



PARAMETER ESTIMATION

METHOD OF MOMENTS AND PERCENTILE MATCHING

- Random sample (X_1, X_2, \dots, X_n) where all n observations came from the **same parametric distribution**, $F(x|\theta)$. θ is a vector (length p) of unknown parameters.
- Let $\mu'_k(\theta) = E(X^k | \theta)$. Using a random sample of independent observations, the empirical estimate of the k th moment is $\tilde{\mu}'_k = \frac{\sum_{j=1}^n x_j^k}{n}$, 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, $F(\pi_g(\theta) | \theta) = g$. 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 15.1 – A **method of moment** estimate of θ is any solution of the p equations $\mu'_k(\theta) = \tilde{\mu}'_k$, $k = 1, 2, \dots, p$.

- Comments:
 - Although definition 15.1 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 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 15.1 – Use the method of moments to estimate parameters for the exponential, gamma and Pareto distributions for Data Set B from chapter 13.

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 $\tilde{\theta} = 1424.4$

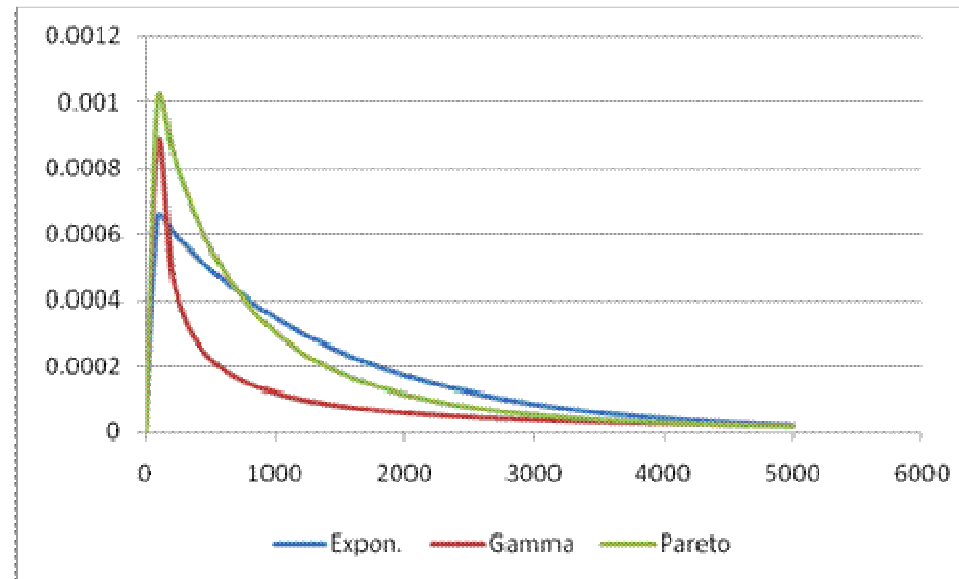
Gamma Distribution: $E(X) = \alpha\theta$, $\alpha > 1$; $E(X^2) = \alpha(\alpha+1)\theta^2$, $\alpha > 2$, then we must solve the system

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

Pareto distribution: $E(X) = \frac{\theta}{\alpha-1}$; $E(X^2) = \frac{2\theta^2}{(\alpha-1)(\alpha-2)}$. The system is

$$\begin{cases} \frac{\theta}{(\alpha-1)} = 1424.4 \\ \frac{2\theta^2}{(\alpha-1)(\alpha-2)} = 13238441.9 \end{cases} \quad \text{and then } \begin{cases} \tilde{\alpha} = 2.442 \\ \tilde{\theta} = 2053.985 \end{cases}$$

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

- 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 15.3 (*type=6* for the *quantile* function)



Definition 15.3 – The smoothed empirical estimate of a percentile is found by

$\hat{\pi}_g = (1-h)x_{(j)} + hx_{(j+1)}$ where $j = \lfloor (n+1)g \rfloor$, $h = (n+1)g - j$, $\lfloor \cdot \rfloor$ indicates the greatest integer function and $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(n)}$ 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 for $1/(n+1) \leq g \leq n/(n+1)$.
 - 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 15.2 – 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: use the median (the parameter is the mean, i.e. a localization parameter). More adequately the idea should lead us to use percentile $1 - e^{-1} \approx 0.6321$ since $\Pr(X < \theta) = 1 - e^{-1}$.

