Statistical Inference Package SIP

Esa Uusipaikka

University of Turku
Department of Statistics
Assistentinkatu 7
FI−20014 TURUN YLIOPISTO
Finland
esa.uusipaikka@utu.fi

Statistical inference is always based on observations from the phenomenon under consideration. The set of observations is the first necessary component of statistical evidence on which the inference relies. The second necessary component is a statistical model. The statistical model is based on the assumption that the observations contain random variation, that is, can be considered to have arisen from some probability distribution.

Statistical inference concerns some characteristic or characteristics of the phenomena from which the observations have arisen. The characteristics of the phenomena under consideration are some functions of the unknown parameter of the statistical model and are called the parameter functions of interest.

Statistical inference consists of statements concerning the unknown value(s) of the interest function(s). Statistical inference differs from other possible modes of inferences in that it always gives measures of uncertainties of the statements made.

The *Statistical Inference Package SIP* package has been developed mainly to assist applied statisticians to make inferences on a real valued parameter function of interest. The package contains functions for calculation of the so−called profile likelihood based intervals and their uncertainties for the unknown value of the real−valued interest function. It includes also functions to make so−called likelihood ratio tests concerning some given hypothesis about the parameters of the statistical model.

■ Introduction

Statistical Inference Package SIP makes it easy for the user to do classical likelihood based statistical inference. It contains procedures for maximum likelihood estimation, likelihood ratio tests of general hypotheses concerning parameters, and profile likelihood based confidence intervals for general interest functions of parameters.

Statistical Inference Package SIP contains large collection of discrete and absolutely continuous univariate distributions and also multivariate distributions. It gives user possibility to form complicated models from the simpler ones.

Statistical Inference Package SIP contains many sophisticated statistical models such as univariate/multivariate linear/non−linear regression model,logistic regression models, Poisson regression models, multinomial regression models etc.

Statistical Inference Package SIP uses a new method for calculation of profile likelihood based confidence intervals for general parameter functions of interest in general parametric statistical models.

Statistical Inference Package SIP gives in addition to the statistical analysis procedures easy access to the powerful tools in MATHEMATICA® for doing mathematics, graphics, programming, and presentation.

■ What is statistical inference?

The fundamental problem towards which the study of statistics is addressed, is that of inference. Some data are observed and we wish to make statements, inferences, about one or more unknown features of the physical system which gave rise to these data ([1], p. 1)

□ Data and it's statistical model

Statistical inference is always based on *observations* from the phenomenon under consideration. Those observations might be the result of a designed experiment or an observational study. The set of observations is the first necessary component of *statistical evidence* on which the inference relies. The second necessary component is a *statistical model*. The statistical model is based on the assumption that the observations contain *random variation*, that is, can be considered to have arisen from some probability distribution. The collection of plausible probability distributions or models forms the statistical model.

