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.24) on page 572.

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 $D = X_1, \ldots, X_n \stackrel{i.i.d}{\sim} N(\mu, \sigma^2)$, 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[22].

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).

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

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 $\widehat{\theta}_2 = 5.42$.

[22] 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 **??** to get the last one:

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

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

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

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[23] 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

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

Differentiating with respect to θ,

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

[23] 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 $\widehat{\theta} = n/\sum_{i=1}^{n} \ln(d_i + 1)$. In R, if the data were in a numeric vector called `d`, the MLE would be `thetahat = 1/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.)

Maximum likelihood for the multivariate normal

Maximum likelihood estimation for the multivariate normal distribution plays an important role in this book. It's a case where closing your eyes and differentiating will get you nowhere.

Expression (A.25) on page 573 gives the likelihood function for the multivariate normal, in a form that is convenient but not obviously similar to a product of multivariate normal densities. Here's how it 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 $(\overline{\mathbf{x}}, \widehat{\boldsymbol{\Sigma}})$.

This way of writing the likelihood function makes maximum likelihood without calculus a lot easier. It's helpful to express the MLE as a theorem.

Theorem A.6 *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. The unique maximum likelihood estimate is $\widehat{\boldsymbol{\mu}} = \overline{\mathbf{x}}$ and $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top$.*

When I am producing proofs for a student audience, I frequently wonder whether I should provide a model of how to write a clean proof, or give a longer proof that is easier to follow. Perhaps because I'm naturally long-winded anyway, I often wind up giving more detail. Here, I will try doing it both ways. The brief one comes first. If you can fill in the gaps without too much effort, great. If necessary or if you wish, look at the second proof.

Proof One Rather than maximizing the likelihood, equivalently minimize

$$-\frac{2}{n}\log\frac{L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{L(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})} = tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) - \log|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}| - p + (\overline{\mathbf{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}).$$

Because $\boldsymbol{\Sigma}$ is positive definite, the last term is nonnegative, and equal to zero if and only if $\boldsymbol{\mu} = \overline{\mathbf{x}}$. Setting $\boldsymbol{\mu} = \overline{\mathbf{x}}$, the task is now to minimize $tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) - \log|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}|$.

The matrix $\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}$ is similar to $\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}$, so the eigenvalues of $\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}$ are real, and positive with probability one. Thus,

$$tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) - \log|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}| = \sum_{j=1}^p \lambda_j - \sum_{j=1}^p \log\lambda_j = \sum_{j=1}^p (\lambda_j - \log\lambda_j).$$

Each term in the sum is positive, and uniquely minimized when $\lambda_j = 1$. So to maximize the likelihood, all the eigenvalues of $\boldsymbol{\Sigma}$ must equal one. By the spectral decomposition (A.9), $\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} = \mathbf{CDC}^\top = \mathbf{CI}_p\mathbf{C}^\top = \mathbf{I}_p$, so that $\boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}}$. ■

Proof Two Rather than maximizing the likelihood, equivalently, (1) Divide the likelihood by a well-chosen expression that is constant with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, (2) Take the natural log, (3) Multiply by $-\frac{2}{n}$, and (4) minimize the result. Using Property 8 of the multivariate normal on page 573,

$$
\begin{aligned}
-\frac{2}{n}\log\frac{L(\boldsymbol{\mu},\boldsymbol{\Sigma})}{L(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}})} &= -\frac{2}{n}\log\frac{|\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\}}{|\widehat{\boldsymbol{\Sigma}}|^{-n/2}(2\pi)^{-np/2}\exp-\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}^{-1})+(\overline{\mathbf{x}}-\overline{\mathbf{x}})^\top\widehat{\boldsymbol{\Sigma}}^{-1}(\overline{\mathbf{x}}-\overline{\mathbf{x}})\right\}} \\
&= -\frac{2}{n}\log\frac{|\boldsymbol{\Sigma}|^{-n/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\}}{|\widehat{\boldsymbol{\Sigma}}|^{-n/2}\exp-\frac{n}{2}\left\{tr(\mathbf{I}_p)+0\right\}} \\
&= -\frac{2}{n}\log\left(\frac{|\boldsymbol{\Sigma}|\exp\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1})+(\overline{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})\right\}}{|\widehat{\boldsymbol{\Sigma}}|\exp\left\{p\right\}}\right)^{-\frac{n}{2}} \\
&= \log\left(\frac{|\boldsymbol{\Sigma}|\exp\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1})+(\overline{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})\right\}}{|\widehat{\boldsymbol{\Sigma}}|e^p}\right) \\
&= \log\frac{|\boldsymbol{\Sigma}|}{|\widehat{\boldsymbol{\Sigma}}|}+tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1})+(\overline{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})-p \\
&= tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1})-\log\frac{|\widehat{\boldsymbol{\Sigma}}|}{|\boldsymbol{\Sigma}|}-p+(\overline{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu}) \\
&= tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1})-\log|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}|-p+(\overline{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})
\end{aligned}
$$

If $\boldsymbol{\Sigma}$ is positive definite, so is $\boldsymbol{\Sigma}^{-1}$. Therefore the last term is nonnegative, and equal to zero if and only if $\overline{\mathbf{x}}-\boldsymbol{\mu}=\mathbf{0} \iff \boldsymbol{\mu}=\overline{\mathbf{x}}$. That is, the function is minimized when $\boldsymbol{\mu}=\overline{\mathbf{x}}$, regardless of what the positive definite matrix $\boldsymbol{\Sigma}$ happens to be.

