# Additional Problems

## Additional Problem 1

Like the

`http://www.stat.umn.edu/geyer/5102/examp/rlike.html#opt`

example of maximum likelihood done by computer except instead of the gamma scale model, we will use the Cauchy location model. The likelihood is given by (6.6.7) on p. 366 of DeGroot and Schervish. For data, use the URL

`http://www.stat.umn.edu/geyer/5102/examp/cauchy.txt`

and for a starting point use the sample median rather than the sample mean, that is, `median(x)` instead of `mean(x)`. The reason for this will become clear later. The sample is a very bad estimate of location for the Cauchy distribution.

## Additional Problem 2

Solve the quadratic equation to prove that the interval (2.18) in the handout does indeed have endpoints (2.19) in the handout.

## Additional Problem 3

Calculate the three kinds of intervals given by equations (2.20), (2.19), and (2.22) in the handout for binomial data with $n = 50$ and $x = 4$. Use 95% for the confidence coefficient.

## Additional Problem 4

Calculate the second and fourth central moments $\mu_2$ and $\mu_4$ in the notation of the handout for the so-called *double exponential* distribution with density

$$f(x) = \tfrac{1}{2} e^{-|x|}, \qquad -\infty < x < \infty$$

(note this distribution is symmetric about zero, so the mean is zero and all odd central moments are zero).

Compare the *correct* asymptotic variance of the sample variance $\mu_4 - \mu_2^2$ Compare the *incorrect* asymptotic variance of the sample variance $2\mu_2^2$ that we would get if we incorrectly assumed the data were normal. (Section 2.10 of the handout).

## Additional Problem 5

Starting with the asymptotic distribution for $S_n^2$ given on p. 16 of the "more on confidence intervals handout" use the delta method to give the asymptotic distribution of $S_n$.

## Additional Problem 6

Using the method of Section 1.2 of the "more on confidence intervals" handout, find an exact 95% confidence interval for the *mean* (not the rate) parameter of an exponential distribution from which it is assumed we have independent and identically distributed data with sample size 15 and sample mean 103.49.

## Additional Problem 7

Using the method of Section 2.9.2 of the "more on confidence intervals" handout, find an asymptotic (approximate, large sample) 95% confidence interval for the mean parameter of a Poisson distribution from which is assumed we have independent and identically distributed data with sample size 50 and sample mean 2.9.

**Hint:** In order to use "plug-in" you need a consistent estimator of the standard deviation of the Poisson distribution. What is the standard deviation and what is its relation to the mean? The sample mean consistently estimates the mean parameter. What does that suggest for a consistent estimator of standard deviation?

## Additional Problem 8

Suppose we have an independent and identically distributed sample from a Geometric($p$) distribution with sample size 30 and sample mean 7.8. Find the maximum likelihood estimate of $p$ and a 95% confidence interval for $p$ based on the MLE and either observed or expected Fisher information.

## Additional Problem 9

Like the example of multiparameter maximum likelihood done by computer

`http://www.stat.umn.edu/geyer/5102/examp/rlike.html#bar`

except instead of the gamma scale-rate model, we will use the Cauchy location-scale model. The probability density function is given by

$$f(x|\theta, \sigma) = \frac{1}{\sigma} \cdot g\left(\frac{x - \theta}{\sigma}\right)$$

where

$$g(z) = \frac{1}{\pi(1 + z^2)}.$$

The R function

```
dcauchy(x, location = theta, scale = sigma)
```

calculates $f(x|\theta, \sigma)$, returning a vector of values if `x` is a vector.

For data, use the URL

`http://www.stat.umn.edu/geyer/5102/examp/cauchy.txt`

Method of moments estimators make no sense for the Cauchy distribution because the Cauchy distribution doesn't have any moments. We have to use estimators based on quantiles instead.

For a starting point for `theta` use the sample median (as we did in Additional Problem 1). This makes sense because $\theta$ is the theoretical median. And for a starting point for the scale parameter `sigma` use half the sample interquartile range, that is, `0.5 * IQR(x)`. This makes sense because the theoretical interquartile range is $2\sigma$.

