3. The following statements are equivalent:

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

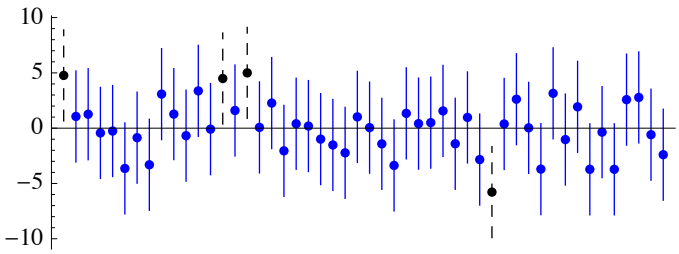
Thus, $P\left(\overline{X} - z_{1-p}\sqrt{\tfrac{\sigma^2}{n}} \le \mu \le \overline{X} - z_p\sqrt{\tfrac{\sigma^2}{n}}\right) = 1 - 2p.$

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

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

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 a 90% confidence interval for μ 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).

[*Question:* If you were to repeat the computer experiment above many times, how many intervals would you expect would contain 0? Why?]

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, $\overline{X}$ is the sample mean and S^2 is the sample variance of a random sample of size n from a normal distribution with mean μ and standard deviation σ.

1. $100(1-\alpha)\%$ *CI for* μ*, when* σ *is known:*

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

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

2. $100(1-\alpha)\%$ *CI for* μ*, when* σ *is estimated:*

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

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* σ^2*, when* μ *is known:*

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

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

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

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

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

5. $100(1-\alpha)\%$ *CI for* σ*:*

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

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 $\overline{X}$,

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

   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 σ^2 when μ 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: $n = 15$, $\overline{x} = 98.1967$

Construct 80% and 90% confidence intervals for μ.

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: $n = 15$, $\overline{x} = 2.90347$, $s^2 = 3.53177$

Construct 90% confidence intervals for σ^2 and σ.

Exercise. As part of a study on body temperatures of healthy adult men and women, 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: $n = 30$, $\overline{x} = 98.4533$, $s = 0.6463$

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

Construct and interpret 95% confidence intervals for μ and σ.

2.4 Method of Moments (MOM) Estimation

The method of moments is a general method for estimating one or more unknown parameters, 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^{th} *moment* of the X distribution is

$$\mu_k = E(X^k), \text{ for } k = 1, 2, \ldots, \text{ whenever the expected value converges.}$$

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

$$\widehat{\mu_k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \text{ for } k = 1, 2, \ldots,$$

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

[*Question:* The k^{th} sample moment is an unbiased and consistent estimator of μ_k, when μ_k exists. Can you see why?]

2.4.2 MOM Estimation Method

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

Then, a *method of moments estimator* of θ is obtained using the following procedure:

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

For example, let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the continuous 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:

$$\mu_1 = \widehat{\mu_1} \quad \Longrightarrow \quad E(X) = \overline{X} \quad \Longrightarrow \quad \frac{b}{2} = \overline{X} \quad \Longrightarrow \quad b = 2\overline{X} \ .$$

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

Exercise. Let $X_1, X_2, \ldots, X_n$ be a random sample from the continuous 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 (slope-parameter model). Let X be a continuous random variable with range equal to the interval $(-1, 1)$ and with density function as follows:

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

where $\alpha \in (-1, 1)$ is an unknown parameter. Find a method of moments estimator of α 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 θ_1 and θ_2), then MOM estimators are obtained using the procedure:

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

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

Since $E(X) = \mu$ and $E(X^2) = Var(X) + (E(X))^2 = \sigma^2 + \mu^2$, 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 $X_1, X_2, \ldots, X_n$ be a random sample from a gamma distribution with parameters α and β, where both parameters are unknown. To find method of moments estimators of α and β we need to solve two equations in the two unknowns.

Since $E(X) = \alpha\beta$ and $E(X^2) = Var(X) + (E(X))^2 = \alpha\beta^2 + (\alpha\beta)^2$, 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 $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with a scalar parameter θ.

