FREQUENTIST ESTIMATION

METHOD OF MOMENTS AND PERCENTILE MATCHING

- Random sample $\left(X_1, X_2, \cdots, X_n\right)$ where all *n* observations came from the **same parametric distribution**, $F(x|\theta)$. θ is a vector (length *p*) of unknown parameters.

- Let $\mu'_k(\theta) = E\left(X^k \mid \theta\right)$. Using a random sample of independent observations, the empirical estimate of the *k*th moment is $\tilde{\mu}'_k = \dfrac{\sum_{j=1}^{n} x_j^k}{n}$, i.e. the *k*th moment of the sample (*k*th empirical moment).

- Let $\pi_g(\theta)$ be the $100g\%$ percentile of the random variable X, that is, $F\left(\pi_g\left(\theta\right)|\theta\right) = g$. If F is continuous this equation will have, at least, one solution. The empirical estimate of this percentile is $\tilde{\pi}_g$, the corresponding percentile of the random variable.

Definition 13.1 – A **method of moment** estimate of θ is any solution of the *p* equations $\mu'_k(\theta) = \tilde{\mu}'_k$, $k = 1, 2, \cdots, p$.

- Comments:
  - Although definition 13.1 can be generalized to consider any set of moments, results are usually better when using the **smallest positive integer moments**.
  - Sometime we must use higher moments to solve the system (for instance $X \sim U(-\theta, \theta)$ cannot be solved using the first moment)
  - It is necessary to check that the relevant population moments exist.
  - There is no guarantee that the equations will have a solution or, if there is a solution, that it will be unique

Example 13.1 – Use the method of moments to estimate parameters for the exponential, gamma and Pareto distributions for Data Set B from chapter 11.

The exponential distribution has one parameter but the Pareto and the Gamma have 2 parameters each, so we will need 2 empirical moments.

$$\tilde{\mu}_1' = \frac{\sum_{j=1}^{20} x_j}{20} = \bar{x} = 1424.4 \quad \text{and} \quad \tilde{\mu}_2' = \frac{\sum_{j=1}^{20} x_j^2}{20} = 13238441.9$$

Exponential distribution: $E(X) = \theta$, then $\tilde{\theta} = 1424.4$

Gamma Distribution: $E(X) = \alpha\theta$ and $E(X^2) = \alpha(\alpha+1)\theta^2$

Then we must solve the system

$$\begin{cases} \alpha\theta = 1424.4 \\ \alpha(\alpha+1)\theta^2 = 13238441.9 \end{cases}$$. The solution is $$\begin{cases} \tilde{\alpha} = 0.181 \\ \tilde{\theta} = \dfrac{1424.4}{\tilde{\alpha}} = 7869.61 \end{cases}$$.

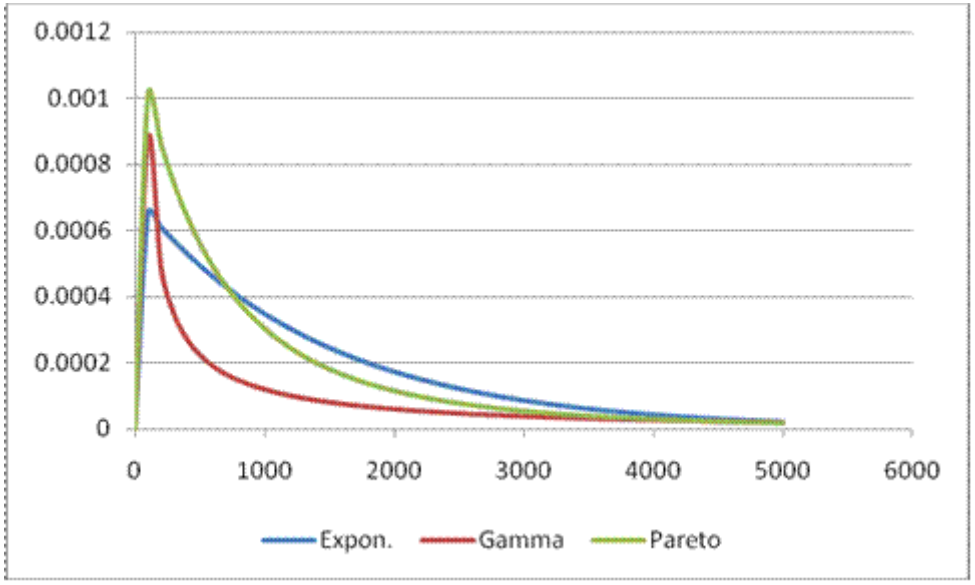
Pareto distribution: $E(X)=\frac{\theta}{\alpha-1}$ and $E(X^2)=\frac{2\theta^2}{(\alpha-1)(\alpha-2)}$ for $\alpha>2$.

The system is then

$$\begin{cases} \dfrac{\theta}{(\alpha-1)}=1424.4 \\ \dfrac{2\theta^2}{(\alpha-1)(\alpha-2)}=13238441.9 \end{cases}$$ and the solution is $\begin{cases} \tilde{\alpha}=2.442 \\ \tilde{\theta}=2053.985 \end{cases}$

- Solutions using R or Excel will be discussed latter
- As we can see (next slide) the choice of a distribution has strong consequences

Estimated distributions



	Exponential	**Gamma**	**Pareto**
$\hat{\Pr}(X > 1000) =$	0.4956	0.2686	0.3796
$\hat{\Pr}(X > 5000) =$	0.0299	0.0850	0.0491
$\hat{\Pr}(X > 50000) =$	5.69×10^{-16}	6.73×10^{-5}	3.73×10^{-4}

Definition 13.2 – A **percentile matching** estimate of θ is any solution of the *p* equations $\pi_{g_k}(\theta) = \hat{\pi}_{g_k}$, $k = 1, 2, \cdots, p$, where $g_1, g_2, \cdots, g_p$ are *p* arbitrarily chosen percentiles. From the definition of percentile, the equations can be written as $F\left(\hat{\pi}_{g_k} \mid \theta\right) = g_k$, $k = 1, 2, \cdots, p$.

- Comments:
  - There is no guarantee that the equations will have a solution or, if there is a solution, that the solution is unique;
  - For discrete random variables percentiles are not always well defined;
  - When using empirical percentiles, i.e. percentiles calculated from the empirical distribution, the situation could be controversial. Most of the time we need an interpolation scheme but, except for the median, there is no "consensual" solution (Hyndman and Fan (1996) present nine different methods and the function *quantile* of the R program allows us to get the percentiles using any of these methods). In this course we will use Definition 13.3 (*type=6* for the *quantile* function)

Definition 13.3 – The smoothed empirical estimate of a percentile is found by $\hat{\pi}_g = (1-h)\,x_{(j)} + h\,x_{(j+1)}$ where $j = \lfloor (n+1)g \rfloor$, $h = (n+1)g - j$, $\lfloor\ \rfloor$ indicates the greatest integer function and $x_{(1)} \le x_{(2)} \le x_{(3)} \le \cdots \le x_{(n)}$ are the order statistics from the sample.

- Comments:
  - Unless the sample has two or more data points with the same values, no two percentiles will have the same value.
  - We can only estimate percentile when $\frac{1}{n+1} \le g \le \frac{n}{n+1}$.
  - The choice of which percentiles to use leads to different estimates. This is a strong point against the percentile matching method except when there is a reason to choose a particular set of percentiles.

Example 13.2 – Use percentile matching to estimate parameters for the exponential and Pareto distribution for Data set B.

Without more information, the choice of the percentiles is quite arbitrary. We will follow *Loss Models*.

Exponential: We can use the median (the parameter is the mean, i.e. a location parameter). As $\Pr(X < \theta) = 1 - e^{-1} \simeq 0.6321$ and we could use this percentile (0.6321).

Sample median: $\hat{\pi}_{0.5} = 0.5 \times 384 + 0.5 \times 457 = 420.5$

We must solve the equation

$$0.5 = F(\hat{\pi}_{0.5} \mid \theta) \Leftrightarrow 0.5 = 1 - \exp(-420.5 / \hat{\theta}) \Leftrightarrow \ln 2 = 420.5 / \hat{\theta} \Leftrightarrow \hat{\theta} = 606.65$$

Pareto: use the 30th and the 80th percentiles.

30th: $j = \lfloor 21 \times 0.3 \rfloor = 6;\ h = 21 \times 0.3 - 6 = 0.3;\ \hat{\pi}_{0.3} = 0.7 \times 161 + 0.3 \times 243 = 185.6$

80th: $j = \lfloor 21 \times 0.8 \rfloor = 16;\ h = 21 \times 0.8 - 16 = 0.8;\ \hat{\pi}_{0.3} = 0.2 \times 1193 + 0.8 \times 1340 = 1310.6$

The equations are

$$\begin{cases} 0.3 = F(185.6 \mid \theta, \alpha) \\ 0.8 = F(1310.6 \mid \theta, \alpha) \end{cases} \Leftrightarrow \begin{cases} 0.7 = \left(\dfrac{\hat{\theta}}{185.6 + \hat{\theta}} \right)^{\hat{\alpha}} \\ 0.2 = \left(\dfrac{\hat{\theta}}{1310.6 + \hat{\theta}} \right)^{\hat{\alpha}} \end{cases} \Leftrightarrow \begin{cases} \ln(0.7) = \hat{\alpha} \ln\left(\dfrac{\hat{\theta}}{185.6 + \hat{\theta}} \right) \\ \ln(0.2) = \hat{\alpha} \ln\left(\dfrac{\hat{\theta}}{1310.6 + \hat{\theta}} \right) \end{cases}$$

That is

$$\begin{cases} \hat{\alpha} = \dfrac{\ln(0.7)}{\ln(\hat{\theta}) - \ln(185.6 + \hat{\theta})} \\ \dfrac{\ln(0.2)}{\ln(0.7)} = \dfrac{\ln(\hat{\theta}) - \ln(1310.6 + \hat{\theta})}{\ln(\hat{\theta}) - \ln(185.6 + \hat{\theta})} \end{cases} \Leftrightarrow \begin{cases} \hat{\alpha} = \dfrac{\ln(0.7)}{\ln(\hat{\theta}) - \ln(185.6 + \hat{\theta})} \\ \dfrac{\ln(0.2)}{\ln(0.7)} - \dfrac{\ln(\hat{\theta}) - \ln(1310.6 + \hat{\theta})}{\ln(\hat{\theta}) - \ln(185.6 + \hat{\theta})} = 0 \end{cases}$$

This system can be solved numerically.