Report the values you obtain for

(a) the MLEs for $\theta$ and $\sigma$.

(b) the observed Fisher information matrix.

(c) 95% confidence intervals for $\theta$ and $\sigma$.

## Additional Problem 10

Suppose the variables $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n$ are independent, and suppose the $X_i$ are identically Exponential($\theta$) distributed and the $Y_i$ are identically Exponential($1/\theta$) distributed.

(a) Find the maximum likelihood estimate when the sample size is $n = 25$ and the sample means are $\overline{X}_n = 3.12$ and $\overline{Y}_n = 0.432$. Give the MLE both as a formula (a function of $\overline{X}_n$ and $\overline{Y}_n$) and numerically.

(b) Calculate both observed and expected Fisher information.

(c) Show that even after the MLE is plugged in for the parameter, observed and expected Fisher information are different, both as formulas (functions of $\overline{X}_n$ and $\overline{Y}_n$) and numerically.

(d) Calculate 95% asymptotic (approximate, large sample) confidence intervals for the parameter $\theta$, one using observed Fisher information, one using expected Fisher information.

## Additional Problem 11

Basically this is Problem 8.6.10 in DeGroot and Schervish. Use the data in their Table 8.1, which can be read into R with the statements

```
calcium <- c( 7, -4, 18, 17, -3, -5, 1, 10,  11, -2)
placebo <- c(-1, 12, -1, -3,  3, -5, 5,  2, -11, -1, -3)
```

(a) Perform a test of the hypotheses stated in Problem 8.6.10 using Welch's approximate test, giving the P-value.

(b) Perform a test of the same hypotheses using the exact $t$-test based on the assumption of equal variances, giving the $P$-value.

(c) Interpret these $P$-values.

(d) Calculate a 95% two-sided confidence interval for the difference of the means of the two groups.

The web page on doing t-tests in R may help.

## Additional Problem 12

For the data in the URL

`http://www.stat.umn.edu/geyer/5102/examp/rob.txt`

calculate the following point estimators

(a) the sample mean

(b) the sample median

(c) the sample 10% trimmed mean

(d) the sample 20% trimmed mean

(e) the median of the Walsh averages (Hodges-Lehmann estimator associated with the Wilcoxon signed rank test)

## Additional Problem 13

For the data in the URL

`http://www.stat.umn.edu/geyer/5102/examp/a13.txt`

calculate confidence intervals for the center of symmetry (we assume the population distribution is symmetric about some point $\theta$ which is the unknown parameter of interest) associated with

(a) the sign test

(b) the Wilcoxon signed rank test

(c) the Student $t$ test

having confidence level above 95% and as close to 95% as you can get (this is what the `wilcox.test` function does by default).

## Additional Problem 14

For the data in the URL

`http://www.stat.umn.edu/geyer/5102/examp/a13.txt`

calculate $P$-values for an upper tailed test about the center of symmetry (we assume the population distribution is symmetric about some point $\theta$ which is the unknown parameter of interest) with null and alternative hypotheses

$$\begin{aligned} H_0 &: \theta = 0 \\ H_1 &: \theta > 0 \end{aligned}$$

for each of the following types of test

(a) the sign test

(b) the Wilcoxon signed rank test

(c) the Student $t$ test

(note: the `t.test` and `wilcox.test` functions do two-tailed tests by default so you must use the optional argument `alternative = "greater"` to do an upper-tailed test).

## Additional Problem 15

For the data in the URL

`http://www.stat.umn.edu/geyer/5102/examp/ds10-9.txt`

which contains two variables `x` and `y`, assume the data follow the simple linear regression model

$$y = \beta_0 + \beta_1 x + \text{error}$$

(a) Calculate the $P$-value for a test with null and alternative hypotheses

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

(b) Interpret the $P$-value. Does the test say the value of the true population regression coefficient $\beta_1$ is statistically significantly different from zero at the 0.05 level?

