Appendix A

Review and Background Material

A.1 Expected Value, Variance and Covariance (Review)

Expected Value Let X be a random variable. If X is continuous, the expected value is defined as

$$E(X) = \int_{-\infty}^{\infty} x\, f_{_X}(x)\, dx.$$

If X is discrete, the formula is

$$E(X) = \sum_{x} x\, p_{_X}(x).$$

Conditional expectation uses these same formulas, only with conditional densities or probability mass functions.

Let $Y = g(X)$. The change of variables formula (a very big Theorem[1]) tells us

$$E(Y) = \int_{-\infty}^{\infty} y\, f_{_Y}(y)\, dy = \int_{-\infty}^{\infty} g(x)\, f_{_X}(x)\, dx \tag{A.1}$$

or, for discrete random variables

$$E(Y) = \sum_{y} y\, p_{_Y}(y) = \sum_{x} g(x)\, p_{_X}(x).$$

One useful function $g(x)$ is the *indicator function* for a set A. $I_A(x) = 1$ if $x \in A$, and $I_A(x) = 0$ if $x \notin A$. The expected value of an indicator function is just a probability

[1]The change of variables formula holds under very general circumstances; see for example Theorem 16.12 in Billingsley's *Probability and measure* [3]. It is extremely convenient and easy to apply, because there is no need to derive the probability distribution of Y. So for example the sets of values where $f_X(x) \neq 0$ and $f_Y(y) \neq 0$ (and therefore the regions over which you are integrating in expression (A.1)) may be different and you don't have to think about it. Furthermore, the function $g(x)$ is almost arbitrary. In particular, it need not be differentiable, a condition you would need if you tried to prove anything for the continuous case with ordinary calculus.

because, for discrete random variables,

$$E(I_A(X)) = \sum_x I_A(x)\, p_X(x) = \sum_{x \in A} p_X(x) = P(X \in A).$$

For continuous random variables, something similar happens; multiplication by $I_A(x)$ erases the density for $x \notin A$, and integration of the product from zero to infinity is just integration over the set A, yielding $P(X \in A)$.

Another useful function is a conditional expectation. If we write the conditional density

$$f_{Y|X}(y|X) = \frac{f_{X,Y}(X, y)}{f_X(X)}$$

with the capital letter X, we really mean it. X is a random variable, not a constant, and for any fixed y, the conditional density is a random variable. The conditional expected value is another random variable $g(x)$:

$$E(Y|X) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|X)\, dy.$$

This may be a strange-looking function, but still it is a function, and one can take its expected value using the change of variables formula A.1.

$$E(E(Y|X)) = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx = \int_{-\infty}^{\infty} E(Y|x)\, f_X(x)\, dx.$$

Provided $|E(Y)| < \infty$, order of integration or summation may be exchanged[2], and we have the double expectation formula:

$$E(Y) = E(E(Y|X)).$$

You will prove a slightly more general and useful version as an exercise.

The change of variables formula (A.1) still holds if $\mathbf{X}$ is a vector, or even if both $\mathbf{X}$ and $\mathbf{Y}$ are vectors, and integration or summation is replaced by multiple integration or summation. So, for example if $\mathbf{X} = (X_1, X_2)^\top$ has joint density $f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,X_2}(x_1, x_2)$ and $g(x_1, x_2) = a_1 x_1 + a_2 x_2$,

$$\begin{aligned} E(a_1X_1 + a_2X_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (a_1x_1 + a_2x_2) f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 \\ &= a_1 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 + a_2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2 f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 \\ &= a_1\, E(X_1) + a_2\, E(X_2). \end{aligned}$$

Using this approach, it is easy to establish the linearity of expected value

$$E\left(\sum_{j=1}^{m} a_j X_j\right) = \sum_{j=1}^{m} a_j E(X_j) \tag{A.2}$$

[2]By Fubini's Theorem. Again, Billingsley's *Probability and measure* [3] is a good source.

and other familiar properties.

The change of variables formula holds if the function of the random vector is just one of the variables. So, for example, since $g(x_1, x_2, \ldots x_p) = x_3$ is one possible function of $x_1, x_2, \ldots x_p$,

$$\begin{aligned} \int \cdots \int x_3 \, f(\mathbf{x}) \, d\mathbf{x} &= \int \cdots \int x_3 \, f(x_1, \ldots x_p) \, dx_1 \cdots dx_p \\ &= E(X_3). \end{aligned}$$

Variance and Covariance Denote $E(X)$ by μ_X. The variance of X is defined as

$$Var(X) = E[(X - \mu_X)^2],$$

and the covariance of X and Y is defined as

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

The *correlation* between X and Y is

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}. \tag{A.3}$$

Exercises A.1

A.1.1) Let $P\{X = x\} = \frac{x}{10}$ for $x = 1, 2, 3, 4$.

(a) Find $E(X)$. Show your work. My answer is 3.

(b) Find $E(X^2)$. Show your work. My answer is 10.

(c) Find $Var(X)$. Show your work. My answer is 1.

2. The random variable x is uniformly distributed on the integers $\{-3, -2, -1, 0, 1, 2, 3\}$, meaning $P(x = -1) = P(x = -2) = \cdots = P(x = 3) = \frac{1}{7}$. Let $y = x^2$.

(a) What is $E(x)$? The answer is a number. Show your work.

(b) Calculate the variance of x. The answer is a number. Show your work.

(c) What is $P(y = -1)$?

(d) What is $P(y = 9)$?

(e) What is the probability distribution of y? Give the y values with their probabilities.

(f) What is $E(y)$? The answer is a number. Did you already do this question?

3. The discrete random variables x and y have joint distribution

	$x = 1$	$x = 2$	$x = 3$
$y = 1$	2/12	3/12	1/12
$y = 2$	2/12	1/12	3/12

(a) What is the marginal distribution of x? List the values with their probabilities.

(b) What is the marginal distribution of y? List the values with their probabilities.

(c) Are x and y independent? Answer Yes or No and show some work.

(d) Calculate $E(x)$. Show your work.

(e) Denote a "centered" version of x by $x_c = x - E(x) = x - \mu_x$.

i. What is the probability distribution of x_c? Give the values with their probabilities.

ii. What is $E(x_c)$? Show your work.

iii. What is the probability distribution of x_c^2? Give the values with their probabilities.

iv. What is $E(x_c^2)$? Show your work.

(f) What is $Var(x)$? If you have been paying attention, you don't have to show any work.

(g) Calculate $E(y)$. Show your work.

(h) Calculate $Var(y)$. Show your work. You may use Question **??** if you wish.

(i) Calculate $Cov(x, y)$. Show your work. You may use Question **??** if you wish.

(j) Let $Z_1 = g_1(x, y) = x + y$. What is the probability distribution of Z_1? Show some work.

(k) Calculate $E(Z_1)$. Show your work.

(l) Do we have $E(x+y) = E(x)+E(y)$? Answer yes or No. Note that the answer *does not require independence.*

(m) Let $Z_2 = g_2(x, y) = xy$. What is the probability distribution of Z_2? List the values with their probabilities. Show some work.

(n) Calculate $E(Z_2)$. Show your work.

(o) Do we have $E(xy) = E(x)E(y)$? Answer yes or No. The connection to independence is established in Question **??**.

4. Here is another joint distribution. The point of this question is that you can have zero covariance without independence.

	$x = 1$	$x = 2$	$x = 3$
$y = 1$	3/12	1/12	3/12
$y = 2$	1/12	3/12	1/12

(a) Calculate $Cov(x, y)$. Show your work. You may use Question **??** if you wish.

(b) Are x and y independent? Answer Yes or No and show some work.

A.1.5) Let $X \sim U(0, \theta)$, meaning for $f(x) = \frac{1}{\theta}$ for $0 < x < \theta$, and zero otherwise.

(a) Find $E(X)$. Show your work. My answer is $\frac{\theta}{2}$.

(b) Find $E(X^2)$. Show your work. My answer is $\frac{\theta^2}{3}$.

(c) Find $Var(X)$. Show your work. My answer is $\frac{\theta^2}{12}$.

A.1.6) Let a be a constant and let X be a random variable, either continuous or discrete (you choose). Use the change of variables formula (A.1) to show that $E(a) = a$.

A.1.7) Use the change of variables formula to prove the linear property given in expression (A.2). If you assume independence, you get a zero.

A.1.8) Let X and Y be discrete random variables, with $E(|h(X)|) < \infty$. Use the change of variables formula to prove $E(h(X)) = E[E(h(X)|Y)]$. Because $E(|h(X)|) < \infty$, Fubini's Theorem says that you are free to exchange order of summation. Is the result of this problem also true for continuous random variables? Why or why not?

A.1.9) Let X and Y be continuous random variables. Prove

$$P(X \in A) = \int_{-\infty}^{\infty} P(X \in A|Y = y)\, f_{_Y}(y)\, dy.$$

This is sometimes called the *Law of Total Probability*. Is it also true for discrete random variables? Why or why not? Hint: use indicator functions.

A.1.10) Let X and Y be continuous random variables. Prove that if X and Y are independent, $E(XY) = E(X)E(Y)$. Draw an arrow to the place in your answer where you use independence, and write "This is where I use independence."

A.1.11) Let X and Y be *discrete* random variables. Prove that if X and Y are independent, $E(XY) = E(X)E(Y)$. Draw an arrow to the place in your answer where you use independence, and write "This is where I use independence."

A.1.12) Let $P(X = 0) = \frac{1}{2}$ and $P(X = -1) = P(X = 1) = \frac{1}{2}$, and let $Y = X^2$.

(a) Find $Cov(X, Y)$.

(b) Are X and Y independent? Answer Yes or No and prove your answer.

(c) Does zero covariance necessarily imply independence? Answer Yes or No.

Below the line, please use only expected value signs, not integrals or summation.

A.1.13) Show that $Cov[X, Y] = E[XY] - \mu_{_X}\mu_{_Y}$.

A.1.14) Show that if the random variables X and Y are independent, $Cov(X, Y) = 0$.

A.1.15) Show that $Var(X) = E[X^2] - \mu_{_X}^2$.

A.1.16) In the following, X and Y are random variables, while a and b are fixed constants. For each pair of statements below, one is true and one is false (that is, not true in general). State which one is true, and prove it. Zero marks if you prove both statements are true, even if one of the proofs is correct.

(a) $Var(aX) = aVar(X)$, or $Var(aX) = a^2Var(X)$.

(b) $Var(aX + b) = a^2Var(X) + b^2$, or $Var(aX + b) = a^2Var(X)$.

(c) $Var(a) = 0$, or $Var(a) = a^2$.

(d) $Cov(aX, bY) = ab\,Cov(X, Y)$, or $Cov(aX, bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$.

(e) $Cov(X + a, Y + b) = Cov(X, Y) + ab$, or $Cov(X + a, Y + b) = Cov(X, Y)$.

(f) $Var(aX + bY) = a^2Var(X) + b^2Var(Y)$, or $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$.

A.1.17) Let X and Y be random variables with $Cov(X, Y) = \sigma_{xy}$, while a and b are fixed constants.

(a) Find $Cov(X + a, Y + b)$

(b) Find $Cov(aX, bY)$.

A.1.18) Let $Y_1, \ldots, Y_n$ be numbers, and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Show

(a) $\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0$

(b) $\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$

(c) The sum of squares $Q_m = \sum_{i=1}^{n}(Y_i - m)^2$ is minimized when $m = \overline{Y}$.

A.1.19) Let $X_1, \ldots, X_n$ be random variables, let $a_1, \ldots, a_n$ be constants, and let $Y = \sum_{i=1}^{n} a_i X_i$. Derive a general formula for $Var(Y)$. Show your work. Now give the useful special case that applies when $X_1, \ldots, X_n$ are independent.

A.1.20) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables (the standard model of a random sample with replacement). Denoting $E(X_i)$ by μ and $Var(X_i)$ by σ^2,

(a) Show $E(\overline{X}) = \mu$; that is, the sample mean is unbiased for μ.

(b) Find $Var(\overline{X})$.

(c) Letting $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sigma^2$, show that $E(S^2) = \sigma^2$. That is, the sample variance is an unbiased estimator of the population variance. Consider adding and subtracting μ.

A.1.21) Let $Y_1, \ldots, Y_n$ be independent random variables with $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$ for $i = 1, \ldots, n$. For this question, please use definitions and familiar properties of expected value, not integrals.

(a) Find $E(\sum_{i=1}^{n} Y_i)$.

(b) Find $Var\left(\sum_{i=1}^{n} Y_i\right)$. Show your work. Draw an arrow to the place in your answer where you use independence, and write "This is where I use independence."

(c) Using your answer to the last question, find $Var(\overline{Y})$.

(d) A statistic T is an *unbiased estimator* of a parameter θ if $E(T) = \theta$. Show that $\overline{Y}$ is an unbiased estimator of μ. This is very quick.

(e) Let $a_1, \ldots, a_n$ be constants and define the linear combination L by $L = \sum_{i=1}^{n} a_i Y_i$. What condition on the a_i values makes L an unbiased estimator of μ?

(f) Is $\overline{Y}$ a special case of L? If so, what are the a_i values?

(g) What is $Var(L)$?

22. In this regression model, the explanatory variables are random. Independently for $i = 1, \ldots, n$, let $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i$, where $E(X_{i,1}) = \mu_1$, $E(X_{i,2}) = \mu_2$, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, ϵ_i is independent of both $X_{i,1}$ and $X_{i,2}$, and

$$cov\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$$

(a) What is $Var(Y_i)$? You may be able to just write down the answer.

(b) What is $Cov(X_{i,1}, Y_i)$? Show your work.

(c) What is $Cov(X_{i,2}, Y_i)$?

A.2 Matrix Calculations

Basic definitions

A matrix is a rectangular array of numbers. They are usually denoted by boldface letters like $\mathbf{A}$, while scalars (1×1 matrices) are lower case in italics, like a, b, c. Matrices are also written by giving their (i, j) element in brackets, like $\mathbf{A} = [a_{i,j}]$.

Let $\mathbf{A} = [a_{i,j}]$ and $\mathbf{B} = [b_{i,j}]$ be $n \times p$ matrices of constants, $\mathbf{C} = [c_{i,j}]$ be $p \times q$, and let u and v be scalars (1×1 matrices). Define

Matrix addition: $\mathbf{A} + \mathbf{B} = [a_{i,j} + b_{i,j}]$. The matrices must have the same number of rows and the same number of columns for addition (or subtraction) to be defined.

Matrix multiplication: $\mathbf{AC} = [\sum_{k=1}^{p} a_{i,k} c_{k,j}]$. Each element of $\mathbf{AC}$ is the inner product of a row of $\mathbf{A}$ and a column of $\mathbf{C}$. Thus, the number of columns in $\mathbf{A}$ must equal the number of rows in $\mathbf{C}$. Even if $q = n$ so that multiplication in both orders is well defined, in general $\mathbf{AC} \neq \mathbf{CA}$.

Scalar multiplication: $u\,\mathbf{A} = [u \cdot a_{i,j}]$

Transposition: $\mathbf{A}^\top = [a_{j,i}]$

Symmetric matrix: A square matrix $\mathbf{D}$ is said to be *symmetric* if $\mathbf{D} = \mathbf{D}^\top$.

Identity matrix: $\mathbf{I}$ is a square matrix with ones on the main diagonal and zeros elsewhere. $\mathbf{IC} = \mathbf{C}$ and $\mathbf{AI} = \mathbf{A}$.

Diagonal matrix: A square matrix $\mathbf{D} = [d_{i,j}]$ is said to be *diagonal* if $d_{i,j} = 0$ for $i \neq j$.

Triangular matrix: A square matrix $\mathbf{D} = [d_{i,j}]$ is said to be *triangular* if $d_{i,j} = 0$ for $i < j$ or $i > j$ (or both, in which case it is also diagonal).

Distributive laws for matrix and scalar multiplication are easy to establish and are left as exercises.

Transpose of a product

The transpose of a product is the product of transposes, in the reverse order: $(\mathbf{AC})^\top = \mathbf{C}^\top\mathbf{A}^\top$.

Linear independence

The idea behind linear independence of a collection of vectors (say, the columns of a matrix) is that none of them can be written as a linear combination of the others. Formally, let $\mathbf{X}$ be an $n \times p$ matrix of constants. The columns of $\mathbf{X}$ are said to be *linearly dependent* if there exists a $p \times 1$ matrix $\mathbf{v} \neq \mathbf{0}$ with $\mathbf{Xv} = \mathbf{0}$. We will say that the columns of $\mathbf{X}$ are linearly *independent* if $\mathbf{Xv} = \mathbf{0}$ implies $\mathbf{v} = \mathbf{0}$.

Row and column rank

The *row rank* of a matrix is the number of linearly independent rows. The *column rank* is the number of linearly independent columns. The rank of a matrix is the minimum of the row rank and the column rank. Thus, the rank of a matrix cannot exceed the minimum of the number of rows and the number of columns.

Matrix Inverse

Let $\mathbf{A}$ and $\mathbf{B}$ be square matrices of the same size. $\mathbf{B}$ is said to be the *inverse* of $\mathbf{A}$ and may be written $\mathbf{B} = \mathbf{A}^{-1}$. The definition is $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$. Thus, there are always two equalities to establish when you are showing that one matrix is the inverse of another. Matrix inverses have the following properties, which may be proved as exercises.

- If a matrix inverse exists, it is unique.
- $\mathbf{A}^{-1\top} = \mathbf{A}^{\top-1}$

- If the scalar $u \neq 0$, $(u\mathbf{A})^{-1} = \frac{1}{u}\mathbf{A}^{-1}$.
- Suppose that the square matrices $\mathbf{A}$ and $\mathbf{B}$ both have inverses. Then $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
- If $\mathbf{A}$ is a $p \times p$ matrix, $\mathbf{A}^{-1}$ exists if and only if the rank of $\mathbf{A}$ equals p.

Sometimes the following formula for the inverse of a 2×2 matrix is useful:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{A.4}$$

In some cases the inverse of the matrix is its transpose. When $\mathbf{A}^\top = \mathbf{A}^{-1}$, the matrix $\mathbf{A}$ is said to be *orthogonal*, because the column (row) vectors are all at right angles (zero inner product). In addition, they all have length one, because the inner product of each column (row) with itself equals one.

Positive definite matrices

The $n \times n$ matrix $\mathbf{A}$ is said to be *positive definite* if

$$\mathbf{v}^\top \mathbf{A}\mathbf{v} > 0 \tag{A.5}$$

for *all* $n \times 1$ vectors $\mathbf{v} \neq \mathbf{0}$. It is called *non-negative definite* (or sometimes positive semi-definite) if $\mathbf{v}^\top \mathbf{A}\mathbf{v} \geq 0$. Positive definiteness is a critical property of variance-covariance matrices, because it says that the variance of any linear combination is greater than zero. See (A.14) on page 130.

Determinants

Let $\mathbf{A} = [a_{i,j}]$ be an $n \times n$ matrix, so that the following applies to square matrices. The *determinant* of $\mathbf{A}$, denoted $|\mathbf{A}|$, is defined as a *sum of signed elementary products*. An elementary product is a product of elements of $\mathbf{A}$ such that there is exactly one element from every row and every column. The "signed" part is determined as follows.

Let S_n denote the set of all permutations of the set $\{1, \ldots, n\}$, and denote such a permutation by $\sigma = (\sigma_1, \ldots, \sigma_n)$. Each permutation may be obtained from $(1, \ldots, n)$ by a finite number of switches of numbers. If the number of switches required is even (this includes zero), let $\text{sgn}(\sigma) = +1$; if it is odd, let $\text{sgn}(\sigma) = -1$. Then,

$$|\mathbf{A}| = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma_i}. \tag{A.6}$$

Some properties of determinants are:

- $|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}|$
- $|\mathbf{A}^\top| = |\mathbf{A}|$

- $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$, and if $|\mathbf{A}| = 0$, $\mathbf{A}^{-1}$ does not exist.
- If $\mathbf{A} = [a_{i,j}]$ is triangular, $|\mathbf{A}| = \prod_{i=1}^{n} a_{i,i}$. That is, for triangular (including diagonal) matrices, the determinant is the product of the elements on the main diagonal.
- Adding a multiple of one row to another row of a matrix, or adding a multiple of a column to another column leaves the determinant unchanged.
- Exchanging any two rows or any two columns of a matrix multiplies the determinant by -1.
- Multiplying a single row or column by a constant multiplies the determinant by that constant, so that $|v\mathbf{A}| = v^n|\mathbf{A}|$

Eigenvalues and eigenvectors

Let $\mathbf{A} = [a_{i,j}]$ be an $n \times n$ matrix, so that the following applies to square matrices. $\mathbf{A}$ is said to have an *eigenvalue* λ and (non-zero) *eigenvector* $\mathbf{x}$ corresponding to λ if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \tag{A.7}$$

Note that λ is a scalar and $\mathbf{x} \neq \mathbf{0}$ is an $n \times 1$ matrix, typically chosen so that it has length one. It is also possible and desirable to choose the eigenvectors so they are mutually perpendicular (the inner product of any two equals zero).

To solve the eigenvalue equation, write

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \Rightarrow \mathbf{A}\mathbf{x} - \lambda\mathbf{x} = \mathbf{A}\mathbf{x} - \lambda\mathbf{I}\mathbf{x} = (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.$$

If $(\mathbf{A} - \lambda\mathbf{I})^{-1}$ existed, it would be possible to solve for $\mathbf{x}$ by multiplying both sides on the left by $(\mathbf{A} - \lambda\mathbf{I})^{-1}$, yielding $\mathbf{x} = \mathbf{0}$. But the definition specifies $\mathbf{x} \neq \mathbf{0}$, so the inverse cannot exist for the definition of an eigenvalue to be satisfied. Since $(\mathbf{A} - \lambda\mathbf{I})^{-1}$ fails to exist precisely when the determinant $|\mathbf{A} - \lambda\mathbf{I}| = 0$, the eigenvalues are the λ values that solve the determinantal equation

$$|\mathbf{A} - \lambda\mathbf{I}| = 0.$$

The left-hand side is a polynomial in λ, called the *characteristic polynomial*. If the matrix $\mathbf{A}$ is real-valued and also symmetric, then all its eigenvalues are guaranteed to be real-valued — a handy characteristic not generally true of solutions to polynomial equations. The eigenvectors can also be chosen to be real, and for our purposes they always will be.