This establishes $\widehat{\boldsymbol{\mu}}=\overline{\mathbf{x}}$. Setting $\boldsymbol{\mu}=\overline{\mathbf{x}}$, the last term vanishes, and the task is now to minimize

$$tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1})-\log|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}| \tag{A.35}$$

over all symmetric and positive definite $\boldsymbol{\Sigma}$.

Recall that the square matrix $\mathbf{B}$ is said to be *similar* to $\mathbf{A}$ if there is an invertible matrix $\mathbf{P}$ with $\mathbf{B}=\mathbf{P}^{-1}\mathbf{AP}$. Similar matrices share important characteristics; for example, they have the same eigenvalues, and the numbers of times each eigenvalue occurs (the multiplicities) are the same for the two matrices.

Choosing $\mathbf{P}=\boldsymbol{\Sigma}^{-1/2}$, write

$$\boldsymbol{\Sigma}^{-1/2}\left(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\Sigma}^{1/2}=\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}.$$

Thus $\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}$ is similar to $\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}$. The matrix $\widehat{\boldsymbol{\Sigma}}$ has an inverse with probability one[24]. Therefore the symmetric matrix $\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}$ has an inverse, is positive definite,

[24] The multivariate normal distribution is continuous, and for that reason, so is the joint distribution of the unique variances and covariances in $\widehat{\boldsymbol{\Sigma}}$. The set of variances and covariances such that one of the columns is a linear combination of others is a set of volume zero in $\mathbb{R}^{p(p+1)/2}$, and hence has probability zero.

and all its eigenvalues are strictly positive. This means the eigenvalues of $\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}$ are positive too, and

$$\begin{aligned} tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) - \log|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}| &= \sum_{j=1}^{p}\lambda_j - \log\prod_{j=1}^{p}\lambda_j \\ &= \sum_{j=1}^{p}\lambda_j - \sum_{j=1}^{p}\log\lambda_j \\ &= \sum_{j=1}^{p}(\lambda_j - \log\lambda_j)\,. \qquad \text{(A.36)} \end{aligned}$$

For $x > 0$, the function $y = x - \log x > 0$, and achieves a unique minimum when $x = 1$. Thus (A.36) can be minimized by choosing $\boldsymbol{\Sigma}$ so that the eigenvalues of $\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}$ all equal one. Such a choice is possible, because $\boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}}$ yields $\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} = \mathbf{I}_p$. The conclusion is that $\widehat{\boldsymbol{\Sigma}}$ is a maximum likelihood estimator of $\boldsymbol{\Sigma}$. Now we will see it is the only one. Let $\boldsymbol{\Sigma}$ be another covariance matrix such that all the eigenvalues of $\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}$ equal one.

The similarity of $\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}$ to $\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}$ means that the eigenvalues of $\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}$ are also all equal to one. Thus by the spectral decomposition theorem (A.9),

$$\begin{aligned} \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} &= \mathbf{CDC}^\top \\ &= \mathbf{CI}_p\mathbf{C}^\top = \mathbf{CC}^\top = \mathbf{I}_p, \end{aligned}$$

because the eigenvectors in the columns of $\mathbf{C}$ are orthonormal. Then,

$$\begin{aligned} & \mathbf{I}_p = \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} \\ \Longleftrightarrow\quad & \boldsymbol{\Sigma}^{1/2}\,\mathbf{I}_p\,\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\right)\boldsymbol{\Sigma}^{1/2} \\ \Longleftrightarrow\quad & \boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}}. \end{aligned}$$

This establishes that with probability one, the likelihood function has a unique maximum at $\boldsymbol{\mu} = \overline{\mathbf{x}}$ and and $\boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}}$. ■

A.6.4 Numerical maximum likelihood

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 *Normal with Mean Equal to Standard Deviation*

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.txt`. Read it. Remember that > is the `R` prompt.

```
> x = scan("https://www.utstat.toronto.edu/brunner/openSEM/data/norm1.data.txt")
Read 50 items
```

Now define a function to compute the log likelihood.

```
loglike1 = function(theta,D) # Data are in a vector called D
    {
    sumdsq = sum(D^2); sumd = sum(D); n = length(D)
    result = -n * log(abs(theta)) - (n/2)*log(2*pi) - sumdsq/(2*theta^2) +
                sumd/theta - n/2
    return(result)
    } # 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,D=x)
[1] -574.2965
> loglike1(-2,D=x)
[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(x); max(x)
[1] -19.28
[1] 10.34
> Theta = -20:10
> cbind(Theta,loglike1(Theta, D=x))
      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, D=x)
> 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 [25]. 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 A.6.2.

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[26]. For this, it is helpful to define a new function, `loglike2`.

```
> mean(x)
[1] -5.051
> loglike2 = function(theta,D)  -loglike1(theta,D)

> 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;

[25] The `nlm` function is good but generic. See *Numerical Recipes* [49] 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, C and Basic languages as well as Fortran. 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.

[26] As described in Section A.6.6, the minus log likelihood plays an important theoretical role in statisdtics, so minimizing the minus log likelihood is something we would probably want to do anyway.

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

```
> nlm(loglike2,mean(x), D=x)$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:

$$\frac{-\sum_{i=1}^n d_i \pm \sqrt{(\sum_{i=1}^n d_i)^2 + 4n\sum_{i=1}^n d_i^2}}{2n} = \frac{1}{2}\left(-\overline{d} \pm \sqrt{\overline{d}^2 + 4\frac{\sum_{i=1}^n d_i^2}{n}}\right), \qquad \text{(A.37)}$$

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

Let's calculate these for the given data.

