A9: Statistics

Sheet 2 — HT 2024

(Lectures 5–8, Notes sections 1.7–3.2)

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

   Assume the individuals in a sample of size $n = 1029$ are independent and that each individual has blood type M with probability $(1-\theta)^2$, type MN with probability $2\theta(1-\theta)$, and type N with probability θ^2. For the following data (Rice, 2007) find the maximum likelihood estimate $\widehat{\theta}$ and use the asymptotic distribution of the MLE to find an approximate 95% confidence interval for θ.

Blood Type	M	MN	N
Frequency	342	500	187

2. Let $X_1, \dots, X_n$ be independent $N(\mu, \sigma^2)$ random variables. Suppose that μ is known, σ is unknown and that we want to estimate $\psi = \log \sigma$.

   (a) Find the maximum likelihood estimator $\widehat{\sigma}$ and the asymptotic normal approximation to the distribution of $\widehat{\sigma}$.

   (b) Use the delta method to find the asymptotic distribution of $\widehat{\psi}$ and hence find an approximate 95% confidence interval for ψ.

   (c) Explain how the interval in (b) can be used to find an approximate confidence interval for σ.

3. A sequence of estimators T_n, $n \geqslant 1$, of a scalar parameter θ is called *consistent* if, for all θ (i.e. whatever the true value of θ), we have that T_n converges in probability to θ as $n \to \infty$.

   Suppose T_n is a sequence of estimators of θ satisfying $\text{bias}(T_n) \to 0$ and $\text{var}(T_n) \to 0$ as $n \to \infty$. Show that T_n is consistent for θ. [*Hint: Chebyshev's inequality.*]

4. 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	1901	40	139	246	157	695	1336	780	1617
145	294	335	203	638	44	562	1354	436	937
33	721	454	30	735	121	76	36	384	38
150	710	667	129	365	280	46	40	9	92
434	402	209	82	736	194	99	599	220	584
759	556	304	83	887	319	375	832	263	460
567	328								

Assuming the data to be a random sample $X_1, \dots, X_n$ 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 if

$$\int_0^a \frac{y^{n-1}e^{-y}}{\Gamma(n)}\, dy = \int_b^\infty \frac{y^{n-1}e^{-y}}{\Gamma(n)}\, dy = 0.025.$$

Obtain Fisher's information for λ and use it to find an approximate 95% confidence interval for λ. The interval given by the exact method above is $(0.0018, 0.0029)$. Verify numerically that your approximate interval is close to this.

5. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with known mean μ and unknown variance σ^2. Three possible confidence intervals for σ^2 are

   (a) $\left(\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{a_1}, \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{a_2} \right)$

   (b) $\left(\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{b_1}, \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{b_2} \right)$

   (c) $\left(\frac{n(\overline{X} - \mu)^2}{c_1}, \frac{n(\overline{X} - \mu)^2}{c_2} \right)$

   where a_1, a_2, b_1, b_2, c_1, c_2 are constants.

   Find values of these six constants which give confidence level 0.90 for each of the three intervals when $n = 10$ and compare the expected widths of the three intervals in this case.

   With $\sigma^2 = 1$, what value of n is required to achieve a 90% confidence interval of expected width less than 2 in cases (b) and (c) above?

   [For a χ^2 with e.g. 6 degrees of freedom, you can use `qchisq(0.05, 6)` to find the 0.05 quantile.]

6. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively, where the parameters μ_1, μ_2, σ^2 are unknown. Let

   $$S^2 = (m + n - 2)^{-1} \left(\sum_{i=1}^{m} (X_i - \overline{X})^2 + \sum_{j=1}^{n} (Y_j - \overline{Y})^2 \right).$$

   Determine the distributions of both

   $$(m + n - 2)S^2/\sigma^2 \quad \text{and} \quad \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S^2(\frac{1}{m} + \frac{1}{n})}}.$$

   Show how to construct a confidence interval for $\mu_1 - \mu_2$.

7. Ten students were asked to guess the width of a lecture room. Their guesses (in metres) were: 10, 11, 12, 13, 15, 16, 17, 18, 19, 25. The actual width of the room was 13.1 m.

   (a) Assuming the data arise from a normal distribution, how would you test whether this distribution has the correct mean? State the appropriate null and alternative hypotheses, and any assumptions you need to make for the hypothesis test to be appropriate.

   (b) Carry out the test you suggested in (i) and state your conclusions.

   (c) Modify your test to test whether the data are from a distribution with a mean value *higher* than the true value and re-state your conclusions.

8. Read in the earthquake data from question 4 and try an exponential Q-Q plot:

```
x <- scan("http://www.stats.ox.ac.uk/~laws/partA-stats/data/quakes.txt")
n <- length(x)
k <- 1:n
plot(-log(1 - k/(n+1)), sort(x), main = "Exponential Q-Q Plot",
     ylab = "Ordered data", xlab = "-log[1 - k/(n+1)]")
abline(0, mean(x))
# abline above plots a line with intercept = 0 and gradient = mean(x)
# - from lecture notes the exponential Q-Q plot should have intercept 0
# and gradient mu if the data are exponential with mean mu
# use ?abline to see the help page for abline
```

Is an exponential model a reasonable assumption for this dataset?