One of the many useful properties of eigenvalues is that **the determinant is the product of the eigenvalues**:

$$|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i$$

Spectral decomposition of symmetric matrices

The *Spectral decomposition theorem* says that every square and symmetric matrix $\mathbf{A} = [a_{i,j}]$ may be written

$$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top, \tag{A.8}$$

where the columns of $\mathbf{P}$ (which may also be denoted $\mathbf{x}_1, \ldots, \mathbf{x}_1$) are the eigenvectors of $\mathbf{A}$, and the diagonal matrix $\mathbf{\Lambda}$ contains the corresponding eigenvalues.

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

Because the eigenvectors are orthonormal, $\mathbf{P}$ is an orthogonal matrix; that is, $\mathbf{P}\mathbf{P}^\top = \mathbf{P}^\top\mathbf{P} = \mathbf{I}$.

The following shows how to get a spectral decomposition from *R*.

```
> help(eigen)
> A = rbind(c(-10,2),
+           c(2,5)) # Symmetric
> eigenA = eigen(A); eigenA
$values
[1]   5.262087 -10.262087

$vectors
          [,1]       [,2]
[1,] 0.1299328  0.9915228
[2,] 0.9915228 -0.1299328

> det(A)
[1] -54
> prod(eigenA$values)
[1] -54

> Lambda = diag(eigenA$values); Lambda
         [,1]      [,2]
[1,] 5.262087   0.00000
[2,] 0.000000 -10.26209

> P = eigenA$vectors; P
          [,1]       [,2]
[1,] 0.1299328  0.9915228
[2,] 0.9915228 -0.1299328
```

```
> P %*% Lambda %*% t(P) # Matrix multiplication
     [,1] [,2]
[1,]  -10    2
[2,]    2    5
```

Another way to express the spectral decomposition is

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^\top, \tag{A.9}$$

where again, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are the eigenvectors of $\mathbf{A}$, and $\lambda_1, \ldots, \lambda_n$ are the corresponding eigenvalues. It's a weighted sum of outer (not inner) products of the eigenvectors; the weights are the eigenvalues.

Continuing the *R* example, here is $\mathbf{x}_1\mathbf{x}_1^\top$. Notice how the diagonal elements add to one, as they must.

```
> eigenA$vectors[,1] %*% t(eigenA$vectors[,1])
           [,1]      [,2]
[1,] 0.01688253 0.1288313
[2,] 0.12883133 0.9831175
```

Reproducing (A.9) for completeness,

```
> prod1 = eigenA$vectors[,1] %*% t(eigenA$vectors[,1])
> prod2 = eigenA$vectors[,2] %*% t(eigenA$vectors[,2])
> eigenA$values[1]
[1] 5.262087
> eigenA$values[1]*prod1 + eigenA$values[2]*prod2
     [,1] [,2]
[1,]  -10    2
[2,]    2    5
> A
     [,1] [,2]
[1,]  -10    2
[2,]    2    5
```

Real symmetric matrices

For a symmetric $n \times n$ matrix $\mathbf{A}$, the eigenvalues are all real numbers, and the eigenvectors can be chosen to be real, perpendicular (inner product zero), and of length one. If a real symmetric matrix is also non-negative definite, as a variance-covariance matrix must be, the following conditions are equivalent:

- Rows linearly independent
- Columns linearly independent

- Rank $= n$
- Positive definite
- Non-singular ($\mathbf{A}^{-1}$ exists)
- All eigenvalues are strictly positive

Trace of a square matrix

The *trace* of a square matrix $\mathbf{A} = [a_{i,j}]$ is the sum of its diagonal elements. Write

$$tr(\mathbf{A}) = \sum_{i=1}^{n} a_{i,i}.$$

Properties like $tr(\mathbf{A}+\mathbf{B}) = tr(\mathbf{A})+tr(\mathbf{B})$ follow immediately from the definition. Perhaps less obvious is the following. Let $\mathbf{A}$ be an $r \times p$ matrix and $\mathbf{B}$ be a $p \times r$ matrix, so that the product matrices $\mathbf{AB}$ and $\mathbf{BA}$ are both defined. These two matrices are not necessarily equal; in fact, they need not even be the same size. But still,

$$tr(\mathbf{AB}) = tr(\mathbf{BA}). \tag{A.10}$$

To see this, write

$$\begin{aligned}
tr(\mathbf{AB}) &= tr\left(\left(\sum_{k=1}^{p} a_{i,k}b_{k,j}\right)\right) \\
&= \sum_{i=1}^{r}\sum_{k=1}^{p} a_{i,k}b_{k,i} \\
&= \sum_{k=1}^{p}\sum_{i=1}^{r} b_{k,i}a_{i,k} \\
&= \sum_{i=1}^{p}\sum_{k=1}^{r} b_{i,k}a_{k,i} \quad \text{(Switching } i \text{ and } k\text{)} \\
&= tr\left(\left(\sum_{k=1}^{r} b_{i,k}a_{k,j}\right)\right) \\
&= tr(\mathbf{BA})
\end{aligned}$$

Notice how the indices of summation i and k have been changed. This is legitimate, because for example $\sum_{i=1}^{r} c_i$ and $\sum_{k=1}^{r} c_k$ both mean $c_1 + \cdots + c_r$.

Also, from the spectral decomposition (A.9), the trace is the sum of the eigenvalues:

$$tr(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i.$$

This follows easily using (A.10), but actually it applies to *any* square matrix; the matrix need not be symmetric.

The *vech* notation

Sometimes, it is helpful to represent the non-redundant elements of a symmetric matrix in the form of a column vector. Let $\mathbf{A} = [a_{i,j}]$ be an $n \times n$ symmetric matrix. $\mathbf{A}$ has $\frac{n(n+1)}{2}$ non-redundant elements: say the main diagonal plus the upper triangular half. Then

$$vech(\mathbf{A}) = \begin{pmatrix} a_{1,1} \\ \vdots \\ a_{1,n} \\ a_{2,2} \\ \vdots \\ a_{2,n} \\ \vdots \\ a_{n,n} \end{pmatrix}.$$

The *vech* operation is distributive: $vech(A + B) = vech(A) + vech(B)$.

Exercises A.2

A.2.1) Which statement is true?

(a) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$

(b) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{BA} + \mathbf{CA}$

(c) Both a and b

(d) Neither a nor b

A.2.2) Which statement is true?

(a) $a(\mathbf{B} + \mathbf{C}) = a\mathbf{B} + a\mathbf{C}$

(b) $a(\mathbf{B} + \mathbf{C}) = \mathbf{B}a + \mathbf{C}a$

(c) Both a and b

(d) Neither a nor b

A.2.3) Which statement is true?

(a) $(\mathbf{B} + \mathbf{C})\mathbf{A} = \mathbf{AB} + \mathbf{AC}$

(b) $(\mathbf{B} + \mathbf{C})\mathbf{A} = \mathbf{BA} + \mathbf{CA}$

(c) Both a and b

(d) Neither a nor b

A.2.4) Which statement is true?

(a) $(\mathbf{AB})^\top = \mathbf{A}^\top\mathbf{B}^\top$

(b) $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$

(c) Both a and b

(d) Neither a nor b

A.2.5) Which statement is true?

(a) $\mathbf{A}^{\top\top} = \mathbf{A}$

(b) $\mathbf{A}^{\top\top\top} = \mathbf{A}^\top$

(c) Both a and b

(d) Neither a nor b

A.2.6) Suppose that the square matrices $\mathbf{A}$ and $\mathbf{B}$ both have inverses. Which statement is true?

(a) $(\mathbf{AB})^{-1} = \mathbf{A}^{-1}\mathbf{B}^{-1}$

(b) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

(c) Both a and b

(d) Neither a nor b

A.2.7) Which statement is true?

(a) $(\mathbf{A}+\mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$

(b) $(\mathbf{A}+\mathbf{B})^\top = \mathbf{B}^\top + \mathbf{A}^\top$

(c) $(\mathbf{A}+\mathbf{B})^\top = (\mathbf{B}+\mathbf{A})^\top$

(d) All of the above

(e) None of the above

A.2.8) Which statement is true?

(a) $tr(\mathbf{A}+\mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B})$

(b) $tr(\mathbf{A}+\mathbf{B}) = tr(\mathbf{B}) + tr(\mathbf{A})$

(c) Both a and b

(d) Neither a nor b

A.2.9) Which statement is true?

(a) $a\,tr(\mathbf{B}) = tr(a\mathbf{B})$.

(b) $tr(\mathbf{B})a = tr(a\mathbf{B})$

(c) Both a and b

(d) Neither a nor b

A.2.10) Which statement is true?

(a) $(a+b)\mathbf{C} = a\mathbf{C} + b\mathbf{C}$

(b) $(a+b)\mathbf{C} = \mathbf{C}a + \mathbf{C}b$

(c) $(a+b)\mathbf{C} = \mathbf{C}(a+b)$

(d) All of the above

(e) None of the above

A.2.11) Let $\mathbf{A}$ and $\mathbf{B}$ be 2×2 matrices. Either

- Prove $\mathbf{AB} = \mathbf{BA}$, or
- Give a numerical example in which $\mathbf{AB} \neq \mathbf{BA}$

A.2.12) Recall that $\mathbf{A}$ symmetric means $\mathbf{A} = \mathbf{A}^\top$. Let $\mathbf{X}$ be an n by p matrix. Prove that $\mathbf{X}^\top\mathbf{X}$ is symmetric.

A.2.13) Recall that an inverse of the matrix $\mathbf{A}$ (denoted $\mathbf{A}^{-1}$) is defined by two properties: $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ and $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. Prove that inverses are unique, as follows. Let $\mathbf{B}$ and $\mathbf{C}$ both be inverses of $\mathbf{A}$. Show that $\mathbf{B} = \mathbf{C}$.

A.2.14) Let $\mathbf{X}$ be an n by p matrix with $n \neq p$. Why is it incorrect to say that $(\mathbf{X}^\top\mathbf{X})^{-1} = \mathbf{X}^{-1}\mathbf{X}^{\top-1}$?

A.2.15) Suppose that the square matrices $\mathbf{A}$ and $\mathbf{B}$ both have inverses. Prove that $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. You have two things to show.

A.2.16) Let $\mathbf{A}$ be a non-singular square matrix. Prove $(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}$.

A.2.17) Using $(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}$, prove that the inverse of a symmetric matrix is also symmetric.

A.2.18) Let $\mathbf{A}$ be a square matrix with the determinant of $\mathbf{A}$ (denoted $|\mathbf{A}|$) equal to zero. What does this tell you about $\mathbf{A}^{-1}$? No proof is necessary here.

A.2.19) Let $\mathbf{a}$ be an $n \times 1$ matrix of constants. How do you know $\mathbf{a}^\top\mathbf{a} \geq 0$?

A.2.20) Let $\mathbf{A}$ be an $n \times p$ matrix of constants. Is it true that $\mathbf{A}^\top\mathbf{A} \geq 0$? Briefly explain.

A.2.21) Let $\mathbf{X}$ be an $n \times p$ matrix of constants. Recall the definition of linear independence. The columns of $\mathbf{X}$ are said to be *linearly dependent* if there exists $\mathbf{v} \neq \mathbf{0}$ with $\mathbf{Xv} = \mathbf{0}$. We will say that the columns of $\mathbf{X}$ are linearly *independent* if $\mathbf{Xv} = \mathbf{0}$ implies $\mathbf{v} = \mathbf{0}$.

(a) Show that if the columns of $\mathbf{X}$ are linearly dependent, then the columns of $\mathbf{X}^\top\mathbf{X}$ are also linearly dependent.

(b) Show that if the columns of $\mathbf{X}$ are linearly dependent, then the *rows* of $\mathbf{X}^\top\mathbf{X}$ are linearly dependent.

(c) Show that if the columns of $\mathbf{X}$ are linearly independent, then the columns of $\mathbf{X}^\top\mathbf{X}$ are also linearly independent. Use $\mathbf{a}^\top\mathbf{a} \geq 0$ the definition of linear independence.

(d) Show that if $(\mathbf{X}^\top\mathbf{X})^{-1}$ exists, then the columns of $\mathbf{X}$ are linearly independent.

(e) Show that if the columns of $\mathbf{X}$ are linearly independent, then $\mathbf{X}^\top\mathbf{X}$ is positive definite. Does this imply the existence of $(\mathbf{X}^\top\mathbf{X})^{-1}$? Locate the rule in the text, and answer Yes or No.

A.2.22) Let $\mathbf{A}$ be a square matrix. Show that

(a) If $\mathbf{A}^{-1}$ exists, the columns of $\mathbf{A}$ are linearly independent.

(b) If the columns of $\mathbf{A}$ are linearly dependent, $\mathbf{A}^{-1}$ cannot exist.

A.2.23) In the following, $\mathbf{A}$ and $\mathbf{B}$ are $n \times p$ matrices of constants, $\mathbf{C}$ is $p \times q$, $\mathbf{D}$ is $p \times n$ and a, b, c are scalars. For each statement below, either prove it is true, or prove that it is not true in general by giving a counter-example. Small numerical counter-examples are best. To give an idea of the kind of proof required for most of these, denote element (i, j) of matrix $\mathbf{A}$ by $[a_{i,j}]$.

(a) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$

(b) $a(\mathbf{B} + \mathbf{C}) = a\mathbf{B} + a\mathbf{C}$

(c) $\mathbf{AC} = \mathbf{CA}$

(d) $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$

(e) $(\mathbf{AC})^\top = \mathbf{C}^\top\mathbf{A}^\top$

(f) $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$

(g) $(\mathbf{AD})^{-1} = \mathbf{A}^{-1}\mathbf{D}^{-1}$

A.2.24) Let $\mathbf{A}$ be a square symmetric matrix, and $\mathbf{A}^{-1}$ exists. Show that $\mathbf{A}^{-1}$ is also symmetric.

A.2.25) The *trace* of a square matrix is the sum of its diagonal elements; we write $tr(\mathbf{A})$. Let $\mathbf{A}$ be $r \times c$ and $\mathbf{B}$ be $c \times r$. Show $tr(\mathbf{AB}) = tr(\mathbf{BA})$.

A.2.26) Recall the *spectral decomposition* of a square symmetric matrix (For example, a variance-covariance matrix). Any such matrix $\mathbf{\Sigma}$ can be written as $\mathbf{\Sigma} = \mathbf{P\Lambda P}^\top$, where $\mathbf{P}$ is a matrix whose columns are the (orthonormal) eigenvectors of $\mathbf{\Sigma}$, $\mathbf{\Lambda}$ is a diagonal matrix of the corresponding (non-negative) eigenvalues, and $\mathbf{P}^\top\mathbf{P} = \mathbf{PP}^\top = \mathbf{I}$.

(a) Let $\mathbf{\Sigma}$ be a square symmetric matrix with eigenvalues that are all strictly positive.

i. What is $\mathbf{\Lambda}^{-1}$?

ii. Show $\mathbf{\Sigma}^{-1} = \mathbf{P\Lambda}^{-1}\mathbf{P}^\top$

(b) Let $\boldsymbol{\Sigma}$ be a square symmetric matrix, and this time some of the eigenvalues might be zero.

i. What do you think $\boldsymbol{\Lambda}^{1/2}$ might be?

ii. Define $\boldsymbol{\Sigma}^{1/2}$ as $\mathbf{P}\boldsymbol{\Lambda}^{1/2}\mathbf{P}^\top$. Show $\boldsymbol{\Sigma}^{1/2}$ is symmetric.

iii. Show $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$.

iv. Show that if the columns of $\boldsymbol{\Sigma}$ are linearly independent, then the columns of $\boldsymbol{\Sigma}^{1/2}$ are also linearly independent. See Question A.2 for the definition of linear independence.

(c) Show that if the symmetric matrix $\boldsymbol{\Sigma}$ is positive definite, then $\boldsymbol{\Sigma}^{-1}$ is also positive definite.

(d) Now return to the situation where the eigenvalues of the square symmetric matrix $\boldsymbol{\Sigma}$ are all strictly positive. Define $\boldsymbol{\Sigma}^{-1/2}$ as $\mathbf{P}\boldsymbol{\Lambda}^{-1/2}\mathbf{P}^\top$, where the elements of the diagonal matrix $\boldsymbol{\Lambda}^{-1/2}$ are the reciprocals of the corresponding elements of $\boldsymbol{\Lambda}^{1/2}$.

i. Show that the inverse of $\boldsymbol{\Sigma}^{1/2}$ is $\boldsymbol{\Sigma}^{-1/2}$, justifying the notation.

ii. Show $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$.

(e) The (square) matrix $\boldsymbol{\Sigma}$ is said to be *positive definite* if $\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a} > 0$ for all vectors $\mathbf{a} \neq \mathbf{0}$. Show that the eigenvalues of a symmetric positive definite matrix are all strictly positive. Hint: the $\mathbf{a}$ you want is an eigenvector.

(f) Let $\boldsymbol{\Sigma}$ be a symmetric, positive definite matrix. Putting together a couple of results you have proved above, establish that $\boldsymbol{\Sigma}^{-1}$ exists.

A.2.27) Using the spectral decomposition (A.9) and $tr(\mathbf{AB}) = tr(\mathbf{BA})$, show that the trace of a square symmetric matrix is the sum of its eigenvalues.

A.3 Random Vectors and Matrices

A *random matrix* is just a matrix of random variables. Their joint probability distribution is the distribution of the random matrix. Random matrices with just one column (say, p) may be called *random vectors*.

Expected Value and Variance-Covariance

Expected Value

The expected value of a matrix is defined as the matrix of expected values. Denoting the $p \times c$ random matrix $\mathbf{X}$ by $[X_{i,j}]$,

$$E(\mathbf{X}) = [E(X_{i,j})].$$

Immediately we have natural properties like

$$\begin{aligned} E(\mathbf{X}+\mathbf{Y}) &= E([X_{i,j}+Y_{i,j}]) \\ &= [E(X_{i,j}+Y_{i,j})] \\ &= [E(X_{i,j})+E(Y_{i,j})] \\ &= [E(X_{i,j})]+[E(Y_{i,j})] \\ &= E(\mathbf{X})+E(\mathbf{Y}). \end{aligned}$$

Let $\mathbf{A}=[a_{i,j}]$ be an $r\times p$ matrix of constants, while $\mathbf{X}$ is still a $p\times c$ random matrix. Then

$$\begin{aligned} E(\mathbf{AX}) &= E\left(\left(\sum_{k=1}^{p} a_{i,k}X_{k,j}\right)\right) \\ &= \left(E\left(\sum_{k=1}^{p} a_{i,k}X_{k,j}\right)\right) \\ &= \left(\sum_{k=1}^{p} a_{i,k}E(X_{k,j})\right) \\ &= \mathbf{A}E(\mathbf{X}). \end{aligned}$$

Similar calculations yield $E(\mathbf{XB})=E(\mathbf{X})\mathbf{B}$, where $\mathbf{B}$ is a matrix of constants. This yields the useful formula

$$E(\mathbf{AXB})=\mathbf{A}E(\mathbf{X})\mathbf{B}. \tag{A.11}$$

Variance-Covariance Matrices

Let $\mathbf{X}$ be a $p\times 1$ random vector with $E(\mathbf{X})=\boldsymbol{\mu}$. The *variance-covariance matrix* of $\mathbf{X}$ (sometimes just called the *covariance matrix*), denoted by $V(\mathbf{X})$, is defined as

$$V(\mathbf{X})=E\left\{(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\top}\right\}. \tag{A.12}$$

The covariance matrix $V(\mathbf{X})$ is a $p \times p$ matrix of constants. To see exactly what it is, suppose $p = 3$. Then

$$\begin{aligned}
V(\mathbf{X}) &= E\left\{ \begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ X_3-\mu_3 \end{pmatrix} \begin{pmatrix} X_1-\mu_1 & X_2-\mu_2 & X_3-\mu_3 \end{pmatrix} \right\} \\
&= E\left\{ \begin{pmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & (X_1-\mu_1)(X_3-\mu_3) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & (X_2-\mu_2)(X_3-\mu_3) \\ (X_3-\mu_3)(X_1-\mu_1) & (X_3-\mu_3)(X_2-\mu_2) & (X_3-\mu_3)^2 \end{pmatrix} \right\} \\
&= \begin{pmatrix} E\{(X_1-\mu_1)^2\} & E\{(X_1-\mu_1)(X_2-\mu_2)\} & E\{(X_1-\mu_1)(X_3-\mu_3)\} \\ E\{(X_2-\mu_2)(X_1-\mu_1)\} & E\{(X_2-\mu_2)^2\} & E\{(X_2-\mu_2)(X_3-\mu_3)\} \\ E\{(X_3-\mu_3)(X_1-\mu_1)\} & E\{(X_3-\mu_3)(X_2-\mu_2)\} & E\{(X_3-\mu_3)^2\} \end{pmatrix} \\
&= \begin{pmatrix} V(X_1) & Cov(X_1,X_2) & Cov(X_1,X_3) \\ Cov(X_1,X_2) & V(X_2) & Cov(X_2,X_3) \\ Cov(X_1,X_3) & Cov(X_2,X_3) & V(X_3) \end{pmatrix}.
\end{aligned}$$

So, the covariance matrix $V(\mathbf{X})$ is a $p \times p$ symmetric matrix with variances on the main diagonal and covariances on the off-diagonals.