1. The *likelihood function* is the joint PDF of the random sample thought of as a function of the parameter θ, 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 λ, then

1. The PDF of X is $f(x) = \lambda e^{-\lambda x}$ when $x > 0$, and $f(x) = 0$ otherwise,
2. The likelihood function for the sample of size n is

$$Lik(\lambda) \;=\; \prod_{i=1}^{n} f(X_i) \;=\; \prod_{i=1}^{n} \lambda e^{-\lambda X_i} \;=\; \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}, \text{ and}$$

3. The log-likelihood function for the sample of size n is

$$\ell(\lambda) = \log\left(\lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}\right) = n\log(\lambda) - \lambda \sum_{i=1}^{n} X_i.$$

2.5.2 ML Estimation Method

The *maximum likelihood estimator* (or *ML estimator*) of θ 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 $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli distribution with success probability p and let $Y = \sum_{i=1}^{n} X_i$ be the sample sum. Then the ML estimator of p is the sample proportion $\widehat{p} = Y/n$. 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 and log-likelihood functions are

$$Lik(p) = \prod_{i=1}^{n} p(X_i) = p^Y (1-p)^{n-Y} \quad \text{and} \quad \ell(p) = Y \log(p) + (n - Y) \log(1 - p).$$

2. The derivative of the log-likelihood function is $\quad \ell'(p) = \dfrac{Y}{p} - \dfrac{(n-Y)}{(1-p)}.$

3. There are three cases to consider:

$\boxed{Y = 0:}$ If no successes are observed, then $Lik(p) = (1-p)^n$ is decreasing on $[0, 1]$. The maximum value occurs at the left endpoint. Thus, $\widehat{p} = 0$.

$\boxed{Y = n:}$ If no failures are observed, then $Lik(p) = p^n$ is increasing on $[0, 1]$. The maximum value occurs at the right endpoint. Thus, $\widehat{p} = 1$.

$\boxed{0 < Y < n:}$ In the usual case, some successes and some failures are observed. Working with the log-likelihood function we see that

$$\ell'(p) = 0 \;\Rightarrow\; \frac{Y}{p} = \frac{(n-Y)}{(1-p)} \;\Rightarrow\; Y(1-p) = p(n-Y) \;\Rightarrow\; p = \frac{Y}{n}.$$

Thus, $\widehat{p} = \frac{Y}{n}$ is the proposed ML estimator.

The second derivative test is generally used to check that the proposed estimator is a maximum. In this case, since

$$\ell\,''(p) = -\frac{Y}{p^2} - \frac{(n-Y)}{(1-p)^2} \quad \text{and} \quad \ell\,''(\widehat{p}) < 0,$$

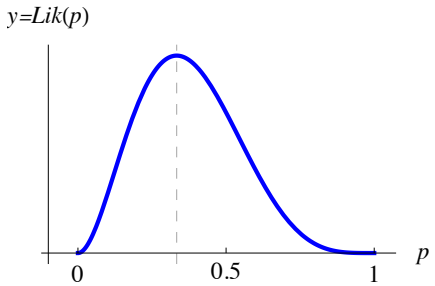
the second derivative test implies that $\widehat{p} = \frac{Y}{n}$ is a maximum.

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

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 range $(-1, 1)$ and with PDF as follows:

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

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

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

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

the log-likelihood simplifies to $\ell(\alpha) = -n\log(2) + \sum_{i=1}^{n}\log(1+\alpha X_i)$, and the first derivative of the log-likelihood function is

$\ell'(\alpha) =$ _____.

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

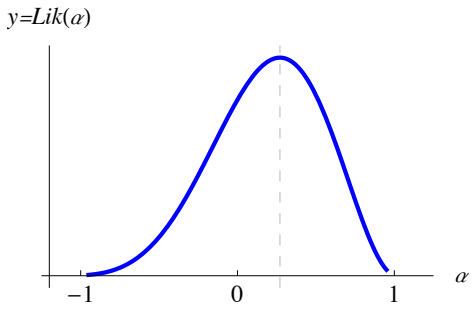
Sample summaries: $n = 16$, $\overline{x} = 0.0964375$

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 α for these data is 0.270822.



Using the work we did on page 21,

a MOM estimate of α for the data above is _____

Exercise: Poisson distribution. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ, where $\lambda > 0$ is unknown. Let $Y = \sum_{i=1}^{n} X_i$ be the sample sum, and assume that $\boxed{Y > 0}$.

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

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

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

2.5.3 Cramér-Rao Lower Bound

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

> ***Theorem (Cramér-Rao Lower Bound).*** Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with scalar parameter θ and let $\widehat{\theta}$ be an unbiased estimator of θ. Under smoothness conditions on the X distribution:
>
> $$Var(\widehat{\theta}) \geq \frac{1}{nI(\theta)}, \quad \text{where} \quad nI(\theta) = -E(\ell''(\theta)).$$
>
> 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 θ.

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 θ is not at the boundary of possible parameter values, and
   (c) The range of X does not depend on θ,

   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. *Cramér-Rao Lower Bound:* The lower bound for the variance given in the theorem is called the *Cramér-Rao lower bound*; it is the reciprocal of the information, $\frac{1}{nI(\theta)}$.