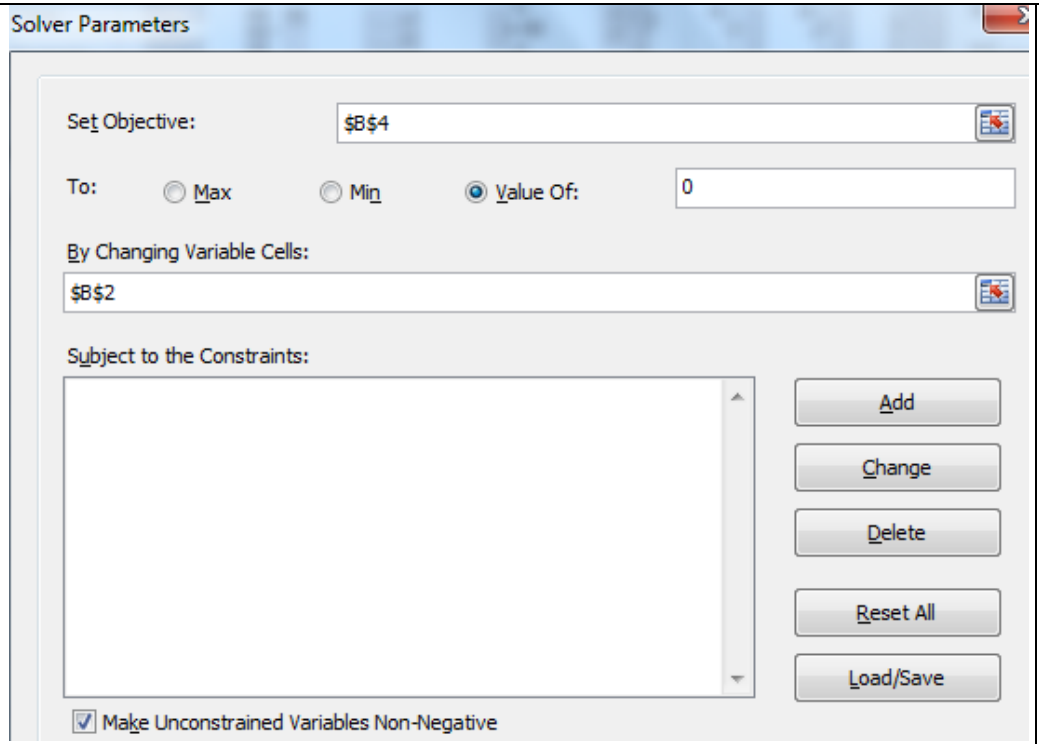
Using Excel's solver we obtain $\hat{\theta} = 715.0315$ for the second equation and, reporting this value in the first equation we get $\hat{\alpha} = 1.545589$ (see next slide) and using R we obtain similar results (slides 11-17)

Of course, the choice of different percentiles leads to different estimates.

Exercise: Use percentiles 0.1 and 0.9, obtain $\hat{\theta}$ and $\hat{\alpha}$, and comment.

Using EXCEL's solver

	A	B
1		
2	Theta	10
3		
4	Equation	2.870072
5		
6	Alpha	0.119952



	A	B
1		
2	Theta	715.0332
3		
4	Equation	-2.02E-06
5		
6	Alpha	1.545592

Using R

We can choose among different approaches. 2 of them are:

- Function **nleqslv** (library nleqslv) to solve systems of nonlinear equations
- Function **nlm** (basic packages) to obtain a solution using a nonlinear minimization

In both cases we need to define the set of equations as a function. Depending on when we abandon the search for an analytical solution, we can define:

```
> fn1=function(x){
+    # x[1]=alpha      x[2]=theta
+    eq1=0.7-(x[2]/(185.6+x[2]))^x[1]
+    eq2=0.2-(x[2]/(1310.6+x[2]))^x[1]
+    return(c(eq1,eq2))
+    }
```

or

```
> fn2=function(x){
+    # x[1]=alpha      x[2]=theta
+    eq1=log(0.7)-x[1]*log(x[2]/(185.6+x[2]))
+    eq2=log(0.2)-x[1]*log(x[2]/(1310.6+x[2]))
+    return(c(eq1,eq2))
+    }
```

or

```
> fn3=function(x){
+    # x=theta
+    eq1=(log(0.2)/log(0.7))-(log(x)-log(1310.6+x))/(log(x)-log(185.6+x))
+    return(eq1)
+    }
```

Using nleqslv

> require(nleqslv)

> **nleqslv(0.5,fn3)** # using fn3 – 0.5 is a first guess for θ

```
$`x`
[1] 715.032          Solution
$fvec                How close to zero?
[1] 4.887363e-09
$termcd              termination code: 1 is OK
[1] 1
$message
[1] "Function criterion near zero"
$scalex
[1] 1
$nfcnt      number of function evaluations, excluding …
[1] 11
$njcnt number of Jacobian evaluations
[1] 1
$iter       number of iterations
[1] 11
```

> **nleqslv(c(1,0.5),fn1)** # using fn1 – 1 and 0.5 are first guesses for α and θ respectively

```
$`x`
[1]   1.54559 715.03199
$fvec
[1] -2.676970e-12 -5.139861e-12
$termcd
[1] 1
$message
[1] "Function criterion near zero"
$scalex
[1] 1 1
$nfcnt
[1] 78
$njcnt
[1] 1
$iter
[1] 54
```

```
> nleqslv(c(1,0.5),fn2)    # using fn2 – 1 and 0.5 are first guesses for α and θ respectively
$`x`
[1]   1.54559 715.03199
$fvec
[1] 1.122602e-12 3.778755e-12
$termcd
[1] 1
$message
[1] "Function criterion near zero"
$scalex
[1] 1 1
$nfcnt
[1] 38
$njcnt
[1] 3
$iter
[1] 29
```

Using nlm

```
> fn1.sq=function(x) return(crossprod(fn1(x),fn1(x)))
> fn2.sq=function(x) return(crossprod(fn2(x),fn2(x)))
> fn3.sq=function(x) return(fn3(x)^2)
```

```
> nlm(fn1.sq,c(1,5))
```

1 and 5 are first guesses for α and θ respectively

```
$`minimum`
[1] 1.130887e-12
$estimate
[1]   1.545571 715.021786
$gradient
[1] -2.820609e-07  6.197286e-10
$code
[1] 1
$iterations
[1] 38
```

```
> nlm(fn2.sq,c(1,5))
  $`minimum`
  [1] 2.188255e-10
  $estimate
  [1]   1.545437 714.922321
  $gradient
  [1] -4.369224e-06  6.098373e-09
  $code
  [1] 2
  $iterations
  [1] 35
> nlm(fn3.sq,5)
  $`minimum`
  [1] 3.826888e-13
  $estimate
  [1] 715.0316
  $gradient
  [1] -1.95615e-11
  $code
  [1] 1
  $iterations
  [1] 15
```

MAXIMUM LIKELIHOOD ESTIMATION

- Why ML estimation?
  - More efficient estimators
  - To cover some annoying cases: An important limitation of moment and percentile matching estimators is that the observations are from the same random variable. If, for instance, half the observations have a deductible of 50 and the other half a deductible of 100 it is not clear to what the sample mean should be equated.
  - More calculus involved
  - Sometimes ML estimators are quite sensitive to "extreme" observations
- To use Maximum Likelihood Estimators
  - We must have a data set with *n* **events,** $A_1, A_2, \cdots, A_n$, where A_j is whatever was observed for the *j*th observation (usually A_j is a value or an interval)
  - The variables $X_1, X_2, \cdots, X_n$ behind the events $A_1, A_2, \cdots, A_n$ do not need to have the same probability distribution but they **must be independent** and **their distribution must depend on the same parameter vector** θ.

- **Definition 13.4** – The **likelihood function** is $L(\theta) = \prod_{j=1}^{n} \Pr(X_j \in A_j \mid \theta)$ and the maximum likelihood estimate of θ is the vector that maximizes the likelihood function.
- Comments:
  - **Notation** – Usually the likelihood function is written as $L(\theta \mid x_1, x_2, \cdots, x_n)$. Because observed data can take many forms, we will write $L(\theta)$ without clarifying the conditioning values.
  - **Independence among events** – As the events $A_1, A_2, \cdots, A_n$ are assumed independent, the likelihood is the probability, given a particular value of θ, of observing what was observed, since $L(\theta) = \prod_{j=1}^{n} \Pr(X_j \in A_j \mid \theta) = \Pr(X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n \mid \theta)$.
  - **Theoretical** – When the probabilistic model is continuous and the observed event is a point, $A_j = x_j$, we know that $\Pr(X_j \in A_j \mid \theta) = 0$ and we will use the density function. The rationale for such a procedure corresponds to interpret the observed value as being in a neighborhood of x_j and to approximate the probability $\Pr(x_j - \varepsilon < X_j < x_j + \varepsilon \mid \theta)$ by means of $2\varepsilon\, f(x_j \mid \theta)$, where $f(x_j \mid \theta)$ is the density function at x_j. Dropping out the multiplicative constants leads to use the density $f(x_j \mid \theta)$ as the contribution to the likelihood function.

- Multiplicative constants that are independent of the elements of the vector θ can be removed from the likelihood function since they will not affect the maximum likelihood estimate. Removing such constants does not change the solution but it will change the value of the likelihood.
- There is no guarantee that the likelihood function has a maximum at eligible parameter values. When maximizing the likelihood function the existence of local maxima can hide the global maximum.
- **Log-likelihood** – In many situations it is easier to use the log-likelihood, that is, to maximize $\ell(\theta) = \ln L(\theta) = \sum_{j=1}^{n} \ln\left(\Pr(X_j \in A_j \mid \theta)\right)$ instead of $L(\theta)$ (as the natural logarithm is a strictly increasing function the solution is unchanged).
- $\ln\left(\Pr(X_j \in A_j \mid \theta)\right)$ is called the **individual contribution** of observation *j* to the log likelihood.
- In many situations numerical methods are needed.

COMPLETE INDIVIDUAL DATA

When there is no truncation and no censoring and the value of each observation is recorded, it is easy to write the log-likelihood function, $\ell(\theta) = \sum_{j=1}^{n} \ln f_{X_j}(x_j \mid \theta)$.