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

We must solve the equation

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

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

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

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

The equations are

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

That is

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

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)

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

Solver Parameters

Set Objective:

To: Max Min Value Of:

By Changing Variable Cells:

Subject to the Constraints:

Make Unconstrained Variables Non-Negative

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



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, \dots, 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, \dots, X_n behind the events A_1, A_2, \dots, 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 15.4** – The **likelihood function** is $L(\theta) = \prod_{j=1}^n \Pr(X_j \in A_j | \theta)$ and the maximum likelihood estimate of θ is the vector that maximizes the likelihood function.
- Comments:
 - **Notation** – Usually the likelihood function is written as $L(\theta | x_1, x_2, \dots, x_n)$. 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, \dots, A_n are assumed independent, the likelihood is the probability, given a particular value of θ , of observing what was observed, since
$$L(\theta) = \prod_{j=1}^n \Pr(X_j \in A_j | \theta) = \Pr(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n | \theta).$$
 - **Theoretical** – When the probabilistic model is continuous and the observed event is a point, $A_j = x_j$, we know that $\Pr(X_j \in A_j | \theta) = 0$ 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 $\Pr(x_j - \varepsilon < X_j < x_j + \varepsilon | \theta)$ by means of $2\varepsilon f(x_j | \theta)$, where $f(x_j | \theta)$ is the density function at x_j . Dropping out the multiplicative constants leads to use the density $f(x_j | \theta)$ 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 $\ell(\theta) = \ln L(\theta) = \sum_{j=1}^n \ln(\Pr(X_j \in A_j | \theta))$ instead of $L(\theta)$ (as the natural logarithm is a strictly increasing function the solution is unchanged).
- $\ln(\Pr(X_j \in A_j | \theta))$ 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, $\ell(\theta) = \sum_{j=1}^n \ln f_{X_j}(x_j | \theta)$.

- **Example 15.4** – 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

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

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

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

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

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

As $\ell''(\theta)\big|_{\theta=\bar{x}} = -n\theta^{-2} < 0$ 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

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

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

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

	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 ; \$B\$1 ; \$B\$2 ; FALSE))						
9	82	-8.18649695							
10	115	-7.914284068							
11	126	-7.84493429							
12	155	-7.69579108							

Solver Parameters

Set Objective:

To: Max Min Value Of:

By Changing Variable Cells:

Subject to the Constraints:

	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.

COMPLETE GROUPED DATA

- We must rectify the likelihood in order to consider the mass probability associated with each group.
- As before, 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 $L(\theta) = \prod_{j=1}^k (F(c_j | \theta) - F(c_{j-1} | \theta))^{n_j}$ and the log likelihood

is $\ell(\theta) = \sum_{j=1}^k n_j \ln(F(c_j | \theta) - F(c_{j-1} | \theta))$

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

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

The log-likelihood is then

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

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 15.6** - 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

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

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

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):

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



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 ($x > 200$, $\alpha > 0$)

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

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

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

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

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

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 15.7 – 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): $(1 - F(5.0)) / (1 - F(0.3))$

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

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

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

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



○ **Comments to Theorem 15.5**

- 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 $\frac{\hat{\theta} - \theta}{I(\theta)^{-1/2}} \sim n(0;1)$

- 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,

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

- 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

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



- 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 $I(\hat{\theta})$. When $I(\hat{\theta})$ is difficult to obtain we can approximate it using the observed information $I(\hat{\theta}) \approx -H(\hat{\theta})$, i.e. using the Hessian matrix of the log likelihood at $\theta = \hat{\theta}$
- **Example 15.9** – 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.

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

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

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

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

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

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

Taking expected values

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

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

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)

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

And we will use the asymptotic covariance matrix

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



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

```
> # Example 15.9 - 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 15.10 – Estimate the covariance matrix in example 15.9 using the observed information

```
> # example 15.10 - 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 with the zeros (Hessian matrix)
> theta.start=c(6,2)
> out=nlm(minuslogliklognorm,theta.start,hessian=TRUE)
Warning messages:
1: In log(theta[2]) : NaNs produced
2: In nlm(minuslogliklognorm, 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 $E(\hat{X}) = \exp(\hat{\mu} + \hat{\sigma}^2 / 2)$. What are the expected value and the (approximate) variance of this estimator?

- **Theorem 15.16 – (Delta method)** Let $\mathbf{X}_n = (X_{1n}, X_{2n}, \dots, X_{kn})^T$ 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 Σ / n , where neither θ nor Σ depend on n . Let g be a function of k variables that is totally differentiable. Let $G_n = g(X_{1n}, X_{2n}, \dots, X_{kn})$. Then G_n is asymptotically normal with mean $g(\theta)$ and variance $(\partial \mathbf{g})^T \Sigma (\partial \mathbf{g}) / n$, where $\partial \mathbf{g}$ is the vector of the first derivatives, that is, $\partial \mathbf{g} = (\partial g / \partial \theta_1, \partial g / \partial \theta_2, \dots, \partial g / \partial \theta_k)^T$ 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 15.12** – Use the delta method to 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 $\hat{\theta} = \bar{X}$ with $E(\hat{\theta}) = \theta$ and $\text{var}(\hat{\theta}) = \theta^2 / n$.

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:

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



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

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

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

- **Example 15.13** – 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

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

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

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

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

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

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?