```
> meanx = mean(x) ; meanxsq = mean(x^2)
> (-meanx + sqrt(meanx^2 + 4*meanxsq) )/2 # Solution 1
[1] 10.3463
> (-meanx - sqrt(meanx^2 + 4*meanxsq) )/2 # Solution 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.37 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 = (-meanx + sqrt(meanx^2 + 4*meanxsq) )/2 ; t1
[1] 10.3463
> t2 = (-meanx - sqrt(meanx^2 + 4*meanxsq) )/2 ; t2
[1] -5.2953
> n = length(x)
> # Now calculate second derivative at t1 and t2
> n/t1^2 - 3*sum(x^2)/t1^4 + 2*sum(x)/t1^3
[1] -0.7061484
> n/t2^2 - 3*sum(x^2)/t2^4 + 2*sum(x)/t2^3
[1] -5.267197
```

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

```
> loglike1(t1, D=x)
[1] -224.9832
> loglike1(t2, D=x)
[1] -155.4413
```

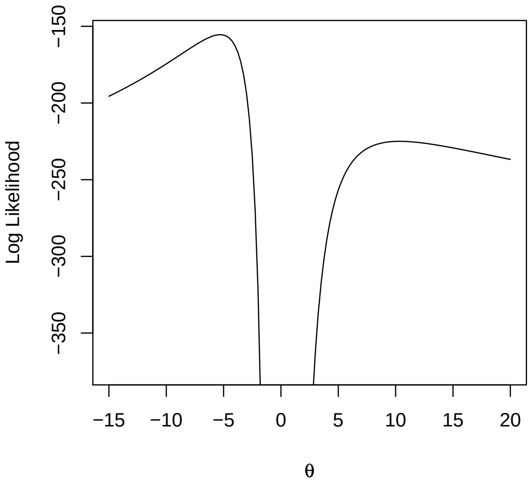
So the maximum we found is higher, which makes sense because it's within the range of the data. But we only found it because we started searching near the correct answer.

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, D=x)
# 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 higher.

Figure A.4: Log Likelihood for Example A.6.1



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, D=x)
$minimum
[1] 283.7589

$estimate
[1] 64.83292

$gradient
[1] 0.701077

$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, D=x)
$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 *The Gamma Distribution*

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.

```
> # Generate data
> set.seed(3201)
> alpha=2; beta=3 # True parameter values
> xx <- round(rgamma(50,shape=alpha, scale=beta),2); xx
 [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  1.79
[15]  4.57  1.40  1.13  6.84  3.21  0.38 11.24  1.72  4.69  1.96  7.87  8.49  5.31  3.40
[29]  5.24  1.64  7.17  9.60  6.97 10.87  5.23  5.53 15.80  6.40 11.25  4.91 12.05  5.44
[43] 12.62  1.81  2.70  3.03  4.09 12.29  3.23 10.94
> c(mean(xx), alpha*beta)  # Sample mean, true expected value.
```

```
[1] 6.8782 6.0000
> c(var(xx), alpha*beta^2) # Sample variance, true variance
[1] 24.90303 18.00000
```

The parameter vector is $\boldsymbol{\theta} = (\alpha, \beta)$, 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))
+      result =  n*a*log(b) + n*lgamma(a) + sumd/b - (a-1)*sumlogd
+      return(result)
+       # 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!

```
> # Method of moments estimates
> momalpha = mean(xx)^2/var(xx); momalpha
[1] 1.899754
> mombeta = var(xx)/mean(xx); mombeta
[1] 3.620574
```

Now for the numerical search. We'll use the `optim` function rather than `nlm`. One advantage of `optim` is that it allows one to set upper and lower bounds, so that the numerical search does not leave the parameter space. Also, `nlm` is older and `optim` has a better reputation — though I've used `nlm` many times, and never had any serious problems with it.

This time, we will request that the `optim` function return the *Hessian* matrix at the place where the search stops (something that `nlm` will do also). As described below, the Hessian yields a multivariable version of the second derivative test.

In the following, notice how the `optim` function assumes that the first argument of the function being minimized is a vector of starting values, and the second argument is the name of the function being minimized. Any other arguments of the function (in this case, the name of the data vector) should be specified by name in the `optim` function call.

```
> startval = c(momalpha,mombeta)
> gammasearch = optim(startval, gmll, lower = c(0,0), method = "L-BFGS-B",
+               hessian=TRUE, datta=xx)
> gammasearch
$par
[1] 1.805930 3.808674

$value
[1] 142.0316

$counts
function gradient
       9        9


$convergence
```

```
[1] 0

$message
[1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"

