1. Independent random variables $X_1, X_2, \ldots, X_n$ have common probability density function $f(x;\theta)$ depending on the unknown parameter θ. Define the terms *likelihood function* for θ, *maximum likelihood estimator* for θ, *Fisher's information* for θ.

What is the connection between Fisher's information and the asymptotic distribution of the maximum likelihood estimator?

Suppose $X_1, X_2, \ldots, X_n$ are independent, each with density

$$f(x;\theta) = \frac{1}{\pi\left[1 + (x-\theta)^2\right]}, \quad x \in \mathbb{R},\ \theta \in \mathbb{R}.$$

Derive an equation which must be satisfied by the maximum likelihood estimator, $\widehat{\theta}$.

Prove that Fisher's information for θ is $\frac{1}{2}n$. Explain how to calculate an approximate 95% confidence interval for θ.

2. Let $\overline{X}$ be the mean of a random sample $X_1, X_2, \ldots, X_n$ from a distribution with mean μ and variance $V(\mu)$. Show that the variance of $h(\overline{X})$ is approximately constant, where

$$h(\mu) = \int^{\mu} V(u)^{-1/2} du.$$

Show that

(i) $V(\mu) = \mu$ for a Poisson distribution;

(ii) $V(\mu) = \mu(1-\mu)/m$ for the proportion X/m of successes when $X \sim B(m,\mu)$;

(iii) $V(\mu) = \mu^2/\nu$ when X has a gamma distribution with mean μ and index ν.

Find variance stabilising transformations for these distributions.

3. The following data are time intervals in days between earthquakes which either registered magnitudes greater than 7.5 on the Richter scale or produced over 1,000 fatalities. Recording starts on 16 December, 1902 and ends on 4 March, 1977, a total period of 27,107 days. There were 63 earthquakes in all, and therefore 62 recorded time intervals.

840	294	454	667	556	304
1901	335	30	129	209	83
40	203	735	365	82	887
139	638	121	280	736	319
246	44	76	46	194	375
157	562	36	40	99	832
695	1354	384	9	599	263
1336	436	38	92	220	460
780	937	150	434	584	567
1617	33	710	402	759	328
145	721				

Assuming the data to be from a random sample $X_1, \ldots, X_n$ drawn from an exponential distribution with parameter λ, obtain the maximum likelihood estimator $\widehat{\lambda}$ of λ and calculate the maximum likelihood estimate.

Given that the moment generating function of a gamma distribution with parameters (n, λ) is

$$M_n(t) = \left(\frac{\lambda}{\lambda - t}\right)^n,$$

show that $Y = \sum_{i=1}^n X_i$ has a gamma distribution. Show that

$$\left(\frac{a}{n\overline{x}}, \frac{b}{n\overline{x}}\right),$$

is an exact 95% central confidence interval for λ, where $\int_0^a \frac{y^{n-1}e^{-y}dy}{\Gamma(n)} = \int_b^\infty \frac{y^{n-1}e^{-y}dy}{\Gamma(n)} = 0.025$.

Obtain Fisher's information for λ and use it to show that

$$(0.0018, 0.0030)$$

is an approximate 95% confidence interval for λ.

4. If S^2 is the variance of a random sample which is drawn from a normal distribution $N(\mu, \sigma^2)$, then cS^2 has a standard distribution for a particular value of the constant c. Write down the distribution and specify the constant c.

How many pairs of positive constants a, b exist such that $P\left(a < cS^2 < b\right) = 0.95$? For any such pair, deduce that the random interval

$$\left(\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right)$$

will contain the variance σ^2 for approximately 95% of a large number of such random samples.

5. One of the data sets you have seen in the lecture notes concerns the silver content of Byzantine coins; it appears in Table 2.3. The silver content for the first and fourth coinages of King Manuel I, Comnemus (1143 – 1180) is reproduced in the table below.

Silver content of coins

First coinage	Fourth coinage
5.9	5.3
6.8	5.6
6.4	5.5
7.0	5.1
6.6	6.2
7.7	5.8
7.2	5.8
6.9	
6.2	

The respective sample means and variances from these data are $\overline{x}_1 = 6.744$, $S_1^2 = 0.295$, $\overline{x}_4 = 5.614$, $S_4^2 = 0.131$.

(i) Assuming the data to be normally distributed, obtain a suitable test statistic for testing the null hypothesis that the mean silver content of the two coinages is the same. What further assumptions are required for the test to be valid?

(ii) Carry out the test with the alternative hypothesis that the mean silver content is different for the two coinages. How would you modify the test if the alternative hypothesis were to be that there is *less* silver in the fourth coinage?

6. (Extension question.) In the theory of random sampling from a distribution involving unknown parameters, a *pivot quantity* is a function of a random sample and of the unknown parameters that has a known standard distribution. Confidence intervals for functions of the unknown parameters may sometimes be calculated by using appropriate pivot quantities. Given independent random samples of n_i ($n_i \geq 2;\ i = 1, 2$) observations from two normal distributions with means μ_i and variances σ_i^2, specify which pivot quantities you would use to obtain confidence intervals for each of the following:

   (i) μ_1 if σ_1^2 is known;

   (ii) σ_1^2;

   (iii) μ_1;

   (iv) σ_1^2 if $\sigma_1^2 = \sigma_2^2$;

   (v) $\mu_1 - \mu_2$ if $\sigma_1^2 = \sigma_2^2$.

   Name the distribution of the pivot quantity in each of the five cases or, alternatively, write down its probability density function. Now use an appropriate pivot quantity to obtain a 99% confidence interval for the difference between the mean silver contents, $\mu_1 - \mu_4$, in question 5. How could you use this to carry out the test in part (ii) of the previous question?

Optional Computer Exercises

1. Referring back to question 3, you will find a file *quakes.txt* containing the earthquake data on the website. If you have downloaded it into a directory called, say, c:\data, you can load it into R with

```
> quakes <- read.table("c:/data/quakes.txt", header = TRUE)
```

Use a probability plot to check the exponential model; are you satisfied that it is plausible? Quantiles for lower and upper tail probability p for a gamma distribution with parameters $(n, 1)$ can be calculated using the functions `qgamma(p, n, lower.tail=TRUE)` and `qgamma(p, n, lower.tail=FALSE)`. Calculate the exact confidence interval suggested in question 3, and compare your result with that for the approximate confidence intrval.

2. The data file *engreek.txt* contains the scores on each of 32 English sentences awarded by native English speaking teachers and by Greek teachers as a result of errors made by Greek-Cypriot learners of English. The scores are assessments made by the teachers of the quantity of errors made by the students. The data are paired for each of 32 sentences, so differences are to be tested for a mean of zero.

If you have downloaded the data set into a directory called, say, c:\data, you can load it into R and obtain a vector of differences with

```
> eg <- read.table("c:/data/engreek.txt", header = TRUE)
> attach(eg)
> diff <- Greek - English
```

Before attempting a t-test, you should check for normality. Try

```
> qqnorm(diff)
```

If you are not entirely happy about the normality, you could try a few transformations. The commands below show how you would try a power $\frac{3}{2}$ transformation, preserving the sign of the difference.

```
> sign <- diff/abs(diff)
> tdiff <- sign*(abs(diff)^(3/2))
> qqnorm(tdiff)
```

When you are satisfied that your data, whether transformed or otherwise, are plausibly normal, perform a t-test of the null hypothesis that the difference is zero (to see how to do this in R, try entering `?t.test`). What should your alternative hypothesis be? What conclusions can you draw?