- The solution is then to define $c = \ell(\hat{\theta}) - 0.5 \times q_\alpha$ (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. **Keep in mind that this result is asymptotic.**
- **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: $\ell(\theta) = \sum_{j=1}^n (-\ln \theta - x_j / \theta) = -n \ln \theta - n \bar{x} / \theta$ 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

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



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 15.15** – In example 15.4, 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

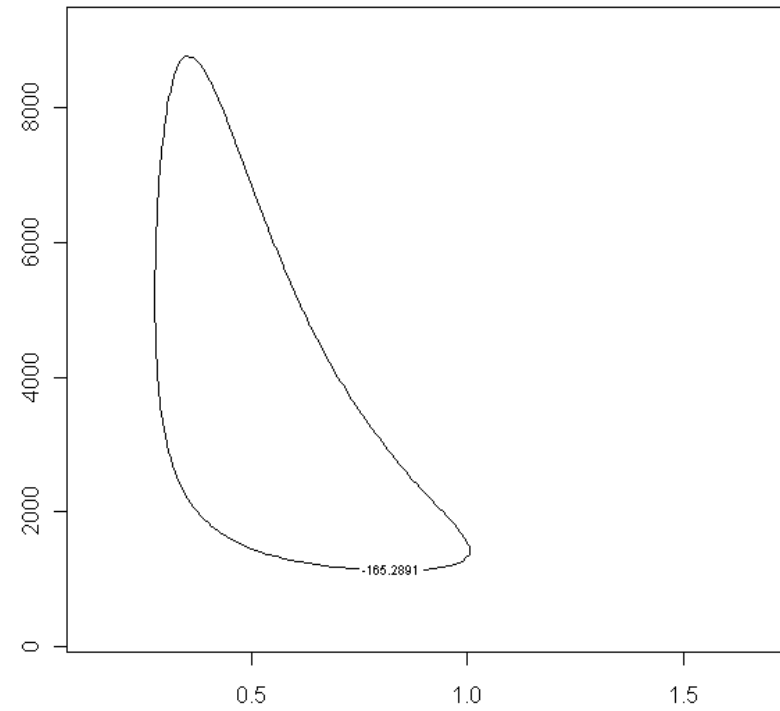
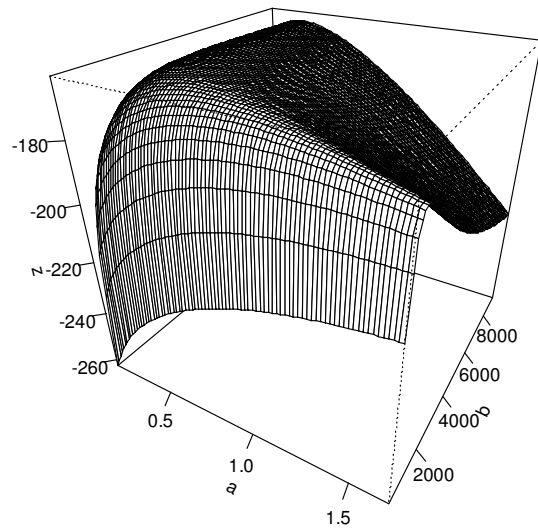
> theta_mv_var=-diag(invH)