$hessian
         [,1]      [,2]
[1,] 36.69402 13.127928
[2,] 13.12793  6.224773

> eigen(gammasearch$hessian)$values
[1] 41.569998  1.348796
```

The `optim` object `gammasearch` is a linked list. The item `par` is the point at which the search stops, so $\widehat{\alpha} = 1.805930$ and $\widehat{\beta} = 3.808674$. The item `value` is the value of the minus log likelihood function where the search stops. The `counts` item is how many times the function and the gradient were evaluated. The gradient is The item `estimate` is The `gradient` is

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

Besides being the direction of steepest incease, it's something that should be zero at the MLE. Optionally, one can provide `optim` with a function that computes the value of the gradient, making the search a lot faster. Otherwise, the partial derivatives are approximated with the slopes of secant lines.

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

$$\mathbf{H} = \left[\frac{\partial^2(-\ell)}{\partial\theta_i\partial\theta_j}\right]. \tag{A.38}$$

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 eigenvalues being positive, it is convenient to check the eigenvalues to determine whether the numerical search has located a minimum. In this case it has. Finally, the value `comvergence=0` means normal termination of the search.

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.

A.6.5 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 *Parameterizing in Terms of Odds Rather than Probability*

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

$$\theta' = \frac{\theta}{1-\theta} = g(\theta). \tag{A.39}$$

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,

$$L'(\theta') = \frac{\theta'^{\sum_{i=1}^n d_i}}{(1+\theta')^n}. \tag{A.40}$$

At this point one could differentiate the log of (A.41) 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.40) 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.6.4

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.

A.6.6 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 [70] in the 1940s. *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* [67] 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:

$$I(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2}\ln f(D_i;\theta)\right]. \tag{A.41}$$

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, Rao and others 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 a small variance is a great property to have 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 the reciprocal (inverse) of the Fisher Information, and many MLEs do that. Subsequently, Wald [70] showed that under some regularity conditions, the variances of maximum likelihood estimators in general attain the Cramér-Rao lower bound as $n \rightarrow \infty$. 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 .

Calculating the expected value in (A.42) 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 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

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \left[-E\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\mathbf{D}_1;\boldsymbol{\theta})\right)\right], \tag{A.42}$$

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

To estimate the Fisher information matrix, one may simply put a hat on $\boldsymbol{\theta}$ in A.43. If calculating the expected values is too much of a pain, one may replace the expected value by a sample mean as well as replacing $\boldsymbol{\theta}$ with $\widehat{\boldsymbol{\theta}}$. The result is

$$\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}}) = \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(-\ln\prod_{q=1}^{n} f(\mathbf{D}_q;\widehat{\boldsymbol{\theta}})\right)\right] = \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(-\ell(\widehat{\boldsymbol{\theta}})\right)\right]. \tag{A.43}$$

$\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}})$ is sometimes loosely called the "expected" Fisher information, and $\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}})$ is sometimes called the "observed" Fisher information, even though it would be more accurate to call it the estimated observed Fisher information. They are both excellent large-sample estimates of $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ in (A.43).

In the one-dimensional case, one divided by the estimated Fisher Information is the (estimated) asymptotic variance of the maximum likelihood estimator. *In the multi-parameter case, the estimated Fisher Information is a matrix, and the corresponding estimated asymptotic variance-covariance matrix is its inverse.* Assume that the true Fisher information matrix is being estimated by $\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}})$, and denote the estimated asymptotic covariance matrix by $\widehat{\mathbf{V}}_n$. In that case we have

$$\widehat{\mathbf{V}}_n = \boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}}_n)^{-1}. \tag{A.44}$$

Now comes the really good part. Comparing Formula (A.44) for the Fisher Information to Formula (A.39) 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*[27].

[27] 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.

So to get a good estimate of the asymptotic covariance matrix, minimize minus the log likelihood, tell the software to give you the Hessian, and calculate its 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,] 36.69402 13.127928
[2,] 13.12793  6.224773
```

and the asymptotic covariance is just

```
> Vhat = solve(gammasearch$hessian); Vhat
           [,1]       [,2]
[1,]  0.1110190 -0.2341369
[2,] -0.2341369  0.6544386
```

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

$$Z = \frac{\widehat{\theta} - \theta}{S_{\widehat{\theta}}} \tag{A.45}$$

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$par[1]; betahat = gammasearch$par[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  5.39 
```

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.46) 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] 2.418794
> pval = 2*(1-pnorm(abs(Z))); pval # Two-sided test
[1] 0.01557207
```

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[28].

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.

[28] 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

A.6.7 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

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

where $\mathbf{L}$ 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

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

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

$$\left[\begin{array}{rrrrrrr} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 0 \end{array}\right] \left[\begin{array}{c} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ 0 \end{array}\right].$$

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

$$\mathbf{L}\mathbf{x} \sim N_r(\mathbf{L}\boldsymbol{\mu}, \mathbf{L}\boldsymbol{\Sigma}\mathbf{L}^\top)$$

and

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

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

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

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 [32] for details.

and

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

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

$$W_n = (\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})^\top (\mathbf{C}\widehat{\mathbf{V}}_n\mathbf{L}^\top)^{-1}(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h}) \sim \chi^2(r), \tag{A.46}$$