As an example consider data consisting of waiting times (in minutes) of 299 consecutive eruptions of the Old Faithful geyser in Yellowstone National Park ([2], p. 350).

```
dataFile = ToFileName[Directory[],"Eruptions.dat"];
dataOnEruptionWaitingTimes = Import[dataFile,"List"]
```

```
{80, 71, 57, 80, 75, 77, 60, 86, 77, 56, 81, 50, 89, 54, 90,
 73, 60, 83, 65, 82, 84, 54, 85, 58, 79, 57, 88, 68, 76, 78,
 74, 85, 75, 65, 76, 58, 91, 50, 87, 48, 93, 54, 86, 53, 78,
 52, 83, 60, 87, 49, 80, 60, 92, 43, 89, 60, 84, 69, 74, 71,
 108, 50, 77, 57, 80, 61, 82, 48, 81, 73, 62, 79, 54, 80,
 73, 81, 62, 81, 71, 79, 81, 74, 59, 81, 66, 87, 53, 80, 50,
 87, 51, 82, 58, 81, 49, 92, 50, 88, 62, 93, 56, 89, 51, 79,
 58, 82, 52, 88, 52, 78, 69, 75, 77, 53, 80, 55, 87, 53, 85,
 61, 93, 54, 76, 80, 81, 59, 86, 78, 71, 77, 76, 94, 75, 50,
 83, 82, 72, 77, 75, 65, 79, 72, 78, 77, 79, 75, 78, 64, 80,
 49, 88, 54, 86, 51, 96, 50, 80, 78, 81, 72, 75, 78, 87, 69,
 55, 83, 49, 82, 57, 84, 57, 84, 73, 78, 57, 79, 57, 90, 62,
 87, 78, 52, 98, 48, 78, 79, 65, 84, 50, 83, 60, 80, 50, 88,
 50, 84, 74, 76, 65, 89, 49, 88, 51, 78, 85, 65, 75, 77, 69,
 92, 68, 87, 61, 81, 55, 93, 53, 84, 70, 73, 93, 50, 87, 77,
 74, 72, 82, 74, 80, 49, 91, 53, 86, 49, 79, 89, 87, 76, 59,
 80, 89, 45, 93, 72, 71, 54, 79, 74, 65, 78, 57, 87, 72, 84,
 47, 84, 57, 87, 68, 86, 75, 73, 53, 82, 93, 77, 54, 96, 48,
 89, 63, 84, 76, 62, 83, 50, 85, 78, 78, 81, 78, 76, 74, 81,
 66, 84, 48, 93, 47, 87, 51, 78, 54, 87, 52, 85, 58, 88, 79}
```

We might assume that these waiting times form a random sample from some exponential distribution. Thus after loading the *Statistical Inference Package SIP*

```
<<StatisticalInference`
```

we define the following statistical model

$\mathcal{M}$ = SamplingModel[ExponentialModel[μ],299]

```
--- SamplingModel ---
```

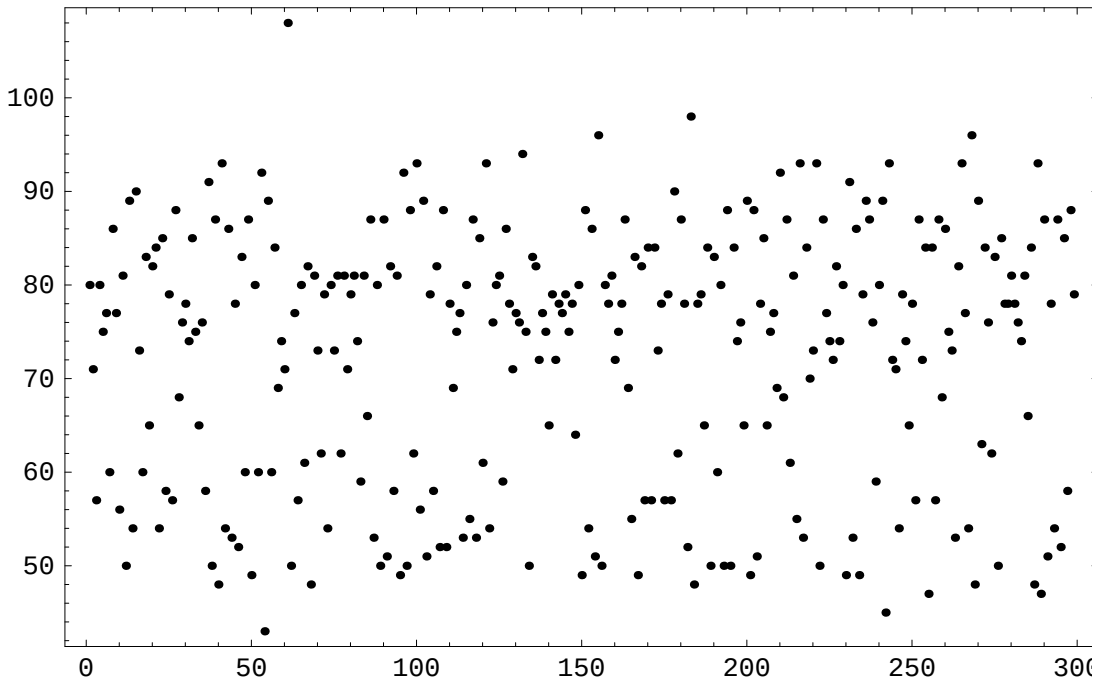
where μ denotes the unknown mean of the exponential distribution.