- **Example 13.4** – Using Data set B, determine the maximum likelihood estimate for an exponential distribution, for a gamma distribution where α is known to equal 2, and for a gamma distribution where both parameters are unknown.

  Exponential distribution

  $f(x \mid \theta) = \theta^{-1} e^{-x/\theta}$, $x > 0$, $\theta > 0$.

  $$\ell(\theta) = \sum_{j=1}^{n} \ln\left(\theta^{-1} e^{-x_j/\theta}\right) = \sum_{j=1}^{n} \left(-\ln\theta - x_j \theta^{-1}\right)$$

  $$\ell'(\theta) = \sum_{j=1}^{n} \left(-\theta^{-1} + x_j \theta^{-2}\right) = -n\theta^{-1} + n\bar{x}\theta^{-2}$$

  $$\ell'(\theta) = 0 \Leftrightarrow 0 = -n\theta^{-1} + n\bar{x}\theta^{-2} \Leftrightarrow \theta = \bar{x}$$

  $$\ell''(\theta) = \sum_{j=1}^{n} \left(\theta^{-2} - 2x_j \theta^{-3}\right) = n\theta^{-2}\left(1 - 2\bar{x}\theta^{-1}\right)$$

  As $\ell''(\theta)\}_{\theta=\bar{x}} = -n\theta^{-2} < 0$ we get $\hat{\theta} = \bar{x} = 1424.4$ (same estimate as with the method of moments)

Gamma distribution with $\alpha = 2$ - similar to the previous case

Gamma distribution with unknown parameters – numerical maximization

$$f(x \mid \alpha, \theta) = \frac{x^{\alpha-1} e^{-x/\theta}}{\theta^{\alpha} \Gamma(\alpha)},\ x > 0,\ \alpha, \theta > 0.$$

$$\ell(\alpha, \theta) = \sum_{j=1}^{n} \ln\left(f(x_j \mid \alpha, \theta) \right) = \sum_{j=1}^{n} \left((\alpha - 1) \ln x_j - \alpha \ln \theta - x_j \theta^{-1} - \ln \Gamma(\alpha) \right)$$

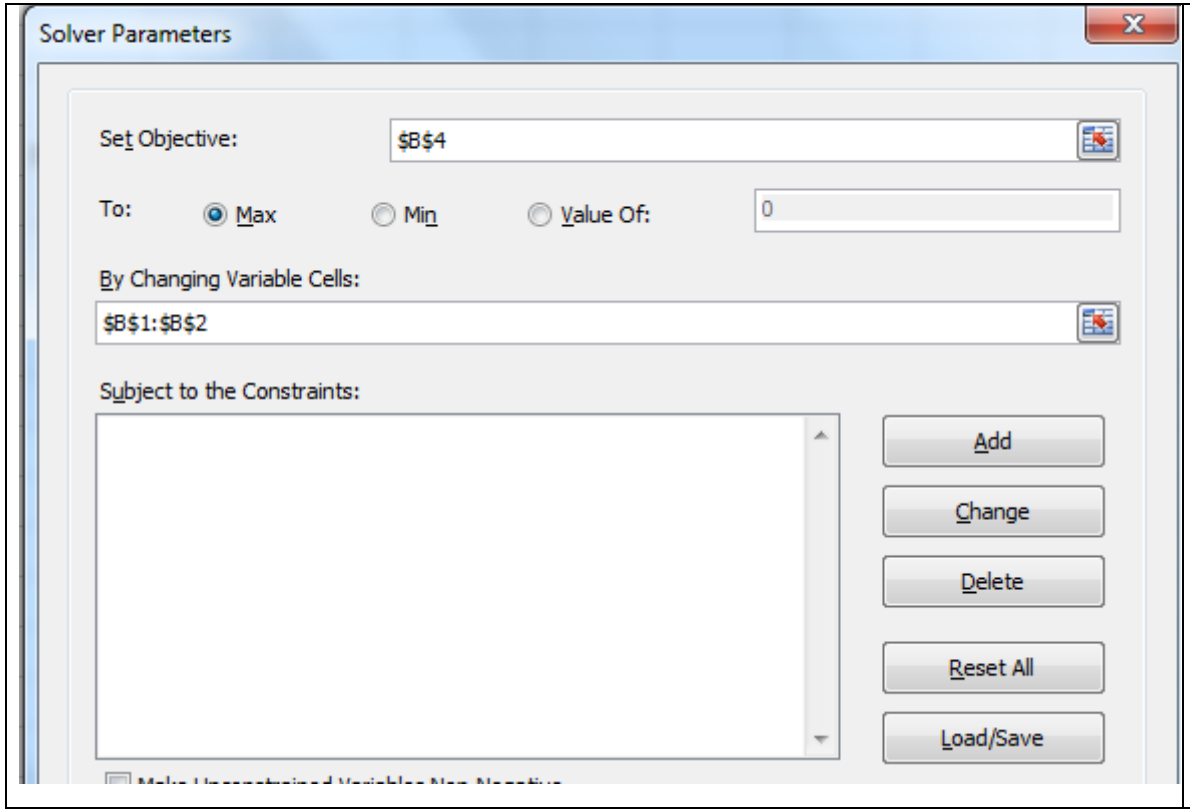
To maximize in order to α requires the derivative of $\ln \Gamma(\alpha)$ which is not an explicit function (we can obtain a solution in order to θ, $\theta = \bar{x} / \alpha$, but the problem remains**).** Consequently, we need to use numerical techniques.

We illustrate the procedure using Microsoft EXCEL solver and R.

EXCEL

E6

	A	B	C	D	E	F	G	H	I
1	alfa	2							
2	theta	500							
3									
4	loglik=	-182.8027631	sum of column ln f(x_j)						
5									
6									
7	x_j	ln f(x_j)							
8	27	-9.187379331	←	LN(GAMMADIST(A8;B1;B2;FALSE))					
9	82	-8.18649695							
10	115	-7.914284068							
11	126	-7.84493429							
12	155	-7.69579108							



	A	B	C
1	alfa	0.556157796	
2	theta	2561.142391	
3			
4	loglik=	-162.2934031	sum
5			
6			
7	x_j	ln f(x_j)	
8	27	-6.307636437	
9	82	-6.822167714	
10	115	-6.98516574	
11	126	-7.030005585	

Then $\hat{\alpha} = 0.55616$ and $\hat{\theta} = 2561.14$. If necessary, we can use a different starting point and/or we can add constraints.

Using R – Two among many solutions.

```
> x=c(27,82,115,126,155,161,243,294,340,384,457,680,855,877,974,
+       1193,1340,1884,2558,15743)
> mean(x)
[1] 1424.4
>
> # 1ST SOLUTION: USE FUNCTION nlm
> # As nlm minimizes a function we introduce minus the log-lik
> minusloglikgamma=function(param,x){
+   alpha=param[1]; theta=param[2]
+   -sum(dgamma(x,shape=alpha,scale=theta,log=TRUE))
+   }
> param.start=c(1,1000) # starting values - important point
> out1=nlm(minusloglikgamma,param.start,x=x)  # Options available
Warning messages:
1: In dgamma(x, shape, scale, log) : NaNs produced
2: In nlm(minusloglikgamma, param.start, x = x) :
  NA/Inf replaced by maximum positive value
>
```

```
>out1
$minimum
[1] 162.2934                          # Minus the log-likelihood
$estimate
[1]    0.556156 2561.146495
$gradient
[1] -8.273560e-05 -6.824815e-09       # Check the convergence
$code
[1] 1                                 # Check the convergence
$iterations
[1] 26
>
> # 2ND SOLUTION: USE FUNCTION maxLik, LIBRARY maxLik
> # As maxLik maximizes a function we introduce the log-lik
> loglikgamma=function(param,x){
+   alpha=param[1]; theta=param[2]
+   sum(dgamma(x,shape=alpha,scale=theta,log=TRUE))
+   }
> # param.start has already been defined
> library(maxLik)
> out2=maxLik(loglikgamma,start=param.start,x=x)
There were 50 or more warnings (use warnings() to see the first 50)
```

```
> out2
Maximum Likelihood estimation
Newton-Raphson maximisation, 22 iterations
Return code 1: gradient close to zero
Log-Likelihood: -162.2934 (2 free parameter(s))
Estimate(s): 0.5562315 2560.365
```

Comments:

- Both functions are based on the Newton-Raphson method;
- We can use the gradient and the Hessian matrix to improve results;
- We can control the process changing some parameters values (tolerance, maximum number of iterations, …);
- Other procedures are available to maximize the log-likelihood.
- Package **fitdistrplus** provides a set of tools to fit distributions using different methods

COMPLETE GROUPED DATA

- We must write the likelihood considering the mass probability associated with each group.
- Let us assume that there are k groups and that group j, with n_j observations, is limited by values c_{j-1} and c_j. The likelihood function is $L(\theta) = \prod_{j=1}^{k} \left(F(c_j \mid \theta) - F(c_{j-1} \mid \theta) \right)^{n_j}$ and the log likelihood is $\ell(\theta) = \sum_{j=1}^{k} n_j \ln\left(F(c_j \mid \theta) - F(c_{j-1} \mid \theta) \right)$
- **Example 13.5 –** From Data Set C, determine the maximum likelihood estimate of an exponential distribution.

  $F(x \mid \theta) = 1 - e^{-x/\theta}$; $F(c_j \mid \theta) - F(c_{j-1} \mid \theta) = e^{-c_{j-1}/\theta} - e^{-c_j/\theta}$

  The log-likelihood is then

  $\ell(\theta) = 99 \times \ln\left(1 - e^{-7500/\theta}\right) + 42 \times \ln\left(e^{-7500/\theta} - e^{-17500/\theta}\right) + \cdots + 3 \times \ln\left(e^{-300000/\theta} - 0\right)$

  Using Microsoft Excel or another numerical procedure to maximize the log-likelihood we get $\hat{\theta} = 29720.77$ and $\ell(\hat{\theta}) = -406.03$.

  Exercise: check the results using EXCEL or R □

TRUNCATED AND CENSORED DATA

- Censored data: Non-censored observations are individual points and censored observations are grouped data.
- Truncated data: More challenging. We must keep in mind that some values of the r.v. cannot be observed.
- Klugman, Panjer and Willmot (*Loss Models*) pointed out that there are two ways to proceed but it is important to underline that these ways correspond to **two different models**. Note that in both situations we only observe the values above the truncation points.