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.47) 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. The function returns a pair of quantities, the Wald test statistic and the p-value.

```
> WaldTest = function(C,thetahat,h=0) # H0: C theta = h
+     {
+     WaldTest = numeric(2)
+     names(WaldTest) = c("W","p-value")
+     dfree = dim(C)[1]
+     W = t(C%*%thetahat-h) %*% solve(C%*%Vhat%*%t(C)) %*% (C%*%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 $\mathbf{L}$ in $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$ 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.47) for the Wald test statistic. Pre-multiplying by $\mathbf{L}$ in $\mathbf{C}\widehat{\mathbf{V}}_n$ picks out row j of $\widehat{\mathbf{V}}_n$, and post-multiplying by $\mathbf{L}^\top$ picks out column j of the result, so that $\mathbf{C}\widehat{\mathbf{V}}_n\mathbf{L}^\top = \widehat{v}_{j,j}$, and inverting it puts it in the denominator. In the numerator, $(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})^\top(\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h}) = (\widehat{\theta}_j - \theta_{j,0})^2$, so that $W_n = Z^2$. Thus, squaring a large-sample Z-test gives a Wald chisquare test with one degree of freedom.

A.6.8 Likelihood Ratio Tests

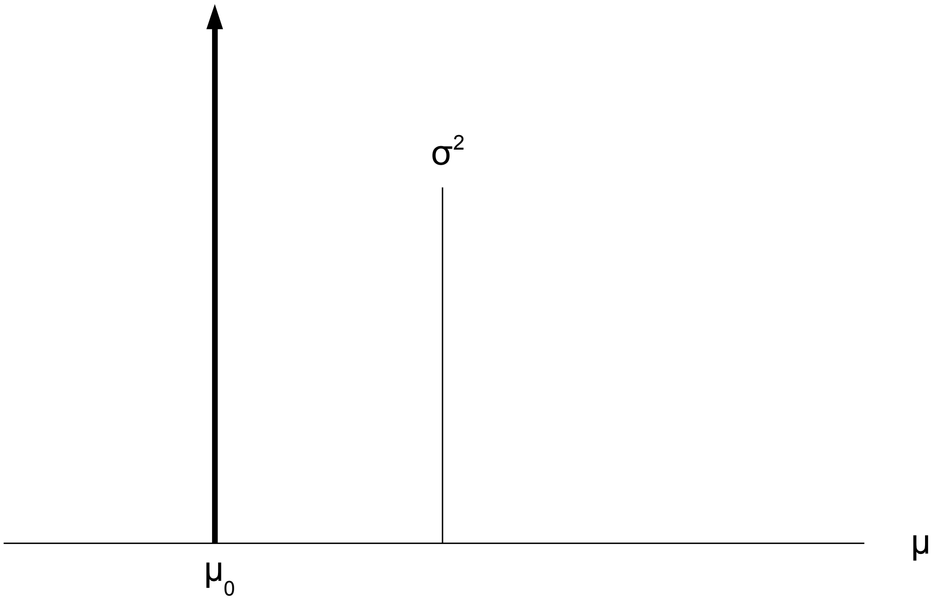
Likelihood ratio tests fall into two categories, exact and large-sample. The main examples of exact likelihood ratio tests include are the standard F-tests and t-tests associated with regression and the analysis of variance for normal data. Here, we concentrate on the large-sample likelihood ratio tests.

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.5.

Figure A.5: Full versus reduced parameter spaces for $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$



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 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

$$\begin{aligned} G^2 &= -2\ln\left(\frac{\max_{\theta\in\Theta_0} L(\theta)}{\max_{\theta\in\Theta} L(\theta)}\right) \\ &= -2\ln\left(\frac{L(\widehat{\theta}_0)}{L(\widehat{\theta})}\right) \\ &= -2\ln L(\widehat{\theta}_0) - [-2\ln L(\widehat{\theta})] \\ &= 2\left(-\ell(\widehat{\theta}_0) - [-\ell(\widehat{\theta})]\right). \end{aligned} \tag{A.47}$$

Thus, the test statistic G^2 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 $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$), then the degrees of freedom equal r, the number or rows in $\mathbf{L}$. 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^2 is $Pr\{X > G^2\}$, 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 $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$, but some are not.

Example A.6.4 *A Non-linear Null Hypothesis*

Suppose you wanted to test $H_0 : \sigma^2 = \mu^2$ based on a normal random sample. The restricted MLE is fairly easy to find numerically (see Example A.6.1), and it seems like the degrees of freedom should equal one because the null hypothesis has one equals sign. Can this be justified formally?

The original proof published in 1938 by Wilks [74] applies to linear null hypotheses, and if you look at high-level textbooks like the *Advanced Theory of Statistics* [67], you will find only Wilks' proof, without modification. A way around this that often works is to use the Invariance Principle of Section A.6.5. 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, and 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 Example A.6.4, 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, anyway). 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.

```
> Gsq = 2 * (gsearch1$minimum-gammasearch$minimum); pval = 1-pchisq(Gsq,df=1)
> Gsq; 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