□ Interest functions

Statistical inference concerns some characteristic or characteristics of the phenomena from which the observations have arisen. *Statistical Inference Package SIP* can cope with situations where the statistical model for the observations is a so−called *parametric model*, that is, where the statistical model as a collection of probability models can be indexed by a real finite dimensional vector. The set of possible values of the index vector is called the parameter space and the generic element of the parameter space is called the *parameter vector* of the statistical model. Sometimes the *parameter* of the statistical model has more complicated structure, but there always exists a one−to−one correspondence between the parameter and the parameter vector. The characteristics of the phenomena under consideration are some functions of the parameter and are called the *parameter functions of interest.*

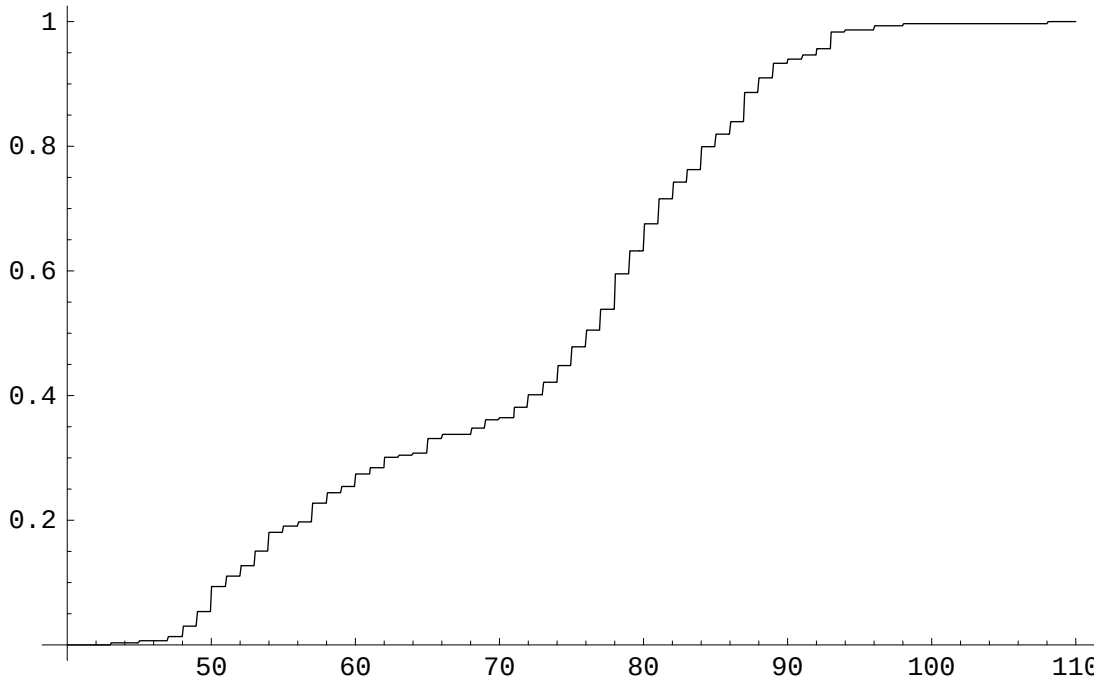
The following plot is a list plot of waiting times.