The matrix of covariances between two random vectors may also be written in a convenient way. Let $\mathbf{X}$ be a $p \times 1$ random vector with $E(\mathbf{X}) = \boldsymbol{\mu}_x$ and let $\mathbf{Y}$ be a $q \times 1$ random vector with $E(\mathbf{Y}) = \boldsymbol{\mu}_y$. The $p \times q$ matrix of covariances between the elements of $\mathbf{X}$ and the elements of $\mathbf{Y}$ is

$$C(\mathbf{X}, \mathbf{Y}) = E\left\{(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{Y} - \boldsymbol{\mu}_y)^\top\right\}. \tag{A.13}$$

The following rule is analogous to $Var(a\,X) = a^2\,Var(X)$ for scalars. Let $\mathbf{X}$ be a $p \times 1$ random vector with $E(\mathbf{X}) = \boldsymbol{\mu}$ and $V(\mathbf{X}) = \boldsymbol{\Sigma}$, while $\mathbf{A} = [a_{i,j}]$ is an $r \times p$ matrix of constants. Then

$$\begin{aligned}
V(\mathbf{AX}) &= E\left\{(\mathbf{AX} - \mathbf{A}\boldsymbol{\mu})(\mathbf{AX} - \mathbf{A}\boldsymbol{\mu})^\top\right\} \\
&= E\left\{\mathbf{A}(\mathbf{X} - \boldsymbol{\mu})\left(\mathbf{A}(\mathbf{X} - \boldsymbol{\mu})\right)^\top\right\} \\
&= E\left\{\mathbf{A}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top\mathbf{A}^\top\right\} \\
&= \mathbf{A}E\{(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top\}\mathbf{A}^\top \\
&= \mathbf{A}V(\mathbf{X})\mathbf{A}^\top \\
&= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top
\end{aligned} \tag{A.14}$$

For scalars, $Var(X + b) = Var(X)$, and the same applies to vectors. Covariances are also unaffected by adding a constant; this amounts to shifting the whole joint distribution by a fixed amount, which has no effect on relationships among variables. So, the following rule is "obvious." Let $\mathbf{X}$ be a $p \times 1$ random vector with $E(\mathbf{X}) = \boldsymbol{\mu}$ and let $\mathbf{b}$ be a $p \times 1$ vector of constants. Then $V(\mathbf{X} + \mathbf{b}) = V(\mathbf{X})$. To see this, note $E(\mathbf{X} + \mathbf{b}) = \boldsymbol{\mu} + \mathbf{b}$ and write

$$\begin{aligned} V(\mathbf{X}+\mathbf{b}) &= E\left\{(\mathbf{X}+\mathbf{b}-(\boldsymbol{\mu}+\mathbf{b}))(\mathbf{X}+\mathbf{b}-(\boldsymbol{\mu}+\mathbf{b}))^\top\right\} \\ &= E\left\{(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^\top\right\} \\ &= V(\mathbf{X}) \end{aligned} \tag{A.15}$$

A similar rule apples to $C(\mathbf{X}+\mathbf{b}, \mathbf{Y}+\mathbf{c})$. A direct calculation is not even necessary, though it is a valuable exercise. Think of stacking $\mathbf{X}$ and $\mathbf{Y}$ one on top of another, to form a bigger random vector. Then,

$$V\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} = \left(\begin{array}{c|c} V(\mathbf{X}) & C(\mathbf{X},\mathbf{Y}) \\ \hline C(\mathbf{X},\mathbf{Y})^\top & V(\mathbf{Y}) \end{array}\right).$$

This is an example of a *partitioned matrix* – a matrix of matrices. At any rate, it is clear from (A.15) that adding a stack of constant vectors to the stack of random vectors has no effect upon the (partitioned) covariance matrix, and in particular no effect upon $C(\mathbf{X}, \mathbf{Y})$.

The Centering Rule

Often, variance and covariance calculations can be simplified by subtracting off expected values first. Letting $E(\mathbf{X}) = \boldsymbol{\mu}_x$, denote the *centered* version of $\mathbf{X}$ by $\overset{c}{\mathbf{X}} = \mathbf{X} - \boldsymbol{\mu}_x$, so that

- $E(\overset{c}{\mathbf{X}}) = \mathbf{0}$
- $V(\mathbf{X}) = E\{(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{X}-\boldsymbol{\mu}_x)^\top\} = E(\overset{c}{\mathbf{X}}\overset{c}{\mathbf{X}}{}^\top)$, and
- $C(\mathbf{X},\mathbf{Y}) = E\left\{(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{Y}-\boldsymbol{\mu}_y)^\top\right\} = E(\overset{c}{\mathbf{X}}\overset{c}{\mathbf{Y}}{}^\top)$

Consider the linear combination $\mathbf{L} = \mathbf{A}_1\mathbf{X}_1 + \cdots + \mathbf{A}_m\mathbf{X}_m + \mathbf{b}$. It may be easily shown that

$$\overset{c}{\mathbf{L}} = \mathbf{A}_1\overset{c}{\mathbf{X}}_1 + \cdots + \mathbf{A}_m\overset{c}{\mathbf{X}}_m\,. \tag{A.16}$$

The Centering Rule says that to calculate variances and covariances of linear combinations, one may simply discard added constants, center all the random vectors, and take expected values of products. Symbolically,

$$V(\mathbf{L}) = E(\overset{c}{\mathbf{L}}\overset{c}{\mathbf{L}}{}^\top) \quad \text{and} \quad C(\mathbf{L}_1,\mathbf{L}_2) = E(\overset{c}{\mathbf{L}}_1\overset{c}{\mathbf{L}}_2{}^\top), \tag{A.17}$$

where the centered linear combinations are given by expressions like (A.16).

To see how useful the Centering Rule can be, $C(\mathbf{AX}+\mathbf{c}, \mathbf{BY}+\mathbf{d})$ will be calculated two ways, first the ordinary way and then with the Centering Rule.

$$\begin{aligned} C(\mathbf{AX}+\mathbf{c},\mathbf{BY}+\mathbf{d}) &= E\{(\mathbf{AX}+\mathbf{c}-(\mathbf{A}\boldsymbol{\mu}_x+\mathbf{c}))(\mathbf{BY}+\mathbf{d}-(\mathbf{B}\boldsymbol{\mu}_y+\mathbf{d}))^\top\} \\ &= E\{(\mathbf{AX}-\mathbf{A}\boldsymbol{\mu}_x)(\mathbf{BY}-\mathbf{B}\boldsymbol{\mu}_y)^\top\} \\ &= E\{\mathbf{A}(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{B}(\mathbf{Y}-\boldsymbol{\mu}_y))^\top\} \\ &= E\{\mathbf{A}(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{Y}-\boldsymbol{\mu}_y)^\top\mathbf{B}^\top\} \\ &= \mathbf{A}E\{(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{Y}-\boldsymbol{\mu}_y)^\top\}\mathbf{B}^\top \\ &= \mathbf{A}C(\mathbf{X},\mathbf{Y})\mathbf{B}^\top \end{aligned}$$

Now with centered variables,

$$\begin{aligned} C(\mathbf{AX}+\mathbf{c},\mathbf{BY}+\mathbf{d}) &= E\{\mathbf{A}\overset{c}{\mathbf{X}}\,(\mathbf{B}\overset{c}{\mathbf{Y}})^\top\} \\ &= E\{\mathbf{A}\,\overset{c}{\mathbf{X}}\overset{c}{\mathbf{Y}}^\top\,\mathbf{B}^\top\} \\ &= \mathbf{A}E\{\overset{c}{\mathbf{X}}\overset{c}{\mathbf{Y}}^\top\}\mathbf{B}^\top \\ &= \mathbf{A}C(\mathbf{X},\mathbf{Y})\mathbf{B}^\top \end{aligned}$$

It is worth pointing out that the centering rule applies to scalar variance-covariance calculations too, since these are special cases of matrix calculations. For example, let $X_1,\ldots,X_n$ be a random sample from a distribution with mean μ and variance σ^2, and consider the task of showing that $Cov(\overline{X},X_j-\overline{X})=0$, which is the key to proving the independence of $\overline{X}$ and S^2 for the normal distribution, and the gateway to the t distribution. Since $\overline{X}$ and $X_j-\overline{X}$ are both linear combinations,

$$\begin{aligned} Cov(\overline{X},X_j-\overline{X}) &= E\left(\overline{\overset{c}{X}}\,(\overset{c}{X}_j-\overline{\overset{c}{X}})\right) \\ &= E\left(\overset{c}{X}_j\overline{\overset{c}{X}}\right)-E\left(\overline{\overset{c}{X}}^2\right) \\ &= E\left(\overset{c}{X}_j\,\frac{1}{n}\sum_{i=1}^n\overset{c}{X}_i\right)-Var(\overline{X}) \\ &= E\left(\frac{1}{n}\sum_{i=1}^n\overset{c}{X}_i\overset{c}{X}_j\right)-\frac{\sigma^2}{n} \\ &= \frac{1}{n}\sum_{i=1}^n E\left(\overset{c}{X}_i\overset{c}{X}_j\right)-\frac{\sigma^2}{n} \\ &= \frac{1}{n}E\left(\overset{c}{X}_j^2\right)+\frac{1}{n}\sum_{i\neq j}E\left(\overset{c}{X}_i\right)E\left(\overset{c}{X}_j\right)-\frac{\sigma^2}{n} \\ &= \frac{1}{n}Var(X_j)+0-\frac{\sigma^2}{n} \\ &= \frac{\sigma^2}{n}-\frac{\sigma^2}{n} \\ &= 0 \end{aligned}$$

This valuable calculation looks worse than it is at first glance, because every little step is shown. It is significantly longer and messier without centering.

The centering notation (which is very non-standard) is helpful at first, but after a while it gets tedious to write the letter c so many times. It is perfectly acceptable to say something like "Assuming the variables have been centered, ..." and just *imagine* the little c characters[3].

[3]Sometimes people write things like "Assuming without loss of generality that all expected values are

Exercises A.3 This exercise set has an unusual feature. *Some of the questions ask you to prove things that are false.* That is, they are not true in general. In such cases, just write "The statement is false," and give a brief explanation to make it clear that you are not just guessing. The explanation is essential for full marks. A small counter-example is always good enough.

A.3.1) Let $\mathbf{X} = [X_j]$ be a random matrix. Show $E(\mathbf{X}^\top) = E(\mathbf{X})^\top$.

A.3.2) Let $\mathbf{X}$ and $\mathbf{Y}$ be random matrices of the same dimensions. Show $E(\mathbf{X} + \mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y})$. Recall the definition $E(\mathbf{Z}) = [E(Z_{i,j})]$.

A.3.3) Let $\mathbf{X}$ be a random matrix, and $\mathbf{B}$ be a matrix of constants. Show $E(\mathbf{XB}) = E(\mathbf{X})\mathbf{B}$. Recall the definition $\mathbf{AB} = [\sum_k a_{i,k}b_{k,j}]$.

A.3.4) Let $\mathbf{X}$ be a $p \times 1$ random vector. Starting with Definition (A.12) on page 129, prove $V(\mathbf{X}) = \mathbf{0}$..

A.3.5) Let the $p \times 1$ random vector $\mathbf{X}$ have expected value $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, and let $\mathbf{A}$ be an $m \times p$ matrix of constants. Prove that the variance-covariance matrix of $\mathbf{AX}$ is either

- $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top$, or
- $\mathbf{A}^2\boldsymbol{\Sigma}$.

Pick one and prove it. Start with the definition of a variance-covariance matrix (A.12) on page 129.

A.3.6) If the $p \times 1$ random vector $\mathbf{X}$ has mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, show $\boldsymbol{\Sigma} = E(\mathbf{X}\mathbf{X}^\top) - \boldsymbol{\mu}\boldsymbol{\mu}^\top$.

A.3.7) Starting with Definition (A.13) on page 130, show $C(\mathbf{X}, \mathbf{Y}) = C(\mathbf{Y}, \mathbf{X})$..

A.3.8) Starting with Definition (A.13) on page 130, show $C(\mathbf{X}, \mathbf{Y}) = E(\mathbf{X}\mathbf{Y}^\top) - \boldsymbol{\mu}_x\boldsymbol{\mu}_y^\top$.

A.3.9) Starting with Definition (A.13) on page 130, show $C(\mathbf{X}, \mathbf{Y}) = \mathbf{0}$..

A.3.10) Let $\mathbf{X}$ be a $p \times 1$ random vector with expected value $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, and let $\mathbf{v}$ be a $p \times 1$ vector of constants.

(a) Let the scalar random variable $Y = \mathbf{v}^\top\mathbf{X}$. What is $Var(Y)$? Use this to prove tell you that *any* variance-covariance matrix must be positive semi-definite. (See the definition on Page 119.)

zero ..." This confused me deeply the first time I saw it as a student, but it's correct. There is no loss of generality because you would get the same answer by letting the expected values be non-zero. But it only applies to variances and covariances of linear combinations.

(b) Using the definition of an eigenvalue (A.7) on Page 120, show that eigenvalues of a variance-covariance matrix cannot be negative[4].

(c) How do you know that the determinant of a variance-covariance matrix must be greater than or equal to zero? The answer is one short sentence.

(d) Let X and Y be scalar random variables. Using what you have shown about the determinant, show $-1 \leq Corr(X,Y) \leq 1$. See the definition of a correlation on Page 113 if necessary. You have just proved the Cauchy-Schwarz inequality using probability tools.

A.3.11) Let the $p \times 1$ random vector $\mathbf{X}$ have mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, and let $\mathbf{c}$ be a $p \times 1$ vector of constants. Find $V(\mathbf{X}+\mathbf{c})$. Show your work, starting with the definition (A.12). Don't use the centering rule yet.

A.3.12) Let $\mathbf{X}$ be a $p \times 1$ random vector with mean $\boldsymbol{\mu}_x$ and variance-covariance matrix $\boldsymbol{\Sigma}_x$, and let $\mathbf{Y}$ be a $q \times 1$ random vector with mean $\boldsymbol{\mu}_y$ and variance-covariance matrix $\boldsymbol{\Sigma}_y$. Recall that $C(\mathbf{X},\mathbf{Y})$ is the $p \times q$ matrix $C(\mathbf{X},\mathbf{Y}) = E\left((\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{Y}-\boldsymbol{\mu}_y)^\top\right)$. Don't use the centering rule yet.

(a) What is the (i,j) element of $C(\mathbf{X},\mathbf{Y})$?

(b) Find an expression for $V(\mathbf{X}+\mathbf{Y})$ in terms of $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_y$ and $C(\mathbf{X},\mathbf{Y})$. Show your work.

(c) Simplify further for the special case where $Cov(X_i,Y_j) = 0$ for all i and j.

(d) Let $\mathbf{c}$ be a $p \times 1$ vector of constants and $\mathbf{d}$ be a $q \times 1$ vector of constants. Find $C(\mathbf{X}+\mathbf{c},\mathbf{Y}+\mathbf{d})$. Show your work.

A.3.13) Prove (A.16). This is the *basis* of the centering rule, so you are not allowed to use the centering rule.

A.3.14) Use the centering rule to show $V(\mathbf{AX}+\mathbf{BY}) = \mathbf{A}V(\mathbf{X})\mathbf{A}^\top + \mathbf{B}V(\mathbf{Y})\mathbf{B}^\top$..

A.3.15) Use the centering rule to find $V(\mathbf{AX}+\mathbf{BY}+\mathbf{c})$. What do you need to specify about the dimensions of the matrices for this to be true?

A.3.16) Write down $V(\mathbf{AX}+\mathbf{BY})$ for the case where $\mathbf{X}$ and $\mathbf{Y}$ are independent. There is no need to show any work.

A.3.17) Use the centering rule to find $C(\mathbf{AX}+\mathbf{c},\mathbf{BX}+\mathbf{d})$. Must $\mathbf{A}$ and $\mathbf{B}$ have the same number of rows?

A.3.18) Let $X_1,\ldots,X_n$ be scalar random variables. Use the centering rule to show

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + \sum_{i \neq j} Cov(X_i,X_j).$$

[4]This property of covariance matrices can sometimes be used to detect problems with the numerical estimation of structural equation models.

A.4 The Multivariate Normal Distribution

The $p \times 1$ random vector $\mathbf{X}$ is said to have a *multivariate normal distribution*, and we write $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if $\mathbf{X}$ has (joint) density

$$f(\mathbf{x}) = \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{A.18}$$

where $\boldsymbol{\mu}$ is $p \times 1$ and $\boldsymbol{\Sigma}$ is $p \times p$ symmetric and positive definite. Positive definite means that for any non-zero $p \times 1$ vector $\mathbf{a}$, we have $\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} > 0$.

- Since the one-dimensional random variable $Y = \sum_{i=1}^p a_i X_i$ may be written as $Y = \mathbf{a}^\top \mathbf{X}$ and $Var(Y) = V(\mathbf{a}^\top \mathbf{X}) = \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}$, it is natural to require that $\boldsymbol{\Sigma}$ be positive definite. All it means is that every non-zero linear combination of $\mathbf{X}$ values has a positive variance.
- $\boldsymbol{\Sigma}$ positive definite is equivalent to $\boldsymbol{\Sigma}^{-1}$ positive definite.

The multivariate normal reduces to the univariate normal when $p = 1$. Other properties of the multivariate normal include the following.

1. $E(\mathbf{X}) = \boldsymbol{\mu}$
2. $V(\mathbf{X}) = \boldsymbol{\Sigma}$
3. If $\mathbf{c}$ is a vector of constants, $\mathbf{X} + \mathbf{c} \sim N(\mathbf{c} + \boldsymbol{\mu}, \boldsymbol{\Sigma})$
4. If $\mathbf{A}$ is a matrix of constants, $\mathbf{AX} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$
5. Linear combinations of multivariate normals are multivariate normal.
6. All the marginals (dimension less than p) of $\mathbf{X}$ are (multivariate) normal, but it is possible in theory to have a collection of univariate normals whose joint distribution is not multivariate normal.
7. For the multivariate normal, zero covariance implies independence. The multivariate normal is the only continuous distribution with this property.
8. The random variable $(\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$ has a chi-square distribution with p degrees of freedom.
9. After a bit of work, the multivariate normal likelihood may be written as

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-np/2} \exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\overline{\mathbf{x}}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})\right\}, \tag{A.19}$$

where $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top$ is the sample variance-covariance matrix (it would be unbiased if divided by $n-1$).

Here how Expression (A.19) above for $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is obtained.

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{n} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\} \\ &= |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-np/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\} \end{aligned}$$

Adding and subtracting $\overline{\mathbf{x}}$ in $\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$, we get

$$\begin{aligned} \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) &= \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}} + \overline{\mathbf{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}} + \overline{\mathbf{x}} - \boldsymbol{\mu}) \\ &= \sum_{i=1}^{n}(\mathbf{a}_i + \mathbf{b})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{a}_i + \mathbf{b}) \\ &= \sum_{i=1}^{n}\left(\mathbf{a}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{a}_i + \mathbf{a}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{b} + \mathbf{b}^\top \boldsymbol{\Sigma}^{-1}\mathbf{a}_i + \mathbf{b}^\top \boldsymbol{\Sigma}^{-1}\mathbf{b}\right) \\ &= \left(\sum_{i=1}^{n}\mathbf{a}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{a}_i\right) + \mathbf{0} + \mathbf{0} + n\,\mathbf{b}^\top \boldsymbol{\Sigma}^{-1}\mathbf{b} \\ &= \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) \;+\; n\,(\overline{\mathbf{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}) \end{aligned}$$

Now, because $\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})$ is a 1×1 matrix, it equals its own trace and we

can use $tr(\mathbf{AB}) = tr(\mathbf{BA})$.

$$\begin{aligned}
\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}}) &= tr\left\{\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}})\right\} \\
&= \sum_{i=1}^{n} tr\left\{(\mathbf{x}_i-\overline{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}})\right\} \\
&= \sum_{i=1}^{n} tr\left\{\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^\top\right\} \\
&= tr\left\{\sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^\top\right\} \\
&= tr\left\{\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^\top\right\} \\
&= n\,tr\left\{\boldsymbol{\Sigma}^{-1}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^\top\right\} \\
&= n\,tr\left(\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\Sigma}}\right),
\end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^\top$ is the sample variance-covariance matrix. Substituting for $\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})$,

$$\begin{aligned}
L(\boldsymbol{\mu},\boldsymbol{\Sigma}) &= |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-np/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right\} \\
&= |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-np/2}\exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\overline{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})\right\}.
\end{aligned}$$

Notice how the multivariate normal likelihood depends on the sample data only through the sufficient statistic $(\overline{\mathbf{X}}, \widehat{\boldsymbol{\Sigma}})$.

Exercises A.4

A.4.1) Let X_1 be Normal(μ_1, σ_1^2), and X_2 be Normal(μ_2, σ_2^2), independent of X_1. What is the joint distribution of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$? What is required for Y_1 and Y_2 to be independent?

A.4.2) Let $\mathbf{X} = (X_1, X_2, X_3)^\top$ be multivariate normal with

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 0 \\ 6 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Let $Y_1 = X_1 + X_2$ and $Y_2 = X_2 + X_3$. Find the joint distribution of Y_1 and Y_2.

A.4.3) Let X_1 be Normal(μ_1, σ_1^2), and X_2 be Normal(μ_2, σ_2^2), independent of X_1. What is the joint distribution of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$? What is required for Y_1 and Y_2 to be independent? Hint: Use matrices.

A.4.4) Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ is multivariate normal with mean zero and covariance matrix $\sigma^2\mathbf{I}_n$, where $\sigma^2 > 0$ is a constant. In the following, it may be helpful to recall that $(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}$.

(a) What is the distribution of $\mathbf{Y}$?

(b) The maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$. What is the distribution of $\hat{\boldsymbol{\beta}}$? Show the calculations.

(c) Let $\widehat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. What is the distribution of $\widehat{\mathbf{Y}}$? Show the calculations.

(d) Let the vector of residuals $\mathbf{e} = (\mathbf{Y} - \widehat{\mathbf{Y}})$. What is the distribution of $\mathbf{e}$? Show the calculations. Simplify both the expected value (which is zero) and the covariance matrix.

A.4.5) Show that if $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $Y = (\mathbf{X} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ has a chi-square distribution with p degrees of freedom.

A.4.6) Write down a scalar version of formula (A.19) for the multivariate normal likelihood, showing that you understand the notation. Then derive your formula from the univariate normal likelihood.

A.4.7) Prove the formula (A.19) for the multivariate normal likelihood. Show all the calculations.

A.4.8) Prove that for *any* positive definite $\boldsymbol{\Sigma}$, the likelihood (A.19) is maximized when $\overline{\mathbf{x}} = \boldsymbol{\mu}$. How do you know this maximum must be unique? Cite the necessary matrix facts from Section A.2 of this Appendix.

A.5 A Bit of Large Sample Theory

For this part, it helps to start by going down to the basement and taking a look at the foundations of the building. There is an underlying sample space Ω, consisting of sample points $\omega \in \Omega$[5]. The specific nature of a point ω in applications depends on what is being observed. For example, if we were observing whether a single individual is male or female, Ω might be $\{F, M\}$. If we selected a pair of individuals and observed their genders in order, Ω might be $\{(F,F), (F,M), (M,F), (M,M)\}$. If we selected n individuals and just *counted* the number of females, Ω might be $\{0, \ldots, n\}$. For limits problems, the points in Ω are infinite sequences.

[5]Throughout most of this book, Ω is a covariance matrix. The symbol will briefly have its usual meaning here, just for the discussion of almost sure convergence

Let $\mathcal{A}$ be a class of subsets of Ω (that is, a set of *events*), and let $\mathcal{P}$ be a probability function that assigns numbers between zero and one inclusive to the elements of $\mathcal{A}$. A *random variable* $X = X(\omega)$ is a function that maps Ω into some other space, typically $\mathbb{R}$ or $\mathbb{R}^k$. Think of taking a measurement: if Ω is a set of students, $X(\omega)$ might be the cumulative grade point average of student ω.