> Gsq = 2 * (gsearch2$minimum-gammasearch$minimum); pval = 1-pchisq(Gsq,df=1)
> Gsq; 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.

```
> Gsq = 2 * (gmll(c(2,3),D)-gammasearch$minimum); pval = 1-pchisq(Gsq,df=1)
> Gsq; 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.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

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

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.

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[29]. 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.6.8

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 $H_0 : \theta = \theta_0$.

[29] 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.

(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 $G^2 = -2\log\frac{\ell(\widehat{\widehat{\theta}})}{\ell(\widehat{\theta})}$.

(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 $G^2 = 11.038$.

(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 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^2? You can do this with a calculator. When I did it a long time ago I got $G^2 = 3.97$.

(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^2, 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^2, 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^2 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^2, 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.

A.6.9 The Bootstrap

Sometimes, the distribution of a statistic or vector of statistics can be tough to figure out. You may not be able to do it at all. Or, maybe you could get an asymptotic answer using the multivariate delta method, but it would be a big job requiring extensive paper and pencil calculations followed by careful programming. The bootstrap, due to Efron [23], is a computer-intensive method that can yield fairly automatic answers in such situations.

Let $\mathbf{x} = (X_1, \ldots, X_n)$ be a random sample from some distribution F. Let $T = T(\mathbf{x})$ be a statistic or vector of statistics. We need to know the distribution of T; an approximate answer will be good enough. You should not turn up your nose at the word "approximate." Bootstrap solutions are approximate in the same sense that a consistent estimator is approximate.

The name "bootstrap" comes from the saying "Pull yourself up by your bootstraps." Figure A.6 shows a pair of boots[30]. The little loops at the back of the boots are the bootstraps; if you hook your fingers in the loops, it's easier to pull your boots on. Pulling yourself up by your bootstraps is physically impossible, but it's a metaphor for getting the job done with the resources you have available, even though it may seem impossible.

Figure A.6: A pair of boots with bootstraps

To appreciate the statistical bootstrap, recall how the idea of a *sampling distribution* is introduced in an elementary statistics class. One does not terrorize the students by referring to functions of a random variable. Instead, the sampling distribution is described as follows. Imagine drawing repeated random samples from the same population. Either the sampling is with replacement, or the population is so large that the distinction between with and without replacement makes no difference. For each sample, calculate the statistic. Make a relative frequency histogram of the values of the statistic. As the number of samples increases, the histogram gets closer and closer to the sampling distribution of the statistic.

So, select a random sample from the population. If the sample size is large, the sample is similar to the population. Sample repeatedly from the sample with replacement; this is called *resampling*. Calculate the statistic for every bootstrap sample. A histogram of the resulting values approximates the shape of the sampling distribution of the statistic.

[30] This photograph was taken by Tarquin. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. For more information, see the entry at the wikimedia site.

To visualize re-sampling, think of writing the n sample data values on marbles, putting the marbles in a jar, and drawing n marbles with replacement. Naturally, there will be some repeats; don't worry about it. In many applications, you will be re-sampling *vectors* of data values, like x_1, x_2, x_3 and x_4. In such cases, keep the values from a given individual together[31]. Think of n strings of beads, with four beads on each string. You randomly sample strings of beads. Of course, in practice all this is done by computer using pseudo-random number generation, but the physical analogy may be helpful as a way of understanding the process.

More formally, let $\mathbf{x} = (X_1, \ldots, X_n)$ be a random sample from some distribution F, possibly a multivariate distribution. $T = T(\mathbf{x})$ is a statistic or a vector of statistics. Form a "bootstrap sample" $\mathbf{x}^*$ by sampling n values from $\mathbf{x}$ *with replacement*. Repeat this process B times, obtaining $\mathbf{x}_1^*, \ldots, \mathbf{x}_B^*$. Calculate the statistic (or vector of statistics) for each bootstrap sample, obtaining $T_1^*, \ldots, T_B^*$. The relative frequencies of $T_1^*, \ldots, T_B^*$ approximate the sampling distribution of T.

It works because the empirical distribution converges to the true distribution function.

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\} \overset{a.s.}{\to} E(I\{X_i \leq x\}) = F(x)$$

Resampling from $\mathbf{x}$ with replacement is the same as simulating a random variable whose distribution is the empirical distribution function $\widehat{F}(x)$. Suppose the distribution function of T is a nice smooth function of F. Then as $n \to \infty$ and $B \to \infty$, bootstrap sample moments and quantiles[32] of $T_1^*, \ldots, T_B^*$ converge to the corresponding moments and quantiles of the distribution of T. If the distribution of $\mathbf{x}$ is discrete and supported on a finite number of points, the technical issues are modest. For continuous distributions with unbounded support it's more challenging, but the conclusions still hold.

Estimating the covariance matrix of a vector of statistics

In structural equation modeling, it is quite common to have a vector of estimators that are known to be consistent and asymptotically multivariate normal. An asymptotic variance-covariance matrix is available provided that the observable data are multivariate normal, but the normality assumption is either doubtful or demonstrably false. So constructing tests and confidence intervals is not routine.

There are two main ways this situation can emerge. In the first scenario, the statistics in question are nice explicit functions of the sample variance-covariance matrix of the observable data. Even when the data are not normally distributed, Theorem A.5 on page 615 establishes that the joint distribution of the sample variances and covariances is

[31] Well, if you were interested in testing independence of x_1 and x_2 from x_3 and x_4, you could put the (x_1, x_2) pairs in one jar and the (x_3, x_4) pairs in another jar, and draw independently from the two jars to assemble a set of four values. This is an example of *bootstrapping under the null hypothesis*, a very nice way to construct tests that make no assumptions about the distribution of the data.

[32] The q quantile of a distribution is the point with q of the distribution at or below it, where $0 \leq q \leq 1$. Quantiles are like percentiles.

asymptotically multivariate normal, and then by the multivariate delta method, differentiable functions of those variances and covariances are approximately multivariate normal too. The asymptotic variances and covariances of the sample variances and covariances – and functions of them – are actually available and can be estimated consistently, but it's a big, unpleasant chore.

In the other scenario, the statistics in question are MLEs, but they are MLEs based on the assumption that the observable data are multivariate normal – an assumption that is questionable or worse. The good news is that by Theorem 5.1 and the "Corollary to Huber's corollary" (Expression 5.4 on page 432) in Chapter 5, these pseudo-MLEs are consistent and have an asymptotic distribution that is multivariate normal. The bad news is that the normal-theory estimates of the asymptotic variance-covariance matrix are incorrect in general, though some exceptions are given in Chapter 5. Again, estimating the right variance-covariance matrix is not out of the question, but it's a big job involving mathematical calculations and computer coding that might never be needed again.

It's a lot easier using the bootstrap. The bootstrap provides a good picture of the sampling distribution of that vector of statistics. The only feature of the sampling distribution that matters is their variance-covariance matrix. Proceed as follows. Draw B bootstrap samples from the sample data, and for each one calculate the vector of statistics. Assemble the results into a sort of data file, with B rows, and one column for each statistic. Calculate the sample variance-covariance matrix of that. The result is an excellent approximation of the asymptotic variance-covariance matrix that's needed for tests and confidence intervals.

Here is an example. In the United States, admission to university is sometimes based partly on the Scholastic Aptitude Test, or SAT. In the old days there were two subtests, Verbal and Math. The data file `openSAT.data.txt`[33] has Verbal score, Math score and first-year grade point average for a sample of 200 students. We first read the data and look at the correlation matrix.