> 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))
```





BAYESIAN ESTIMATION

- Until now discussion about estimation has assumed a frequentist approach, namely:
 - The parameter of the population distribution is unknown but fixed (not random);
 - The inference procedures are based not only on the observed sample but also on the population of samples that could have been observed.
- The Bayesian approach assumes that our lack of knowledge about the parameters value should be translated using probability distributions (consequently **unknown parameters are treated as random variables**) and that **only the observed data** (and not the population of samples) **is relevant** to make statistical inference.

Bayesian Inference

$$\left. \begin{array}{l} \text{Model } f_{X|\Theta}(x|\theta) \\ \text{Sample } (x_1, x_2, \dots, x_n) \end{array} \right\} \rightarrow \left. \begin{array}{l} \text{Model distribution } f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) \\ \text{Prior distribution } \pi(\theta) \end{array} \right\} \begin{array}{l} \text{Bayes} \\ \rightarrow \\ \text{Theorem} \end{array} \text{Posterior distribution } \pi_{\Theta|\mathbf{x}}(\theta|\mathbf{x})$$



Model distribution

- **Theoretical model for the population:** for instance Bernoulli, normal,
- **Instead of considering a random sample** $\mathbf{X} = (X_1, X_2, \dots, X_n)$ – usually the sampling process generates i.i.d. observations – we look at the observed sample $\mathbf{x} = (x_1, x_2, \dots, x_n)$
- **Definition 15.9** – The **model distribution** is the probability distribution for the data as collected given a particular value for the parameter. Its pdf is denoted by $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$, where vector notation for \mathbf{x} is used to remind us that all the data appear here. Also note that this is identical to the **likelihood function**, and so that name may also be used at times.
- **Comments:**
 - If the observations are i.i.d., then $L(\theta|\mathbf{x}) = f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = \prod_{i=1}^n f_{\mathbf{X}|\theta}(x_i|\theta)$.
 - Only the likelihood of the **observed sample** is relevant



Prior distribution

- **Definition 15.7** – The **prior distribution** is a probability distribution over the space of possible parameter values. It is denoted by $\pi(\theta)$ and represents our opinion concerning the relative chances that various values of θ are the true value of the parameter.
- **Comments:**
 - The existence of a **prior** for θ (scalar or vector) is the core of **Bayesian inference**. From a theoretical point of view it raises important questions about the concept of probability.
 - From a practical point of view, the determination of the prior is a **major problem** of Bayesian methods. In many situations we have some insights about possible parameter values but the main difficulty is translating this knowledge into a probability distribution.
 - Due to the difficulty of finding a prior, we often use an **improper prior** distribution (vague prior) or we take advantage of **conjugate priors**.
- **Definition 15.8** – An **improper prior distribution** is one for which the probabilities (or probability density function) are nonnegative but their sum (integral) is infinite.



- **Comments:**
 - The improper prior is one possible solution when we have minimal knowledge about the parameter behavior.
 - Universal agreement on the best way to construct a vague (or non-informative) prior does not exist.
 - However the use of the improper prior $\pi(\theta) = 1/\theta, \theta > 0$ as a **vague prior** for a scale parameter is quite consensual.