ListPlot[*dataOnEruptionWaitingTimes*,Axes−>False,Frame−>True,PlotRange−>All];



In the list plot there is an indication of a gap in the vicinity of the waiting time 65, which is confirmed by the empirical cumulative distribution function.

Plot[Evaluate[DistributionFunction[EmpiricalModel[*dataOnEruptionWaitingTimes*],y,{ }]],{y,40,110}];



Thus a more plausible model for the waiting times is a sampling model from a mixture of two distributions, for example, two normal distributions with unknown means and standard deviations. The following expression defines the statistical model

$\mathcal{M}$ = SamplingModel[MixtureModel[NormalModel[μ,σ],2],299]

```
--- SamplingModel ---
```

with parameter of the form $\{\{p, 1-p\}, \{\mu_1, \sigma_1\}, \{\mu_2, \sigma_2\}\}$, where p denotes the unknown mixing probability, $\{\mu_1, \sigma_1\}$ the unknown mean and standard deviation of the first component of the mixture, and $\{\mu_2, \sigma_2\}$ the unknown mean and standard deviation of the second component of the mixture. Thus the parameter vector of the model consists five unknown real numbers.

In this example we might be interested in the probability p of the first component of the mixture, in the odds $\frac{p}{1-p}$, difference of the means $\mu_1 - \mu_2$, ratio of the variances $\frac{\sigma_1^2}{\sigma_2^2}$ etc.

□ Statements and their uncertainties

Statistical inference consists of *statements* concerning the unknown value(s) of the interest function(s). The statement might assert that the unknown values are equal to some given known values, or that the unknown values belong to a given subset or collection of subsets of their plausible values under the statistical model. Statistical inference differs from other possible modes of inferences in that it always gives *measures of uncertainties* of the statements made. These measures of uncertainties are a necessary component of statistical inference and arise because of the assumed randomness contained in the observations. They describe the reliability of the inference.

In case of waiting times of eruptions of Old Faithful the statement might be that the mixing probability p is less than 1/2 or that the difference of the means belongs to some given interval of the real line. In the following *Statistical Inference Package SIP* functions that can be used to generate statements about the unknown parameter and their uncertainties are considered.

■ Statistical inference in *SIP*

□ Calculation of uncertainty in inference

The uncertainties of statements made in the statistical inference are derived from the probability distributions of the statistical model. In case of *likelihood based inference*, on which *Statistical Inference Package SIP* relies, these uncertainties come from certain asymptotic distributional results of very general nature. From this generality follows that the same method for deriving inferential statements and their uncertainties can be applied in a very large collection of statistical models.

□ Profile likelihood based confidence interval

Consider first the case where interest of the analysis concentrates on some real valued function of the parameter. Then a region, usually an interval, of the real line is sought such that the observations with their statistical model support points inside the interval more than points outside the interval. Thus the result of the statistical inference is a statement that the unknown value of the interest function belongs to an interval of the real line and a measure of uncertainty of this statement. The *Statistical Inference Package SIP* function ProfileInterval calculates so−called *profile likelihood based confidence intervals*, which provide just this kind statistical inference.

In case of waiting times of eruptions of Old Faithful the 95%−level profile likelihood based confidence interval for the the probability of the first component of the mixture is given by the following expressions.

```
ℐ=ProfileInterval[ℳ,dataOnEruptionWaitingTimes,p,{{p,1−p},{{μ1,σ1},{μ2,σ2}}}]
```

```
--- ProfileIntervalModel ---    Convergence: True
```

```
ConfidenceLimits[ℐ]
```

```
{0.249478, 0.368547}
```

This means that the observations and their statistical model, in other words the statistical evidence, supports the statement that the unknown value of the probability of the first component of the mixture belongs to the interval (0.249,0.369) and the uncertainty of this statement is 0.05 (= 1 − 0.95). Thus we can with a rather small 'risk' or with a rather large 'reliability' state that p is an element of the interval (0.249,0.369). Let us check this crudely by assuming that all observations less than 65 minutes come from the first component of the mixture and the rest from the second component of the mixture.

```
Length[Select[dataOnEruptionWaitingTimes, (# < 65) &]] /
  Length[dataOnEruptionWaitingTimes] // N
```

```
0.307692
```

Clearly this crude way of assessing the probability p is successful in this case only because of the nature of data and never should replace the calculation of the confidence interval, but

in this case it in a simple way shows the agreement between the data and the calculated confidence interval.

