MAST90105: Lab and Workshop Problems for Week 8

The Lab and Workshop this week covers problems arising from Module 5, Section 2 onwards and the first part of Module 6. In the lab this week, you will enter some commands from this sheet into R-studio and see the results. The problems will not be assigned to groups this week.

1 Lab

Introduction

Data: In this lab we will estimate distributions for the Tasmanian Rainfall data wthat are the maximum daily rainfall in each year from 1995 to 2014 recorded by 34 weather stations in Tasmania. Data source: Australian Bureau of Meteorology. (`http://www.bom.gov.au/`)

Goals: (i) Estimate parameters using maximum likelihood; (ii) estimate parameters using method of moments (iii) estimate relative error in maximum likelihood and method of moments estimators by simulation.

Both R and Mathematica will be used to accomplish these goals.

Task 1 -Reading the data in

From last week, you know that in R can import the data with the command:

```
tasmania=read.csv("L:/MAST90105MethodsofMathematicalStatistics/EditedRainfall.csv")
```

Please note that if you cut and paste this command from this file, you must paste it into *one* line of code in RStudio. RStudio will not accept file names spread across more than one line and you will not be able to read in the file.

Another tip before you start the commands is to open the File menu, and go to New File and then R script. If you cut and paste the commands from this file into the R Script, you can select and use the Run button to execute code. If you save the file to your desktop, for example as "Rainfall.R", you can then save it on a USB stick - if you have one. Or you can email it to yourself before closing the session. This script can then be opened from RStudio.

There is a copy of this file "EditedRainfall.csv" on the LMS in the Workshop/Labs area, so you can work on this at home if you transfer the file to your computer and change the address of the file to the appropriate one on your computer. Or you could save it on a USB stick from the L: drive or you could email a copy of the file to yourself.

You can create variables for with the following commands:

```
s1 = tasmania[, 2]  # create a vector for station 1 (Burnie)
s2 = tasmania[, 3]  # create a vector for station 2 (Cape Grim)
```

In Mathematica, the command to read in the data and create the variables $s1, s2$ is:

```
tasmania=Import["L:\\MAST90105MethodsofMathematicalStatistics\\EditedRainfall.csv"];
s1 = tasmania[[2 ;;, 2]]; s2 = tasmania[[3;;,3]
```

Several aspects of the Mathematica import of data merit explanation:

- Whilst R uses forward slash, /, in file names, Mthematica uses the Windows version of \ and uses two of them.
- By using the "Insert" Menu item, "File Path ...", you can navigate directly to the item in the Lab-Materials folder so avoid typing and or cut and paste.
- When the data is Imported into Mathematica, it is stored as a list whose items are lists. Each list is a row of the ".csv" file.
- The first list is the first row, so it has the names of the columns in the ".csv" file: Year, Burnie (Round Hill), Cape Grim (Woolnorth) etc.
- The second list is the second row - this has the data for the year 1995: 1995, 64.8, 49 etc
- To access the columns, the part function "[[ ]]" is used.
- "[[2,3]]" refers to the second row and third column, ie is the Cape Grim rainfall for 1995 which is 49.
- "[[2;;,2]]" refers to all rows from 2 onwards and the second column, so this is the rainfall for 1995 to 2014 for Burnie.

Task 2 - Maximum likelihood estimation

In this task, we will look at maximum likelihood estimation. In simple cases the MLE is available in closed form. For example, if observations were iid from $N(\mu, \sigma^2)$, then we already know that the MLE for μ and σ^2 are $\hat{\mu} = \overline{X}$ and $\hat{\sigma}^2 = (n-1)S^2/n$. Thus, the ML estimates for the extreme rainfall data can be quickly computed as

```
mu.hat = mean(s1)
mu.hat
n = length(s1)
sigma.hat = sqrt((n - 1) * var(s1)/n)
sigma.hat
```