The 2.5% and 97.5% quantiles for a gamma distribution with parameters $(n, 1)$ can be calculated as follows.

```
a <- qgamma(0.025, n)
b <- qgamma(0.975, n)
```

That is, the function `qgamma(p, n)` calculates the pth quantile of a gamma distribution with shape parameter n and rate parameter 1.

Calculate the exact confidence interval of question 4:

```
c(a, b) / sum(x)
```

Also use R to check that the approximate 95% confidence interval for λ obtained using Fisher's information is as given in question 4 – you might have obtained one of two possible approx intervals:

```
# approx interval using lambda.hat +/- 1.96*I(lambda.hat)^{-1/2}
xbar <- mean(x)
c(1 - 1.96/sqrt(n), 1 + 1.96/sqrt(n)) / xbar

# second approx interval from substituting I(lambda) = n/lambda^2
# and then solving the inequalities
# i.e. not replacing lambda by lambda.hat in order to estimate a variance
c(1/(1 + 1.96/sqrt(n)), 1/(1 - 1.96/sqrt(n))) / xbar
```

9. Following the previous question, for data that are exponential with parameter λ there are three possible confidence intervals for λ – one based on the gamma distribution plus two approximation possibilities. These three are all different, but numerically they are almost the same in question 4 where $n = 62$.

Use the code below to investigate how the three differ when n is small, e.g. $n = 10$. What do you conclude?

```
# investigate how the three intervals perform in small samples,
# e.g. n = 10, using data generated from an exponential, parameter 1

# generate the sample, calculate and plot the three intervals
# repeat m times, e.g. m = 33 giving 99 intervals in total

# copy-and-paste the following chunk into R, you don’t need to work out
# the details of what all the plotting commands are doing

# ---begin chunk---
n <- 10
a <- qgamma(0.025, n)
b <- qgamma(0.975, n)
m <- 33

plot(1, 1, type = "n", yaxt = "n", xlim = c(0, 5), ylim = c(0, 4*m),
     xlab = "lambda", ylab = "",
     main = paste("95% CIs: samples of size", n, "from exponential, parameter 1"))
abline(v = 1)
legend("topright", c("interval1", "interval2", "interval3"),
       lty = 1, lwd = 2, col = c(1, "orange2", "steelblue2"))

for (i in 1:m) {
  x <- rexp(n)
  ci1 <- c(a, b) / sum(x)
  ci2 <- c(1 - 1.96/sqrt(n), 1 + 1.96/sqrt(n)) / mean(x)
  ci3 <- c(1/(1 + 1.96/sqrt(n)), 1/(1 - 1.96/sqrt(n))) / mean(x)
  lines(ci1, rep(4*i-1, 2), lwd = 2)
  lines(ci2, rep(4*i-2, 2), lwd = 2, col = "orange")
  lines(ci3, rep(4*i-3, 2), lwd = 2, col = "steelblue")
}
# ---end chunk---

# x <- rexp(n) generates a sample of size n
# use ?rexp to see the help page for rexp - when no rate parameter is
# given, rate = 1 is the default, hence vertical line on the plot at
# the true value lambda = 1
# the three intervals behave differently in small samples
# try repeating with larger n, e.g. n = 20, 50
# - you only need to change the first line n <- 10 to a different value
# at n = 50 the three intervals are close, especially intervals 1 & 2
# (and n = 62 for the data in question 3)
```

10. To do question 7 you need the sample mean $\overline{x}$ and sample standard deviation s:

```
x <- c(10, 11, 12, 13, 15, 16, 17, 18, 19, 25)
mean(x)
sd(x)
```

Use the functions `qt` and/or `pt` to determine the significance (or otherwise) of the test statistic in question 7:

The pth quantile of a t_r-distribution can be calculated using `qt(p, r)`, so e.g. the 97.5% quantile of a t_4-distribution can be found using: `qt(0.975, 4)`

Alternatively, the cdf of a t_r-distribution at y can be calculated using `pt(y, r)`, so e.g. the probability that a t_4 random variable is less than 1.96 is given by: `pt(1.96, 4)`

```
# test statistic
tobs <- sqrt(10)*(mean(x) - 13.1)/sd(x)

# two-sided p-value
2*(1 - pt(tobs, df = 9))

# one-sided p-value
1 - pt(tobs, df = 9)

# can check using t.test, see ?t.test
# - by default it assumes two-sided, and also uses a method for unequal variances
# hence we want var.equal = TRUE
t.test(x, mu = 13.1, var.equal = TRUE)

# one-sided
t.test(x, mu = 13.1, alternative = "greater", var.equal = TRUE)

# or could compare tobs to the quantiles
qt(0.975, df = 9)
qt(0.95, df = 9)
```