Suppose the random variable X maps Ω into the set of real numbers $\mathbb{R}$. Then X induces a probability measure on a class[6] $\mathcal{B}$ of subsets of $\mathbb{R}$, by means of

$$Pr\{X \in B\} = \mathcal{P}(\{\omega \in \Omega : X(\omega) \in B\})$$

for $B \in \mathcal{B}$.

Suppose we have a sample of data $X_1(\omega), \ldots, X_n(\omega)$, and we calculate a function of the sample data $T = T(X_1, \ldots, X_n)$. For example T could be a *statistic* like the sample mean $\overline{X}$. It is helpful to write $T = T_n(\omega)$, to indicate that T is a random variable (a function from Ω into $\mathbb{R}$) that depend upon the sample size n.

Frequently it is useful to let $n \to \infty$, because when the sequence $T_1, T_2, \ldots$ converges, it is an indication of what happens when the sample is large enough. But this is not just a sequence of numbers; it is a sequence of functions. Several different types of convergence are meaningful.

Modes of Convergence

Throughout, let $T_1, T_2, \ldots$ be a sequence of random variables, and let T be another random variable. It is quite possible and often useful for $T = T(\omega)$ to be a constant — that is, a constant function of ω. In that case T is a "degenerate" random variable, with $P\{T = c\} = 1$ for some constant c.

Almost Sure Convergence

We say that T_n converges *almost surely* to T, and write $T_n \overset{a.s.}{\to}$ if

$$\mathcal{P}\{\omega : \lim_{n\to\infty} T_n(\omega) = T(\omega)\} = 1.$$

That is, except possibly for $\omega \in A$ with $\mathcal{P}(A) = 0$, $T_n(\omega)$ converges to the random variable $T(\omega)$ like an ordinary limit, and all the usual rules apply — for example, the limit of a continuous function is the continuous function of the limit, L'Hôpital's rule and so on. Almost sure convergence is also called *convergence with probability one*, or sometimes *strong convergence*.

Almost sure convergence may be the most technically "advanced" mode of convergence, but it is also perhaps the easiest to work with, because you treat the sequence $T_1, T_2, \ldots$ like numbers, find the limit, and then mention that the result applies "except possibly on a set of probability zero."

The main entry point to establishing almost sure convergence is the *Strong Law of Large Numbers,* which involves almost sure convergence to a constant. Let $X_1, \ldots X_n$ be

[6] I'm thinking of the Borel σ-algebra, but there is no need to go that far.

independent and identically distributed random variables with expected value μ. Denote the sample mean as usual by $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. The Strong Law of Large Numbers (SLLN) says

$$\overline{X}_n \stackrel{a.s.}{\to} \mu. \tag{A.20}$$

The only condition required for this to hold is the existence of the expected value.

Let $X_1, \ldots X_n$ be independent and identically distributed random variables; let X be a general random variable from this same distribution, and $Y = g(X)$. The change of variables formula (A.1) can be combined with the Strong Law of Large Numbers to write

$$\frac{1}{n}\sum_{i=1}^n g(X_i) = \frac{1}{n}\sum_{i=1}^n Y_i \stackrel{a.s.}{\to} E(Y) = E(g(X)). \tag{A.21}$$

This means that sample moments converge almost surely to population moments:

$$\frac{1}{n}\sum_{i=1}^n X_i^k \stackrel{a.s.}{\to} E(X^k)$$

It even yields rules like

$$\frac{1}{n}\sum_{i=1}^n U_i^2 V_i W_i^3 \stackrel{a.s.}{\to} E(U^2 V W^3).$$

Convergence in Probability

We say that T_n converges *in probability* to T, and write $T_n \stackrel{P}{\to} T$ if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P\{|T_n - T| < \epsilon\} = 1.$$

Convergence in probability is implied by almost sure convergence, so corresponding to the Strong Law of Large Numbers is the Weak Law of Large Numbers (WLLN). Let $X_1, \ldots X_n$ be independent and identically distributed random variables with expected value μ. Then the sample mean converges in probability to μ:

$$\overline{X}_n \stackrel{P}{\to} \mu. \tag{A.22}$$

A change of variables rule like expression (A.21) holds, and sample moments converge in probability to population moments. These rules follow from the corresponding facts about almost sure convergence.

Another way of establishing convergence in probability to a constant without using the definition is the *Variance Rule*. Let θ be a constant. Then if $\lim_{n\to\infty} E(T_n) = \theta$ and $\lim_{n\to\infty} Var(T_n) = 0$, it follows that $T_n \stackrel{P}{\to} \theta$. But convergence in probability does not imply the conditions of the Variance Rule.

Convergence in Distribution

Denote the cumulative distribution functions of $T_1, T_2, \ldots$ by $F_1(t), F_2(t), \ldots$ respectively, and denote the cumulative distribution function of T by $F(t)$. We say that T_n converges *in distribution* to T, and write $T_n \xrightarrow{d} T$ if for every point t at which F is continuous,

$$\lim_{n\to\infty} F_n(t) = F(t).$$

The main entry point to convergence in distribution is the *Central Limit Theorem*. Let $X_1, \ldots X_n$ be independent and identically distributed random variables with mean μ and variance σ^2. Then

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim N(0,1).$$

In applications, the sample standard deviation may be substituted for σ, and the result still holds.

A useful tool is provided by the univariate *delta method*[7]. Let $\sqrt{n}(X_n - \theta) \xrightarrow{d} X$, and let $g(x)$ be a function with $g'(\theta) \neq 0$ and $g''(x)$ continuous at $x = \theta$. Then

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)X.$$

In particular, $\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} Y \sim N(0, g'(\mu)^2\sigma^2)$.

Connections among the Modes of Convergence

- $T_n \xrightarrow{a.s.} T \Rightarrow T_n \xrightarrow{P} T \Rightarrow T_n \xrightarrow{d} T$.
- If a is a constant, $T_n \xrightarrow{d} a \Rightarrow T_n \xrightarrow{P} a$.

Sometimes we say the distribution of the sample mean is approximately normal, or asymptotically normal. This is justified by the Central Limit Theorem, but it does *not* mean that $\overline{X}_n$ converges in distribution to a normal random variable. The Law of Large Numbers says that $\overline{X}_n$ converges almost surely (and in probability) to a constant, μ. This means $\overline{X}_n$ converges to μ in distribution as well. So why would we say that for large n, the sample mean is approximately $N(\mu, \frac{\sigma^2}{n})$?

What we have is $Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim N(0,1)$. So,

$$\begin{aligned} Pr\{\overline{X}_n \leq x\} &= Pr\left\{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq \frac{\sqrt{n}(x-\mu)}{\sigma}\right\} \\ &= Pr\left\{Z_n \leq \frac{\sqrt{n}(x-\mu)}{\sigma}\right\} \approx \Phi\left(\frac{\sqrt{n}(x-\mu)}{\sigma}\right), \end{aligned}$$

[7]The delta method is named after the way it is proved; it uses Taylor's theorem, and the "delta" part is connected to the definition of a derivative. We will just use it.

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

Now suppose that Y is *exactly* $N(\mu, \frac{\sigma^2}{n})$. Then,

$$\begin{aligned} Pr\{Y \leq x\} &= Pr\left\{\frac{\sqrt{n}(Y-\mu)}{\sigma} \leq \frac{\sqrt{n}(x-\mu)}{\sigma}\right\} \\ &= Pr\left\{Z \leq \frac{\sqrt{n}(x-\mu)}{\sigma}\right\} = \Phi\left(\frac{\sqrt{n}(x-\mu)}{\sigma}\right). \end{aligned}$$

So we see that the Central Limit Theorem tells us to calculate probabilities for $\overline{X}_n$ just as we would if $\overline{X}_n$ had a distribution that was exactly normal with expected value μ and variance $\frac{\sigma^2}{n}$. This the justification for saying that the sample mean is "asymptotically normal," and writing $\overline{X}_n \dot{\sim} N(\mu, \frac{\sigma^2}{n})$. Here are three additional remarks.

- Quantities like $\frac{1}{n}\sum_{i=1}^n X_i^2$ and $\frac{1}{n}\sum_{i=1}^n X_i Y_i$ and so on are asymptotically normal too, because they are just sample means.
- The delta method says that smooth functions of the sample mean are asymptotically normal.
- All this generalizes nicely to the multivariate case.

Consistency

For this application, $T_1, T_2, \ldots$ are not just random variables: They are *statistics*[8] that estimate some parameter θ. The statistic T_n is said to be *consistent* for θ if $T_n \xrightarrow{P} \theta$ for all $\theta \in \Theta$.

Let us take a closer look at this important concept. Using the definition of convergence in probability, saying that T_n is consistent for θ means that for any tiny positive constant ϵ, no matter *how* tiny,

$$\lim_{n\to\infty} P\{|T_n - \theta| < \epsilon\} = 1.$$

So, take an arbitrarily small interval around the true parameter value. For any given sample size n, a certain amount of the probability distribution of T_n falls between $\theta - \epsilon$ and $\theta + \epsilon$. Consistency means that in the limit, *all* the probability falls in this interval, no matter how small the interval is. Basically, consistency is saying that for a large enough sample size, the statistic (estimator) will probably be close to parameter it is estimating — regardless of how strict your definitions of "probably" and "close" might be.

Even better than ordinary consistency is *strong consistency*, which means $T_n \xrightarrow{a.s.} \theta$. Instead of saying T_n will probably be close to θ, strong consistency says that for a large enough sample size, the probability that it *will* be close equals one. Because almost sure convergence implies convergence in probability, strong consistency implies ordinary consistency.

[8] A statistic is a function of the sample data that does not depend functionally upon any unknown parameter. That is, symbol for the parameter does not appear in the formula for the statistic.

One last remark is that while consistency is an important property in an estimator, in a way it is the least we should expect. Consistency means that with an infinite amount of data, we would know the truth. If this is *not* the case, something is seriously wrong[9].

Exercises A.5

A.5.1) Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with density

$$f(x;\theta) = \frac{1}{\theta^{1/2}\sqrt{2\pi}}\, e^{-\frac{x^2}{2\theta}},$$

where the parameter $\theta > 0$. Propose a reasonable estimator for the parameter θ, and use the Law of Large Numbers to show that your estimator is consistent.

A.5.2) Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ, and let X be a general random variable with that distribution. You know that $E(X) = Var(X) = \lambda$; there is no need to prove it.

From the Strong Law of Large Numbers, it follows immediately that $\overline{X}_n$ is strongly consistent for λ. Let

$$\widehat{\lambda} = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n-4}.$$

Is $\widehat{\lambda}$ also consistent for λ? Answer Yes or No and prove your answer.

A.5.3) Let $X_1, \ldots, X_n$ be a random sample from a Binomial distribution with parameters 3 and θ. That is,

$$P(X_i = x_i) = \binom{3}{x_i}\theta^{x_i}(1-\theta)^{3-x_i},$$

for $x_i = 0, 1, 2, 3$. Find a reasonable estimator of θ, and prove that it is strongly consistent. Where you get your estimator does not really matter, but please state how you thought of it.

A.5.4) Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with density

$$f(x;\tau) = \frac{\tau^{1/2}}{\sqrt{2\pi}}\, e^{-\frac{\tau x^2}{2}},$$

where the parameter $\tau > 0$. Let

$$\widehat{\tau} = \frac{n}{\sum_{i=1}^{n} X_i^2}.$$

Is $\widehat{\tau}$ consistent for τ? Answer Yes or No and prove your answer. Hint: You can just write down $E(X^2)$ by inspection. This is a very familiar distribution; have confidence!

[9] In structural equation models, a parameter that is not identifiable cannot be estimated consistently. This is why model identification is such an important topic.

A.5.5) Independently for $i = 1, \ldots, n$, let

$$Y_i = \beta X_i + \epsilon_i,$$

where $E(X_i) = E(\epsilon_i) = 0$, $Var(X_i) = \sigma_x^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and ϵ_i is independent of X_i. Let

$$\widehat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$

Is $\widehat{\beta}$ consistent for β? Answer Yes or No and prove your answer.

A.5.6) Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with $\alpha = \beta = \theta > 0$. That is, the density is

$$f(x; \theta) = \frac{1}{\theta^\theta \Gamma(\theta)} e^{-x/\theta} x^{\theta - 1},$$

for $x > 0$. Let $\widehat{\theta} = \overline{X}_n$. Is $\widehat{\theta}$ consistent for θ? Answer Yes or No and prove your answer.

A.5.7) Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ. Show that $T_n = \frac{1}{n+400} \sum_{i=1}^n X_i$ is consistent for μ.

A.5.8) Let $X_1, \ldots, X_n$ be a random sample from a distribution with expected value μ and variance σ_x^2. Independently of $X_1, \ldots, X_n$, let $Y_1, \ldots, Y_n$ be a random sample from a distribution with the same expected value μ and variance σ_y^2. Let Let $T_n = \alpha \overline{X}_n + (1 - \alpha)\overline{Y}_n$, where $0 \leq \alpha \leq 1$. Is T_n always a consistent estimator of μ? Answer Yes or No and show your work.

A.5.9) Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Prove that the sample variance $S^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$ is consistent for σ^2.

A.5.10) Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a bivariate distribution with $E(X_i) = \mu_x$, $E(Y_i) = \mu_y$, $Var(X_i) = \sigma_x^2$, $Var(Y_i) = \sigma_y^2$, and $Cov(X_i, Y_i) = \sigma_{xy}$. Show that the sample covariance $S_{xy} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$ is a consistent estimator of σ_{xy}.

Convergence of random vectors

Almost all applied problems are multi-parameter, and that certainly applies to the ones in this book. Parameter estimates are usually random vectors. It is very convenient that in terms of convergence, the multivariate case is very similar to the univariate case just discussed. The is based on material in Thomas Ferguson's beautiful little book *A course in large sample theory*, which is highly recommended. All quantities in boldface are vectors in $\mathbb{R}^m$ unless otherwise indicated.

1. Definitions

- ⋆ $\mathbf{T}_n \xrightarrow{a.s.} \mathbf{T}$ means $P\{\omega : \lim_{n\to\infty} \mathbf{T}_n(\omega) = \mathbf{T}(\omega)\} = 1$.
- ⋆ $\mathbf{T}_n \xrightarrow{P} \mathbf{T}$ means $\forall \epsilon > 0$, $\lim_{n\to\infty} P\{||\mathbf{T}_n - \mathbf{T}|| < \epsilon\} = 1$.
- ⋆ $\mathbf{T}_n \xrightarrow{d} \mathbf{T}$ means for every continuity point $\mathbf{t}$ of $F_{\mathbf{T}}$, $\lim_{n\to\infty} F_{\mathbf{T}_n}(\mathbf{t}) = F_{\mathbf{T}}(\mathbf{t})$.

2. $\mathbf{T}_n \xrightarrow{a.s.} \mathbf{T} \Rightarrow \mathbf{T}_n \xrightarrow{P} \mathbf{T} \Rightarrow \mathbf{T}_n \xrightarrow{d} \mathbf{T}$.
3. If $\mathbf{a}$ is a vector of constants, $\mathbf{T}_n \xrightarrow{d} \mathbf{a} \Rightarrow \mathbf{T}_n \xrightarrow{P} \mathbf{a}$.
4. Strong Law of Large Numbers (SLLN): Let $\mathbf{X}_1, \ldots \mathbf{X}_n$ be independent and identically distributed random vectors with finite first moment, and let $\mathbf{X}$ be a general random vector from the same distribution. Then $\overline{\mathbf{X}}_n \xrightarrow{a.s.} E(\mathbf{X})$.
5. Central Limit Theorem: Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. random vectors with expected value vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then $\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu})$ converges in distribution to a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$.
6. Slutsky Theorems for Convergence in Distribution:
   (a) If $\mathbf{T}_n \in \mathbb{R}^m$, $\mathbf{T}_n \xrightarrow{d} \mathbf{T}$ and if $f : \mathbb{R}^m \to \mathbb{R}^q$ (where $q \leq m$) is continuous except possibly on a set C with $P(\mathbf{T} \in C) = 0$, then $f(\mathbf{T}_n) \xrightarrow{d} f(\mathbf{T})$.
   (b) If $\mathbf{T}_n \xrightarrow{d} \mathbf{T}$ and $(\mathbf{T}_n - \mathbf{Y}_n) \xrightarrow{P} 0$, then $\mathbf{Y}_n \xrightarrow{d} \mathbf{T}$.
   (c) If $\mathbf{T}_n \in \mathbb{R}^d$, $\mathbf{Y}_n \in \mathbb{R}^k$, $\mathbf{T}_n \xrightarrow{d} \mathbf{T}$ and $\mathbf{Y}_n \xrightarrow{P} \mathbf{c}$, then
   $$\begin{pmatrix} \mathbf{T}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathbf{T} \\ \mathbf{c} \end{pmatrix}$$
7. Slutsky Theorems for Convergence in Probability:
   (a) If $\mathbf{T}_n \in \mathbb{R}^m$, $\mathbf{T}_n \xrightarrow{P} \mathbf{T}$ and if $f : \mathbb{R}^m \to \mathbb{R}^q$ (where $q \leq m$) is continuous except possibly on a set C with $P(\mathbf{T} \in C) = 0$, then $f(\mathbf{T}_n) \xrightarrow{P} f(\mathbf{T})$.
   (b) If $\mathbf{T}_n \xrightarrow{P} \mathbf{T}$ and $(\mathbf{T}_n - \mathbf{Y}_n) \xrightarrow{P} 0$, then $\mathbf{Y}_n \xrightarrow{P} \mathbf{T}$.
   (c) If $\mathbf{T}_n \in \mathbb{R}^d$, $\mathbf{Y}_n \in \mathbb{R}^k$, $\mathbf{T}_n \xrightarrow{P} \mathbf{T}$ and $\mathbf{Y}_n \xrightarrow{P} \mathbf{Y}$, then
   $$\begin{pmatrix} \mathbf{T}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{P} \begin{pmatrix} \mathbf{T} \\ \mathbf{Y} \end{pmatrix}$$
8. Delta Method (Theorem of Cramér, Ferguson p. 45): Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be such that the elements of $\dot{\mathrm{g}}(\mathbf{x}) = \left[\frac{\partial g_i}{\partial x_j}\right]_{k\times d}$ are continuous in a neighborhood of $\boldsymbol{\theta} \in \mathbb{R}^d$. If $\mathbf{T}_n$ is a sequence of d-dimensional random vectors such that $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{T}$, then $\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} \dot{\mathrm{g}}(\boldsymbol{\theta})\mathbf{T}$. In particular, if $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{T} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, then $\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} \mathbf{Y} \sim N(\mathbf{0}, \dot{\mathrm{g}}(\boldsymbol{\theta})\boldsymbol{\Sigma}\dot{\mathrm{g}}(\boldsymbol{\theta})')$.

In the multivariate delta method, the matrix $\dot{\mathrm{g}}(\boldsymbol{\theta})$ ia the Jacobian of the transformation g. It says that smooth functions of asymptotically normal random variables are also asymptotically normal.

A.6 Estimation and inference

A.6.1 Statistical Models

A *statistical model* is a set of assertions that partly specify the probability distribution of the observable data. The specification may be direct or indirect. As an example of direct specification, let $X_1, \ldots, X_n$ be a random sample from a normal distribution with expected value μ and variance σ^2. As an example of indirect specification, let $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$ for $i = 1, \ldots, n$, where

$\beta_0, \ldots, \beta_k$ are unknown constants. x_{ij} are known constants.
$\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables.
σ^2 is an unknown constant.

Statistical models leave something unknown. Otherwise, they are probability models. The unknown part of the model for the data is called the *parameter*. Usually, parameters are numbers or vectors of numbers – unknown constants. They are usually denoted by θ or $\boldsymbol{\theta}$ or other Greek letters.

The *parameter space* is the set of values that can be taken on by the parameter, and will be denoted by Θ, with $\theta \in \Theta$. For the normal random sample example, the parameter space is $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$. For the regression example given above, $\Theta = \{(\beta_0, \ldots, \beta_k, \sigma^2) : -\infty < \beta_j < \infty, \sigma^2 > 0\}$.

Parameters need not be numbers. For example, let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with unknown distribution function $F(x)$. The parameter is the unknown distribution function $F(x)$, and the parameter space is a space of distribution functions. We may be interested only in a *function* of the parameter, like

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx$$

The rest of $F(x)$ is just a nuisance parameter.

We will use the following framework for parameter estimation and statistical inference. The data are $D_1, \ldots, D_n$ (the letter D stands for data). The distribution of these independent and identically distributed random variables depends on the parameter θ, which is an element of the parameter space Θ. That is,

$$D_1, \ldots, D_n \overset{i.i.d.}{\sim} P_\theta,\ \theta \in \Theta.$$

Both the data values and the parameter may be vectors, even though they are not written in boldface.

To give one more example, the data vector could be $D = \mathbf{X}_1, \ldots \mathbf{X}_n$, a vector of independent multivariate normals of dimension p. The parameter space is $\{\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^p$, and $\boldsymbol{\Sigma}$ is a $p \times p$ symmetric positive definite matrix. P_θ is the joint distribution function of $\mathbf{X}_1, \ldots \mathbf{X}_n$, with joint density

$$f(\mathbf{x}_1, \ldots \mathbf{x}_n) = \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $f(\mathbf{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal density (A.18) on page 135.

For the model $D \sim P_\theta, \theta \in \Theta$, we don't know θ. We never know θ. All we can do is guess. We will estimate θ (or a function of θ) based on the observable data. Let T denote an *estimator* of θ (or a function of θ): $T = T(D)$ For example, if $D = X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, the usual estimator is $T = (\overline{X}, S^2)$. For an ordinary fixed-x multiple regression model, $T = (\widehat{\boldsymbol{\beta}}, MSE)$. In these and in all other cases, T is a *statistic*, a random variable or vector that can be computed from the data without knowing the values of any unknown parameters.

How do we get a recipe for T? Guess? It's good to be systematic. Lots of methods are available. We will consider two: Method of moments and Maximum Likelihood.

A.6.2 Method of Moments Estimation

The following is based on a random sample like $(X_1, Y_1), \ldots, (X_n, Y_n)$. Moments are quantities like $E\{X_i\}$, $E\{X_i^2\}$, $E\{X_iY_i\}$, $E\{W_iX_i^2Y_i^3\}$, and so on. *Central* moments are moments of *centered* random variables, such as

$$E\{(X_i - \mu_x)^2\}$$

$$E\{(X_i - \mu_x)(Y_i - \mu_y)\}$$

$$E\{(X_i - \mu_x)^2(Y_i - \mu_y)^3(Z_i - \mu_z)^2\}$$

These are all *population* moments. Sample moments are analogous to population moments, and are natural estimators.