Probability theory says that a better model for maxima is the extreme value (EV) distribution, or Gumbel distribution, with pdf

$$f(x;\mu,\sigma) = \frac{1}{\sigma}\exp\left\{-\frac{x-\mu}{\sigma}\right\}\exp\left[-\exp\left\{-\frac{x-\mu}{\sigma}\right\}\right], \quad x \in \mathbb{R},\ \mu \in \mathbb{R},\ \sigma > 0.$$

The log-likelihood function for this model is

$$\ell(\mu, \sigma) = -n \log \sigma - \sum_i \frac{x_i - \mu}{\sigma} - \sum_i \exp\left\{-\frac{x_i - \mu}{\sigma}\right\}.$$

Unfortunately, as it is often the case in practice, there is no closed form solution for the MLE and maximisation has to be carried out numerically.

There are number of routines available to maximize the log-likelihood function. One simple approach is given by the function `fitdistr` in the library `MASS`, which can be used to fit a number of models (see `help` for available distributions). Since the Gumbel distribution is not in the default list of distributions we must write its pdf before using `fitdistr`.

```
library(MASS)
# you may need to
# install.packages('evd',repos='https://cloud.r-project.org')
# evd contains functions pgumbel, qgumbel, dgumbel,
# rgumbel for the cdf, quantiles, pdf and random
# numbers for the Gumbel distribution the argument
# names are loc and scale
library(evd)
# fits the Gumbel distribution
(gumbel.fit = fitdistr(x = s1, densfun = dgumbel, start = list(loc = 50,
    scale = 10), lower = 1e-04))
```

Note that `fitdistr` requires 3 main arguments: data, a pdf (or pmf) and a starting point used by the optimisation algorithm. Choosing suitable starting parameter values is crucial to ensure a good solution. If preliminary estimates - such as the Method of Moments estimators in Task 3 - are available you should consider them as starting points.

The numbers in brackets below the estimates are estimates of the standard deviations of the estimators. These numbers are obtained from the theory of maximum likelihood estimators with large sample sizes. In Task 4, these standard deviations will be estimated, as in lectures, by simulation.

For the Normal, log-Normal, geometric, exponential and Poisson distributions `fitdistr` uses closed-form MLEs. For example

```
(normal.fit = fitdistr(x = s1, densfun = "normal"))
```

For all other distributions, `fitdistr` maximises numerically the likelihood function using the optimisation routine `optim`.

In practical applications, we frequently write the negative log-likelihood from scratch and use directly `optim` or some other optimisation routine, depending on the likelihood at hand. Most numerical optimisation routines assume that you want to minimise a function. Note that maximising the log-likelihood ℓ is equivalent to minimise the negative log-likelihood function $-\ell$ and write

```
log.like = function(theta) # log-likelihood
{
loc  = theta[1]; scale = theta[2]
out = sum(log(dgumbel(s1, loc=loc, scale=scale)))
return(out)
}
# fnscale = -1 turns the optimisation from minimisation to maximisation
fit = optim(c(50,10),log.like,lower=0.0001,method="L-BFGS-B",
control = list(fnscale=-1)) # fits MLEs
theta.hat = fit$par # returns estimates
theta.hat
```

The arguments in `optim` are the vector of starting values and the function to be minimised. A number of other tuning parameters can be used depending on the optimisation method specified in optim. The function returns a number of useful outputs, including final parameter estimates and value of the minimised negative log-likelihood, and information about convergence (to see them type `fit`). A detailed explanation can be found in `Value` in `help(optim)`. Various classic optimisation algorithms are available within this function (see `Details` and references in `help(optim)`). When the parameters' are known to be in some interval, optimisation can be carried out using the method `method="L-BFGS-B"` for `optim`, or the function `nlminb`, which allow us to specify upper and lower bound for each parameter.