  First model – We want to estimate the distribution of the truncated values;

  Second model – We want to estimate the model behind the values without truncation (more interesting case);

- **Example 13.6 -** Assume the values in Data Set B had been truncated from below at 200. Using both methods estimate the value of α for a Pareto distribution with $\theta = 800$ known. Then use the model to estimate the cost per payment with deductibles of 0, 200 and 400.

  As data has been truncated at 200 we only consider observations above 200 (14 observations)

First model – Shift the data by subtracting 200. In this model we will consider that the shifted data follow a Pareto distribution with unknown α and $\theta = 800$. The density and the log-likelihood are

$$f(x \mid \alpha, \theta = 800) = \frac{\alpha 800^{\alpha}}{(800+x)^{\alpha+1}},\ x>0,\ \alpha>0 \qquad \ell(\alpha) = \sum_{j=1}^{n}\left(\ln\alpha + \alpha\ln 800 - (\alpha+1)\ln(800+x_j)\right)$$

$$\ell'(\alpha) = \frac{n}{\alpha} + n\times\ln 800 - \sum_{j=1}^{n}\ln(800+x_j) \qquad \ell'(\alpha)=0 \Leftrightarrow \alpha = \frac{n}{-n\times\ln 800 + \sum_{j=1}^{n}\ln(800+x_j)}$$

We get $\hat{\alpha} = 1.348191$. Then, using this setup our estimate is that**, when a deductible of 200 is in force, the cost per payment follows a Pareto distribution with** $\hat{\alpha} = 1.348191$ **and** $\theta = 800$. The expected value of a payment is 2297.59 = 800/(1.348191-1).

Because data have been shifted it is not possible to estimate the cost with no deductible.

For a deductible of 400, we have to impose a new deductible of 200 in our shifted data. The expected cost per payment is given by (theorem 8.3):

$$E(X-200 \mid X>200) = \frac{E(X) - E(X \wedge 200)}{1-F(200)}$$

Using Loss Models' appendix we get

$$E(X)=\frac{\theta}{\alpha-1} \text{ and } E(X\wedge 200)=\frac{\theta}{\alpha-1}\left(1-\left(\frac{\theta}{200+\theta}\right)^{\alpha-1}\right)$$

Then

$$E(X-200\,|\,X>200)=\frac{E(X)-E(X\wedge 200)}{1-F(200)}=\frac{\frac{800}{0.348191}\times\left(\frac{800}{200+800}\right)^{0.348191}}{\left(\frac{800}{200+800}\right)^{1.348191}}\approx 2871.90$$

Second model – The purpose is to fit a model for the original population, knowing that data were truncated at 200. The density of the observed values is now ($x > 200$, $\alpha > 0$)

$$g(x \mid \alpha, \theta = 800) = \frac{f(x \mid \alpha, \theta = 800)}{1 - F(200 \mid \alpha, \theta = 800)} = \frac{\frac{\alpha 800^{\alpha}}{(800+x)^{\alpha+1}}}{\frac{800^{\alpha}}{(800+200)^{\alpha}}} = \frac{\alpha 1000^{\alpha}}{(800+x)^{\alpha+1}}$$

Note that the values x_j are the original ones (except those below 200 that are not observed).

$$\ell(\alpha) = \sum_{j=1}^{n} \left(\ln \alpha + \alpha \ln 1000 - (\alpha+1) \ln(800 + x_j) \right)$$

$$\ell'(\alpha) = \sum_{j=1}^{n} \left(\frac{1}{\alpha} + \ln 1000 - \ln(800 + x_j) \right) = \frac{n}{\alpha} + n \times \ln 1000 - \sum_{j=1}^{n} \ln(800 + x_j)$$

$$\ell'(\alpha) = 0 \Leftrightarrow \frac{n}{\alpha} = -n \times \ln 1000 + \sum_{j=1}^{n} \ln(800 + x_j) \Leftrightarrow \alpha = \frac{n}{-n \times \ln 1000 + \sum_{j=1}^{n} \ln(800 + x_j)}$$

We get $\hat{\alpha} = 1.538166$, i.e. the **cost per payment without deductible follows a Pareto distribution with** $\hat{\alpha} = 1.538166$ and $\theta = 800$.

The introduction of a deductible of 200 originates an expected cost per payment given by

$$\frac{E(X)-E(X\wedge 200)}{1-F(200)}=\frac{\frac{800}{0.538166}\times\left(\frac{800}{200+800}\right)^{0.538166}}{\left(\frac{800}{200+800}\right)^{1.538166}}\approx 1858.16$$

As it is natural (we are using a different set of hypothesis), this value is different from that obtained with the first model. Note also that we can estimate that only $0.7095=1-F(200|\hat{\alpha},\theta)$ of the claims are reported.

The introduction of a deductible of 400 originates an expected cost per payment given by

$$\frac{E(X)-E(X\wedge 400)}{1-F(400)}=\frac{\frac{800}{0.538166}\times\left(\frac{800}{400+800}\right)^{0.538166}}{\left(\frac{800}{400+800}\right)^{1.538166}}\approx 2229.80$$

Example 13.7 – Determine Pareto and gamma models for the time to death for Data Set D2.

In Data Set D2 we faced 4 different situations:

	Situation	Contribution to the likelihood	Meaning of x
1	Subjects are observed from time *d*=0 and died at time x (observed during the period of the study). No truncation nor censoring.	$f(x\|\theta)$	Time of death
2	Subjects are observed at time *d*=0 and didn't die during the period of the study. No truncation but censoring.	$1-F(x\|\theta)$	Time of censoring
3	Subjects are observed from time *d*>0 (truncation) and died at time *x* (no censoring)	$\frac{f(x\|\theta)}{1-F(d\|\theta)}$	Time of death
4	Subjects are observed at time *t*>0 (truncation) and didn't die during the period of the study (censoring)	$\frac{1-F(x\|\theta)}{1-F(d\|\theta)}$	Time of censoring

It is straightforward to write the contributions to the likelihood (or to the log-likelihood). For instance:

Obs 1 – $d=0$ (no truncation); $x=0.1$ (censoring): $1-F(0.1)$

Obs 4 – $d=0$ (no truncation); $x=0.8$ (no censoring): $f(0.8)$

Obs 31 – $d=0.3$ (truncation); $x=5$ (censoring): $\left(1-F(5.0)\right)/\left(1-F(0.3)\right)$

Obs 33 – $d=1.0$ (truncation); $x=4.1$ (no censoring): $f(4.1)/\left(1-F(1.0)\right)$

Sometimes it is useful to get a single expression for all the situations. Using *d*=0 for the no truncation situation and noting that $F(0|\theta)=0$ we can rewrite the contribution to the likelihood from cases 1 and 2 as $\frac{f(x|\theta)}{1-F(d|\theta)}$ and $\frac{1-F(x|\theta)}{1-F(d|\theta)}$ respectively (with *d*=0 for both cases). Then we define a dummy variable, *v*, assuming value 1 when the x value corresponds to a death (0 otherwise) and we write the likelihood as

$$L(\theta)=\prod_{j=1}^{n}\frac{(1-v_j)\times(1-F(x_j|\theta)+v_j\times f(x_j|\theta)}{1-F(d_j|\theta)}$$

and the log likelihood as $\ell(\theta)=\sum_{j=1}^{n}\left(\ln\left((1-v_j)\times(1-F(x_j|\theta)+v_j\times f(x_j|\theta)\right)-\ln\left(1-F(d_j|\theta)\right)\right)$.

Now you can compute a solution using EXCEL or R. Exercise: Do it using EXCEL

gamma model (using R).

```
> d=c(rep(0,30),0.3,0.7,1.0,1.8,2.1,2.9,2.9,3.2,3.4,3.9)
> x=c(0.1,0.5,0.8,0.8,1.8,1.8,2.1,2.5,2.8,2.9,2.9,3.9,4.0,4.0,4.1,4.8,4.8,4.8,
+     rep(5.0,12),5.0,5.0,4.1,3.1,3.9,5.0,4.8,4.0,5.0,5.0)
> v=c(rep(0,3),1,rep(0,5),1,1,0,1,0,0,1,rep(0,16),1,1,rep(0,3),1,0,0)
>
> minusloglikgamma1=function(theta){
+   -sum(log((1-v)*(1-pgamma(x,shape=theta[1],scale=theta[2],log=FALSE))+
+         v*dgamma(x,shape=theta[1],scale=theta[2],log=FALSE))-
+         log(1-pgamma(d,shape=theta[1],scale=theta[2],log=FALSE)))
+   }
>
> theta.start=c(3,2)
> out=nlm(minusloglikgamma1,theta.start)
> out
$minimum
[1] 28.52685
$estimate
[1] 2.616737     3.311384
$gradient
```

```
[1] 1.026956e-05 3.390297e-06
$code
[1] 1
$iterations
[1] 14
```

The solution is then $\hat{\alpha} = 2.616737$ **and** $\hat{\theta} = 3.311384$**.**

Pareto model

```
> minusloglikPareto1=function(theta){
+   -sum(log((1-d)*(x+theta[2])^(-theta[1])+d*(x+theta[2])^(-theta[1]-1))-
+       theta[1]*log(1+theta[2]))
+   }
> theta.start=c(3,2)
> outPareto=nlm(minusloglikPareto1,theta.start)
Error in nlm(loglikPareto1, theta.start) :
  non-finite value supplied by 'nlm'
In addition: There were 50 or more warnings (use warnings() to see the first 50)
>
```

We are unable to find a solution in this set up.

VARIANCE AND INTERVAL ESTIMATION

- It is not easy to determine the variance of the maximum likelihood estimators. In most situations we need to approximate the variance which can be done when "mid regularity conditions" are verified. There are many ways to write those conditions.
- **Theorem 15.5** – Assume that the pdf (pf in the discrete case) $f(x|\theta)$ satisfies the following for θ in an interval containing the true value (replace integrals by sums for discrete variables):
  - i. $\ln f(x|\theta)$ is three times differentiable with respect to θ.
  - ii. $\int \frac{\partial}{\partial \theta} f(x|\theta)\,dx = 0$ - *This formula implies that the derivative may be taken outside the integral and so we are just differentiating the constant 1* (the main idea is that we can swap the derivation with the integration - the limits of the integral cannot be functions of θ).
  - iii. $\int \frac{\partial^2}{\partial \theta^2} f(x|\theta)\,dx = 0$ - *This formula is the same concept for the second derivative*

iv. $-\infty < \int f(x|\theta) \frac{\partial^2}{\partial \theta^2} \ln f(x|\theta)\, dx < 0$ - *This inequality establishes that the indicated integral exists and that the expected value of the second derivative of the log likelihood is negative.*

v. There exists a function $f(x|\theta)$ such that

$$\int H(x)\, f(x|\theta)\, dx < \infty \text{ with } \left| \int \frac{\partial^3}{\partial \theta^3} \ln f(x|\theta)\, dx \right| < H(x).$$

This inequality guaranties that the population is not overpopulated with regards to extreme values.