- **Definition 15.23** – A prior distribution is said to be a conjugate prior distribution for a given model if the resulting posterior distribution is from the same family as the prior (but perhaps with different parameters).



Bayes Theorem: How to obtain the posterior distribution?

- **Definition 15.12** – The **posterior distribution** is the conditional probability distribution of the parameters, given the observed data. It is denoted $\pi_{\Theta|X}(\theta | \mathbf{x})$.
- **Theorem 15.14** – (Part a) The posterior distribution can be computed as

$$\pi_{\Theta|X}(\theta | \mathbf{x}) = \frac{f_{X|\Theta}(\mathbf{x} | \theta) \times \pi(\theta)}{\int f_{X|\Theta}(\mathbf{x} | \theta) \times \pi(\theta) d\theta}$$

- **Comment:**

- This is the central purpose of Bayesian analysis: The posterior distribution tells us how our opinion has changed once we observed the data (compared with the prior).
- In most situations we determine the posterior up to a normalizing constant. This constant can be determined using the condition $\int \pi_{\Theta|X}(\theta | \mathbf{x}) d\theta = 1$ but it is obtained more easily when the posterior belongs to a known family of distributions. In such cases we identify the core of the family and then we get the constant (using for instance Appendix A or B of the book).
- Remember Bayes's formula: Partition $\{A_1, A_2, \dots\}$, event B , then $P(A_i | B) = \frac{P(B | A_i) \times P(A_i)}{\sum_i P(B | A_i) \times P(A_i)}$



The predictive distribution

- **Definition 15.13** – The predictive distribution is the conditional distribution of a new observation y given the data \mathbf{x} . It is denoted $f_{Y|\mathbf{X}}(y|\mathbf{x})$
- **Theorem 15.14** – (Part b) The predictive distribution can be computed as

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \int f_{Y|\theta}(y|\theta) \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x}) d\theta$$

Where $f_{Y|\theta}(y|\theta)$ is the pdf of the new observation, given the parameter value.



Example 15.17 – The following amounts were paid on a hospital liability policy

125 132 141 107 133 319 126 104 145 223.

The amount of a single payment has the single-parameter Pareto distribution with $\theta = 100$ and α unknown. The prior is a gamma distribution with parameters $\alpha = 2$ and $\theta = 1$. Determine all of the relevant Bayesian quantities.

Prior: $\pi(\alpha) = \alpha e^{-\alpha}, \alpha > 0$

This means that $\alpha \sim \gamma(2,1), E(\alpha) = \text{var}(\alpha) = 2$

Likelihood:

$$L(\alpha | \mathbf{x}) = \prod_{i=1}^n f(x_i | \alpha) = \prod_{i=1}^n \frac{\alpha 100^\alpha}{x_i^{\alpha+1}} \quad x_i > 100$$

$$= \alpha^{10} \prod_{i=1}^{10} \frac{100^\alpha}{x_i^\alpha} \prod_{i=1}^{10} \frac{1}{x_i} = \alpha^{10} \times 0.022346^\alpha \times \frac{1}{\prod_{i=1}^{10} x_i} \propto \alpha^{10} \times 0.022346^\alpha$$



Posterior:

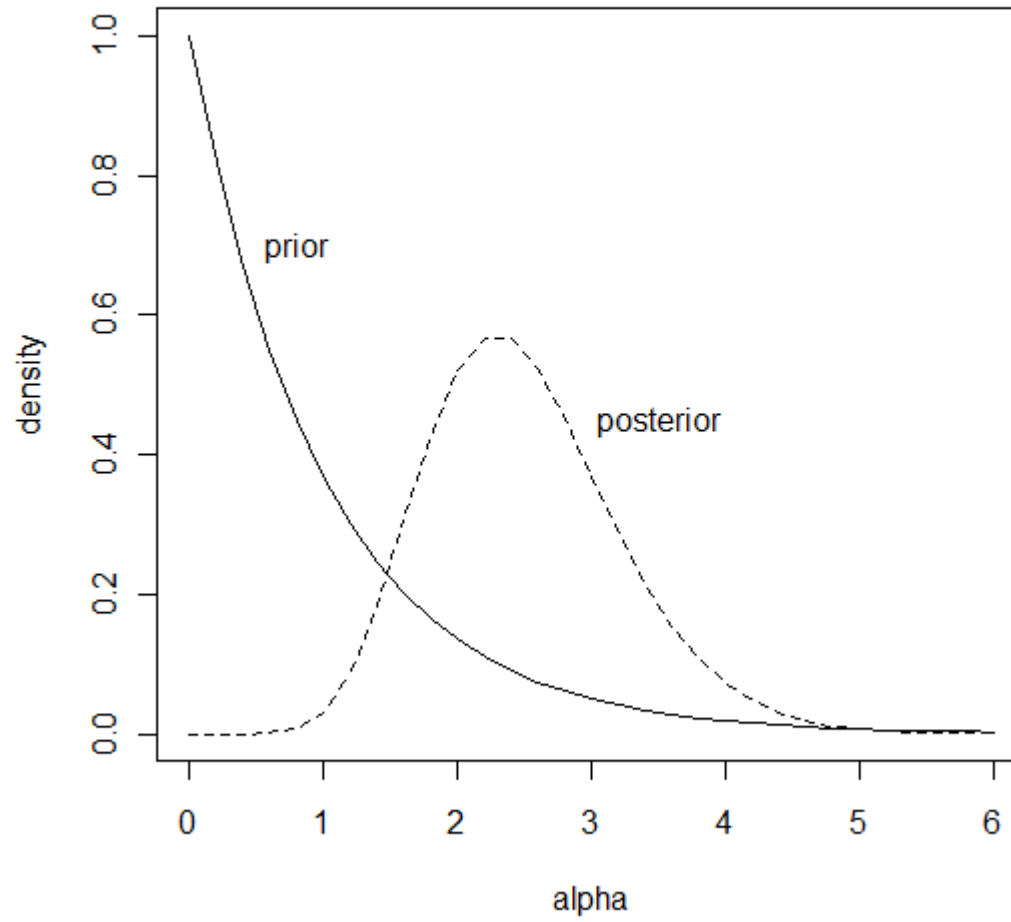
$$\begin{aligned}\pi_{\mathbf{A}|\mathbf{X}}(\alpha | \mathbf{x}) &\propto L(\alpha | \mathbf{x}) \times \pi(\alpha) \propto \alpha^{10} \times 0.022346^\alpha \times \alpha \times e^{-\alpha} \\ &= \alpha^{11} \times \exp(\alpha \ln 0.022346) \times e^{-\alpha} = \alpha^{11} e^{-\alpha(1 - \ln 0.022346)} \\ &= \alpha^{11} e^{-4.80112\alpha} \quad \alpha > 0\end{aligned}$$

We get the *core* of a gamma distribution with parameters 12 and $1/4.80112$ and then we know the normalizing constant which is $4.80112^{12} / \Gamma(12) = 3.757995$. As the posterior belongs to the same family of the prior we said that we are using a conjugate prior for this model.

The point here is that the observed samples leads us to change our believes about α from a $\gamma(2,1)$ to a $\gamma(12,0.20828)$ and now $E(\alpha | \mathbf{x}) = 2.49942$ and $\text{var}(\alpha | \mathbf{x}) = 0.52059$.

We can draw both densities on the same graph to visualize the differences:

```
> x=seq(0,6,by=0.2)
> plot(x,dgamma(x,shape=1,scale=1),type="l",ylab="density",xlab="alpha")
> y=dgamma(x,shape=12,scale=0.208285) # posterior
> lines(x,y,type="l",lty=2)
> text(3.5,0.45,"posterior"); text(0.8,0.7,"prior")
```





Predictive:

$$\begin{aligned}
 f_{Y|X}(y|\mathbf{x}) &= \int_0^\infty f_{Y|A}(y|\alpha) \pi_{A|X}(\alpha|\mathbf{x}) d\alpha = \int_0^\infty \frac{\alpha 100^\alpha}{y^{\alpha+1}} 3.757995 \alpha^{11} e^{-\alpha/0.20828} d\alpha \\
 &= \frac{3.757995}{y} \int_0^\infty \alpha^{12} e^{-\alpha/0.20828 + \alpha \ln 100 - \alpha \ln y} d\alpha = \frac{3.757995}{y} \int_0^\infty \alpha^{12} e^{-\alpha(1/0.20828 - \ln 100 + \ln y)} d\alpha \\
 &= \frac{3.757995}{y} \int_0^\infty \alpha^{12} e^{-\alpha(0.195951 + \ln y)} d\alpha \quad y > 100
 \end{aligned}$$

The integrand is the *core* of a gamma density function with parameters 13 and $1/(0.195951 + \ln y)$.

Then we can use the usual normalizing constant to calculate the integral. We get

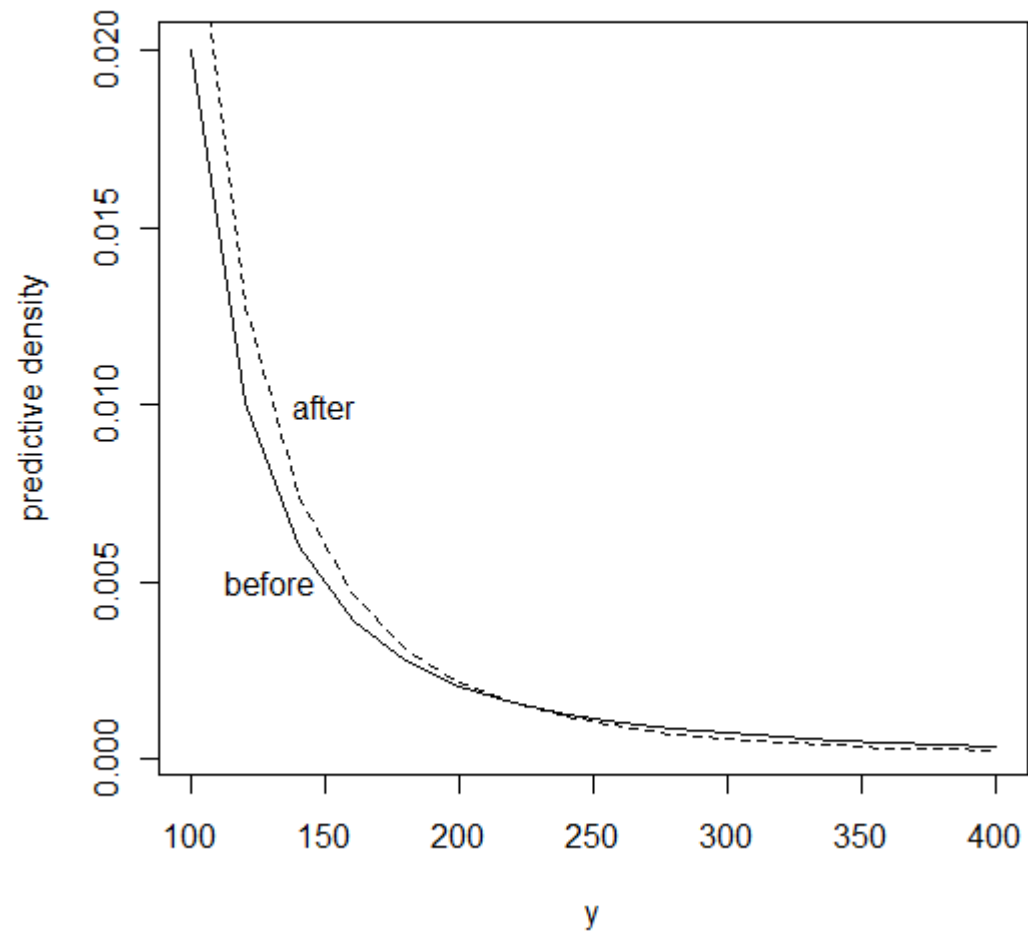
$$f_{Y|X}(y|\mathbf{x}) = \frac{\Gamma(13) \times 3.757995}{(0.195951 + \ln y)^{13} \times y} = \frac{3.757995 \times 12!}{(0.195951 + \ln y)^{13} \times y}, \quad y > 100$$

The density does not look familiar but it can be proved that $\ln Y - \ln 100$ has a Pareto distribution.