Population moment	Sample moment
$E\{X_i\}$	$\frac{1}{n}\sum_{i=1}^n X_i$
$E\{X_i^2\}$	$\frac{1}{n}\sum_{i=1}^n X_i^2$
$E\{X_iY_i\}$	$\frac{1}{n}\sum_{i=1}^n X_iY_i$
$E\{(X_i - \mu_x)^2\}$	$\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$
$E\{(X_i - \mu_x)(Y_i - \mu_y)\}$	$\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)$
$E\{(X_i - \mu_x)(Y_i - \mu_y)^2\}$	$\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)^2$

The method of moments is based on estimating population moments by the corresponding sample moments. For the model $D \sim P_\theta$ with $\theta \in \Theta$, the population moments are a function of θ. The procedure is to first find θ as a function of the population moments, and then estimate θ with that function of the *sample* moments.

Let m denote a vector of population moments, and let $\widehat{m}$ denote the corresponding vector of sample moments. First, find $m = g(\theta)$. Then solve for θ, obtaining $\theta = g^{-1}(m)$.

Let $\widehat{\theta} = g^{-1}(\widehat{m})$. It doesn't matter if you solve first or put hats on first[10].

For example, suppose $X_1, \ldots, X_n \overset{i.i.d}{\sim} U(0, \theta)$. That is, the data are a random sample from a uniform distribution on $(0, \theta)$, so that the model density is $f(x) = \frac{1}{\theta}$ for $0 < x < \theta$. First, find the moment (expected value).

$$\begin{aligned} E(X_i) &= \int_0^\theta x \frac{1}{\theta}\, dx \\ &= \frac{1}{\theta} \int_0^\theta x\, dx \\ &= \frac{1}{\theta} \frac{x^2}{2}\Big|_0^\theta = \frac{1}{2\theta}(\theta^2 - 0) \\ &= \frac{\theta}{2} \end{aligned}$$

So $m = \frac{\theta}{2} \Leftrightarrow \theta = 2m$, and $\widehat{\theta} = 2\overline{X}$.

Sample problem Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on $(0, \theta)$. Estimate θ by the Method of Moments for the following data. Your answer is a number. Show some work. Data: `4.09 0.13 0.84 3.83 2.13 4.67 4.61 0.40 4.19 0.71`.

Answer $\overline{X} = 2.56$ so $\widehat{\theta} = 2\overline{X} = 2 * 2.56 = 5.12$.

Method of moments estimators are not unique. What moments you use are up to you.

$$E(X_i^2) = \frac{1}{\theta} \int_0^\theta x^2\, dx = \frac{\theta^2}{3}$$

So set $m = \frac{\theta^2}{3} \Leftrightarrow \theta = \sqrt{3m}$, and

$$\widehat{\theta} = \sqrt{\frac{3}{n} \sum_{i=1}^n X_i^2},$$

which is not equal to $2\overline{X}$. Presumably estimates based on lower-order moments are better in some sense, but I don't know the details.

To compare the two estimates $\widehat{\theta}_1 = 2\overline{X}$ and $\widehat{\theta}_2 = \sqrt{\frac{3}{n}\sum_{i=1}^n X_i^2}$ for the numerical example,

```
x     4.09   0.13   0.84    3.83   2.13    4.67    4.61   0.40  4.19   0.71
x^2  16.7281 0.0169 0.7056 14.6689 4.5369 21.8089 21.2521 0.16 17.5561 0.5041
```

yielding $\widehat{\theta}_1 = 5.12$ and $\widehat{\theta}_2 = 5.42$.

[10] For most models the function g is well behaved, with continuous mixed partial derivatives. In that case the multivariate delta method from the end of Section A.5 guarantees that $\widehat{\theta}$ is asymptotically multivariate normal even when the data are definitely not normal. This yields distribution-free tests and confidence intervals with surprisingly little effort.

Method of Moments estimator for the normal Let $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$. From the moment-generating function or a textbook, $E(X_i) = \mu$ and $E(X_i^2) = \sigma^2 + \mu^2$. Solving for the parameters, $\mu = E(X_i)$ and $\sigma^2 = E(X_i^2) - (E(X_i))^2$. The Method of Moments estimators are $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$.

A regression example Independently for $i = 1, \ldots, n$, let $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where

- $E(X_i) = \mu_x$, $Var(X_i) = \sigma_x^2$
- $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_\epsilon^2$
- X_i and ϵ_i are independent.

The distributions of X_i and ϵ_i are unknown, so they are part of the parameter. The parameter is $(\beta_0, \beta_1, F_\epsilon(\epsilon), F_x(x))$. As mentioned earlier, there is no conceptual problem with parameters that are functions (infinite-dimensional) instead of just real numbers or vectors.

We want to estimate β_0 and β_1, a two-dimensional *function* of the parameter. First, calculate some moments.

$$\begin{array}{ll} E(X_i) = \mu_x & Var(X_i) = \sigma_x^2 \\ E(Y_i) = \beta_0 + \beta_1\mu_x & Cov(X_i, Y_i) = \beta_1\sigma_x^2 \end{array}$$

Use the Centering Rule on Page 131 to get the last one:

$$\begin{aligned} Cov(X_i, Y_i) &= E(\overset{c}{X}_i \overset{c}{Y}_i) \\ &= E\{\overset{c}{X}_i\,(\beta_1\,\overset{c}{X}_i + \epsilon_i)\} \\ &= E\{\beta_1\,\overset{c}{X}_i^{\,2} + \overset{c}{X}_i\,\epsilon_i)\} \\ &= \beta_1 E\{\overset{c}{X}_i^{\,2}\} + E\{\overset{c}{X}_i\}E\{\epsilon_i\} \\ &= \beta_1\sigma_x^2 \end{aligned}$$

Putting hats on first (optional), we solve $\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1\overline{X}$ and $\widehat{\sigma}_{xy} = \widehat{\beta}_1\widehat{\sigma}_x^2$ for $\widehat{\beta}_0$ and $\widehat{\beta}_1$, obtaining

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x^2} = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2} \text{ and} \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1\overline{X} \end{aligned}$$

These happen to be the same as the least-squares estimates.

Since $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are nice differentiable functions of various quantities that are essentially sample means, the multivariate delta method from the end of Section A.5 implies that the asymptotic joint distribution of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is bivariate normal. This holds regardless of the distributions of X_i and ϵ_i, provided only that their moments exist, and opens the door to distribution-free tests and confidence intervals. The story for multiple regression is almost exactly the same. The only requirement is a sample large enough for the Central Limt Theorem to work.

A.6.3 Maximum Likelihood Estimation

The idea behind maximum likelihood is to estimate the unknown parameter by the quantity that makes the probability of obtaining the observed data as large as possible. This probability is represented[11] by the likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(d_i;\theta),$$

where $f(d_i;\theta)$ is the density or probability mass function evaluated at d_i.

Let $\widehat{\theta}$ denote the usual Maximum Likelihood Estimate (MLE). That is, it is the parameter value for which the likelihood function is greatest, over all $\theta \in \Theta$. Because the log is an increasing function, maximizing the likelihood is equivalent to maximizing the log likelihood, which will be denoted

$$\ell(\theta) = \ln L(\theta).$$

In elementary situations where the support of the distribution does not depend on the parameter, you get the MLE by closing your eyes, differentiating the log likelihood, setting the derivative to zero, and solving for θ. Then if you are being careful, you carry out the second derivative test; if $\ell''(\widehat{\theta}) < 0$, the log likelihood is concave down at your answer, and you have found the maximum. Here is an example, useful mostly to clarify ideas and serve as a contrast to more realistic cases.

Example Let $D_1, \ldots, D_n$ be a random sample (independent and identically distributed random variables) from a distribution with density $f(y) = \frac{\theta}{(d+1)^{\theta+1}}$ for $d > 0$, where the unknown parameter θ is strictly greater than zero. The log likelihood is

$$\begin{aligned}
\ell(\theta) &= \ln \prod_{i=1}^{n} \frac{\theta}{(d_i+1)^{\theta+1}} \\
&= \sum_{i=1}^{n} \left(\ln\theta - (\theta+1)\ln(d_i+1)\right) \\
&= n\ln\theta - (\theta+1)\sum_{i=1}^{n} \ln(d_i+1)
\end{aligned}$$

Differentiating with respect to θ,

$$\begin{aligned}
\ell'(\theta) &= \frac{n}{\theta} - \sum_{i=1}^{n} \ln(d_i+1) \stackrel{\text{set}}{=} 0 \\
&\Rightarrow \quad \theta = \frac{1}{n}\sum_{i=1}^{n} \ln(d_i+1).
\end{aligned}$$

[11] If the data are discrete, the likelihood function is exactly the probability of observing the data that actually were observed. In the continuous case the likelihood function is approximately proportional to the probability of observing a data vector that falls into a small region surrounding the vector (point) that was observed.

Carrying out the second derivative test,

$$\ell''(\theta) = -n\theta^{-2} = -\frac{n}{\theta^2} < 0,$$

so the log likelihood function is concave down and we have located a maximum. This justifies writing $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \ln(d_i + 1)$. In R, if the data were in a numeric vector called `d`, the MLE would be `thetahat = mean(log(d+1))`.

Some Very Basic Math

If the calculations in that last example seemed obvious, you can skip this section.

I have noticed that a major obstacle for many students when doing maximum likelihood calculations is a set of basic mathematical operations they actually know. But the mechanics are rusty, or the notation used in Statistics is troublesome. So, with sincere apologies to those who don't need this, here are some basic rules.

- The distributive law: $a(b + c) = ab + ac$. You may see this in a form like
$$\theta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \theta x_i$$
- Power of a product is the product of powers: $(ab)^c = a^c\, b^c$. You may see this in a form like
$$\left(\prod_{i=1}^{n} x_i\right)^{\alpha} = \prod_{i=1}^{n} x_i^{\alpha}$$
- Multiplication is addition of exponents: $a^b a^c = a^{b+c}$. You may see this in a form like
$$\prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n \exp(-\theta \sum_{i=1}^{n} x_i)$$
- Powering is multiplication of exponents: $(a^b)^c = a^{bc}$. You may see this in a form like
$$(e^{\mu t + \frac{1}{2}\sigma^2 t^2})^n = e^{n\mu t + \frac{1}{2} n\sigma^2 t^2}$$
- Log of a product is sum of logs: $\ln(ab) = \ln(a) + \ln(b)$. You may see this in a form like
$$\ln \prod_{i=1}^{n} x_i = \sum_{i=1}^{n} \ln x_i$$
- Log of a power is the exponent times the log: $\ln(a^b) = b\, \ln(a)$. You may see this in a form like
$$\ln(\theta^n) = n \ln \theta$$
- The log is the inverse of the exponential function: $\ln(e^a) = a$. You may see this in a form like
$$\ln\left(\theta^n \exp(-\theta \sum_{i=1}^{n} x_i)\right) = n \ln \theta - \theta \sum_{i=1}^{n} x_i$$

Exercises A.6.3

1. Choose the correct answer.
   (a) $\prod_{i=1}^{n} e^{x_i} =$
      i. $\exp(\prod_{i=1}^{n} x_i)$
      ii. e^{nx_i}
      iii. $\exp(\sum_{i=1}^{n} x_i)$
   (b) $\prod_{i=1}^{n} \lambda e^{-\lambda x_i} =$
      i. $\lambda e^{-\lambda^n x_i}$
      ii. $\lambda^n e^{-\lambda n x_i}$
      iii. $\lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i)$
      iv. $\lambda^n \exp(-n\lambda \sum_{i=1}^{n} x_i)$
      v. $\lambda^n \exp(-\lambda^n \sum_{i=1}^{n} x_i)$
   (c) $\prod_{i=1}^{n} a_i^b =$
      i. na^b
      ii. a^{nb}
      iii. $(\prod_{i=1}^{n} a_i)^b$
   (d) $\prod_{i=1}^{n} a^{b_i} =$
      i. na^{b_i}
      ii. a^{nb_i}
      iii. $\sum_{i=1}^{n} a^{b_i}$
      iv. $a^{\prod_{i=1}^{n} b_i}$
      v. $a^{\sum_{i=1}^{n} b_i}$
   (e) $\left(e^{\lambda(e^t-1)}\right)^n =$
      i. $ne^{\lambda(e^t-1)}$
      ii. $e^{n\lambda(e^t-1)}$
      iii. $e^{\lambda(e^{nt}-1)}$
      iv. $e^{n\lambda(e^t-n)}$
   (f) $\left(\prod_{i=1}^{n} e^{-\lambda x_i}\right)^2 =$
      i. $e^{-2n\lambda x_i}$
      ii. $e^{-2\lambda \sum_{i=1}^{n} x_i}$
      iii. $2e^{-\lambda \sum_{i=1}^{n} x_i}$

2. True, or False?
   (a) $\sum_{i=1}^{n} \frac{1}{x_i} = \frac{1}{\sum_{i=1}^{n} x_i}$
   (b) $\prod_{i=1}^{n} \frac{1}{x_i} = \frac{1}{\prod_{i=1}^{n} x_i}$

(c) $\frac{a}{b+c} = \frac{a}{b} + \frac{a}{c}$

(d) $\ln(a+b) = \ln(a) + \ln(b)$

(e) $e^{a+b} = e^a + e^b$

(f) $e^{a+b} = e^a e^b$

(g) $e^{ab} = e^a e^b$

(h) $\prod_{i=1}^{n}(x_i + y_i) = \prod_{i=1}^{n} x_i + \prod_{i=1}^{n} y_i$

(i) $\ln(\prod_{i=1}^{n} a_i^b) = b\sum_{i=1}^{n} \ln(a_i)$

(j) $\sum_{i=1}^{n}\prod_{j=1}^{n} a_j = n\prod_{j=1}^{n} a_j$

(k) $\sum_{i=1}^{n}\prod_{j=1}^{n} a_i = \sum_{i=1}^{n} a_i^n$

(l) $\sum_{i=1}^{n}\prod_{j=1}^{n} a_{i,j} = \prod_{j=1}^{n}\sum_{i=1}^{n} a_{i,j}$

3. Simplify as much as possible.

(a) $\ln\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}$

(b) $\ln\prod_{i=1}^{n}\binom{m}{x_i}\theta^{x}(1-\theta)^{m-x_i}$

(c) $\ln\prod_{i=1}^{n}\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

(d) $\ln\prod_{i=1}^{n}\theta(1-\theta)^{x_i-1}$

(e) $\ln\prod_{i=1}^{n}\frac{1}{\theta}e^{-x_i/\theta}$

(f) $\ln\prod_{i=1}^{n}\frac{1}{\beta^{\alpha}\Gamma(\alpha)}e^{-x_i/\beta}x_i^{\alpha-1}$

(g) $\ln\prod_{i=1}^{n}\frac{1}{2^{\nu/2}\Gamma(\nu/2)}e^{-x_i/2}x_i^{\nu/2-1}$

(h) $\ln\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

(i) $\prod_{i=1}^{n}\frac{1}{\beta-\alpha}I(\alpha \leq x_i \leq \beta)$ (Express in terms of the minimum and maximum y_1 and y_n.)

Numerical maximum likelihood

Differentiating and setting the derivative to zero is all very well, and you will be asked to do it in some of the exercises. But in this course, as in much of applied statistics, you will find that you can write the log likelihood and differentiate it easily enough, but when you set the derivatives to zero, you obtain a set of equations that are impossible to solve explicitly. This means that the problem needs to be solved numerically. That is, you use a computer to calculate the value of the log likelihood for a set of parameter values, and you search until you have found the biggest one.

But how do you search? It's easy in one or two dimensions, but structural equation models can easily involve dozens, scores or even hundreds of parameters. It's a bit like being dropped by helicopter onto a mountain range, and asked to find the highest peak blindfolded. All you can do is walk uphill. The gradient is the direction of steepest

increase, so walk that way. How big a step should you take? That's a good question. When you come to a place where the surface is level, or approximately level, stop. How level is level enough? That's another good question. Once you find a "level" place, you can check to see if the surface is concave down there. If so, you're at a maximum. Is it the global maximum (the real MLE), or just a local maximum? It's usually impossible to tell for sure. You can get the helicopter to drop you in several different places fairly far apart, and if you always arrive at the same maximum you will feel more confident of your answer. But it could still be just a local maximum that is easy to reach. The main thing to observe is that where you start is *very* important. Another point is that for realistically big problems, you need high-grade, professionally written software.

The following example is one that you can do by hand, though maybe not with your eyes closed. But it will serve to illustrate the basic ideas of numerical maximum likelihood.

Example A.6.1

Let $D_1, \ldots, D_n$ be a random sample from a normal distribution with mean θ and variance θ^2. A sample of size 50 yields:

```
  5.85 -15.02 -13.24  -1.63  -0.07  -2.40  -3.02  -3.19  -5.16   0.79  -1.03 -10.69
-12.96  -4.55   0.57  -7.94  -6.80   2.95  -9.01  -9.33 -11.93  -7.13  10.34  -1.01
 -4.18  -1.30  -7.56  -1.25  -4.64  -4.88  -4.06  -1.91  -1.81  -6.92 -13.27  -5.52
  4.40 -12.17  -4.55  -5.82  -0.81 -19.28  -4.97  -7.78  -5.07  -5.45  -4.27  -4.98
 -9.56  -9.33
```

Find the maximum likelihood estimate of θ. You only need an approximate value; one decimal place of accuracy will do.

Again, this is a problem that can be solved explicitly by differentiation, and the reader is invited to give it a try before proceeding. Have the answer? Is it still the same day you started? Now for the numerical solution. First, write the log likelihood as

$$\begin{aligned} \ell(\theta) &= \ln \prod_{i=1}^{n} \frac{1}{|\theta|\sqrt{2\pi}} e^{-\frac{(d_i-\theta)^2}{2\theta^2}} \\ &= -n \ln |\theta| - \frac{n}{2} \ln(2\pi) - \frac{\sum_{i=1}^{n} d_i^2}{2\theta^2} + \frac{\sum_{i=1}^{n} d_i}{\theta} - \frac{n}{2}. \end{aligned}$$

We will do this in `R`. The data are in a file called `norm1.data`. Read it. Remember that $>$ is the `R` prompt.

```
> D <- scan("norm1.data")
Read 50 items
```

Now define a function to compute the log likelihood.

```
loglike1 <- function(theta) # Assume data are in a vector called D
    {
    sumdsq <- sum(D^2); sumd <- sum(D); n <- length(D)
    loglike1 <- -n * log(abs(theta)) - (n/2)*log(2*pi) - sumdsq/(2*theta^2) +
                sumd/theta - n/2
    loglike1 # Return value of function
    } # End definition of function loglike1
```

Just to show how the function works, compute it at a couple of values, say $\theta = 2$ and $\theta = -2$.

```
> loglike1(2)
[1] -574.2965
> loglike1(-2)
[1] -321.7465
```

Negative values of the parameter look more promising, but it is time to get systematic. The following is called a *grid search*. It is brutal, inefficient, and usually effective. It is too slow to be practical for large problems, but this is a one-dimensional parameter and we are only asked for one decimal place of accuracy. Where should we start? Since the parameter is the mean of the distribution, it should be safe to search within the range of the data. Start with widely spaced values and then refine the search. All we are doing is to calculate the log likelihood for a set of (equally spaced) parameter values and see where it is greatest. After all, that is the *idea* behind the MLE.

```
> min(D); max(D)
[1] -19.28
[1] 10.34
> Theta <- -20:10
> cbind(Theta,loglike1(Theta))
      Theta
 [1,]   -20  -211.5302
 [2,]   -19  -208.6709
 [3,]   -18  -205.6623
 [4,]   -17  -202.4911
 [5,]   -16  -199.1423
 [6,]   -15  -195.6002
 [7,]   -14  -191.8486
 [8,]   -13  -187.8720
 [9,]   -12  -183.6580
[10,]   -11  -179.2022
[11,]   -10  -174.5179
[12,]    -9  -169.6565
[13,]    -8  -164.7513
[14,]    -7  -160.1163
```

```
[15,]    -6  -156.4896
[16,]    -5  -155.6956
[17,]    -4  -162.7285
[18,]    -3  -193.8796
[19,]    -2  -321.7465
[20,]    -1 -1188.0659
[21,]     0        NaN
[22,]     1 -1693.1659
[23,]     2  -574.2965
[24,]     3  -362.2463
[25,]     4  -289.0035
[26,]     5  -256.7156
[27,]     6  -240.6729
[28,]     7  -232.2734
[29,]     8  -227.8888
[30,]     9  -225.7788
[31,]    10  -225.0279
```

First, we notice that at $\theta = 0$, the log likelihood is indeed Not a Number. For this problem, the parameter space is all the real numbers except zero – unless one wants to think of a normal random variable with zero variance as being degenerate at μ; that is, $P(D = \mu) = 1$. (In his case, what would the data look like?)

But the log likelihood is greatest around $\theta = -5$. We are asked for one decimal place of accuracy, so,

```
> Theta <- seq(from=-5.5,to=-4.5,by=0.1)
> Loglike <- loglike1(Theta)
> cbind(Theta,Loglike)
      Theta   Loglike
 [1,]  -5.5 -155.5445
 [2,]  -5.4 -155.4692
 [3,]  -5.3 -155.4413
 [4,]  -5.2 -155.4660
 [5,]  -5.1 -155.5487
 [6,]  -5.0 -155.6956
 [7,]  -4.9 -155.9136
 [8,]  -4.8 -156.2106
 [9,]  -4.7 -156.5950
[10,]  -4.6 -157.0767
[11,]  -4.5 -157.6665
> thetahat <- Theta[Loglike==max(Loglike)]
>             # Theta such that Loglike is the maximum of Loglike
> thetahat
[1] -5.3
```

To one decimal place of accuracy, the maximum is at $\theta = -5.3$. It would be easy to refine the grid and get more accuracy, but that will do. This is the last time we will see our friend the grid search, but you may find the approach useful in homework.

Now let's do the search in a more sophisticated way, using R's `nlm` (non-linear minimization) function. [12] The `nlm` function has quite a few arguments; try `help(nlm)`. The ones you always need are the first two: the name of the function, and a starting value (or vector of starting values, for multiparameter problems).

Where should we start? Since the parameter equals the expected value of the distribution, how about the sample mean? It is often a good strategy to use Method of Moment estimators as starting values for numerical maximum likelihood. Method of Moments estimation is reviewed in Section **??**.