Then the following results hold:

i. As $n \to \infty$, the probability that the likelihood equation ($L'(\theta) = 0$) has a solution goes to 1.

ii. As $n \to \infty$, the distribution of the mle $\hat{\theta}_n$ converges to a normal distribution with mean θ and variance such that $I(\theta)\,\mathrm{var}(\hat{\theta}_n) \to 1$ where

$$I(\theta) = -nE\left(\frac{\partial^2}{\partial \theta^2} \ln f(X|\theta) \right) = nE\left(\frac{\partial}{\partial \theta} \ln f(X|\theta) \right)^2$$

Comments to Theorem 13.5

- The quantity $I(\theta)$ is called Fisher's information (of the entire sample = $n\mathfrak{I}(\theta)$ in "Review of …")
- The second statement can be written as $\dfrac{\hat{\theta}-\theta}{I(\theta)^{-1/2}} \overset{\circ}{\sim} n(0;1)$
- The theorem assumes an i.i.d. sample. A more general version of the result can be established and uses the log-likelihood function, that is,

$$I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\ell(\theta \mid X_1, X_2, \cdots, X_n)\right) = E\left(\frac{\partial}{\partial\theta}\ell(\theta \mid X_1, X_2, \cdots, X_n)\right)^2$$

- If there is more than one parameter, the result can be generalized and the maximum likelihood estimators will follow an asymptotic multidimensional normal distribution. $I(\theta)$ is now a matrix with (*r*,*s*) element given by

$$I(\theta)_{r,s} = -E\left(\frac{\partial^2}{\partial\theta_r\,\partial\theta_s}\ell(\theta \mid X_1, X_2, \cdots, X_n)\right)$$

- The inverse of Fisher's information matrix is the Cramér-Rao lower bound for the variance of unbiased estimators of θ, that is to say, no unbiased estimator is asymptotically more accurate than the maximum likelihood estimator.
- When Fisher's information matrix depends on θ we estimate it using $I\left(\hat{\theta}\right)$. When $I\left(\hat{\theta}\right)$ is difficult to obtain we can approximate it using the observed information $I(\hat{\theta}) \approx -H\left(\hat{\theta}\right)$, i.e. using the Hessian matrix of the log likelihood at $\theta = \hat{\theta}$

- **Example 13.9** – Estimate the covariance matrix of the mle for the lognormal distribution. Then apply this result for Data set B.

  Note: When using the lognormal it is usually more adequate to take logarithms of the observed values and to use the normal (gaussian) distribution.

$$L(\mu,\sigma) = \prod_{j=1}^{n} \frac{1}{x_j\,\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(\ln x_j - \mu\right)^2}{2\sigma^2}\right)$$

$$\ell(\mu,\sigma) = \sum_{j=1}^{n} \left(-\ln x_j - \ln\sigma - \ln\left(\sqrt{2\pi}\right) - \frac{\left(\ln x_j - \mu\right)^2}{2\sigma^2}\right)$$

$$\frac{\partial \ell}{\partial \mu} = \sum_{j=1}^{n}\left(-2(-1)\frac{\left(\ln x_j - \mu\right)}{2\sigma^2}\right) = \sum_{j=1}^{n}\left(\frac{\ln x_j - \mu}{\sigma^2}\right)$$

$$\frac{\partial \ell}{\partial \sigma} = \sum_{j=1}^{n}\left(-\frac{1}{\sigma} - (-2)\frac{\left(\ln x_j - \mu\right)^2}{2\sigma^3}\right) = \sum_{j=1}^{n}\left(-\frac{1}{\sigma} + \frac{\left(\ln x_j - \mu\right)^2}{\sigma^3}\right)$$

$$\frac{\partial^2 \ell}{\partial \mu^2} = \sum_{j=1}^{n}\left(\frac{-1}{\sigma^2}\right) = -\frac{n}{\sigma^2} \qquad \frac{\partial^2 \ell}{\partial \mu \partial \sigma} = \sum_{j=1}^{n}(-2)\left(\frac{\ln x_j - \mu}{\sigma^3}\right) = -2\sum_{j=1}^{n}\left(\frac{\ln x_j - \mu}{\sigma^3}\right)$$

$$\frac{\partial^2 \ell}{\partial \sigma^2} = \sum_{j=1}^{n}\left(\frac{1}{\sigma^2} + (-3)\frac{\left(\ln x_j - \mu\right)^2}{\sigma^4}\right) = \frac{n}{\sigma^2} - 3\sum_{j=1}^{n}\left(\frac{\left(\ln x_j - \mu\right)^2}{\sigma^4}\right)$$

Taking expected values

$$E\left(\frac{\partial^2 \ell}{\partial \mu^2}\right) = -\frac{n}{\sigma^2} \qquad E\left(\frac{\partial^2 \ell}{\partial \mu \partial \sigma}\right) = -2\sum_{j=1}^{n}\frac{E\left(\ln X_j\right) - \mu}{\sigma^3} = 0$$

$$E\left(\frac{\partial^2 \ell}{\partial \sigma^2}\right) = \frac{n}{\sigma^2} - 3\sum_{j=1}^{n}\frac{E\left(\ln X_j - \mu\right)^2}{\sigma^4} = \frac{n}{\sigma^2} - 3\sum_{j=1}^{n}\frac{\sigma^2}{\sigma^4} = -\frac{2n}{\sigma^2}$$

Fisher's information matrix and lower bound

$$I(\mu,\sigma)=\begin{bmatrix} n/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{bmatrix} \text{ and } I(\mu,\sigma)^{-1}=\begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/2n \end{bmatrix}$$

As the information matrix depends on the parameter σ we must estimate the matrix. First we estimate μ and σ (for this purpose only the estimation of σ is necessary)

$$\begin{cases} \dfrac{\partial \ell}{\partial \mu}=0 \\ \dfrac{\partial \ell}{\partial \sigma}=0 \end{cases} \Leftrightarrow \begin{cases} \sum_{j=1}^{n}\left(\dfrac{\ln x_j-\mu}{\sigma^2}\right)=0 \\ \sum_{j=1}^{n}\left(-\dfrac{1}{\sigma}+\dfrac{(\ln x_j-\mu)^2}{\sigma^3}\right)=0 \end{cases} \Leftrightarrow \begin{cases} \hat{\mu}=\dfrac{\sum_{j=1}^{n}\ln x_j}{n} \\ \hat{\sigma}=\sqrt{\dfrac{\sum_{j=1}^{n}(\ln x_j-\hat{\mu})^2}{n}} \end{cases}$$

And we will use the asymptotic covariance matrix

$$\hat{\text{var}}(\hat{\mu},\hat{\sigma})=I(\hat{\mu},\hat{\sigma})^{-1}=\begin{bmatrix} \hat{\sigma}^2/n & 0 \\ 0 & \hat{\sigma}^2/2n \end{bmatrix}$$

Now using Data Set B we get (**Note that the number of observations is too low to use an asymptotic approximation**)

```
> # Example 13.9 - solution following the book
> x=c(27,82,115,126,155,161,243,294,340,384,457,680,855,877,974,1193,1340,1884,2558,15743)
> n=length(x);  mu=sum(log(x))/n;  sig2=sum((log(x)-mu)^2)/n; sig=sqrt(sig2)
> mu; sig2; sig
[1] 6.137878
[1] 1.930456
[1] 1.389408
> I=matrix(c(n/sig2,0,0,2*n/sig2),nrow=2,byrow=TRUE)
> I
         [,1]     [,2]
[1,] 10.36025  0.00000
[2,]  0.00000 20.72049
> mat_V=solve(I)
> mat_V
          [,1]      [,2]
[1,] 0.0965228 0.0000000
[2,] 0.0000000 0.0482614
```

Example 13.10 – Estimate the covariance matrix in example 13.9 using the observed information

```
> # example 13.10 - Following the book
> sig3=sig2*sig; sig4=sig2*sig2;
> H=matrix(c(-n/sig2,-(2/sig3)*sum(log(x)-mu),-(2/sig3)*sum(log(x)-mu),
n/sig2-(3/sig4)*sum((log(x)-mu)^2)),nrow=2,byrow=TRUE)
> H
             [,1]             [,2]
[1,] -1.036025e+01  -3.973669e-15
[2,] -3.973669e-15  -2.072049e+01
> matV_H=solve(-H)
> matV_H
             [,1]             [,2]
[1,]  9.652279e-02  -1.851064e-17
[2,] -1.851064e-17   4.826140e-02
>
> #using numerical optimization
>
> minuslogliklognorm=function(theta){
+   -sum(-log(x)-log(theta[2])-0.5*log(2*pi)-0.5*(( (log(x)-theta[1]) / theta[2] )^2))
+   }
```

```
> # Be aware of the starting point!
> # Numerical optimization could be erroneous  (Hessian matrix)
> theta.start=c(6,2)
> out=nlm(minusloglikllognorm,theta.start,hessian=TRUE)
Warning messages:
1: In log(theta[2]) : NaNs produced
2: In nlm(minusloglikllognorm, theta.start, hessian = TRUE) :
  NA/Inf replaced by maximum positive value
> out
$minimum
[1] 157.7139
$estimate
[1] 6.137875 1.389408
$gradient
[1] -2.713500e-06 -2.659279e-07
$hessian
             [,1]          [,2]
[1,] 10.360257841 -0.004526871
[2,] -0.004526871 20.710188098
```

```
$code
[1] 1
$iterations
[1] 7
> HH=out$hessian          # HH is the hessian of minus the log likelihood, i.e. HH is equal to
                          minus the hessian of the likelihood
> solve(HH)               # inverse of HH
             [,1]         [,2]
[1,] 9.652270e-02  2.109811e-05
[2,] 2.109811e-05  4.828542e-02
```

Estimation of a function of the parameters

- What can we do when our interest is about a function of the parameters?