Note that in the function ProfileInterval the expressions p, $\mu 1$, $\sigma 1$, $\mu 2$, and $\sigma 2$ must be symbols and subscripted forms cannot be used unless these subsripted forms have been declared to be treated as symbols with the Symbolize function from the *Notation* package.

This calculates 95%−level profile likelihood based confidence interval for the probability of the ratio of variances.

```
ℐ = ProfileInterval[ℳ, dataOnEruptionWaitingTimes,
  σ1^2/σ2^2, {{p, 1 - p}, {{μ1, σ1}, {μ2, σ2}}}]
--- ProfileIntervalModel ---    Convergence: True
```

```
ConfidenceLimits[ℐ]
```

```
{0.243271, 0.763639}
```

□ Likelihood ratio test

Sometimes the researcher has theoretical reasons to believe that the value(s) of some given interest function(s) have given properties. To assure that the researcher has enough evidence for her belief before she maintains that her belief is 'true' the researcher is often obliged to perform a *significance test*. The significance test involves a *statistical hypothesis*, which is an assumption about the model parameter that is an 'opposite' of the belief of the researcher such that if the assumption is 'true', then the belief is 'false'. The *Statistical Inference Package SIP* function LRTest calculates *likelihood ratio tests*, which show how much the statistcal evidence, that is, the obervations and their statistical model support the research hypothesis.

In case of the waiting times of eruptions of Old Faithful assume that there are some theoretical reasons to maintain that the variances of the components of the mixture are not equal. To study, is there enough evidence for this statement in the observations and in their statistical model, first a statistical hypothesis must be constructed. In this case the natural statistical hypothesis is the assumption, that the variances are equal, that is, $H: \sigma_1^2 = \sigma_2^2$ or equivalently $H: \sigma_1 = \sigma_2$.

This calculates the likelihood ratio test for the hypothesis of equal variances.

```
𝒯 = LRTest[ℳ, dataOnEruptionWaitingTimes,
  {p, μ1, μ2, σ}, {{p, 1 - p}, {{μ1, σ}, {μ2, σ}}}]
--- LRTestModel ---
```

The observed significance level is

```
ObservedSignificanceLevel[𝒯]
```

```
0.00384683
```

The result tells us that the uncertainty of the statement "the variances of the mixture components are different" has uncertainty 0.0038 and so there is very strong evidence for the fact that the variances are not equal.

In fact a better way to study the relation of the variances is to calculate profile likelihood based confidence interval of their unknown ratio. This was done in the previous subsection and the 95%−level profile likelihood based confidence interval for ratio is (0.243,0.764), meaning that with small risk we can state that the variance of the first component of the mixture is smaller than the variance of the second component and that approximately

$\frac{1}{4}\,\sigma_2^2 < \sigma_1^2 < \frac{3}{4}\,\sigma_2^2$ with small uncertainty. Clearly this inference is in agreement with qualitative impression we get from the list plot of waiting times, but the amount of the difference is hard to jugde from the plot alone.

□ Maximum likelihood estimate

Because the calculation of profile likelihood based confidence intervals and likelihood ratio tests requires the calculation (*restricted*) *maximum likelihood estimate* of the parameter of the statistical model, *Statistical Inference Package SIP* contains a function MLEFit, which calculates so−called *fitted model.* The fitted model includes that value of the parameter which is most supported by the statistical evidenc. This value of the parameter is the point in the parameter space that maximizes the observed likelihood function, that is, the probability of the actual observations considered as a function of the parameter. It is, however, important to realise that the maximum likelihood estimate alone does not produce statistical inference, because it does not contain any measure of the uncertainty. Thus maximum likelihood estimate should never be used without giving proper statistical inference in the form of confidence region (interval) or significance test.