```
> sat = read.table("https://www.utstat.toronto.edu/brunner/openSEM/data/openSAT.data.txt")
> head(sat)
  VERBAL MATH  GPA
1    578  567 2.68
2    474  653 2.51
3    546  657 1.95
4    664  686 2.81
5    600  619 2.79
6    488  738 2.36
> cor(sat)
          VERBAL      MATH       GPA
VERBAL 1.0000000 0.2751041 0.3224927
MATH   0.2751041 1.0000000 0.1941086
GPA    0.3224927 0.1941086 1.0000000
```

These correlations are not too impressive, but remember that the students were admitted largely on the basis of having high SAT scores, so this is an example of how restricted

[33] This is a reconstructed data set based on a Minitab data set. I believe the Minitab data set is a cleaned-up version of real data from Penn State University.

range can weaken an observed correlation. Verbal score appears to be more highly correlated with GPA than Math score, but is the difference statistically significant? This is a meaningful but non-standard question.

By Theorem A.5 and the multivariate delta method, the asymptotic distribution of the sample correlation coefficients is multivariate normal and centered on the true correlations. For a Wald test and a confidence interval, all we need is an estimate of the covariance matrix.

Now we'll follow the recipe. Put the row numbers in a "jar." Sample from the jar with replacement, putting the rows into a bootstrap data set. Calculate the correlations. Do this B times, saving the results in an array that will be called `bootdata`.

```
> # Bootstrap the correlations
> n = dim(sat)[1] # Sample size is the number of rows in the data file
> set.seed(9999) # Set random number seed so results can be duplicated.
> jar = 1:n; B = 1000
> bootdata = matrix(NA,B,3)
> colnames(bootdata) = c(’Verbal-Math’,’Verbal-GPA’,’Math-GPA’)
> for(j in 1:B)
+
+     rowz = sample(jar,size=n,replace=TRUE)
+     xstar = sat[rowz,]
+     kor = cor(xstar)
+     bootdata[j,1] = kor[1,2] # Correlation of Verbal with Math
+     bootdata[j,2] = kor[1,3] # Correlation of Verbal with GPA
+     bootdata[j,3] = kor[2,3] # Correlation of   Math with GPA
+      # Next bootstrap sample
> head(bootdata)
     Verbal-Math Verbal-GPA  Math-GPA
[1,]   0.3020368  0.3171977 0.2320282
[2,]   0.3589208  0.2834930 0.2247893
[3,]   0.1572560  0.3590254 0.2988522
[4,]   0.1989407  0.3582051 0.0998772
[5,]   0.3165621  0.3644107 0.2394445
[6,]   0.2808987  0.2934830 0.1626899
```

The estimated covariance matrix we need is just the sample covariance matrix of these bootstrapped statistics.

```
> Vhat = var(bootdata); Vhat # Asymptotic covariance matrix
             Verbal-Math   Verbal-GPA    Math-GPA
Verbal-Math 0.0044099830 0.0002516633 0.001059281
Verbal-GPA  0.0002516633 0.0037209355 0.001182263
Math-GPA    0.0010592808 0.0011822628 0.004240506
```

To test for difference between the two correlations, we'll use the `Wtest` function. The present application isn't quite a Wald test strictly speaking, but the theory applies.

```
> # Now use it
> # Test H0: Corr(Verbal,GPA) = Corr(Math,GPA)
> source("http://www.utstat.utoronto.ca/~brunner/Rfunctions/Wtest.txt")
> # function(L,Tn,Vn,h=0) # H0: L theta = h
> LL = cbind(0,1,-1)
> estcorr = c(corsat[1,2],corsat[1,3],corsat[2,3])
> Wtest(L=LL,Tn=estcorr,Vn=Vhat)
         W         df    p-value
2.94491891 1.00000000 0.08614802
```

So the difference between is not statistically significant at the 0.05 level. How about a confidence interval?

```
> # 95 percent CI for Corr(Verbal,GPA) - Corr(Math,GPA)
> estdiff = corsat[1,3]-corsat[2,3]; estdiff # Estimated difference between correlations
[1] 0.128384
> sediff = as.numeric(sqrt( LL %*% Vhat %*% t(LL) ))
> CI = c(estdiff - 1.96*sediff, estdiff + 1.96*sediff); round(CI,4)
[1] -0.0182  0.2750
```

Observe that the confidence interval includes zero, which must happen since the hypothesis of zero difference was not rejected. It absolutely *must* happen because squaring the z statistic corresponding to the confidence interval yields the Wald chi-square.

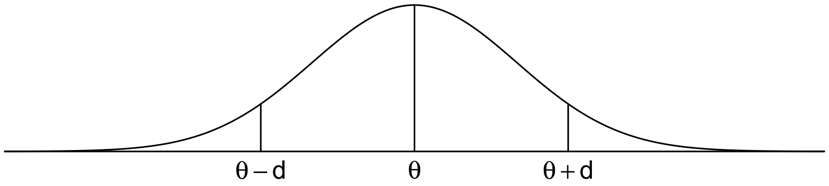
Bootstrapping MLEs In structural equation modeling it is common practice to estimate the model parameters with normal theory maximum likelihood, even if there is no particular reason to believe that the data are normally distributed. Fortunately, almost regardless of the distribution of the sample data, the resulting estimators are consistent by Theorem 5.1, and have asymptotically normal distributions by Corollary 5.4 on page 432. The normal theory estimates of the variances and covariances of the estimators might not be correct (see Chapter 5), but that problem is neatly solved by bootstrapping the pseudo-MLE's and estimating their variance-covariance matrix, exactly as in the example above. In `lavaan`, the `se="bootstrap"` option does the trick. Here are a couple of examples.