  Example: Assume that our interest, in the last couple of examples, was about the expected value of *X*, that is $E(X) = \exp(\mu + \sigma^2/2)$. The point estimator is easy to obtain, using the invariance property of the mle, and we get $\hat{E}(X) = \exp(\hat{\mu} + \hat{\sigma}^2/2)$. What are the expected value and the (approximate) variance of this estimator?

- **Theorem 13.16** – (**Delta method**) Let $\mathbf{X}_n = (X_{1n}, X_{2n}, \cdots, X_{kn})^T$ be a multidimensional variable of dimension *k* based on a sample of size *n.* Assume that $\mathbf{X}$ is asymptotically normal with mean θ and covariance matrix $\boldsymbol{\Sigma}/n$, where neither θ nor $\boldsymbol{\Sigma}$ depend on n. Let g be a function of k variables that is totally differentiable. Let $G_n = g(X_{1n}, X_{2n}, \cdots, X_{kn})$. Then G_n is asymptotically normal with mean $g(\theta)$ and variance $(\partial \mathbf{g})^T \boldsymbol{\Sigma} (\partial \mathbf{g})/n$, where $\partial \mathbf{g}$ is the vector of the first derivatives, that is, $\partial \mathbf{g} = (\partial g/\partial \theta_1, \partial g/\partial \theta_2, \cdots, \partial g/\partial \theta_k)^T$ and it is to be evaluated at θ, the true parameters of the original random variable.

- **Comments**:
  - There are several presentations of the delta method
  - **When** $k=1$, **the theorem reduces to the following statement**: Let $\hat{\theta}$ be an estimator of θ that has an asymptotic normal distribution with mean θ and variance σ^2/n. Then $g(\hat{\theta})$ has an asymptotic normal distribution with mean $g(\theta)$ and variance $g'(\theta)^2 \times (\sigma^2/n)$.
- **Example 13.12** – Use the delta method the approximate the variance of the mle of the probability that an observation from an exponential distribution exceeds 200. Apply this result to Data Set B.

  As it is well known, the mle estimator of θ is $\hat{\theta}=\overline{X}$ with $E\left(\hat{\theta}\right)=\theta$ and $\text{var}\left(\hat{\theta}\right)=\theta^2/n$.

  We want to estimate $\Pr(X>200)=e^{-200/\theta}=g(\theta)$

  $\hat{\Pr}(X>200)=g(\hat{\theta})=e^{-200/\hat{\theta}}$

  Delta method:

  $$E\left(g(\hat{\theta})\right)\approx g(\theta)=e^{-200/\theta} \quad \text{and} \quad \text{var}\left(g(\hat{\theta})\right)\approx g'(\theta)^2\,\text{var}(\hat{\theta})=\left(\frac{200}{\theta^2}e^{-200/\theta}\right)^2\frac{\theta^2}{n}=\frac{200^2e^{-400/\theta}}{n\theta^2}$$

Application to Data Set B: $n = 20$; Estimate: $\hat{\theta} = 1424.4$

$$\hat{\Pr}(X > 200) = g(\hat{\theta}) = e^{-200/\hat{\theta}} = 0.8690 \qquad \hat{\text{var}}\left(g(\hat{\theta})\right) \approx \frac{200^2\, e^{-400/1424.4}}{20 \times 1424.4^2} = 0.000744402$$

95% Confidence Interval: $0.8690019 \mp 1.645 \times 0.02728373$, that is (0.8241; 0.9139)

- **Example 13.13 –** Construct a 95% confidence interval for the mean of a lognormal population using Data set B. Compare this to the more traditional confidence interval based on the sample mean

  Note that the sample size is too small to use asymptotic results!

  Usual method

  $\bar{x} \pm 1.96 \times s / \sqrt{n}$, i.e. $1424.4 \pm 1.96 \times 3435.04 / \sqrt{20}$, that is (-81.07, 2929.87).

  Note that this interval includes values that are not admissible ($E(X) = g(\theta) > 0$).

  Delta method

$$\theta = \begin{bmatrix} \mu \\ \sigma \end{bmatrix} \qquad g(\mu, \sigma) = \exp(\mu + \sigma^2 / 2) \qquad \partial \mathbf{g} = \begin{bmatrix} \frac{\partial g}{\partial \mu} \\ \frac{\partial g}{\partial \sigma} \end{bmatrix} = \begin{bmatrix} g(\mu, \sigma) \\ \sigma\, g(\mu, \sigma) \end{bmatrix}$$

$$\hat{\theta} = \begin{bmatrix} \hat{\mu} \\ \hat{\sigma} \end{bmatrix} \qquad \text{var}(\hat{\theta}) = \frac{\Sigma}{n} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/(2n) \end{bmatrix} = \frac{\sigma^2}{n}\begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$$ (see example 15.9)

$$\text{var}\left(g(\hat{\theta})\right) \approx (\partial \mathbf{g})^T \boldsymbol{\Sigma} (\partial \mathbf{g})/n = \begin{bmatrix} g(\mu,\sigma) & \sigma\, g(\mu,\sigma) \end{bmatrix}\left(\frac{\sigma^2}{n}\right)\begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} g(\mu,\sigma) \\ \sigma\, g(\mu,\sigma) \end{bmatrix}$$

$$= \left(\frac{\sigma^2}{n}\right)\begin{bmatrix} g(\mu,\sigma) & \sigma\, g(\mu,\sigma)/2 \end{bmatrix}\begin{bmatrix} g(\mu,\sigma) \\ \sigma\, g(\mu,\sigma) \end{bmatrix} = \left(\frac{\sigma^2}{n}\right)\left(g(\mu,\sigma)^2 + \frac{\sigma^2}{2} g(\mu,\sigma)^2\right)$$

$$= \left(\frac{\sigma^2}{n}\right) \times \left(1 + \frac{\sigma^2}{2}\right) \times \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

From example 15.9 we know that the mle estimates are $\hat{\mu} = 6.1379$ and $\hat{\sigma} = 1.3894$. Then

$$\hat{\text{var}}\left(g(\hat{\theta})\right) \approx \left(\frac{\hat{\sigma}^2}{n}\right) \times \left(1 + \frac{\hat{\sigma}^2}{2}\right) \times \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right) = 280444$$

The 95% confidence interval is then $1215.75 \mp 1.96 \times \sqrt{280444}$, that is, (177.79; 2253.71)

NON-NORMAL CONFIDENCE INTERVALS

- In the previous section the confidence intervals are based on 2 assumptions:
  1. The normal distribution is a reasonable approximation for the true distribution of the maximum likelihood estimators (large samples);
  2. When there is more than one parameter, the construction of separate confidence intervals is an acceptable procedure.
- We will see an alternative procedure (the result is still asymptotic) which let us built confidence regions to answer to point 2.
- The new procedure to define confidence intervals is based on the likelihood ratio tests (to be formally presented in chapter 16 of *Loss Models*).
- The idea is to include in the confidence interval (region) the values of θ with a greater likelihood, i.e. our likelihood interval will be defined as $\{\theta : \ell(\theta) \geq c\}$ with $c \leq \ell(\hat{\theta})$ to guarantee that the interval is not empty.

- The question is how to define c in such a way that the procedure produces a $100(1-\alpha)\%$ confidence region?

  A case by case solution can be searched for (and for some cases a solution founded) or we can use an **asymptotic result** using $c = \ell(\hat{\theta}) - 0.5 \times q_{\alpha}$ (be aware of a typo in the book – $c = \ell(\hat{\theta}) - 0.5 \times q_{\alpha/2}$ instead of the correct solution) where q_{α} is the $1-\alpha$ quantile of a chi-square distribution with degrees of freedom equal to the number of estimated parameters.

- **Example 13.14** – Use this method to construct a 95% confidence interval for the parameter of an exponential distribution. Compare the answer to the normal approximation, using Data Set B.

  Exponential distribution: $\ell(\theta) = \sum_{j=1}^{n} \left(-\ln \theta - x_j / \theta \right) = -n \ln \theta - n\bar{x} / \theta$ and $\hat{\theta} = \bar{x}$.

  Data Set B: $n = 20, \bar{x} = 1424.4,$

  Normal approximation

  $$\ell'(\theta) = -\frac{n}{\theta} + \frac{n\bar{x}}{\theta^2}; \quad \ell''(\theta) = \frac{n}{\theta^2} - \frac{2n\bar{x}}{\theta^3}; \; I(\theta) = -E\left(\frac{n}{\theta^2} - \frac{2n\bar{X}}{\theta^3} \right) = -\left(\frac{n}{\theta^2} - \frac{2n}{\theta^2} \right) = \frac{n}{\theta^2}; \quad I(\theta)^{-1} = \frac{\theta^2}{n}$$

  The confidence interval is $\bar{x} \mp 1.96 \times \bar{x} / \sqrt{n}$, that is, (800.129; 2048.67)

Non – normal approximation

$\ell(\hat{\theta}) = -n\ln\bar{x} - n;\ \ q_{0.05} = 3.841$ (we are estimating 1 parameter)

The interval is given by

$$-n\ln\theta - n\bar{x}/\theta \geq -n\ln\bar{x} - n - 0.5\times 3.841 \Leftrightarrow \ln\theta + \bar{x}/\theta \leq \ln\bar{x} + 1 + 1.9205/n$$

which has to be solved numerically ($\ln\bar{x} + 1 + 1.9205/20 = 8.35753$). Using EXCEL's solver we get the interval (946.788; 2285.246)

Comment: To be rigorous we need to prove that the equation $\ln\theta + \bar{x}/\theta = \ln\bar{x} - 1 - 1.9205/n$ has only 2 roots and that the inequality is strict between the roots.

Challenging question: are you able to prove that?

- **Example 13.15** – In example 13.4, the mle for a gamma model for Data Set B were $\hat{\alpha} = 0.55616$ and $\hat{\theta} = 2561.1$. Determine a 95% confidence region for the true values.