In Mathematica, estimation can be carried out using the commands:

$$\begin{aligned}\text{maxLikeDistNormal} =& \text{EstimatedDistribution}[\text{s1},\\ & \text{NormalDistribution}[\mu,\sigma],\text{AccuracyGoal}\to\infty]\\ \text{maxLikeDistExtreme} =& \text{EstimatedDistribution}[\text{s1},\\ & \text{ExtremeValueDistribution}[\mu,\sigma],\text{AccuracyGoal}\to\infty]\end{aligned}$$

The Accuracy Goal ensures maximum numerical accuracy.

Next, carry out a visual check on the fitted models. The following plots the fitted normal and Gumbel models over the histogram of the data

```
# Write fitted normal and Gumbel pdfs
pdf1 = function(x) {
    dnorm(x, mean = mu.hat, sd = sigma.hat)
}
pdf2 = function(x) {
    dgumbel(x, loc = theta.hat[1], scale = theta.hat[2])
}

# Plot data and fitted models
hist(s1, freq = FALSE, col = "gray", main = NULL, xlab = "x",
    xlim = c(0, 100))
curve(pdf1, from = 0, to = 100, col = 2, lty = 2, lwd = 2,
```

```
    add = TRUE)
curve(pdf2, from = 0, to = 100, col = 1, lty = 1, lwd = 2,
    add = TRUE)
```

In Mathematica, the plot can be obtained from:

$$\begin{aligned}&\text{Show}[\text{Histogram}[\text{s1}, \{10\}, \text{PDF}],\\ &\text{Plot}[\{\text{PDF}[\text{maxLikeDistNormal}, x], \text{PDF}[\text{maxLikeDistExtreme}, x]\}, \{x, 0, 100\},\\ &\text{PlotLegends} \to \text{Expressions}], \text{PlotRange} \to \text{All}]\end{aligned}$$

The command `Show` is to put the histogram and the two curves on the same graph. The argument $\{10\}$ in the histogram gives 10 bins and so makes the histograms comparable. It is necesary to add $\{x, 0, 100\}$ to the Plot again for comparability. The `PlotRange` option in Show ensures that the x-axis and y-axis are as large as necessary.

From the plots, which distribution appears to be a better fit?

Finally, the value of the maximised log-likelihood function $\ell(\hat{\theta}) = \log L(\hat{\theta})$ is a useful statistic describing the goodness-of-fit of models and can be used for model comparison (provided that the models under exam have the same number of parameters). The largest the likelihood function, the better the model fits the data at hand.

There are a number of ways to obtain this information. For example, it can be extracted from the output of `fitdistr` as

```
normal.fit$loglik
gumbel.fit$loglik
```

or it can be computed from scratch by evaluating directly a log-likelihood function

```
log.like(theta.hat)  # Gumbel log-likelihood
```

In Mathematica, they can be obtained from:

$$\begin{aligned}&\{\text{LogLikelihood}[\text{maxLikeDistExtreme}, \text{s1}],\\ &\text{LogLikelihood}[\text{maxLikeDistNormal}, \text{s1}]\}\end{aligned}$$

The log-likelihood value for the Gumbel model is larger than that obtained from the normal model, which suggests that the Gumbel model is more apropriate for the the extreme rainfall data in Burnie.

Task 3 - Method of moments estimation

In this task we compute point estimates using the method of moments (MM). Assume that the data are generated by $X \sim Gumbel(\mu, \sigma)$ with pdf given in Task 2. Mathematica will give the mean and variance using the commands:

$$\{\text{Mean}[\text{ExtremeValueDistribution}[\mu, \sigma]], \quad \text{Variance}[\text{ExtremeValueDistribution}[\mu, \sigma]]\}$$

The output shows that

$$E(X) = \gamma\sigma + \mu$$
$$Var(X) = \frac{\pi^2\sigma^2}{6}$$

where $\gamma = 0.577215$ is Euler's gamma constant, which is the limit as $n \to \infty$ of the difference between the harmonic series $1 + 1/2 + 1/3 + \cdots + n$ and the logarithm of n. The connection between the Euler constant and the mean can be established through differentiating the Γ function with respect to the shape parameter, but note that R cannot give the formulae only the numerical values in contrast to Mathematica. In R, the constant γ is available as `-digamma(1)`, where `digamma` is the derivative of the Γ function.