In the case of waiting times of eruptions of Old Faithful the likelihood function calculated from the first ten observations has the form

```
LikelihoodFunction[SamplingModel[MixtureModel[NormalModel[μ, σ], 2], 10],
 Take[dataOnEruptionWaitingTimes, 10], {{p, 1 - p}, {{μ1, σ1}, {μ2, σ2}}}]
```

$$\left(\frac{e^{-\frac{(56-\mu 1)^2}{2\sigma 1^2}}\,p}{\sqrt{2\pi}\,\sigma 1}+\frac{e^{-\frac{(56-\mu 2)^2}{2\sigma 2^2}}\,(1-p)}{\sqrt{2\pi}\,\sigma 2}\right)\left(\frac{e^{-\frac{(57-\mu 1)^2}{2\sigma 1^2}}\,p}{\sqrt{2\pi}\,\sigma 1}+\frac{e^{-\frac{(57-\mu 2)^2}{2\sigma 2^2}}\,(1-p)}{\sqrt{2\pi}\,\sigma 2}\right)$$
$$\left(\frac{e^{-\frac{(60-\mu 1)^2}{2\sigma 1^2}}\,p}{\sqrt{2\pi}\,\sigma 1}+\frac{e^{-\frac{(60-\mu 2)^2}{2\sigma 2^2}}\,(1-p)}{\sqrt{2\pi}\,\sigma 2}\right)\left(\frac{e^{-\frac{(71-\mu 1)^2}{2\sigma 1^2}}\,p}{\sqrt{2\pi}\,\sigma 1}+\frac{e^{-\frac{(71-\mu 2)^2}{2\sigma 2^2}}\,(1-p)}{\sqrt{2\pi}\,\sigma 2}\right)$$
$$\left(\frac{e^{-\frac{(75-\mu 1)^2}{2\sigma 1^2}}\,p}{\sqrt{2\pi}\,\sigma 1}+\frac{e^{-\frac{(75-\mu 2)^2}{2\sigma 2^2}}\,(1-p)}{\sqrt{2\pi}\,\sigma 2}\right)\left(\frac{e^{-\frac{(77-\mu 1)^2}{2\sigma 1^2}}\,p}{\sqrt{2\pi}\,\sigma 1}+\frac{e^{-\frac{(77-\mu 2)^2}{2\sigma 2^2}}\,(1-p)}{\sqrt{2\pi}\,\sigma 2}\right)^2$$
$$\left(\frac{e^{-\frac{(80-\mu 1)^2}{2\sigma 1^2}}\,p}{\sqrt{2\pi}\,\sigma 1}+\frac{e^{-\frac{(80-\mu 2)^2}{2\sigma 2^2}}\,(1-p)}{\sqrt{2\pi}\,\sigma 2}\right)^2\left(\frac{e^{-\frac{(86-\mu 1)^2}{2\sigma 1^2}}\,p}{\sqrt{2\pi}\,\sigma 1}+\frac{e^{-\frac{(86-\mu 2)^2}{2\sigma 2^2}}\,(1-p)}{\sqrt{2\pi}\,\sigma 2}\right)$$

and the fitted model and maximum likelihood estimate from all the observations is given by the following expressions.

```
ℱ = MLEFit[ℳ, dataOnEruptionWaitingTimes]
```

```
--- FittedModel ---    Convergence: True
```