2.5.4 Efficient Estimators; Minimum Variance Unbiased Estimators

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

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

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

Example: Bernoulli distribution. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli distribution with success probability p and let $Y = \sum_{i=1}^{n} X_i$ be the sample sum.

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

To see this,

1. From the work we did starting on page 25, we know that $\widehat{p} = \frac{Y}{n}$ 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 $nI(p) = \dfrac{n}{p(1-p)}$:

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

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

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

$$\begin{aligned}
E(\widehat{p}) &= E\left(\frac{Y}{n}\right) = \frac{1}{n}\,E(Y) = \frac{1}{n}(nE(X)) = \frac{1}{n}(np) = p \\
Var(\widehat{p}) &= Var\left(\frac{Y}{n}\right) = \frac{1}{n^2}\,Var(Y) = \frac{1}{n^2}(nVar(X)) = \frac{1}{n^2}(np(1-p)) = \frac{p(1-p)}{n}.
\end{aligned}$$

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

Exercise: Poisson Distribution. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ and let $Y = \sum_{i=1}^{n} X_i$ be the sample sum. Assume that $\boxed{Y > 0}$.

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

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 $\widehat{\theta}_n$ be the ML estimator of θ based on a random sample of size n from the X distribution, and
>
> $$Z_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{1/(nI(\theta))}}, \quad \text{where } nI(\theta) \text{ is the information.}$$
>
> Under smoothness conditions on the X distribution,
>
> $$\lim_{n\to\infty} P(Z_n \le x) = \Phi(x) \quad \text{for every real number } x,$$
>
> where $\Phi(\cdot)$ is the CDF 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 θ 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 $\ell(\theta) = \log(Lik(\theta))$ be the log-likelihood function, and $\widehat{\theta}$ be the ML estimator.
2. Using the second order Taylor polynomial of $\ell(\theta)$ expanded around $\widehat{\theta}$, we get
$$\ell(\theta) \approx \ell(\widehat{\theta}) + \ell'(\widehat{\theta})(\theta - \widehat{\theta}) + \frac{\ell''(\widehat{\theta})}{2}(\theta - \widehat{\theta})^2 = \ell(\widehat{\theta}) + \frac{\ell''(\widehat{\theta})}{2}(\theta - \widehat{\theta})^2 \quad \text{since } \ell'(\widehat{\theta}) = 0.$$
3. Since $Lik(\theta) = e^{\ell(\theta)} = \exp(\ell(\theta))$, we get
$$Lik(\theta) \approx \exp\left(\ell(\widehat{\theta}) + \frac{\ell''(\widehat{\theta})}{2}(\theta - \widehat{\theta})^2\right) = Lik(\widehat{\theta})e^{-(\theta-\widehat{\theta})^2/(2(-1/\ell''(\widehat{\theta})))}.$$
The rightmost expression has the approximate form of the density function of a normal distribution with mean $\widehat{\theta}$ 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 θ has the following form:

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

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

Exercise: Bernoulli Distribution. Let $Y = \sum_{i=1}^{n} X_i$ be the sample sum of a random sample of size n from a Bernoulli distribution with parameter p, where $p \in (0,1)$.

(a) Assume that the sample size n is large and that $0 < Y < n$. Use Fisher's theorem to find the form of an approximate $100(1-\alpha)\%$ confidence interval for p.[1]

(b) Public health officials in Florida conducted a study to determine the level of resistance to the antibiotic penicillin in individuals diagnosed with a strep infection. They chose a simple random sample of 1714 individuals from this population, and tested cultures taken from these individuals. In 973 cases, the culture showed partial or complete resistance to the antibiotic. Use your answer to part (a) to construct an approximate 99% confidence interval for the proportion p of all individuals who would exhibit partial or complete resistance to penicillin when diagnosed with a strep infection. Interpret your interval.

[1] *Note:* The approximation is good when both $E(Y) = np > 10$ and $E(n-Y) = n(1-p) > 10$.

Exercise: Poisson Distribution. Let $Y = \sum_{i=1}^{n} X_i$ be the sample sum of a random sample of size n from a Poisson distribution with parameter λ, where $\lambda > 0$.