Let $\overline{X} = \sum_i X_i/n$ and $\hat{\sigma}^2 = \sum_i (X_i - \bar{X})^2/n$ be the first two sample central moments. Note that the method of moments can either use central moments or non-central moments. Because the variance formula for the Gumbel distribution only involves the parameter σ, it makes sense to use central moments here.

Show that the resulting sentral method of of moment estimators for μ and σ are

$$\widetilde{\sigma} = \frac{\sqrt{6}\hat{\sigma}}{\pi}, \quad \widetilde{\mu} = \overline{X} - \gamma\widetilde{\sigma}.$$

The estimators are then easily computed in both Mathematica and R. In Mathematica, both central moment estimators and moment estimators (without centering ie using the mean of the squares rather than the variance) are easily computed by altering the same command used for estimating the maximum likelihood estimators as follows:

$$\begin{aligned}\text{MoMCentralDistExtreme} =& \text{EstimatedDistribution}[\text{s1}, \text{ExtremeValueDistribution}[\mu, \sigma],\\ & \text{AccuracyGoal} \to \infty,\\ & \text{ParameterEstimator} \to \text{MethodOfCentralMoments}]\\ \text{MoMNonCentralDistExtreme} =& \text{EstimatedDistribution}[\text{s1}, \text{ExtremeValueDistribution}[\mu, \sigma],\\ & \text{AccuracyGoal} \to \infty,\\ & \text{ParameterEstimator} \to \text{MethodOfMoments}]\end{aligned}$$

In R:

```
(sigma.tilde = sqrt(6) * sigma.hat/pi)
(mu.tilde = mu.hat + digamma(1) * sigma.tilde)
```

Compare MM and ML estimates for this model, as well as those obtained in Mathematica and R. Note that the MM estimates are quite close to the maximum likelihood estimates obtained in Task 2, but the MM estimates are cheap from a computational viewpoint and can be used as the starting point for finding the maximum likelihood estimates.

Task 4 - Error in estimation

As in lectures, the possible error in estimation can be assessed through simulation. In this case, unlike in lectures, the parameters are not known, but the estimates can be used - either the Maximum Likelihood or the Method of Moments estimators. To get reasonable accuracy, 1000 samples will be generated from the extreme value distribution with the estimated parameters from the data, the maximum likelihood and method of moments estimators found for each sample, and the means and standard deviations of the estimators obtained. In Mathematica this can be done using the `Table` command to generate samples and their standard deviations as follows:

$$\begin{aligned}
\text{TableMuSigMLE} =& \text{Table}[\text{s1r} = \text{RandomReal}[\text{maxLikeDistExtreme}, \text{Length}[\text{s1}]];\\
&\{\{\mu,\sigma\}/.\,\text{FindDistributionParameters}[\text{s1r}, \text{ExtremeValueDistribution}[\mu,\sigma],\\
&\text{ParameterEstimator} \to \text{MethodOfCentralMoments}],\\
&\{\mu,\sigma\}/.\,\text{FindDistributionParameters}[\text{s1r}, \text{ExtremeValueDistribution}[\mu,\sigma],\\
&\text{ParameterEstimator} \to \text{MaximumLikelihood}]\}, 1000];\\
\text{MeanStDevMLE} =& \{\text{Mean}[\text{TableMuSigMLE}], \text{StandardDeviation}[\text{TableMuSigMLE}]\}\\
\text{TableMuSigMoM} =& \text{Table}[\text{s1r} = \text{RandomReal}[\text{MoMCentralDistExtreme}, \text{Length}[\text{s1}]];\\
&\{\{\mu,\sigma\}/.\,\text{FindDistributionParameters}[\text{s1r}, \text{ExtremeValueDistribution}[\mu,\sigma],\\
&\text{ParameterEstimator} \to \text{MethodOfCentralMoments}],\\
&\{\mu,\sigma\}/.\,\text{FindDistributionParameters}[\text{s1r}, \text{ExtremeValueDistribution}[\mu,\sigma],\\
&\text{ParameterEstimator} \to \text{MaximumLikelihood}]\}, 1000];\\
\text{MeanStDevMoM} =& \{\text{Mean}[\text{TableMuSigMoM}], \text{StandardDeviation}[\text{TableMuSigMoM}]\}
\end{aligned}$$