The maximum likelihod estimate is

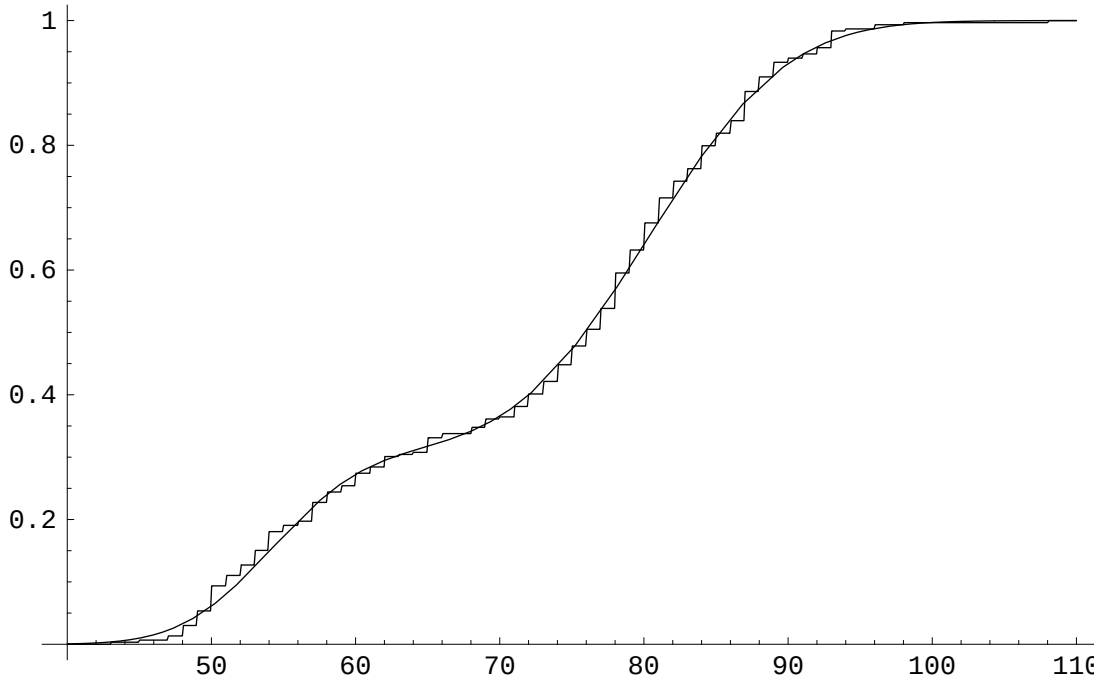
```
MatrixForm /@ MLParameterEstimate[ℱ] // N
```

$$\left\{\begin{pmatrix}0.308046\\0.691954\end{pmatrix},\ \begin{pmatrix}54.2254 & 4.97619\\80.3679 & 7.50195\end{pmatrix}\right\}$$

This means that the most supported value for the probability of the first component of the mixture is 0.31 and the most supported values for the means and standard deviations of the mixture components are (54.2, 5.00) for the first component and (80.4, 7.50) for the second component, respectively.

This shows the fitted and empirical cumulative distribution functions in same graph.

```
Plot[Evaluate[{DistributionFunction[
    EmpiricalModel[dataOnEruptionWaitingTimes], y, {}],
   DistributionFunction[MixtureModel[NormalModel[μ, σ], 2],
    y, MLParameterEstimate[ℱ]]}], {y, 40, 110}];
```



The plot shows that the agreement between the observations and the fitted model is good.

■ *Statistical Inference Package SIP* Overview

This section describes with real examples those features of *Statistical inference Package SIP*, which make the package a unique package for statistical inference.