  Gamma distribution

  - $\ell(\alpha,\theta) = \sum_{j=1}^{n}\left((\alpha-1)\ln x_j - \frac{x_j}{\theta} - \alpha\ln\theta - \ln\Gamma(\alpha)\right) = (\alpha-1)\sum_{j=1}^{n}\ln x_j - \frac{n\bar{x}}{\theta} - n\alpha\ln\theta - n\ln\Gamma(\alpha)$
  - $\ell(\hat{\alpha},\hat{\theta}) = -162.2934$
  - $c = \ell(\hat{\alpha},\hat{\theta}) - 0.5\times q_{\alpha} = -165.2889$ (using a $\chi^2_{(2)}$)

  We must solve the inequality

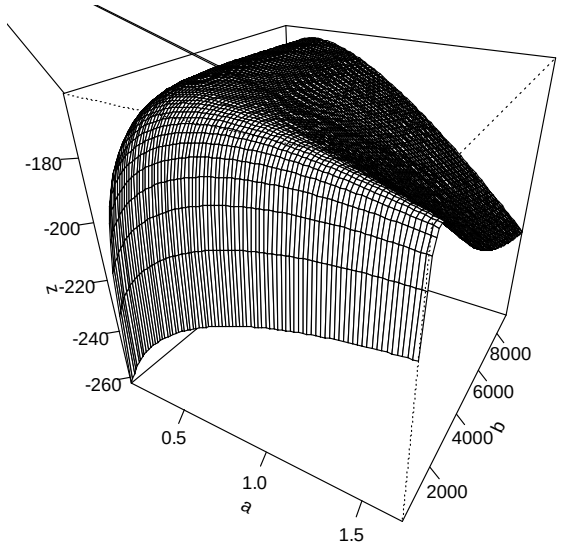
  $$122.7576\times(\alpha-1) - \frac{28488}{\theta} - 20\alpha\ln\theta - 20\ln\Gamma(\alpha) \geq -165.2889$$

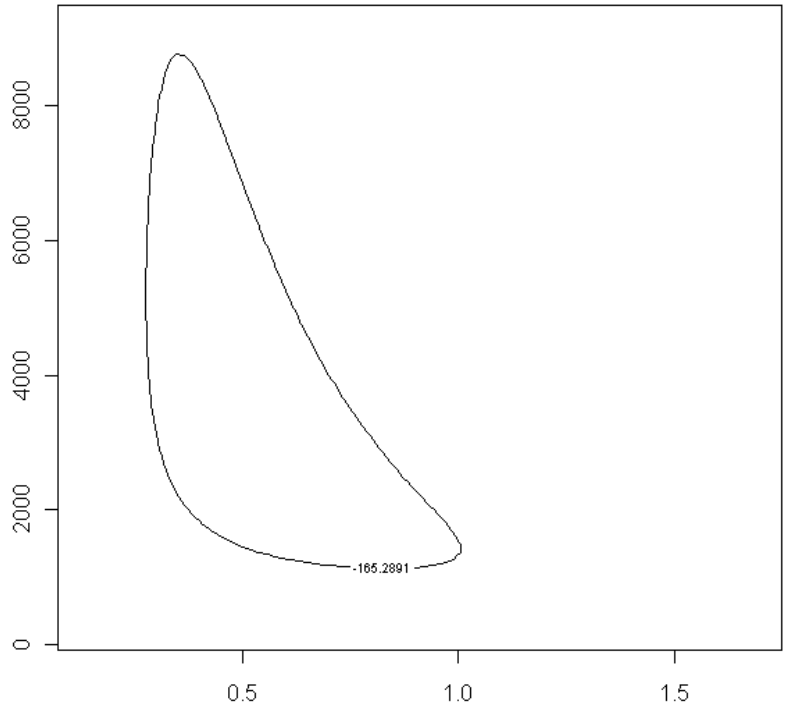
```
> x=c(27,82,115,126,155,161,243,294,340,384,457,680,855,877,974,1193,1340,1884,2558,15743)
>
> minusloglikgamma=function(theta){
+   -sum(dgamma(x,shape=theta[1],scale=theta[2],log=TRUE))
+   }
>
```

```
> loglikgamma=function(a,b){
+   sum(dgamma(x,shape=a,scale=b,log=TRUE))
+   }
>
> theta.start=c(mean(x)*mean(x)/var(x),var(x)/mean(x))
> out=nlm(minusloglikgamma,theta.start,hessian=TRUE)
> out
$minimum
[1] 162.2934

$estimate
[1]    0.556157 2561.146543

$gradient
[1] -6.110668e-06  4.771822e-10
$hessian
             [,1]        [,2]
[1,] 82.442844018 7.808613e-03
[2,]  0.007808613 1.695060e-06
$code
```

```
[1] 1
$iterations
[1] 35

> # Independent confidence intervals
> theta_mv=out$estimate
> invH=solve(-out$hessian) # The function is minus the loglikelihood
> theta_mv_var=-diag(invH)
> linf=theta_mv-1.96*sqrt(theta_mv_var); lsup=theta_mv+1.96*sqrt(theta_mv_var)
> linf; lsup;
[1]   0.2686390 555.9871246
[1]    0.843675 4566.305962
>
> # Confidence region
> q=qchisq(0.05,2,lower.tail=FALSE)
> cc=-out$minimum-0.5*q   # The function is minus the loglikelihood
>
> a=seq(.5*linf[1],2*lsup[1],(2*lsup[1]-.5*linf[1])/81)
> b=seq(.5*linf[2],2*lsup[2],(2*lsup[2]-.5*linf[2])/81)
>
```

```
> z=array(0,dim=c(length(a),length(b)))
> for(i in 1:length(a)) {
+   for(j in 1:length(b)) {
+     z[i,j]=loglikgamma(a[i],b[j])
+     }
+   }
> persp(a,b,z,theta=30,phi=30,ticktype="detailed")
> contour(a,b,z,level=c(cc))
```


-180
-200
-220
-240
-260
z
0.5
1.0
1.5
a
2000
4000
6000
8000
b


8000
6000
4000
2000
0
0.5
1.0
1.5
-165.2891

Appendix 1 – Example 15.5 using R

Example 15.5 – From Data Set C, determine the maximum likelihood estimate of an exponential distribution.

$$F(x|\theta)=1-e^{-x/\theta};\ \ F(c_j|\theta)-F(c_{j-1}|\theta)=e^{-c_{j-1}/\theta}-e^{-c_j/\theta}$$

The log-likelihood is then

$$\ell(\theta)=99\times\ln\left(1-e^{-7500/\theta}\right)+42\times\ln\left(e^{-7500/\theta}-e^{-17500/\theta}\right)+\cdots+3\times\ln\left(e^{-300000/\theta}-0\right)$$

Using Microsoft Excel or another numerical procedure to maximize the log-likelihood we get $\hat{\theta}=29720.77$ and $\ell(\hat{\theta})=-406.03$.

Using R we get

```
> n=c(99,42,29,28,17,9,3)
> linf=c(0,7500,17500,32500,67500,125000,300000)
> lsup=c(7500,17500,32500,67500,125000,300000,Inf)
>
> loglikgroupedexp=function(theta){
+   -sum(n*log(pexp(lsup,rate=1/theta[1])-
pexp(linf,rate=1/theta[1]) ) )
```

```
+    }
> theta.start=c(10000)
> out=nlm(loglikgroupedexp,theta.start)
> out
$minimum
[1] 406.0267
$estimate
[1] 29720.75
$gradient
[1] -1.692637e-09
$code
[1] 1
$iterations
[1] 10
```

Appendix 1a – Example 15.14

Example 15.14 – Use this method to construct a 95% confidence interval for the parameter of an exponential distribution. Compare the answer to the normal approximation, using Data Set B.

Exponential distribution: $\ell(\theta)=\sum_{j=1}^{n}\left(-\ln\theta - x_j/\theta\right)=-n\ln\theta - n\bar{x}/\theta$ and $\hat{\theta}=\bar{x}$.

Data Set B: $n=20, \bar{x}=1424.4$,

Normal approximation

$$\ell'(\theta)=-\frac{n}{\theta}+\frac{n\bar{x}}{\theta^2};\quad \ell''(\theta)=\frac{n}{\theta^2}-\frac{2n\bar{x}}{\theta^3};\quad I(\theta)=-E\left(\frac{n}{\theta^2}-\frac{2n\bar{X}}{\theta^3}\right)=-\left(\frac{n}{\theta^2}-\frac{2n}{\theta^2}\right)=\frac{n}{\theta^2};\quad I(\theta)^{-1}=\frac{\theta^2}{n}$$

The confidence interval is $\bar{x}\mp 1.96\times\bar{x}/\sqrt{n}$, that is, (800.129; 2048.67)

Non – normal approximation

$\ell(\hat{\theta})=-n\ln\bar{x}-n;\quad q_{0.05}=3.841$ (we are estimating 1 parameter)

The interval is given by

$$-n\ln\theta - n\bar{x}/\theta \geq -n\ln\bar{x}-n-0.5\times 3.841 \Leftrightarrow \ln\theta+\bar{x}/\theta \leq \ln\bar{x}+1+1.9205/n$$

which has to be solved numerically ($\ln \bar{x} + 1 + 1.9205/20 = 8.35753$). Using EXCEL's solver we get the interval (946.788; 2285.246)

Comment: To be rigorous we need to prove that the equation $\ln\theta + \bar{x}/\theta = \ln \bar{x} + 1 + 1.9205/n$ has only 2 roots and that the inequality is strict between the roots.

Challenging question: are you able to prove that?

Let us define $\Psi(\theta) = \ln\theta + \bar{x}/\theta$ and calculate $\Psi'(\theta) = \frac{1}{\theta} - \frac{\bar{x}}{\theta^2} = \frac{\theta - \bar{x}}{\theta^2}$. It is clear that $\Psi(\theta)$ is decreasing for $\theta < \bar{x}$ and increasing for $\theta > \bar{x}$. Consequently there is at most 2 points where $\Psi(\theta) = a$, one for $\theta < \bar{x}$ and the other for $\theta > \bar{x}$ assuming that $a > \Psi(\bar{x}) = \ln \bar{x} + 1$ which is always the case ($a = \ln \bar{x} + 1 + 1.9205/n$)

As $\Psi(\theta) \to \infty$ when $\theta \to \infty$ and $\Psi(\theta) \to \infty$ when $\theta \to 0^+$ we can conclude that there is exactly 2 points.

The first limit is obvious and to calculate the second one note that $\Psi(\theta) = \frac{\theta \ln\theta + \bar{x}}{\theta}$ and, as $\theta \ln\theta \to 0$, when $\theta \to 0^+$ we get the result.