## Additional Problem 16

For the data in the URL

`http://www.stat.umn.edu/geyer/5102/examp/ds10-9.txt` which contains two variables `x` and `y`, assume the pairs $(X_i, Y_i)$ are independent and identically bivariate normal distributed with correlation

$$\rho = \operatorname{cor}(X_i, Y_i)$$

(a) Calculate the $P$-value for a test with null and alternative hypotheses

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

(b) Interpret the $P$-value. Does the test say the value of the true correlation coefficient $\rho$ is statistically significantly different from zero at the 0.05 level?

## Additional Problem 17

For the data in the URL

`http://www.stat.umn.edu/geyer/5102/examp/ds10-9.txt`

which contains two variables `x` and `y`, assume the data follow the simple linear regression model

$$y = \beta_0 + \beta_1 x + \text{error}$$

(a) Calculate the $P$-value for a test with null and alternative hypotheses

$$H_0 : \beta_1 = 0.6$$
$$H_1 : \beta_1 \neq 0.6$$

(b) Interpret the $P$-value. Does the test say the value of the true population regression coefficient $\beta_1$ is statistically significantly different from 0.6 at the 0.05 level?

**Note:** This is exactly the same as Additional Problem 15 (word for word) except that the hypothesized value of the regression coefficient is 0.6 rather than zero.

## Additional Problem 18

For the data in the URL

`http://www.stat.umn.edu/geyer/5102/examp/ds10-9.txt`

which contains two variables `x` and `y`, assume the data follow the simple linear regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \text{error}$$

(a) Calculate the $P$-value for a test with null and alternative hypotheses

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

(b) Interpret the $P$-value. Does the test say the value of the true population regression coefficient $\beta_2$ is statistically significantly different from zero at the 0.05 level?

**Note:** This is exactly the same as Additional Problem 15 except that it is about the quadratic regression model rather than the simple linear model and the test is about $\beta_2$ rather than about $\beta_1$.

## Additional Problem 19

For the data in the URL

`http://www.stat.umn.edu/geyer/5102/examp/sally.txt`

which contains two variables `x` and `y`, it is clear from the scatter plot produced by `plot(x, y)` that a simple linear regression will not fit the data (no statistics needed, the points are obviously nowhere near a straight line).

From the scatter plot curves up at both ends, it is clear that a polynomial of even degree is needed for the regression function (assuming we restrict our consideration to polynomials), because a polynomial of odd degree would go up at one end and down at the other.

(a) Fit the following three regression models:

- The **quadratic** model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \text{error}$$

- The **quartic** (fourth degree) model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \text{error}$$

- The sixth degree model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \beta_6 x^6 + \text{error}$$

Report the regression coefficients for each model.

(b) Perform a test in which the quadratic model is the little model and the quartic model is the big model. Report the $F$ statistic and the $P$-value for the $F$ test for model comparison. Interpret the $P$-value. Which model does this test tell you to use?

(c) Perform a test in which the forth degree model is the little model and the sixth degree model is the big model. Report the $F$ statistic and the $P$-value for the $F$ test for model comparison. Interpret the $P$-value. Which model does this test tell you to use?

(d) Make a scatter plot of the data points, with the estimated regression function plotted for all three models on the same plot (use `lty = 2`, `lty = 3`, and so forth to distinguish the lines). Hand in the plot. Comment on the differences between the curves and the relation to the results of the $F$ tests.

## Additional Problem 20

Modify the example calculating the MSE of an estimator by simulation making two changes. Use the $t$ distribution with 2.5 degrees of freedom for the distribution of the data (instead of the standard Cauchy distribution in the example) and use the 20for the point estimator (instead of the median in the example). Provide both a point estimate and a confidence interval for the actual true MSE.

## Additional Problem 21

Modify the percentile bootstrap confidence interval example making two changes. Make the parameter to be estimated the interquartile range of the population and the point estimator of this parameter the interquartile range of the data (calculated by the `IQR` function in R).