One characteristic that `nlm` shares with most optimization routines is that it likes to *minimize* rather than maximizing. So we will minimize the negative of the log likelihood function. For this, it is necessary to define a new function, `loglike2`.

```
> mean(D)
[1] -5.051
> loglike2 <- function(theta) { loglike2 <- -loglike1(theta); loglike2 }
> nlm(loglike2,mean(D))
$minimum
[1] 155.4413

$estimate
[1] -5.295305

$gradient
[1] -1.386921e-05

$code
[1] 1

$iterations
[1] 4
```

By default, `nlm` returns a list with four elements; `minimum` is the value of the function at the point where it reaches its minimum, `estimate` is the value at which the minimum was located; that's the MLE. `Gradient` is the slope in the direction of greatest increase; it should be near zero. `Code` is a diagnosis of how well the optimization went; the value of 1 means everything seemed okay. See `help(nlm)` for more detail.

We could have gotten just the MLE with

[12] The `nlm` function is good but generic. See Numerical Recipes for a really good discussion of routines for numerically minimizing a function. They also provide source code. The *Numerical Recipes* books have versions for the Pascal, Fortran and Basic languages as well as C. This is a case where a book definitely delivers more than the title promises. It may be a cookbook, but it is a very good cookbook written by expert chemists.

```
> nlm(loglike2,mean(D))$estimate
[1] -5.295305
```

That's the answer, but the numerical approach misses some interesting features of the problem, which can be done with paper and pencil in this simple case. Differentiating the log likelihood separately for $\theta < 0$ and $\theta > 0$ to get rid of the absolute value sign, and then re-uniting the two cases since the answer is the same, we get

$$\ell'(\theta) = -\frac{n}{\theta} + \frac{\sum_{i=1}^n d_i^2}{\theta^3} - \frac{\sum_{i=1}^n d_i}{\theta^2}.$$

Setting $\ell'(\theta) = 0$ and re-arranging terms, we get

$$n\theta^2 + (\sum_{i=1}^n d_i)\theta - (\sum_{i=1}^n d_i^2) = 0.$$

Of course this expression is not valid at $\theta = 0$, because the function we are differentiating is not even defined there. The quadratic formula yields two solutions:

$$\frac{-\sum_{i=1}^n d_i \pm \sqrt{(\sum_{i=1}^n d_i)^2 + 4n\sum_{i=1}^n d_i^2}}{2n} = \frac{1}{2}\left(-\overline{d} \pm \sqrt{\overline{d}^2 + 4\frac{\sum_{i=1}^n d_i^2}{n}}\right), \tag{A.23}$$

where $\overline{d}$ is the sample mean.

Let's calculate these for the given data.

```
> meand <- mean(D) ; meandsq <- sum(D^2)/length(D)
> (-meand + sqrt(meand^2 + 4*meandsq) )/2
[1] 10.3463
> (-meand - sqrt(meand^2 + 4*meandsq) )/2
[1] -5.2953
```

The second solution is the one we found with the numerical search. What about the other one? Is it a minimum? Maximum? Saddle point? The second derivative test will tell us. The second derivative is

$$\ell''(\theta) = \frac{n}{\theta^2} - \frac{3\sum_{i=1}^n d_i^2}{\theta^4} + \frac{2\sum_{i=1}^n d_i}{\theta^3}.$$

Substituting A.23 into this does not promise to be much fun, so we will be content with a numerical answer for this particular data set. Call the first root `t1` and the second one (our MLE) `t2`.

```
> t1 <- (-meand + sqrt(meand^2 + 4*meandsq) )/2 ; t1
[1] 10.3463
> t2 <- (-meand - sqrt(meand^2 + 4*meandsq) )/2 ; t2
[1] -5.2953
> n <- length(D)
```

```
> # Now calculaate second derivative at t1 and t2
> n/t1^2 - 3*sum(D^2)/t1^4 + 2*sum(D)/t1^3
[1] -0.7061484
> n/t2^2 - 3*sum(D^2)/t2^4 + 2*sum(D)/t2^3
[1] -5.267197
```

The second derivative is negative in both cases; they are both local maxima! Which peak is higher?

```
> loglike1(t1)
[1] -224.9832
> loglike1(t2)
[1] -155.4413
```

So the maximum we found is higher, which makes sense because it's within the range of the data. But we only found it because we started searching near the correct answer.

Let's plot the log likelihood function, and see what this thing looks like. We know that because the natural log function goes to minus infinity as its (positive) argument approaches zero, the log likelihood plunges to $-\infty$ at $\theta = 0$. A plot would look like a giant icicle and we would not be able to see any detail where it matters. So we will zoom in by limiting the range of the y axis. Here is the `R` code.

```
Theta <- seq(from=-15,to=20,by=0.25); Theta <- Theta[Theta!=0]
Loglike <- loglike1(Theta)
# Check where to break off the icicle
max(Loglike); Loglike[Theta==-3];  Loglike[Theta==3]

plot(Theta,Loglike,type='l',xlim=c(-15,20),ylim=c(-375,-155),
    xlab=expression(theta),ylab="Log Likelihood")
    # This is how you get Greek letters.
```

Here is the picture. You can see the local maxima around $\theta = -5$ and $\theta = 10$, and also that the one for negative θ is a higher.

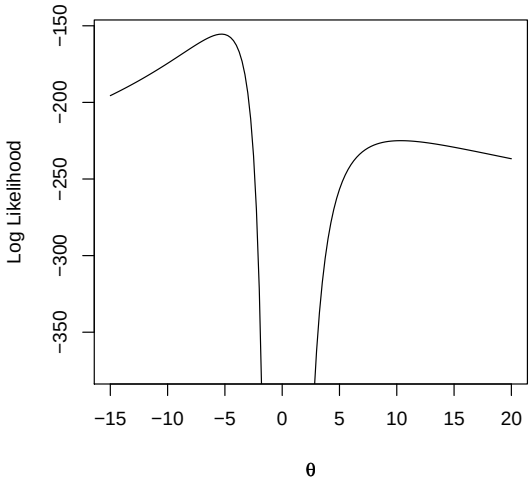
Presumably we would have reached the bad answer if we had started the search in a bad place. Let's try starting the search at $\theta = +3$.

```
> nlm(loglike2,3)
$minimum
[1] 283.7589

$estimate
[1] 64.83292

$gradient
[1] 0.701077
```

Figure A.1: Log Likelihood for Example A.6.1




```
$code
[1] 4

$iterations
[1] 100
```

What happened?! The answer is way off, nowhere near the positive root of 10.3463. And the `minimum` (of *minus* the log likelihood) is over 283, when it would have been 224.9832 at $\theta = 10.3463$.

What happened was that the slope of the function was very steep at our starting value of $\theta = 3$, so `nlm` took a huge step in a positive direction. It was too big, and landed in a nearly flat place. Then `nlm` wandered around until it ran out of its default number of iterations (notice `iterations`=100). The exit `code` of 4 means maximum number of iterations exceeded.

It should be better if we start close to the answer, say at $\theta = 8$.

```
> nlm(loglike2,8)
$minimum
[1] 224.9832

$estimate
[1] 10.34629

$gradient
```

```
[1] -4.120564e-08

$code
[1] 1

$iterations
[1] 6
```

That's better. The moral of this story is clear. Good starting are *very* important.

Now let us look at an example of a multi-parameter problem where an explicit formula for the MLE is impossible, and numerical methods are required.

Example A.6.2

Let $D_1, \ldots, D_n$ be a random sample from a Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$. The probability density function is

$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta} x^{\alpha-1}$$

for $x > 0$, and zero otherwise. Here is a random sample of size $n = 50$. For this example, the data are simulated using `R`, with known parameter values $\alpha = 2$ and $\beta = 3$. The seed for the random, number generator is set so the pseudo-random numbers can be recovered if necessary.

```
> set.seed(3201); alpha=2; beta=3
> D <- round(rgamma(50,shape=alpha, scale=beta),2); D
 [1] 20.87 13.74  5.13  2.76  4.73  2.66 11.74  0.75 22.07 10.49  7.26  5.82 13.08
[14]  1.79  4.57  1.40  1.13  6.84  3.21  0.38 11.24  1.72  4.69  1.96  7.87  8.49
[27]  5.31  3.40  5.24  1.64  7.17  9.60  6.97 10.87  5.23  5.53 15.80  6.40 11.25
[40]  4.91 12.05  5.44 12.62  1.81  2.70  3.03  4.09 12.29  3.23 10.94
> mean(D); alpha*beta
[1] 6.8782
[1] 6
> var(D); alpha*beta^2
[1] 24.90303
[1] 18
```

The parameter vector $\boldsymbol{\theta} = (\alpha, \beta)$, and the parameter space Θ is the first quadrant of $\mathbb{R}^2$.

$$\Theta = \{(\alpha, \beta) : \alpha > 0, \beta > 0\}$$

The log likelihood is

$$\begin{aligned}
\ell(\alpha,\beta) &= \ln \prod_{i=1}^{n} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} e^{-d_i/\beta} d_i^{\alpha-1} \\
&= \ln \left(\beta^{-n\alpha}\, \Gamma(\alpha)^{-n} \exp(-\frac{1}{\beta}\sum_{i=1}^{n} d_i) \left(\prod_{i=1}^{n} d_i \right)^{\alpha-1} \right) \\
&= -n\alpha \ln \beta - n \ln \Gamma(\alpha) - \frac{1}{\beta}\sum_{i=1}^{n} d_i + (\alpha-1)\sum_{i=1}^{n} \ln d_i.
\end{aligned}$$

The next step would be to partially differentiate the log likelihood with respect to α and β, set both partial derivatives to zero, and solve two equations in two unknowns. But even if you are confident that the gamma function is differentiable (it is), you will be unable to solve the equations. It has to be done numerically.

Define an `R` function for the minus log likelihood. Notice the `lgamma` function, a direct numerical approximation of $\ln \Gamma(\alpha)$. The plan is to numerically minimize the minus log likelihood function over all (α, β) pairs, for this particular set of data values.

```
> # Gamma minus log likelihood: alpha=a, beta=b
> gmll <- function(theta,datta)
+      {
+       a <- theta[1]; b <- theta[2]
+       n <- length(datta); sumd <- sum(datta); sumlogd <- sum(log(datta))
+       gmll <- n*a*log(b) + n*lgamma(a) + sumd/b - (a-1)*sumlogd
+      gmll
+      } # End function gmll
```

Where should the numerical search start? One approach is to start at reasonable estimates of α and β — estimates that can be calculated directly rather than by a numerical approximation. As in Example A.6.1, Method of Moments estimators are a convenient, high-quality choice.

For a gamma distribution, $E(D) = \alpha\beta$ and $Var(D) = \alpha\beta^2$. So,

$$\alpha = \frac{E(D)^2}{Var(D)} \quad \text{and} \quad \beta = \frac{Var(D)}{E(D)}.$$

Replacing population moments by sample moments and writing $\widetilde{\alpha}$ and $\widetilde{\beta}$ for the resulting Method of Moments estimators, we obtain

$$\widetilde{\alpha} = \frac{\overline{D}^2}{S_D^2} \quad \text{and} \quad \widetilde{\beta} = \frac{S_D^2}{\overline{D}},$$

where $\overline{D}$ is the sample mean and S_D^2 is the sample variance. For these data, the Method of Moments estimates are reasonably close to the correct values of $\alpha = 2$ and $\beta = 3$, but they are not perfect. Parameter estimates are not the same as parameters!

```
> momalpha <- mean(D)^2/var(D); momalpha
[1] 1.899754
> mombeta <- var(D)/mean(D); mombeta
[1] 3.620574
```

Now for the numerical search. This time, we will request that the `nlm` function return the *Hessian* at the place where the search stops. The Hessian is defined as follows. Suppose we are minimizing a function $g(\theta_1, \ldots, \theta_k)$ – say, a minus log likelihood. The Hessian is a $k \times k$ matrix of mixed partial derivatives. It may be written in terms of its (i, j) element s

$$\mathbf{H} = \left[\frac{\partial^2 g}{\partial \theta_i \partial \theta_j} \right]. \tag{A.24}$$

In the following, notice how the `nlm` function assumes that the first argument of the function being minimized is a vector of arguments over which we should minimize, and any other arguments (in this case, the name of the data vector) can be specified by name in the `nlm` function call.

```
> gammasearch = nlm(gmll,c(momalpha,mombeta),hessian=T,datta=D); gammasearch
$minimum
[1] 142.0316

$estimate
[1] 1.805930 3.808674

$gradient
[1] 2.847002e-05 9.133932e-06

$hessian
         [,1]      [,2]
[1,] 36.68932 13.127271
[2,] 13.12727  6.222282

$code
[1] 1

$iterations
[1] 6

> eigen(gammasearch$hessian)$values
[1] 41.565137  1.346466
```

The `nlm` object `gammasearch` is a linked list. The item `minimum` is the value of the minus log likelihood function where the search stops. The item `estimate` is the point at which

the search stops, so $\widehat{\alpha} = 1.805930$ and $\widehat{\beta} = 3.808674$. The `gradient` is

$$\left(-\frac{\partial \ell}{\partial \alpha}, -\frac{\partial \ell}{\partial \beta}\right)^{\top}.$$

Besides being the direction of steepest decrease, it's something that should be zero at the MLE. And indeed it is, give or take a bit of numerical inaccuracy.

The Hessian at the stopping place is in `gammasearch$hessian`. The Hessian is the matrix of mixed partial derivatives defined by

$$\mathbf{H} = \left[\frac{\partial^2(-\ell)}{\partial \theta_i \partial \theta_j}\right].$$

The rules about Hessian matrices are

- If the second derivatives are continuous, $\mathbf{H}$ is symmetric.
- If the gradient is zero at a point and $|\mathbf{H}| \neq 0$
  - If $\mathbf{H}$ is positive definite, there is a local minimum at the point.
  - If $\mathbf{H}$ is negative definite, there is a local maximum at the point.
  - If $\mathbf{H}$ has both positive and negative eigenvalues, the point is a saddle point.

The `eigen` command returns a linked list; one item is an array of the eigenvalues, and the other is the eigenvectors in the form of a matrix. Since for real symmetric matrices, positive definite is equivalent to all positive eigenvalues, it is convenient to check the eigenvalues to determine whether the numerical search has located a minimum. In this case it has. Finally, `code=1` means normal termination of the search, and `iterations=6` means the function took 6 steps downhill to reach its target.

It is very helpful to have the true parameter values $\alpha = 2$ and $\beta = 3$ for this example. $\widehat{\alpha} = 1.8$ seems pretty close, while and $\widehat{\beta} = 3.8$ seems farther off. This is a reminder of how informative confidence intervals and tests can be.

The Invariance Principle

The Invariance Principle of maximum likelihood estimation says that *the MLE of a function is that function of the MLE*. An example comes first, followed by formal details.

Example A.6.3

Let $D_1, \ldots, D_n$ be a random sample from a Bernoulli distribution (1=Yes, 0=No) with parameter $\theta, 0 < \theta < 1$. The parameter space is $\Theta = (0, 1)$, and the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \theta^{d_i}(1-\theta)^{1-d_i} = \theta^{\sum_{i=1}^{n} d_i}(1-\theta)^{n-\sum_{i=1}^{n} d_i}.$$

Differentiating the log likelihood with respect to θ, setting the derivative to zero and solving yields the usual estimate $\widehat{\theta} = \overline{d}$, the sample proportion.

Now suppose that instead of the probability, we write this model in terms of the *odds* of $D_i = 1$, a re-parameterization that is often useful in categorical data analysis. Denote the odds by θ'. The definition of odds is

$$\theta' = \frac{\theta}{1-\theta} = g(\theta). \tag{A.25}$$

As θ ranges from zero to one, θ' ranges from zero to infinity. So there is a new parameter space: $\theta' \in \Theta' = (0, \infty)$.

To write the likelihood function in terms of θ', first solve for θ, obtaining

$$\theta = \frac{\theta'}{1+\theta'} = g^{-1}(\theta').$$

The likelihood in terms of θ' is then

$$\begin{aligned} L(g^{-1}(\theta')) &= \theta^{\sum_{i=1}^n d_i}(1-\theta)^{n-\sum_{i=1}^n d_i} \\ &= \left(\frac{\theta'}{1+\theta'}\right)^{\sum_{i=1}^n d_i}\left(1-\frac{\theta'}{1+\theta'}\right)^{n-\sum_{i=1}^n d_i} \\ &= \left(\frac{\theta'}{1+\theta'}\right)^{\sum_{i=1}^n d_i}\left(\frac{1+\theta'-\theta'}{1+\theta'}\right)^{n-\sum_{i=1}^n d_i} \\ &= \frac{\theta'^{\sum_{i=1}^n d_i}}{(1+\theta')^n}. \end{aligned}$$

Note how re-parameterization changes the functional form of the likelihood function. The general formula is $L'(\theta') = L(g^{-1}(\theta'))$. For this example,

$$L'(\theta') = \frac{\theta'^{\sum_{i=1}^n d_i}}{(1+\theta')^n}. \tag{A.26}$$

At this point one could differentiate the log of (A.26) with respect to θ', set the derivative to zero, and solve for θ'. The point of the invariance principle is that this is unnecessary. The maximum likelihood estimator of $g(\theta)$ is $g(\widehat{\theta})$, so one need only look at (A.25) and write

$$\widehat{\theta'} = \frac{\widehat{\theta}}{1-\widehat{\theta}} = \frac{\overline{d}}{1-\overline{d}}\ .$$

It is often convenient to parameterize a statistical model in more than one way. The invariance principle can save a lot of work in practice, because it says that you only have to maximize the likelihood function once. It is useful theoretically too.

In Example A.6.3, the likelihood function has only one maximum and the function g linking θ' to θ' is one-to-one, which is why we can write g^{-1}. This is the situation where the invariance principle is clearest and most useful. Here is a proof.

Let the parameter $\theta \in \Theta$, and re-parameterize by $\theta' = g(\theta)$. The new parameter space is $\Theta' = \{\theta' : \theta' = g(\theta), \theta \in \Theta\}$. The function $g : \Theta \to \Theta'$ is one-to-one, meaning that there exists a function g^{-1} such that $g^{-1}(g(\theta)) = \theta$ for all $\theta \in \Theta$. Suppose the likelihood function $L(\theta)$ has a unique maximum at $\widehat{\theta} \in \Theta$, so that for all $\theta \in \Theta$ with $\theta \neq \widehat{\theta}$, $L(\widehat{\theta}) > L(\theta)$. For every $\theta \in \Theta$,

$$L(\theta) = L(g^{-1}(g(\theta))) = L(g^{-1}(\theta')) = L'(\theta')$$

Maximizing $L'(\theta')$ over $\theta' \in \Theta'$ yields $\widehat{\theta'}$ satisfying $L'(\widehat{\theta'}) \geq L'(\theta')$ for all $\theta' \in \Theta'$. The invariance principle says $\widehat{\theta'} = g(\widehat{\theta})$.

Let $\theta_0 = g^{-1}(\widehat{\theta'})$ so that $g(\theta_0) = \widehat{\theta'}$. The objective is to show that this value $\theta_0 \in \Theta$ equals $\widehat{\theta}$. Suppose on the contrary that $\theta_0 \neq \widehat{\theta}$. Then because the maximum of $L(\theta)$ over Θ is unique, $L(\widehat{\theta}) > L(\theta_0)$. Therefore,

$$\begin{aligned} & L(g^{-1}(g(\widehat{\theta}))) > L(g^{-1}(g(\theta_0))) \\ \Rightarrow \quad & L'(g(\widehat{\theta})) > L'(g(\theta_0)) \\ \Rightarrow \quad & L'(g(\widehat{\theta})) > L'(\widehat{\theta'}). \end{aligned}$$

Since $g(\widehat{\theta}) \in \Theta'$, this contradicts $L'(\widehat{\theta'}) \geq L'(\theta')$ for all $\theta' \in \Theta'$, showing $\widehat{\theta} = \theta_0$. Not leaving anything to the imagination, we then have $g(\widehat{\theta}) = g(\theta_0) = \widehat{\theta'}$.

This concludes the proof, but it may be useful to establish the "obvious" fact that uniqueness of the maximum over Θ implies uniqueness of the maximum over Θ'. If $\widehat{\theta'_1}$ and $\widehat{\theta'_2}$ are two points in Θ' with $L'(\widehat{\theta'_1}) \geq L'(\theta')$ and $L'(\widehat{\theta'_2}) \geq L'(\theta')$ for all $\theta' \in \Theta'$, the preceding argument shows that $g(\widehat{\theta}) = \widehat{\theta'_1}$ and $g(\widehat{\theta}) = \widehat{\theta'_2}$. Because function values are unique, this can only happen if $\widehat{\theta'_1} = \widehat{\theta'_2}$

Exercises ??

A.6.1) For each of the following distributions, derive a general expression for the Maximum Likelihood Estimator (MLE). Carry out the second derivative test to make sure you have a maximum. (What is the relationship of this to the Hessian?) Then use the data to calculate a numerical estimate.

(a) $p(x) = \theta(1-\theta)^x$ for $x = 0, 1, \ldots$, where $0 < \theta < 1$. Data: `4, 0, 1, 0, 1, 3, 2, 16, 3, 0, 4, 3, 6, 16, 0, 0, 1, 1, 6, 10`. Answer: 0.2061856

(b) $f(x) = \frac{\alpha}{x^{\alpha+1}}$ for $x > 1$, where $\alpha > 0$. Data: `1.37, 2.89, 1.52, 1.77, 1.04, 2.71, 1.19, 1.13, 15.66, 1.43` Answer: 1.469102

(c) $f(x) = \frac{\tau}{\sqrt{2\pi}} e^{-\frac{\tau^2 x^2}{2}}$, for x real, where $\tau > 0$. Data: `1.45, 0.47, -3.33, 0.82, -1.59, -0.37, -1.56, -0.20` Answer: 0.6451059

(d) $f(x) = \frac{1}{\theta} e^{-x/\theta}$ for $x > 0$, where $\theta > 0$. Data: `0.28, 1.72, 0.08, 1.22, 1.86, 0.62, 2.44, 2.48, 2.96` Answer: 1.517778

A.6.2) The univariate normal density is

$$f(y|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(y-\mu)^2}{\sigma^2}}$$

(a) Show that the univariate normal likelihood may be written

$$L(\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp -\frac{n}{2\sigma^2}\left\{\widehat{\sigma}^2 + (\overline{y}-\mu)^2\right\},$$

where $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$. Hint: Add and subtract $\overline{y}$.

(b) How does this expression allow you to see *without differentiating* that the MLE of μ is $\overline{y}$?

A.6.3) Let $X_1, \ldots, X_5$ be a random sample from a Gamma distribution with parameters $\alpha > 0$ and $\beta = 1$. That is, the density is

$$f(x;\alpha) = \frac{1}{\Gamma(\alpha)} e^{-x} x^{\alpha-1}$$

for $x > 0$, and zero otherwise.