Appendix 2 – Proof of theorem 8.7

Assuming that X is a continuous random variable with pdf given by $f(x)$, let us show that

$$E(X-d \mid X>d)=\frac{E(X)-E(X\wedge d)}{1-F(d)}$$

$$E(X)=\int_{-\infty}^{+\infty} x\, f(x)\,dx \qquad E(X\wedge d)=\int_{-\infty}^{d} x\, f(x)\,dx+\int_{d}^{+\infty} d\; f(x)\,dx$$

$$E(X-d \mid X>d)=\int_{d}^{+\infty}(x-d)\, f_{X>d}(x)\,dx=\int_{d}^{+\infty}(x-d)\frac{f(x)}{1-F(d)}dx$$

$$=\frac{1}{1-F(d)}\left(\int_{d}^{+\infty} x\, f(x)\,dx-\int_{d}^{+\infty} d\; f(x)\,dx\right)$$

$$=\frac{1}{1-F(d)}\left(\int_{-\infty}^{d} x\, f(x)\,dx+\int_{d}^{+\infty} x\, f(x)\,dx-\int_{-\infty}^{d} x\, f(x)\,dx-\int_{d}^{+\infty} d\; f(x)\,dx\right)$$

$$=\frac{1}{1-F(d)}\left(\int_{-\infty}^{+\infty} x\, f(x)\,dx-\int_{-\infty}^{d} x\, f(x)\,dx-\int_{d}^{+\infty} d\; f(x)\,dx\right)$$

$$=\frac{1}{1-F(d)}\left(E(X)-E(X\wedge d)\right)$$

Appendix 3 - Approximate second derivative

Basic idea – As it is well known $\frac{df(x)}{dx} = \lim_{\varepsilon \to 0} \frac{f(x+\varepsilon) - f(x)}{\varepsilon}$. An approximation for this expression at point x can be obtained by $\frac{df(x)}{dx} \approx \frac{\Delta f(x)}{dx} \approx \frac{f(x+h/2) - f(x-h/2)}{h}$ using a small value $h > 0$. If we want to approximate $\frac{\partial^2 f(x_1, x_2, x_3)}{\partial x_1 \, \partial x_2}$ we will have

$$\frac{\partial^2 f(x_1, x_2, x_3)}{\partial x_1 \, \partial x_2} = \frac{\partial}{\partial x_2}\left(\frac{\partial f(x_1, x_2, x_3)}{\partial x_1}\right) \approx \frac{\partial}{\partial x_2}\left(\frac{f(x_1 + h_1/2, x_2, x_3) - f(x_1 - h_1/2, x_2, x_3)}{h_1}\right)$$

$$\approx \frac{1}{h_2}\left\{\left(\frac{f(x_1 + h_1/2, x_2 + h_2/2, x_3) - f(x_1 + h_1/2, x_2 - h_2/2, x_3)}{h_1}\right) - \left(\frac{f(x_1 - h_1/2, x_2 + h_2/2, x_3) - f(x_1 - h_1/2, x_2 - h_2/2, x_3)}{h_1}\right)\right\}$$

$$= \frac{1}{h_1 h_2}\left(f(x_1 + h_1/2, x_2 + h_2/2, x_3) - f(x_1 + h_1/2, x_2 - h_2/2, x_3) - f(x_1 - h_1/2, x_2 + h_2/2, x_3) + f(x_1 - h_1/2, x_2 - h_2/2, x_3)\right)$$

$$\frac{\partial^2 f(x_1,x_2,x_3)}{\partial x_1^2} = \frac{\partial}{\partial x_1}\left(\frac{\partial f(x_1,x_2,x_3)}{\partial x_1}\right) \approx \frac{\partial}{\partial x_1}\left(\frac{f(x_1+h_1/2,x_2,x_3)-f(x_1-h_1/2,x_2,x_3)}{h_1}\right)$$

$$\approx \frac{1}{h_1}\left\{\left(\frac{f(x_1+h_1,x_2,x_3)-f(x_1,x_2,x_3)}{h_1}\right)-\right.$$

$$\left.\left(\frac{f(x_1,x_2,x_3)-f(x_1-h_1,x_2-h_2/2,x_3)}{h_1}\right)\right\}$$

$$= \frac{1}{h_1^2}\left(f(x_1+h_1,x_2,x_3)-2\times f(x_1,x_2,x_3)+f(x_1-h_1,x_2,x_3)\right)$$

$$\frac{\partial^2 f(x_1,x_2,x_3)}{\partial x_1\,\partial x_2} = \frac{\partial}{\partial x_1}\left(\frac{\partial f(x_1,x_2,x_3)}{\partial x_2}\right) \approx \frac{\partial}{\partial x_1}\left(\frac{f(x_1,x_2+h_2/2,x_3)-f(x_1,x_2-h_2/2,x_3)}{h_2}\right)$$

$$\approx \frac{1}{h_2}\left\{\left(\frac{f(x_1+h_1/2,x_2+h_2/2,x_3)-f(x_1-h_1/2,x_2+h_2/2,x_3)}{h_1}\right)-\right.$$

$$\left.\left(\frac{f(x_1+h_1/2,x_2-h_2/2,x_3)-f(x_1-h_1/2,x_2-h_2/2,x_3)}{h_1}\right)\right\}$$

$$= \frac{1}{h_1 h_2}\left(f(x_1+h_1/2,x_2+h_2/2,x_3)-f(x_1+h_1/2,x_2-h_2/2,x_3)\right.$$

$$\left.-f(x_1-h_1/2,x_2+h_2/2,x_3)+f(x_1-h_1/2,x_2-h_2/2,x_3)\right)$$

The generalization for the second derivative of a function of *k* variables is straightforward.

Appendix 4 – Example 15.11

Example 15.11 – Repeat example 15.10 using approximate derivatives and assuming that there are 15 significant digits.

```
> # example 15.10 - Following the book
> # we redefine the loglikelihood to get more tractable expressions
> logliklognorm1=function(a,b){-sum(-log(x)-log(b)-0.5*log(2*pi)-0.5* (( (log(x)-a) / b )^2))}
>
> h1=mu/10^5; h2=sig/10^5;
> h1; h2
[1] 6.137878e-05
[1] 1.389408e-05
>
> # Do not forget that the likelihood function gives minus loglikelihood!
> d_mu_2=-(1/h1^2)*(logliklognorm1(mu+h1,sig)-2*logliklognorm1(mu,sig)+
+              logliklognorm1(mu-h1,sig))
> d_sig_2=-(1/h2^2)*(logliklognorm1(mu,sig+h2)-2*logliklognorm1(mu,sig)+
+              logliklognorm1(mu,sig-h2))
>
```

```
> d_mu_sig=-(1/(h1*h2))*(logliklognorm1(mu+h1/2,sig+h2/2)-logliklognorm1(mu+h1/2,sig-h2/2)-
+                   logliklognorm1(mu-h1/2,sig+h2/2)+logliklognorm1(mu-h1/2,sig-h2/2))
>
> d_mu_2; d_sig_2; d_mu_sig;
[1] -10.36024
[1] -20.72028
[1] 0
>
```

Appendix 5 – Likelihood ratio test

This is a common presentation of the likelihood ratio test, different from the generalization presented in the next chapter. A brief presentation can be seen in Wasserman (2004) – "All of statistics, A concise course in statistical inference" – or a more detailed one in Casella and Berger (2002) – "Statistical Inference, 2nd edition". Let us follow Wasserman (2004).

Definition

Consider testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$. The likelihood ratio statistic is

$$\lambda = 2 \times \ln\left(\frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)}\right) = 2 \times \ln\left(\frac{L(\hat{\theta})}{L(\hat{\theta}_0)}\right)$$

Where $\hat{\theta}$ is the mle estimator of θ and $\hat{\theta}_0$ is the mle of $\hat{\theta}$ when θ is restricted to be in Θ_0.

Theorem

Suppose that $\theta=\left(\theta_1,\theta_2,\cdots,\theta_q,\theta_{q+1},\cdots,\theta_r\right)$. Let $\Theta_0=\left\{\theta:\left(\theta_{q+1},\theta_{q+2},\cdots,\theta_r\right)=\left(\theta_{0,q+1},\theta_{0,q+2},\cdots,\theta_{0,r}\right)\right\}$, i.e. the last $r-q$ values are fixed. Let λ be the likelihood ratio test statistic. Under $H_0:\theta\in\Theta_0$, $\lambda\left(X_1,X_2,\cdots,X_n\right)\overset{\circ}{\sim}\chi^2_{(r-q)}$ where $r-q$ is the dimension of Θ minus the dimension of Θ_0. The rejection area will be given by the observed values of λ greater than the adequate percentile of the chi-square distribution, that is, $W=\left\{\lambda:\lambda>q_\alpha\right\}$. Alternatively, the p-value for the test is $\Pr\left(\chi^2_{(r-q)}>\lambda_{obs}\right)$.

Comments

- Remember that the distribution is asymptotic
- For example, if $\theta=\left(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5\right)$ and the null hypothesis is $H_0:\theta_3=\theta_5=0.2$ then the limiting distribution has 3 degrees of freedom.
- A common situation is, for instance, when and the null is $H_0:\theta_1=\theta_{01}\wedge\theta_2=\theta_{02}$ versus $H_0:\theta_1\neq\theta_{01}\vee\theta_2\neq\theta_{02}$. The likelihood ratio is $\lambda=2\times\ln\left(\frac{L(\hat{\theta}_1,\hat{\vartheta}_2)}{L(\theta_{01},\theta_{02})}\right)=2\times\left(\ell(\hat{\theta}_1,\hat{\vartheta}_2)-\ell(\theta_{01},\theta_{02})\right)$

and we use a chi-square distribution with 2 degrees of freedom. Note that we only need to obtain the mle and to calculate the log likelihood at two points.

Appendix 6 – How to define c for likelihood ratio confidence intervals

Using appendix 5 it is straightforward to obtain this value. Remember that we wish to include all the θ such that $\lambda(\theta \mid x_1, x_2, \cdots, x_n) \leq q_\alpha$. But

$$\lambda(\theta \mid x_1, x_2, \cdots, x_n) \leq q_\alpha \Leftrightarrow 2 \times \left(\ell(\hat{\theta}) - \ell(\theta) \right) < q_\alpha \Leftrightarrow \ell(\hat{\theta}) - \ell(\theta) \leq 0.5 \times q_\alpha$$

$$\Leftrightarrow \ell(\theta) \geq \ell(\hat{\theta}) - 0.5 \times q_\alpha$$

Then $c = \ell(\hat{\theta}) - 0.5 \times q_\alpha$