(a) Assume that the sample size n is large and that $Y > 0$. Use Fisher's theorem to find the form of an approximate $100(1-\alpha)\%$ confidence interval for λ.[2]

(b) Skydiving is a popular sport, with roughly 250 thousand jumps made per month in the United States alone. The popularity of the sport is due, in part, to safety improvements that have been in effect for more than a decade. In spite of these improvements, fatalities still occur. Public health officials recently published a table listing the number of skydiving fatalities over the last 60 months (5 years):

Number of cases, x:	0	1	2	3	4	5	6	7	*Total:*
Number of months:	1	7	19	11	9	6	4	3	60

Use your answer to part (a) to construct an approximate 95% confidence interval for the mean monthly rate λ of skydiving fatalities. Interpret your interval.[3]

[2] *Note:* The approximation is good when $E(Y) = n\lambda > 100$.

[3] *Question:* What additional assumptions are you making here?

Exercise: Exponential distribution. Let $Y = \sum_{i=1}^{n} X_i$ be the sample sum of a random sample of size n from an exponential distribution with parameter λ.

(a) Demonstrate that the ML estimator of λ is $\widehat{\lambda} = n/Y$.

(b) Demonstrate that $nI(\lambda) = n/\lambda^2$.

(*c*) Assume that the sample size n is large. Use Fisher's theorem to find the form of an approximate $100(1-\alpha)\%$ confidence interval for λ.

(*d*) A university computer lab has 128 computers, which are used continuously from time of installation until they break. The average time to failure for these 128 computers was 1.523 years. Assume this information is a summary of a random sample from an exponential distribution. Use your answer to part (*c*) to construct an approximate 90% confidence interval for the exponential parameter λ. Interpret your interval.

2.5.7 Multinomial Experiments

Consider a multinomial experiment with k outcomes, each of whose probabilities is a function of the scalar parameter θ:

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

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^{th} outcome in n independent trials of the experiment, for $i = 1, 2, \ldots, k$.

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

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

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

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

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

(a) Set up and simplify the likelihood and log-likelihood functions.

(*b*) Find the ML estimator, $\widehat{\theta}$. Check that your estimator corresponds to a maximum.

(*c*) Find the information, $nI(\theta)$. In addition, report the Cramer-Rao lower bound.

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

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:*

$$AA, \ Aa, \text{ and } aa.$$

4. Let θ 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 θ, the estimated proportions of each blood type (MM, MN, NN) in the population, and an approximate 90% confidence interval for θ.

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 under smoothness conditions, and let

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

1. To find the ML estimators, we need to solve the following system of partial derivatives for the unknown parameters θ_1 and θ_2:

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

   where $\ell_i()$ is the partial derivative of the log-likelihood function with respect to θ_i.

2. To check that $(\widehat{\theta_1}, \widehat{\theta_2})$ is a maximum using the second derivative test, we need to check that the following two conditions hold:

   $$d_1 = \ell_{11}(\widehat{\theta_1}, \widehat{\theta_2}) < 0 \text{ and } d_2 = \ell_{11}(\widehat{\theta_1}, \widehat{\theta_2})\ell_{22}(\widehat{\theta_1}, \widehat{\theta_2}) - (\ell_{12}(\widehat{\theta_1}, \widehat{\theta_2}))^2 > 0,$$

   where $\ell_{ij}()$ is the second partial derivative of the log-likelihood function, where you differentiate first with respect to θ_i and next with respect to θ_j.

Exercise: Normal distribution. Let X_1, X_2, ..., X_n be a random sample of size n from a normal distribution with mean μ and variance σ^2.

(*a*) Write $L(\mu, \sigma^2)$ and $\ell(\mu, \sigma^2)$. Simplify each function as much as possible.

(*b*) Find simplified forms for the first and second partial derivatives of $\ell(\mu, \sigma^2)$.

(*c*) Demonstrate that the ML estimators of μ and σ^2 are

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

and check that you have a maximum.

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

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

The partial derivatives of the log-likelihood with respect to α and β are as follows:

$\ell_1(\alpha,\beta)$ = _____

$\ell_2(\alpha,\beta)$ = _____

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.

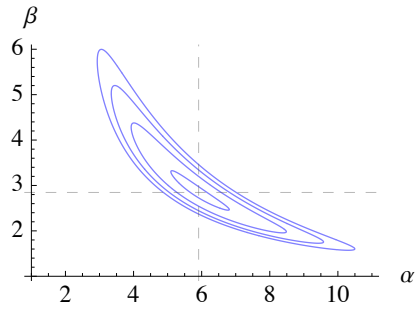
For example, assume the following data are the values of a random sample from the X distribution:

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 log-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 (α,β) is $(5.91, 2.85)$.



Note that, for these data, the MOM estimate of (α,β) is $(5.17, 3.25)$.