```

> y=seq(100, 400, by=20)
> yy1=2*(y^(-1))*((1+log(y/100))^(-3));
> plot(y, yy1, type="l", ylab="predictive density", xlab="y")
> yy2=3.757995*factorial(12)*(y^(-1))*((0.195951+log(y))^(-13));
> lines(y, yy2, type="l", lty=2)
> text(130, 0.005, "before"); text(150, 0.010, "after")

```





From a Bayesian point of view the analysis is complete when we specified the posterior distribution which quantifies our knowledge about θ after the observation of the sample. However, for practical purposes point estimation and/or “confidence interval” are, most of the time, needed. The problem is how to sum up a distribution in one point or using an interval. For point estimation the usual Bayesian solution is to use a loss function.

Bayesian point estimation

- **Definition 15.15** – A **loss function** $l_j(\hat{\theta}_j, \theta_j)$ describes the penalty paid by the investigator when $\hat{\theta}_j$ is the estimate and θ_j is the true value of the j th parameter.
- **Comment:** The loss function is random since it depends on θ_j .
- **Definition 15.16** – The **Bayes estimate** for a given loss function is the one that minimizes the expected loss, given the posterior distribution of the parameter in question.
- **Definition 15.17** – For **squared-error** loss, the loss function is (all subscripts are dropped for convenience) $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. For **absolute loss** it is $l(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$. For **zero-one loss** it is $l(\hat{\theta}, \theta) = 0$ if $\hat{\theta} = \theta$ and 1 otherwise.
- **Comment:** Strictly speaking, Definition 15.17 defines the loss functions up to a multiplicative constant.



- **Theorem 15.18** – For squared-loss, the Bayes estimate is the mean of the posterior distribution; for absolute loss it is the median and for zero-one loss it is the mode.

Challenging question: Prove the theorem for the squared loss function (easier) and other functions.

- **Comments:**
 - There is no guarantee that the posterior's mean exists (or the mode) or that the median is unique.
 - When no otherwise specified, the term Bayes estimate refers to the posterior mean (squared-loss function).
- **Example 15.18** – Determine the three estimates of α (example 15.17 continued)
The posterior is a gamma distribution with parameters 12 and 0.20828. Then $E(\alpha | \mathbf{x}) = 2.49942$, the mode is $11 \times 0.20828 = 2.29132$ and the median has to be determined numerically (2.430342).
- Sometimes the expected value of the predictive distribution is of interest. We can calculate it using the predictive and it can be shown that $E(Y | \mathbf{x}) = \int y f_{Y|\mathbf{x}}(y | \mathbf{x}) dy = \int \pi_{\Theta|\mathbf{x}}(\theta | \mathbf{x}) E(Y | \theta) d\theta$ (see *Loss Models*).



Bayesian HPD credibility set

- **Definition 15.19** – The points $a < b$ define a $100 \times (1 - \alpha)\%$ **credibility interval** for θ_j , provided that $\Pr(a \leq \theta_j \leq b) \geq 1 - \alpha$.
- **Comments:**
 - The term credibility is used to underline the differences between the frequentist (confidence interval) and the Bayesian approaches. This term has no relation with credibility theory.
 - The inequality is due to discrete distribution
 - Definition 15.19 does not produce a unique solution for the credibility interval. Usually we look for the shortest interval.

- **Theorem 15.20** – If the posterior random variable $\theta_j | \mathbf{x}$ is continuous and unimodal, then the $100 \times (1 - \alpha)$ credibility interval with the smallest width, $b - a$, is the unique solution to