Several aspects of the Mathematica simulations merit explanation:

- `RandomReal` produces the random numbers from the extreme value distribution.
- The Table command ensures that the commands inside its square brackets are repeated 1000 times.
- The ; delays execution of the command to find the sample so that the same sample is used with both moment and maximum likelihood estimation to increase the accuracy of comparisons.
- The around the two `FindDistributionParameters` commands means that the output of each of the 1000 repititions has two components in a list, one for moment estimation and one for maximum likelihood estimation.
- The $\{\mu,\sigma\}/.$ replaces μ,σ with the output of `FindDistributionParameters` (this output is a rule rather than the value).
- `TableMuSig` is a list with 1000 entries.
- `MeanStDevMLE` is a list with 2 entries the first consisting of means and the second of standard deviations.

In R, only commands for the estimates starting with the maximum likelihood estimates for s1 as the simulating distribution are given. Repeat these, with the necessary changes, substituting the method of moments estimators for the parameters of simulation:

```
# function to simulate from the gumbel distribution
# with the same sample size as s1 and mle estimates
# for the paramters
rgum <- function(x) {
    rgumbel(length(s1), loc = gumbel.fit$estimate[1],
        scale = gumbel.fit$estimate[2])
}
# simulate 1000 samples producing a list
samples <- lapply(1:1000, rgum)
# function to find the maximum likelihood estimates
fit.mle.gum <- function(x) {
    fitdistr(x, densfun = dgumbel, start = list(loc = 50,
        scale = 10), lower = 1e-04)
}
# function to find the method of moments estimators
fit.mom.gum <- function(x) {
    mu.hat = mean(x)
    sigma.hat = sqrt(var(x) * (length(x) - 1)/length(x))
    sigma.tilde = sqrt(6) * sigma.hat/pi
    c(sigma.tilde = sigma.tilde, mu.tilde = mu.hat +
        digamma(1) * sigma.tilde)
}
# get lists of 1000 mle and mom estimates
gumbelsamples.mle.fit <- lapply(samples, fit.mle.gum)
gumbelsamples.mom.fit <- lapply(samples, fit.mom.gum)
# extract vectors of nle and mom location and scale
# estimates
loc.mle.samples <- sapply(gumbelsamples.mle.fit, function(x) x$estimate[1])
loc.mom.samples <- sapply(gumbelsamples.mom.fit, function(x) x[2])
scale.mle.samples <- sapply(gumbelsamples.mle.fit,
    function(x) x$estimate[2])
scale.mom.samples <- sapply(gumbelsamples.mom.fit,
    function(x) x[1])
# print a list of means of the estimates
lapply(list(mean.mle.loc = loc.mle.samples, mean.mom.loc = loc.mom.samples,
    mean.mle.scale = scale.mle.samples, mean.mom.scale = scale.mom.samples),
    mean)
# print a list of sd's of the estimates
lapply(list(sd.mle.loc = loc.mle.samples, sd.mom.loc = loc.mom.samples,
    sd.mle.scale = scale.mle.samples, sd.mom.scale = scale.mom.samples),
    sd)
```

Comment on your results comparing the maximum likelihood estimates with the method of moments estimates, particularly on the standard deviations of the estimates.