```
boot = lavaan(fullmod, data=X, se="bootstrap")
fit3 = cfa(model3,data=simdat, se="bootstrap")
```

Quantile Bootstrap Confidence Intervals An alternative to normal-theory confidence intervals are the quantile confidence intervals, which use more information about the exact shape of the sampling distribution out on the tails. Suppose T_n is a consistent estimator of θ, and the distribution of T_n is approximately symmetric around θ. Then the lower $(1-\alpha)100\%$ confidence limit for θ is the $\alpha/2$ sample quantile of $T_1^*, \ldots, T_B^*$, and the upper limit is the $1-\alpha/2$ sample quantile. For example, the 95% confidence interval ranges from the 2.5th to the 97.5th percentile of $T_1^*, \ldots, T_B^*$.

Symmetry is a requirement that is often ignored when computing quantile bootstrap intervals. The distribution of T_n symmetric about θ means for all $d > 0$, $P\{T_n > \theta + d\} = P\{T_n < \theta - d\}$. See Figure A.7.

Figure A.7: A symmetric distribution



Select d so that $P\{T_n > \theta + d\} = P\{T_n < \theta - d\}$ equals $\alpha/2$. Then

$$\begin{aligned} 1 - \alpha &= P\{\theta - d < T_n < \theta + d\} \\ &= P\{T_n - d < \theta < T_n + d\} \end{aligned}$$

To use this result, an estimate of d is required.

There are two natural estimates. Letting $Q_{\alpha/2}$ denote the true $\alpha/2$ quantile of the distribution of T_n,

$$1 - \alpha = P\{\theta - d < T_n < \theta + d\} = P\{Q_{\alpha/2} < T_n < Q_{1-\alpha/2}\}.$$

The estimates should satisfy

$$\begin{array}{rclcrcl} \widehat{\theta} - \widehat{d}_1 & = & \widehat{Q}_{\alpha/2} & \Rightarrow & \widehat{d}_1 & = & T_n - \widehat{Q}_{\alpha/2} \\ \widehat{\theta} + \widehat{d}_2 & = & \widehat{Q}_{1-\alpha/2} & \Rightarrow & \widehat{d}_2 & = & \widehat{Q}_{1-\alpha/2} - T_n, \end{array}$$

where T_n has been used to estimate θ, and $\widehat{Q}_{\alpha/2}$ and $\widehat{Q}_{1-\alpha/2}$ are the bootstrap quantiles.

Then, take $1 - \alpha = P\{T_n - d < \theta < T_n + d\}$ and plug in the estimates of d_1 and d_2. Using $\widehat{d}_1$ on the left yields

$$T_n - \widehat{d}_1 = T_n - (T_n - \widehat{Q}_{\alpha/2}) = \widehat{Q}_{\alpha/2}$$

Using $\widehat{d}_2$ on the right yields

$$T_n + \widehat{d}_2 = T_n + (\widehat{Q}_{1-\alpha/2} - T_n) = \widehat{Q}_{1-\alpha/2},$$

so that the $(1 - \alpha)100\%$ bootstrap quantile confidence interval is

$$\left(\widehat{Q}_{\alpha/2}, \widehat{Q}_{1-\alpha/2}\right). \tag{A.48}$$

There are indications that the coverage of this interval can approach $1 - \alpha$ faster with increasing sample size than a confidence interval based on the central limit theorem. See Chapter 22 of Efron and Tibshirani [24].

To test hypotheses like $H_0 : \theta = \theta_0$, one can simply check whether the $(1 - \alpha)100\%$ quantile confidence interval for θ includes θ_0, and reject the null hypothesis at significance level α if it does.

Justifying the Assumption of Symmetry All this depends on the statistic T_n having a distribution that is approximately symmetric. When the distribution of the estimator is not symmetric about the parameter being estimated, quantile confidence intervals are unjustified and often quite inaccurate. Ignoring this point has led to confusion ans suspicion about the bootstrap, especially among non-statisticians. So how does one justify the assumption of symmetry, particularly when the distribution of T_n is elusive? The easiest answer is asymptotic normality. Smooth functions of asymptotic normals are asymptotically normal, and this includes maximum likelihood estimators as well as functions of the sample moments. Of course the normal distribution is symmetric, and this justifies the use of quantile confidence intervals. Here is an illustration using the SAT data.

```
> # Now a quantile confidence interval
> difcorr = bootdata[,2]-bootdata[,3]
> difcorr = sort(difcorr)
> # 0.025 * 1000 = 25, so go midway between number 25 and number 26,
> # And midway between number 974 and 975
> LowerQuant = (difcorr[25]+difcorr[26])/2
> UpperQuant = (difcorr[974]+difcorr[975])/2
> qCI = c(LowerQuant,UpperQuant) # 95% Quantile interval
> round(qCI,4)
[1] -0.0281  0.2704
```

This confidence interval is very similar to the one directly based on asymptotic normality. Again, it provides no evidence that the correlation between Verbal SAT and first-year GPA is different from the correlation between Math SAT and first-year GPA.