The five data values are 2.06, 1.08, 0.96, 1.32, 1.53. Find an approximate numerical value of the maximum likelihood estimate of α. Your final answer is one number. For this question you will hand in a one-page printout. On the back, you will write a brief explanation of what you did.

A.6.4) For each of the following distributions, try to derive a general expression for the Maximum Likelihood Estimator (MLE). Then, use R's `nlm` function to obtain the MLE numerically for the data supplied for the problem. The data are in a separate HTML document, because it saves a lot of effort to copy and paste rather than typing the data in by hand, and PDF documents can contain invisible characters that mess things up. NOTE! Put them here as well as in assignment HTML document.

(a) $f(x) = \frac{1}{\pi[1+(x-\theta)^2]}$ for x real, where $-\infty < \theta < \infty$.

```
-3.77  -3.57 4.10 4.87  -4.18  -4.59  -5.27  -8.33 5.55  -4.35  -0.55 5.57
-34.78 5.05 2.18 4.12  -3.24 3.78  -3.57 4.86
```

For this one, try at least two different starting values and *plot the minus log likelihood function!*

(b) $f(x) = \frac{1}{2}e^{-|x-\theta|}$ for x real, where $-\infty < \theta < \infty$.

```
3.36 0.90 2.10 1.81 1.62 0.16 2.01 3.35 4.75 4.27 2.04
```

(c) $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ for $0 < x < 1$, where $\alpha > 0$ and $\beta > 0$.

```
0.45 0.42 0.38 0.26 0.43 0.24 0.32 0.50 0.44 0.29 0.45 0.29 0.29 0.32 0.30
0.32 0.30 0.38 0.43 0.35 0.32 0.33 0.29 0.20 0.46 0.31 0.35 0.27 0.29 0.46
0.43 0.37 0.32 0.28 0.20 0.26 0.39 0.35 0.35 0.24 0.36 0.28 0.32 0.23 0.25
0.43 0.30 0.43 0.33 0.37
```

If you are getting a lot of warnings, maybe it's because the numerical search is leaving the parameter space. If so and if you are using R, try `help(nlminb)`.

For each distribution, be able to state (briefly) why differentiating the log likelihood and setting the derivative to zero does not work. For the computer part, bring to the quiz one sheet of printed output for each of the 3 distributions. The three sheets should be separate, because you may hand only one of them in. Each printed page should show the following, *in this order*.

- Definition of the function that computes the likelihood, or log likelihood, or minus log likelihood or whatever.
- How you got the data into R – probably a `scan` statement.
- Listing of the data for the problem.
- The `nlm` statement and resulting output.

A.6.5) Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ is multivariate normal with mean zero and covariance matrix $\sigma^2\mathbf{I}_n$, with $\sigma^2 > 0$ an unknown constant.

(a) What is the distribution of $\mathbf{Y}$? There is no need to show any work.

(b) Assuming that the columns of $\mathbf{X}$ are linearly independent, show that the maximum likelihood estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y$. Don't use derivatives. The trick is to add and subtract $\hat{\boldsymbol{\beta}}$, distribute the expected value, and simplify. Does your answer apply for any value of σ^2? Why or why not?

(c) Given the MLE of $\boldsymbol{\beta}$, find the MLE of σ^2. Show your work. This time you may differentiate.

Interval Estimation and Testing

All the tests and confidence intervals here are based on large-sample approximations, primarily the Central Limit Theorem. See Section A.5 for basic definitions and results. They are valid as the sample size $n \to \infty$, but frequently perform well for samples that are only fairly large. How big is big enough? This is a legitimate question, and the honest answer is that it depends upon the distribution of the data. In practice, people often just apply these tools almost regardless of the sample size, because nothing better is available. Some do it with their eyes closed, some squint, and some have their eyes wide open.

The basic result comes from the research of Abraham Wald (give a source) in the 1950s. *As the sample size n increases, the distribution of the maximum likelihood estimator* $\widehat{\boldsymbol{\theta}}_n$ *approaches a multivariate normal* with expected value $\boldsymbol{\theta}$ and variance-covariance matrix $\mathbf{V}_n(\boldsymbol{\theta})$. It is quite remarkable that anyone could figure this out, given that it includes cases like the Gamma, where no closed-form expressions for the maximum likelihood estimators are possible. The theorem in question is not true for every distribution, but it is true if the distribution of the data is not too strange. The precise meaning of "not too

strange" is captured in a set of technical conditions called *regularity conditions*. Volume 2 of *Kendall's advanced theory of statistics* [14] is a good textbook source for the details.

If $\boldsymbol{\theta}$ is a $k \times 1$ matrix, then $\mathbf{V}_n(\boldsymbol{\theta})$ is a $k \times k$ matrix, called the *asymptotic covariance matrix* of the estimators. It's not too surprising that it depends on the parameter $\boldsymbol{\theta}$, and it also depends on the sample size n. Using the asymptotic covariance matrix, it is possible to construct a variety of useful tests and confidence intervals.

Fisher Information

The fact that $\mathbf{V}_n(\boldsymbol{\theta})$ depends on the unknown parameter will present no problem; substituting $\widehat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$ yields an *estimated* asymptotic covariance matrix. So consider the form of the matrix $\mathbf{V}$.

Think of a one-parameter maximum likelihood problem, where we differentiate the log likelihood, set the derivative to zero and solve for θ; the solution is $\widehat{\theta}$. The log likelihood will be concave down at $\widehat{\theta}$, but the exact way it looks will depend on the distribution as well as the sample size. In particular, it could be almost flat at $\widehat{\theta}$, or it could be nearly a sharp peak, with extreme downward curvature. In the latter case, clearly the log likelihood is more informative about θ. It contains more information. One of the many good ideas of R. A. F. Fisher was that the second derivative reflects curvature, and and can be viewed as a measure of the information provided by the sample data. It is called the *Fisher Information* in his honour.

Now with increasing sample size, nearly all log likelihood functions acquire more and more downward curvature at the MLE. This makes sense – more data provide more information. But how about the information from just one observation? If you look at the second derivative of the log likelihood function,

$$\frac{\partial^2 \ell}{\partial \theta^2} = \frac{\partial^2}{\partial \theta^2} \ln \prod_{i=1}^{n} f(d_i;\theta) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln f(d_i;\theta),$$

you see that it is the sum of n quantities. Each observation is contributing a piece to the downward curvature. But how much? Well, it depends on the particular data value x_i. But the data are a random sample, so in fact the contribution is a random quantity: $\frac{\partial^2}{\partial \theta^2} \ln f(X_i;\theta)$. How about the information one would *expect* an observation to contribute? Okay, take the expected value. Finally, note that because the curvature is down at the MLE, the quantity we are discussing is negative. But we want to call this "information," and it would be nicer if it were a positive number, so higher values meant more information. Okay, multiply by -1. This leads to the definition of the Fisher Information in a single observation:

$$I(\theta) = E\left[-\frac{\partial^2}{\partial \theta^2} \ln f(D_i;\theta)\right]. \tag{A.27}$$

The information is the same for $i = 1, \ldots, n$, and the Fisher Information in the entire sample is just $nI(\theta)$.

It was clear that Fisher was onto something good, because for many problems where the variance of $\widehat{\theta}$ can be calculated exactly, it is one divided by the Fisher Information. Subsequently Cramér and Rao discovered the *Cramér-Rao Inequality*, which says that for *any* statistic T that is an unbiased estimator of θ,

$$Var(T) \geq \frac{1}{nI(\theta)}.$$

That's impressive, because to have a small variance is a great property in an estimator; it means precise estimation. The Cramér-Rao inequality tells us that in terms of variance, one cannot do better than an unbiased estimator whose variance equals hte reciprocal (inverse) of the Fisher Information, and many MLEs do that. Subsequently, Wald[13] showed that under some regularity conditions, the variances of maximum likelihood estimators in general attain the Cramér-Rao lower bound as $n \to \infty$. Thus, to learn the asymptotic variance of $\widehat{\theta}$, you do not need an explicit formula for $\widehat{\theta}$. All you need is the Fisher Information. Also, in terms of variance nothing can beat maximum likelihood estimation, at least for large samples. So if the distribution of the data is known so you can write down the likelihood, it is difficult to justify any method of estimation other than maximum likelihood.

Calculating the expected value in (A.27) is often not too hard because taking the log and differentiating twice results in some simplification; it's a source of many fun homework problems. But still it can be a chore, especially for multiparameter problems, which will be taken up shortly. For larger sample sizes, the Law of Large Numbers (Section A.5) guarantees that the expected value can be approximated quite well by a sample mean, so that

$$I(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2}\ln f(D_1;\theta)\right] \approx \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial^2}{\partial\theta^2}\ln f(D_i;\theta).$$

This is sometimes called the *observed* Fisher Information.

Multiplying the observed Fisher Information by n to get the approximate information in the entire sample yields

$$\sum_{i=1}^{n} -\frac{\partial^2}{\partial\theta^2}\ln f(D_i;\theta) = \frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n} -\ln f(D_i;\theta) = \frac{\partial^2}{\partial\theta^2}\left(-\ln\prod_{i=1}^{n} f(D_i;\theta)\right).$$

That's just the second derivative of the minus log likelihood.

The parameter θ is unknown, so to get the *estimated* Fisher Information in the whole sample, substitute $\widehat{\theta}$. The result is

$$\frac{\partial^2}{\partial\theta^2}\left(-\ln\prod_{i=1}^{n} f(D_i;\widehat{\theta})\right).$$

That's the second derivative of minus the log likelihood, evaluated at the maximum likelihood estimate. And, it's a function of the sample data that is not a function of any

[13] Need a reference

unknown parameters; in other words it is a statistic. If you have already carried out the second derivative test to check that you really had a maximum, all you need to do to estimate the variance of $\widehat{\theta}$ is take the reciprocal of the second derivative and multiply by -1. It is truly remarkable how neatly this all works out.

Generalization to the multivariate case is very natural. Now the parameter is $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$ and the Fisher Information *Matrix* is a $k \times k$ matrix of (expected) mixed partial derivatives, defined by

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \left[-E\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} f(\mathbf{D}_1; \boldsymbol{\theta}) \right) \right],$$

where the boldface $\mathbf{D}_i$ is an acknowledgement that the data might also be multivariate.

In the estimated observed Fisher Information evaluated at the MLE (which will simply be called the "Fisher Information Matrix" unless other wise noted), expected value is replaced by a sample mean and $\boldsymbol{\theta}$ is replaced by $\widehat{\boldsymbol{\theta}}$. The formula is

$$\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}}) = \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \left(-\ln \prod_{q=1}^{n} f(\mathbf{D}_q; \widehat{\boldsymbol{\theta}}) \right) \right] = \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \left(-\ell(\widehat{\boldsymbol{\theta}}) \right) \right]. \tag{A.28}$$

In the one-dimensional case, one over the estimated Fisher Information is the (estimated) asymptotic variance of the maximum likelihood estimator. *In the multi-parameter case, the Fisher Information is a matrix, and the estimated asymptotic variance-covariance matrix is its inverse.* Denoting the estimated asymptotic covariance matrix by $\widehat{\mathbf{V}}_n$, we have

$$\widehat{\mathbf{V}}_n = \boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}}_n)^{-1}. \tag{A.29}$$

Now comes the really good part. Comparing Formula (A.28) for the Fisher Information to Formula (A.24) for the Hessian, we see that they are exactly the same. And *the Hessian evaluated at $\widehat{\boldsymbol{\theta}}$ is a by-product of the numerical search for the MLE*[14].

So to get the asymptotic covariance matrix, minimize minus the log likelihood, tell the software to give you the Hessian, and calculate the inverse by computer. The theoretical story may be a bit long here, but what you have to do in practice is quite simple.

Continuing with the Gamma distribution Example A.6.2, the Hessian is

```
> gammasearch$hessian
         [,1]      [,2]
[1,] 36.68932 13.127271
[2,] 13.12727  6.222282
```

and the asymptotic covariance is just

[14] At least for generic numerical minimization routines like `R`'s `nlm`. Some specialized methods like iterative proportional fitting of log-linear models and Fisher scoring (iteratively re-weighted least squares) for generalized linear models maximize the likelihood indirectly and do not require calculation of the Hessian.

```
> Vhat = solve(gammasearch$hessian); V
           [,1]       [,2]
[1,]  0.1111796 -0.2345577
[2,] -0.2345577  0.6555638 .
```

The diagonal elements of $\widehat{\mathbf{V}}$ are the estimated variances of the sampling distributions of $\widehat{\alpha}$ and $\widehat{\beta}$ respectively, and their square roots are the standard errors.

```
> SEalphahat = sqrt(Vhat[1,1]); SEbetahat = sqrt(Vhat[2,2])
```

In general, let θ denote an element of the parameter vector, let $\widehat{\theta}$ be its maximum likelihood estimator, and let the standard error of $\widehat{\theta}$ be written $S_{\widehat{\theta}}$. Then Wald's Central Limit Theorem for maximum likelihood estimators tells us that

$$Z = \frac{\widehat{\theta} - \theta}{S_{\widehat{\theta}}} \tag{A.30}$$

has an approximate standard normal distribution. In particular, for the Gamma example

$$Z_1 = \frac{\widehat{\alpha} - \alpha}{S_{\widehat{\alpha}}} \quad \text{and} \quad Z_2 = \frac{\widehat{\beta} - \beta}{S_{\widehat{\beta}}}$$

may be treated as standard normal.

Confidence Intervals

These quantities may be used to produce both tests and confidence intervals. For example, a 95% confidence interval for the parameter θ is obtained as follows.

$$\begin{aligned} 0.95 &\approx Pr\{-1.96 \leq Z \leq 1.96\} \\ &= Pr\left\{-1.96 \leq \frac{\widehat{\theta} - \theta}{S_{\widehat{\theta}}} \leq 1.96\right\} \\ &= Pr\left\{\widehat{\theta} - 1.96\, S_{\widehat{\theta}} \leq \theta \leq \widehat{\theta} + 1.96\, S_{\widehat{\theta}}\right\} \end{aligned}$$

This could also be written $\widehat{\theta} \pm 1.96\, S_{\widehat{\theta}}$.

If you are used to seeing confidence intervals with a $\sqrt{n}$ and wondering where it went, recall that $S_{\overline{X}} = \frac{S}{\sqrt{n}}$. The $\sqrt{n}$ is also present in the confidence interval for θ, but it is embedded in $S_{\widehat{\theta}}$.

Here are the 95% confidence intervals for the Gamma distribution example:

```
> alphahat = gammasearch$estimate[1]; betahat = gammasearch$estimate[2]
> Lalpha = alphahat - 1.96*SEalphahat; Ualpha = alphahat + 1.96*SEalphahat
> Lbeta = betahat - 1.96*SEbetahat; Ubeta = betahat + 1.96*SEbetahat
> cat("\nEstimated alpha = ",round(alphahat,2),"  95 percent CI from ",
+     round(Lalpha,2)," to ",round(Ualpha,2), "\n\n")
```

```
Estimated alpha =  1.81   95 percent CI from  1.15  to  2.46

> cat("\nEstimated beta = ",round(betahat,2),"  95 percent CI from ",
+     round(Lbeta,2)," to ",round(Ubeta,2), "\n\n")

Estimated beta =  3.81   95 percent CI from  2.22  to  5.4
```

Notice that while the parameter estimates may not seem very accurate, the 95% confidence intervals do include the true parameter values $\alpha = 2$ and $\beta = 3$.

Z-tests

The standard normal variable in (A.30) can be used to form a Z-test of $H_0 : \theta = \theta_0$ using

$$Z = \frac{\widehat{\theta} - \theta_0}{S_{\widehat{\theta}}}.$$

So for example, suppose the data represent time intervals between events occurring in time, and we wonder whether the events arise from a Poisson process. In this case the distribution of times would be exponential, which means $\alpha = 1$. To test this null hypothesis at the 0.05 level,

```
> Z = (alphahat-1)/SEalphahat; Z
[1] 2.417046
> pval = 2*(1-pnorm(abs(Z))); pval # Two-sided test
[1] 0.01564705
```

So, the null hypothesis is rejected, and because the value is positive, the conclusion is that the true value of α is greater than one[15].

When statistical software packages display this kind of large-sample Z-test, they usually just divide $\widehat{\theta}$ by its standard error, testing the null hypothesis $H_0 : \theta = 0$. For parameters like regression coefficients, this is usually a good generic choice.

[15]The following basic question arises from time to time. Suppose a null hypothesis is rejected in favour of a two-sided alternative. Are we then "allowed" to look at the sign of the test statistic and conclude that $\theta < \theta_0$ or $\theta > \theta_0$, or must we just be content with saying $\theta \neq \theta_0$? The answer is that directional conclusions are theoretically justified as well as practically desirable. Think of splitting up the two-sided level α test (call it the *overall test*) into two one-sided tests with significance level $\alpha/2$. The null hypotheses of these tests are $H_{0,a} : \theta \leq \theta_0$ and $H_{0,b} : \theta \geq \theta_0$. Exactly one of these null hypotheses will be rejected if and only if the null hypothesis of the overall test is rejected, so the set of two one-sided tests is fully equivalent to the overall two-sided test. And directional conclusions from the one-sided tests are clearly justified.

On a deeper level, notice that the null hypothesis of the overall test is the intersection of the null hypotheses of the one-sided tests, and its critical region (rejection region) is the union of the critical regions of the one-sided tests. This makes the two one-sided tests a set of *union-intersection multiple comparisons*, which are always simultaneously protected against Type I error at the significance level of the overall test. Performing the two-sided test and then following up with a one-sided test is very much like following up a statistically significant ANOVA with Scheffeé tests. Indeed, Scheffé tests are another example of union-intersection multiple comparisons. See [7] for details.

A.6.4 Wald Tests

The approximate multivariate normality of the MLE can be used to construct a larger class of hypothesis tests for *linear* null hypotheses. A linear null hypothesis sets a collection of linear combinations of the parameters to zero. Suppose $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$ is a $k \times 1$ vector. A linear null hypothesis can be written

$$H_0 : \mathbf{C}\boldsymbol{\theta} = \mathbf{h},$$

where $\mathbf{C}$ is an $r \times k$ matrix of constants, with rank r, $r \leq k$. As an example let $\boldsymbol{\theta} = (\theta_1, \ldots \theta_7)^\top$, and the null hypothesis is

$$\theta_1 = \theta_2, \quad \theta_6 = \theta_7, \quad \frac{1}{3}(\theta_1 + \theta_2 + \theta_3) = \frac{1}{3}(\theta_4 + \theta_5 + \theta_6).$$

This may be expressed in the form $\mathbf{C}\boldsymbol{\theta} = \mathbf{h}$ as follows:

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Recall from Section A.4 of this appendix that if $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\mathbf{C}$ is an $r \times k$ constant matrix of rank r, then

$$\mathbf{C}\mathbf{X} \sim N_r(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top)$$

and

$$(\mathbf{C}\mathbf{X} - \mathbf{C}\boldsymbol{\mu})^\top (\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top)^{-1} (\mathbf{C}\mathbf{X} - \mathbf{C}\boldsymbol{\mu}) \sim \chi^2(r).$$

Similar facts hold asymptotically — that is approximately, as the sample size n approaches infinity. Because (approximately) $\widehat{\boldsymbol{\theta}}_n \sim N_k(\boldsymbol{\theta}, \widehat{\mathbf{V}}_n)$,

$$\mathbf{C}\widehat{\boldsymbol{\theta}}_n \sim N_r(\mathbf{C}\boldsymbol{\theta}, \mathbf{C}\widehat{\mathbf{V}}_n\mathbf{C}^\top)$$

and

$$(\mathbf{C}\widehat{\boldsymbol{\theta}}_n - \mathbf{C}\boldsymbol{\theta})^\top (\mathbf{C}\widehat{\mathbf{V}}_n\mathbf{C}^\top)^{-1} (\mathbf{C}\widehat{\boldsymbol{\theta}}_n - \mathbf{C}\boldsymbol{\theta}) \sim \chi^2(r).$$

So, if $H_0 : \mathbf{C}\boldsymbol{\theta} = \mathbf{h}$ is true, we have the Wald test statistic

$$W_n = (\mathbf{C}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})^\top (\mathbf{C}\widehat{\mathbf{V}}_n\mathbf{C}^\top)^{-1} (\mathbf{C}\widehat{\boldsymbol{\theta}}_n - \mathbf{h}) \sim \chi^2(r), \tag{A.31}$$

where again,

$$\widehat{\mathbf{V}}_n = \boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}})^{-1} = \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j}\left(-\ell(\widehat{\boldsymbol{\theta}})\right)\right]^{-1}.$$

Here is a test of $H_0 : \alpha = \beta$ for the Gamma distribution example. A little care must be taken to ensure that the matrices in (A.31) are the right size.

```
> #  H0: C theta = 0 is that alpha = beta <=> alpha-beta=0
> # Name C is used by R
> CC = rbind(c(1,-1)); is.matrix(CC); dim(CC)
[1] TRUE
[1] 1 2
> thetahat = as.matrix(c(alphahat,betahat)); dim(thetahat)
[1] 2 1
> W = t(CC%*%thetahat) %*% solve(CC%*%Vhat%*%t(CC)) %*% CC%*%thetahat
> W = as.numeric(W) # it was a 1x1 matrix
> pval2 = 1-pchisq(W,1)
> cat("Wald Test:  W = ", W, ", p = ", pval2, "\n")
Wald Test:  W =  3.245501 , p =  0.07161978
```

We might as well define a function to do Wald tests in general. In the function `WaldTest`, the null hypothesis is $\mathbf{L}\boldsymbol{\theta} = \mathbf{h}$, but that's just because the name `C` is used by `R` for contrasts. The function returns a pair of quantities, the Wald test statistic and the p-value.

```
> WaldTest = function(L,thetahat,h=0) # H0: L theta = h
+     {
+     WaldTest = numeric(2)
+     names(WaldTest) = c("W","p-value")
+     dfree = dim(L)[1]
+     W = t(L%*%thetahat-h) %*% solve(L%*%Vhat%*%t(L)) %*% (L%*%thetahat-h)
+     W = as.numeric(W)
+     pval = 1-pchisq(W,dfree)
+     WaldTest[1] = W; WaldTest[2] = pval
+     WaldTest
+     } # End function WaldTest
```

Here is the same test of $H_0 : \alpha = \beta$ done immediately above, just to test out the function. Notice that the default value of $\mathbf{h}$ in $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$ is zero, so it does not have to be specified. The matrix `CC` has already been created, and the computed values are the same as before, naturally.

```
> WaldTest(CC,as.matrix(c(alphahat,betahat)))
         W    p-value
3.24550127 0.07161978
```

Here is a test of $H_0 : \alpha = 2, \beta = 3$, which happen to be the true parameter values. The null hypothesis is not rejected.

```
> C2 = rbind(c(1,0),
+           c(0,1) )
> WaldTest(C2,as.matrix(c(alphahat,betahat)),c(2,3))
        W   p-value
1.3305497 0.5141322
```

Finally, here is a test of $H_0 : \alpha = 1$, which was done earlier with a Z-test.

```
> WaldTest(t(c(1,0)),as.matrix(c(alphahat,betahat)),1)
         W    p-value
5.84210645 0.01564708
> Z; pval
[1] 2.417045
[1] 0.01564708
> Z^2
[1] 5.842106
```

The results of the Wald and Z tests are identical, with $W_n = Z^2$. In general, suppose the matrix $\mathbf{C}$ in $H_0 : \mathbf{C}\boldsymbol{\theta} = \mathbf{h}$ has just a single row, and that row contains one 1 in position j and all the rest zeros. Take a look at Formula (A.31) for the Wald test statistic. Pre-multiplying by $\mathbf{C}$ in $\mathbf{C}\widehat{\mathbf{V}}_n$ picks out row j of $\widehat{\mathbf{V}}_n$, and post-multiplying by $\mathbf{C}^\top$ picks out column j of the result, so that $\mathbf{C}\widehat{\mathbf{V}}_n\mathbf{C}^\top = \widehat{v}_{j,j}$, and inverting it puts it in the denominator. In the numerator, $(\mathbf{C}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})^\top(\mathbf{C}\widehat{\boldsymbol{\theta}}_n - \mathbf{h}) = (\widehat{\theta}_j - \theta_{j,0})^2$, so that $W_n = Z^2$. Thus, squaring a large-sample Z-test gives a Wald chisquare test with one degree of freedom.