2 Workshop

1. a. A random sample $X_1, \ldots, X_n$ of size n is taken from a Poisson distribution with mean $\lambda > 0$.

   i. Show the maximum likelihood estimator of λ is $\hat{\lambda} = \bar{X}$.

   ii. Suppose with $n = 40$ we observe 5 zeros, 7 ones, 12 twos, 9 threes, 5 fours, 1 five, and 1 six. What is the maximum likelihood estimate of λ.

   b. Find the maximum likelihood estimator, $\hat{\theta}$, if $X_1, \cdots, X_n$ is a random sample from the following probability density function:

   $$f(x;\theta) = (1/2)\exp(-|x-\theta|), -\infty < x < \infty, 0 < \theta < \infty$$

   This involves minimizing $\sum_{i=1}^{n} |x_i - \theta|$*, which is difficult. Try* $n = 5$ *and a sample 6.1, -1.1, 3.2, 0.7, 1.7.Then deduce the MLE.*

2. Let $f(x;\theta) = \theta x^{\theta-1}$, $0 < x < 1$, $\theta \in \Omega = \{\theta : 0 < \theta < \infty\}$ and let $X_1, \ldots, X_n$ denote a random sample from this distribution. Note that

   $$\int_0^1 x\theta\, x^{\theta-1} dx = \frac{\theta}{\theta+1}$$

   a. Sketch the p.d.f. of X for $\theta = 1/2$ and $\theta = 2$.

   b. Show that $\hat{\theta} = -n/\ln\left(\prod_{i=1}^{n} X_i\right)$ is the maximum likelihood estimator of θ.

   c. For each of the following three sets of observations from this distribution compute the maximum likelihood estimates and the methods of moments estimates.

X	Y	Z
0.0256	0.9960	0.4698
0.3051	0.3125	0.3675
0.0278	0.4374	0.5991
0.8971	0.7464	0.9513
0.0739	0.8278	0.6049
0.3191	0.9518	0.9917
0.7379	0.9924	0.1551
0.3671	0.7112	0.0710
0.9763	0.2228	0.2110
0.0102	0.8609	0.2154

($\sum_{i=1}^n \ln(x_i) = -18.2063$, $\sum_{i=1}^n \ln(y_i) = -4.5246$, $\sum_{i=1}^n \ln(z_i) = -10.42968$, $\sum_{i=1}^n x_i = 3.7401$, $\sum_{i=1}^n y_i = 7.0592$, $\sum_{i=1}^n z_i = 4.6368$.)

3. Let $X_1, \ldots, X_n$ be a random sample from the exponential distribution whose p.d.f. is $f(x;\theta) = (1/\theta)\exp(-x/\theta)$, $0 < x < \infty$, $0 < \theta < \infty$.

   a. Show that $\bar{X}$ is an unbiased estimator of θ.

   b. Show that the variance of $\bar{X}$ is θ^2/n. What is a good estimate of θ if a random sample of size 5 yielded the values 3.5, 8.1, 0.9, 4.4 and 0.5?

4. Let $X_1, \ldots, X_n$ be a random sample from a distribution having finite variance σ^2. Show that

$$S^2 = \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{n-1}$$

   is an unbiased estimator of σ^2. HINT: Write

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

   and compute $E(S^2)$.

5. Let $X_1, \ldots, X_n$ be a random sample of size n from the distribution with p.d.f. $f(x;\theta) = (1/\theta)x^{(1-\theta)/\theta}$, $0 < x < 1$.

a. Show the mean of X is $E(X) = 1/(1+\theta)$.

b. Show the maximum likelihood estimator of θ is

$$\hat{\theta} = -\frac{1}{n}\sum_{i=1}^{n} \ln X_i$$

c. Is the MLE unbiased? (You can do this using integration by parts or using Mathematica combined with rules about expectation of the sample mean.)

d. Show the method of moments estimator of θ is

$$\widetilde{\theta} = \frac{1-\bar{X}}{\bar{X}}.$$