$$\int_a^b \pi_{\theta_j | \mathbf{x}}(\theta_j | \mathbf{x}) d\theta = 1 - \alpha \text{ and } \pi_{\theta_j | \mathbf{x}}(b | \mathbf{x}) = \pi_{\theta_j | \mathbf{x}}(a | \mathbf{x})$$

This interval is a special case of a highest posterior density (HPD).

- **Comment:** The posterior cannot have any local maximum except the mode which is unique.

- **Example 15.20** – Determine the shortest 95% credibility interval for the parameter α (example 15.17 continued)

Let us use EXCEL's solver to determine the interval

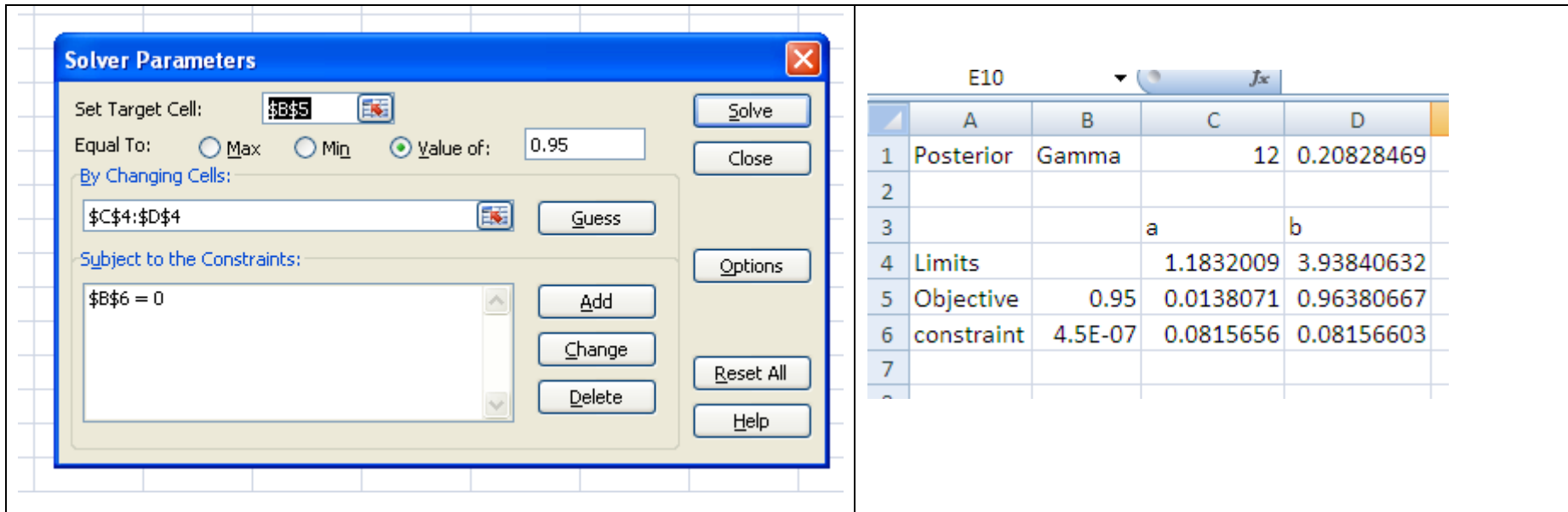
	A	B	C	D
1	Posterior	Gamma	12	0.20828469
2				
3			a	b
4	Limits		0.5	5
5	Objective	0.99748	8.473E-06	0.99748392
6	constraint	0.00672	0.0001664	0.00688849
7				
8				

In cells C4 and D4 we put two initial values for the limits of the interval

In cells C5 and D5 we calculate the value of the distribution function at points a and b respectively. Cell B5 contains the probability of the interval.

In cells C6 and D6 we calculate the value of the posterior density function at points a and b respectively. Cell B6 contains the difference between the density at point b and the density at point a .

Now we want to get the value $1 - \alpha = 0.95$ in cell (5,B) and the value 0 in cell (6,B).



The image shows the Excel Solver Parameters dialog box on the left and a data table on the right. The dialog box is configured with the following settings:

- Set Target Cell: $\$B\5
- Equal To: Max Min Value of: 0.95
- By Changing Cells: $\$C\$4:\$D\4
- Subject to the Constraints: $\$B\$6 = 0$

The data table on the right is as follows:

	A	B	C	D
1	Posterior	Gamma	12	0.20828469
2				
3			a	b
4	Limits		1.1832009	3.93840632
5	Objective	0.95	0.0138071	0.96380667
6	constraint	4.5E-07	0.0815656	0.08156603
7				

The credibility interval is then (1.1832009; 3.93840632)

To get an approximate solution we place a probability of 0.025 at each end \rightarrow (1.29148; 4.09947).

- **Definition 15.21** – For any posterior distribution the $100 \times (1 - \alpha) \%$ HPD credibility set is the set of parameter values C such that $\Pr(\theta_j \in C) \geq 1 - \alpha$ and $C = \{\theta_j : \pi_{\theta_j | \mathbf{x}}(\theta_j | \mathbf{x}) \geq c\}$ for some c , where c is the largest value for which the previous inequality holds.
- Sometimes computing posterior probabilities is difficult but computing posterior moments is easier. We can them using the Bayesian central limit theorem.



- **Theorem 15.22 – Bayesian central limit theorem** – If $\pi(\theta)$ and $f_{\mathbf{X}|\Theta}(x|\theta)$ are both twice differentiable in the elements of θ and other commonly satisfied assumptions hold, then the posterior distribution of Θ given $\mathbf{X} = \mathbf{x}$ is asymptotically normal.
- **Comment:** The “commonly satisfied assumptions” are like those presented with Theorem 15.5
- **Example 15.21** – Construct a 95% credibility interval for α using the Bayesian central limit theorem (example 15.17 continued).

The posterior is $\gamma(12, 0.20828)$ and then $E(\alpha|\mathbf{x}) = 2.49942$ and $\text{var}(\alpha|\mathbf{x}) = 0.52059$. The credibility interval is then $2.49942 \pm 1.96 \times \sqrt{0.52059}$, i.e. (1.085238, 3.913594). Note that the method is not appropriate for this example as the sample size is far from large.

$$L(\theta|\mathbf{x}) = \prod_{j=1}^n f_{\mathbf{X}_j|\Theta}(x_j|\theta) = \prod_{j=1}^n \frac{p(x_j) e^{r(\theta)x_j}}{q(\theta)} = \frac{e^{r(\theta)\sum x_j} \prod_{j=1}^n p(x_j)}{q(\theta)^n}$$

$$\text{With } k^* = k + n \text{ and } \mu^* = \frac{\sum x_j + \mu k}{k + n} = \frac{k}{k + n} \times \mu + \frac{n}{k + n} \times \bar{x}.$$

- **Computational issues:** Bayesian analysis proceeds by taking integrals (or sums) and most of the time numerical integration is needed.