A.6.5 Likelihood Ratio Tests

Likelihood ratio tests fall into two categories, exact and large-sample. The main examples of exact likelihood ratio tests include are the standard F-tests and t-tests associated with regression and the analysis of variance for normal data. Here, we concentrate on the large-sample likelihood ratio tests.

Consider the following hypothesis-testing framework. The data are $D_1, \ldots, D_n$. The distribution of these independent and identically distributed random variables depends on the parameter θ, and we are testing a null hypothesis H_0.

$$
\begin{array}{l}
D_1, \ldots, D_n \overset{i.i.d.}{\sim} P_\theta,\ \theta \in \Theta, \\
H_0 : \theta \in \Theta_0 \text{ v.s. } H_A : \theta \in \Theta \cap \Theta_0^c,
\end{array}
$$

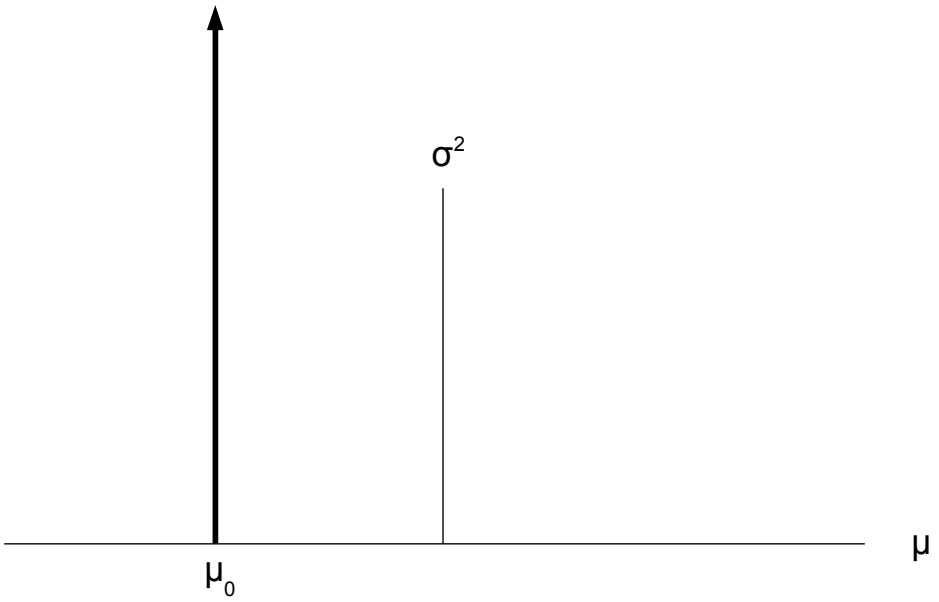
For example, let $D_1, \ldots, D_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. The null hypothesis is $H_0 : \mu = \mu_0$ v.s. versus $H_A : \mu \neq \mu_0$. The full parameter space is $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$ and the restricted parameter space is $\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0\}$. The full and restricted parameter spaces are shown in Figure A.2.

In general, the data have likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(d_i; \theta),$$

where $f(d_i; \theta)$ is the density or probability mass function evaluated at d_i. Let $\widehat{\theta}$ denote the usual Maximum Likelihood Estimate (MLE). That is, it is the parameter value for which the likelihood function is greatest, over all $\theta \in \Theta$. Let $\widehat{\theta}_0$ denote the *restricted* MLE. The restricted MLE is the parameter value for which the likelihood function is greatest, over

Figure A.2: Full versus reduced parameter spaces for $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$



all $\theta \in \Theta_0$. This MLE is *restricted* by the null hypothesis $H_0 : \theta \in \Theta_0$. It should be clear that $L(\widehat{\theta}_0) \leq L(\widehat{\theta})$, so that the *likelihood ratio*.

$$\lambda = \frac{L(\widehat{\theta}_0)}{L(\widehat{\theta})} \leq 1.$$

The likelihood ratio will equal one if and only if the overall MLE $\widehat{\theta}$ is located in Θ_0. In this case, there is no reason to reject the null hypothesis.

Suppose that the likelihood ratio is strictly less than one. If it's a *lot* less than one, then the data are a lot less likely to have been observed under the null hypothesis than under the alternative hypothesis, and the null hypothesis is questionable. This is the basis of the likelihood ratio tests.

If λ is small (close to zero), then $\ln(\lambda)$ is a large negative number, and $-2\ln\lambda$ is a large positive number.

Tests will be based on

$$\begin{aligned} G &= -2\ln\left(\frac{\max_{\theta\in\Theta_0} L(\theta)}{\max_{\theta\in\Theta} L(\theta)}\right) \\ &= -2\ln\left(\frac{L(\widehat{\theta}_0)}{L(\widehat{\theta})}\right) \\ &= -2\ln L(\widehat{\theta}_0) - [-2\ln L(\widehat{\theta})] \\ &= 2\left(-\ell(\widehat{\theta}_0) - [-\ell(\widehat{\theta})]\right). \end{aligned} \tag{A.32}$$

Thus, the test statistic G is the *difference* between two -2 log likelihood functions. This means that to carry out a test, you can minimize $-\ell(\theta)$ twice, first over all $\theta \in \Theta$, and then over all $\theta \in \Theta_0$. The test statistic is the difference between the two minimum values, multiplied by two.

If the null hypothesis is true, then the test statistic G has, if the sample size is large, an approximate chisquare distribution, with degrees of freedom equal to the difference of the *dimension* of Θ and Θ_0. For example, if the null hypothesis is that 4 elements of θ equal zero, then the degrees of freedom are equal to 4. If the null hypothesis imposes r linearly independent linear restrictions on θ (as in $H_0 : \mathbf{C}\boldsymbol{\theta} = \mathbf{h}$), then the degrees of freedom equal r, the number or rows in $\mathbf{C}$. Another way to obtain the degrees of freedom is by counting the equal signs in the null hypothesis.

The p-value associated with the test statistic G is $Pr\{X > G\}$, where X is a chisquare random variable with r degrees of freedom. If $p < \alpha$, we reject H_0 and call the results "statistically significant." The standard choice is $\alpha = 0.05$.

Many null hypotheses are linear statements of the form $H_0 : \mathbf{C}\boldsymbol{\theta} = \mathbf{h}$, but some are not. To take a simple example, suppose you wanted to test $H_0 : \sigma^2 = \mu^2$ based on a normal random sample. It seems like the degrees of freedom should equal one, but can this be justified formally?

The original proof published in 1938 by Wilks [16] applies to linear null hypotheses, and if you look at high-level textbooks like the *Advanced Theory of Statistics* [14], you will find only Wilks' proof, without modification. A way around this that often works is to use the Invariance Principle on Page 164. Suppose the null hypothesis is that one or more non-linear functions of θ equal zero. If you can make those functions part of a function that is one-to-one, then re-parameterize. Your null hypothesis is now a linear null hypothesis in the new paraameter space. Wilks' theorem applies, and you are done. Furthermore, you don't have to literally re-parameterize. A glance at the proof of the Invariance Principle confirms that the likelihood ratio test statistic is the same under the original and re-parameterized models. Thus, the degrees of freedon equals he number of equals signs in the null hypothesis, period.

For the example of $H_0 : \sigma^2 = \mu^2$, let $\theta_1' = \sigma^2 - \mu^2$ and $\theta_2' = \mu$. The function is one-to-one, because $\mu = \theta_2'$ and $\sigma^2 = \theta_1' + \theta_2'^2$. The null hypothesis is $H_0 : \theta_1' = 0$. That's is a linear null hypothesis, so by Wilks' Theorem, the test statistic has a chi-squared distribution with $df = 1$.

Sometimes this lovely trick does not work. In a regression, it is easy to test the null hypothesis that β_1 and β_2 are both zero; this is a linear null hypothesis. But suppose that you want to test the null hypothesis that β_1 *or* β_2 (or maybe both) are equal to zero. This is reasonable and attractive, because the alternative is that they are both non-zero, and it would be nice to have a single test for this. The null hypothesis is $H_0 : \beta_1\beta_2 = 0$, which is non-linear. Furthermore, any function that yields $\theta_1' = \beta_1\beta_2 = 0$ can't be one-to-one, because recovering β_1 or β_2 would potentially involve dividing by zero. Thus, while it would be perfectly possible to obtain the restricted MLE $\widehat{\theta}_0$ numerically and calculate the likelihood ratio statistic, its distribution under the null hypothesis is mysterious (to me). So, transforming a non-linear null hypothesis into a linear one by a one-to-one re-parameterization is a method that often works, but not always.

To illustrate the likelihood ratio tests, consider (one last time) the Gamma distribution Example A.6.2. For comparison, the likelihood ratio method will be used test the same three null hypotheses that were tested earlier using Wald tests. They are

- $H_0 : \alpha = 1$
- $H_0 : \alpha = \beta$
- $H_0 : \alpha = 2, \beta = 3$

For $H_0 : \alpha = 1$, the restricted parameter space is $\Theta_0 = \{(\alpha, \beta) : \alpha = 1, \beta > 0\}$. Because the Gamma distribution with $\alpha = 1$ is exponential, the restricted MLE is $\widehat{\theta}_0 = (1, \overline{d})$. It is more informative, though, to use numerical methods.

To maximize the likelihood function (or minimize minus the log likelihood) over Θ_0, it might be tempting to impose the restriction on θ, simplify the log likelihood, and write the code for a new function to minimize. But this strategy is *not* recommended. It's time consuming, and mistakes are possible. In the `R` work shown below, notice how the function `gmll1` is just a "wrapper" for the unrestricted minus log likelihood function `gmll`. It is a function of β (and the data, of course), but all it does is call `gmll` with α set to one and β free to vary.

```
> gmll1 <- function(b,datta) # Restricted gamma minus LL with alpha=1
+     { gmll1 <- gmll(c(1,b),datta)
+       gmll1
+     } # End of function gmll1
> mean(D) # Resticted MLE of beta, just to check
[1] 6.8782
```

The next step is to invoke the nonlinear minimization function `nlm`. The second argument is a (vector of) starting value(s). Starting the search at $\beta = 1$ turns out to be unfortunate.

```
> gsearch1 <- nlm(gmll1,1,datta=D); gsearch1
$minimum
[1] 282.6288

$estimate
[1] 278.0605

$gradient
[1] 0.1753689

$code
[1] 4

$iterations
[1] 100
```

The answer `g1search$estimate=278.0605` is way off the correct answer of $\overline{d} = 6.8782$, it took 100 steps, and the exit code of 4 means the function ran out of the default number of iterations. Starting at the unrestricted $\widehat{\beta}$ works better.

```
> gsearch1 <- nlm(gmll1,betahat,datta=D); gsearch1
$minimum
[1] 146.4178

$estimate
[1] 6.878195

$gradient
[1] -1.768559e-06

$code
[1] 1

$iterations
[1] 7
```

That's better. Good starting values are important! Now the test statistic is easy to calculate.

```
> G = 2 * (gsearch1$minimum-gammasearch$minimum); pval = 1-pchisq(G,df=1)
> G; pval
[1] 8.772448
[1] 0.003058146
```

Let us carry out the other two tests, and then compare the Wald and likelihood ratio test results together in a table.

For $H_0 : \alpha = \beta$, the restricted parameter space is $\Theta_0 = \{(\alpha, \beta) : \alpha = \beta > 0\}$.

```
> gmll2 <- function(ab,datta) # Restricted gamma minus LL with alpha=1
+     { gmll2 <- gmll(c(ab,ab),datta)
+       gmll2
+     } # End of function gmll2
> abstart = (alphahat+betahat)/2
> gsearch2 <- nlm(gmll2,abstart,datta=D); gsearch2
Warning messages:
1: NaNs produced in: log(x)
2: NA/Inf replaced by maximum positive value
$minimum
[1] 144.1704

$estimate
```

```
[1] 2.562369

$gradient
[1] -4.991384e-07

$code
[1] 1

$iterations
[1] 4

> G = 2 * (gsearch2$minimum-gammasearch$minimum); pval = 1-pchisq(G,df=1)
> G; pval
[1] 4.277603
[1] 0.03861777
```

This seems okay; it only took 4 iterations and the exit code of 1 is a clean bill of health. But the warning messages are a little troubling. Probably they just indicate that the search tried a negative parameter value, outside the parameter space. The `R` function `nlminb` does minimization with bounds. Let's try it.

```
> gsearch2b <- nlminb(start=abstart,objective=gmll2,lower=0,datta=D); gsearch2b
$par
[1] 2.562371

$objective
[1] 144.1704

$convergence
[1] 0

$message
[1] "relative convergence (4)"

$iterations
[1] 5

$evaluations
function gradient
       7        8
```

Since `nlminb` gives almost the same restricted $\widehat{\alpha} = \widehat{\beta} = 2.5624$ (and no warnings), the warning messages from `nlm` were probably nothing to worry about.

Finally, for $H_0 : \alpha = 2, \beta = 3$ the restricted parameter space Θ_0 is a single point and no optimization is necessary. All we need to do is calculate the minus log likelihood there.

Table A.1: Tests on data from a gamma distribution with $\alpha = 2$ and $\beta = 3$

$n = 50$				
	Wald		Likelihood Ratio	
H_0	χ^2	p-value	χ^2	p-value
$\alpha = 1$	5.8421	0.0156	8.7724	0.0031
$\alpha = \beta$	3.2455	0.0762	4.2776	0.0386
$\alpha = 2, \beta = 3$	1.3305	0.5141	2.2692	0.1320
$n = 200$				
$\alpha = 1$	34.1847	5.01e-09	58.2194	2.34e-14
$\alpha = \beta$	0.9197	0.3376	0.9664	0.3256
$\alpha = 2, \beta = 3$	1.5286	0.4657	1.2724	0.2593

```
> G = 2 * (gmll(c(2,3),D)-gammasearch$minimum); pval = 1-pchisq(G,df=1)
> G; pval
[1] 2.269162
[1] 0.1319713
```

The top panel of Table A.1 shows the Wald and likelihood ratio tests that have been done on the Gamma distribution data. But this is $n = 50$, which is not a very large sample. In the lower panel, the same tests were done for a sample of $n = 200$, formed by adding another 150 cases to the original data set. The results are typical; the χ^2 values are much closer except where they are far out on the tails, and both test lead to the same conclusions (though not always to the truth).

Like the Wald tests, likelihood ratio tests are very flexible and are distributed approximately as chi-square under the null hypothesis for large samples. In fact, they are *asymptotically equivalent* under H_0, meaning that if the null hypothesis is true, the difference between the likelihood ratio statistic and the Wald statistic goes to zero in probability as the sample size approaches infinity.

Since the Wald and likelihood ratio tests are equivalent, does it matter which one you use? The answer is that usually, Wald tests and likelihood ratio tests lead to the same conclusions and their numerical values are close. But the tests are only equivalent as $n \to \infty$. When there is a meaningful difference, the likelihood ratio tests usually perform better, especially in terms of controlling Type I error rate for relatively small sample sample sizes.

Table A.2 below contains the most extreme example I know. For a particular structural equation model with normal data (details don't matter for now), ten thousand data sets were randomly generated so that the null hypothesis was true. This was done for several sample sizes: $n = 50, 100, 250, 500$ and $1,000$. Using each of the 50,000 resulting data sets, the null hypothesis was tested with a Wald test and a likelihood ratio test at the $\alpha = 0.05$ significance level. If the asymptotic results held, we would expect both tests to reject H_0 500 times at each sample size.

Table A.2: Wald versus likelihood ratio: Type I error in 10,000 simulated datasets

	n				
Test	50	100	250	500	1000
Wald	1180	1589	1362	0749	0556
Likelihood Ratio	0330	0391	0541	0550	0522

So for this deliberately nasty example, the Wald test requires $n = 1,000$ before it settles down to something like the theoretical 0.05 significance level. The likelihood ratio test needs $n = 250$, and for smaller sample sizes it is conservative, with a Type I error rate somewhat *lower* than 0.05[16]. In general, when the Wald and likelihood ratio tests have a contest of this sort, it is usually a draw. When there is a winner, it is always the likelihood ratio test, but the margin of victory is seldom as large as this.

Exercises A.6.5

A.6.1) Let $Y_1, \ldots, Y_n$ be a random sample from a distribution with density $f(y) = \frac{1}{\theta}e^{-\frac{y}{\theta}}$ for $y > 0$, where the parameter $\theta > 0$. We are interested in testing $H_0 : \theta = \theta_0$.

(a) What is Θ?

(b) What is Θ_0?

(c) What is Θ_1?

(d) Derive a general expression for the large-sample likelihood ratio statistic $G = -2\log\frac{\ell(\widehat{\widehat{\theta}})}{\ell(\widehat{\theta})}$.

(e) A sample of size $n = 100$ yields $\overline{Y} = 1.37$ and $S^2 = 1.42$. One of these quantities is unnecessary and just provided to irritate you. Well, actually it's a mild substitute for reality, which always provides you with a huge pile of information you don't need. Anyway, we want to test $H_0 : \theta = 1$. You can do this with a calculator. When I did it a long time ago I got $G = 11.038$.

(f) At $\alpha = 0.05$, the critical value of chisquare with one degree of freedom is 3.841459. Do you reject H_0? Answer Yes or No.

A.6.2) The label on the peanut butter jar says peanuts, partially hydrogenated peanut oil, salt and sugar. But we all know there is other stuff in there too. In the United States, the Food and Drug administration requires that a shipment of peanut butter be rejected if it contains an average of more than 8 rat hairs per pound (well, I'm not sure if it's exactly 8, but let's pretend). There is very good reason to assume

[16]This suggests that the power will not be wonderful for smaller sample sizes, in this example. But keeping Type I error rates below 0.05 is the first priority.

that the number of rat hairs per pound has a Poisson distribution with mean λ, because it's easy to justify a Poisson process model for how the hairs get into the jars. We will test $H_0 : \lambda = \lambda_0$.

(a) What is Θ?

(b) What is Θ_0?

(c) What is Θ_1?

(d) Derive a general expression for the large-sample likelihood ratio statistic.

(e) We sample 100 1-pound jars, and observe a sample mean of $\overline{Y} = 8.57$. Should we reject the shipment? We want to test $H_0 : \lambda = 8$. What is the value of G? You can do this with a calculator. When I did it a long time ago I got $G = 3.97$.

(f) Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

(g) Do you reject the shipment of peanut butter? Answer Yes or No.

A.6.3) The normal distribution has density

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}.$$

Find an explicit formula for the MLE of $\theta = (\mu, \sigma^2)$. This example is in practically every mathematical statistics textbook, so the full solution is available. But please try it yourself first.

A.6.4) Write an `R` function that performs a large-sample likelihood ratio test of $H_0 : \sigma^2 = \sigma_0^2$ for data from a single normal random sample. The function should take the sample data and σ_0^2 as input, and return 3 values: G, the degrees of freedom, and the p-value. Run your function on the data in `var.dat`, testing $H_0 : \sigma^2 = 2$; see link to the data on the course web page.

For this question, you need to bring a printout with a listing of your function (showing how it is defined), and also part of an R session showing execution of the function, and the resulting output.

A.6.5) For k samples from independent normal distributions, the usual one-way analysis of variance tests equality of means assuming equal variances. Now you will construct a large-sample likelihood ratio test for equality of means, except that you will *not* assume equal variances. Write an `R` function to do it.

Input to the function should be the sample data, in the form of a matrix. The first column should contain group membership (the explanatory variable). It is okay to assume that the unique values in this column are the integers from 1 to k. The second column should contain values of the normal random variates – the response variable.

The function should return 3 values: G, the degrees of freedom, and the p-value. Run your function on the sample in `kars.dat`; see link to the data on the course web page. This data set shows country of origin and gas mileage for a sample of automobiles.

A.6.6) Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from a multivariate normal population with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Using the MLEs

$$\widehat{\boldsymbol{\mu}} \;=\; \overline{\mathbf{X}} \text{ and } \widehat{\boldsymbol{\Sigma}} \;=\; \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\top,$$

derive the large-sample likelihood ratio test G for testing whether the components of the random vectors $\mathbf{X}_i$ are independent. That is, we want to test whether $\boldsymbol{\Sigma}$ is diagonal. It is okay to use material from the class notes without proof.

A.6.7) Using `R`, write a program to compute the test you derived in the preceding question. Your program should return 3 values: G, the degrees of freedom, and the p-value. Run it on the sample in `fourvars.dat`; see link to the data on the course web page. Bring a printout listing your program and illustrating the run on `fourvars.dat`. Of course it would be nice if your program were general, but it is not required. Note that for this problem, numerical maximum likelihood is not needed. Both your restricted and your unrestricted MLEs can and should be in explicit form.