□ More general statistical distributions and models

Statistical inference Package SIP contains *30* functions for defining *univariate/multivariate and discrete/ continuous distributions*. It contains also *19* functions for defining various *statistical models*. These include *sampling model*, *submodel*, *regression models* (8 functions), *models for stochastic processes* (3 functions), and *hierarchical models* (4 functions). Statistical model functions accept as arguments any statistical distributions and models. This recursive way of defining statistical models in *Statistical inference Package SIP* allows users to *generate and analyse very complicated models.*

As an example consider a sample of 345 subjects has been classified according to their blood group, obtaining the following frequency table ([3], p. 97). The colums of the table contain values of blood group and counts.

```
dataFile = ToFileName[Directory[],"BloodGroups.dat"];
TableForm[dataOnBloodGroups = Import[dataFile],TableAlignments->Right]
```

```
 A     150
 B      29
AB       6
 O     160
```

```
bloodGroupCounts = dataOnBloodGroups[[All,2]];
n                = Plus @@ bloodGroupCounts;
```

Denoting by p and q the allele frequencies of alleles A and B, respectively, and assuming Hardy–Weinberg equilibrium the statistical model for the counts can be defined as a submodel of the multinomial distribution as follows.

```
𝒟 = MultinomialModel[4,n]
```

```
--- MultinomialModel ---
```

```
ℳ = Submodel[𝒟,{p,q},{p^2+2p(1−p−q),q^2+2q(1−p−q),2p q,1−(p^2+2p(1−p−q))−(q^2+2q(1−p−q))−2p q
```

```
--- Submodel ---
```

The log–likelihood function has the following form.

```
l = Select[LogLikelihoodFunction[ℳ,bloodGroupCounts,{p,q}],(!NumericQ[#])&]
```

```
150 Log[p^2 + 2 p (1 - p - q)] + 6 Log[2 p q] +
 160 Log[1 - p^2 - 2 p (1 - p - q) - 2 p q - 2 (1 - p - q) q - q^2] +
 29 Log[2 (1 - p - q) q + q^2]
```

□ Very general regression models

Statistical inference Package SIP contains a function for defining regression models in which the responses can have any statistical distributions or models such that *any parameter or parameters of those can depend linearly or nonlinearly on the fixed explanatory variables and unknown regression parameters*.

As an example consider leaf springs dataset containing free height measurements of leaf springs in the unloaded condition with 8 inches as target value. The measurements have been done under low and high values of 5 treatments with three repeats ([4]). The columns of the following table contain values of oil temperature (–/+), transfer (–/+), heating (–/+), furnace (–/+), hold down (–/+), and three height measurements.

```
dataFile = ToFileName[Directory[],"LeafSprings.dat"];
TableForm[dataOnLeafSprings = Import[dataFile],TableAlignments->Right]
```

–	–	–	–	–	7.78	7.78	7.81
–	–	–	+	+	8.15	8.18	7.88
–	–	+	–	+	7.5	7.56	7.5
–	–	+	+	–	7.59	7.56	7.75
–	+	–	–	+	7.94	8.	7.88
–	+	–	+	–	7.69	8.09	8.06
–	+	+	–	–	7.56	7.62	7.44
–	+	+	+	+	7.56	7.81	7.69
+	–	–	–	–	7.5	7.25	7.12
+	–	–	+	+	7.88	7.88	7.44
+	–	+	–	+	7.5	7.56	7.5
+	–	+	+	–	7.63	7.75	7.56
+	+	–	–	+	7.32	7.44	7.44
+	+	–	+	–	7.56	7.69	7.62
+	+	+	–	–	7.18	7.18	7.25
+	+	+	+	+	7.81	7.5	7.59

```
o = oilTemperature     = dataOnLeafSprings[[All,1]];
d = transferTime       = dataOnLeafSprings[[All,2]];
c = heatingTime        = dataOnLeafSprings[[All,3]];
b = furnaceTemperature = dataOnLeafSprings[[All,4]];
e = holdDownTime       = dataOnLeafSprings[[All,5]];
heightMeasurements     = dataOnLeafSprings[[All,{6,7,8}]];
```

The values of height measurements form the response

```
resp = heightMeasurements;
n    = Length[resp];
```

and the vectors for the effects of oil temperature (O), transfer time (D), heating time (C), furnace temperature (B), and 'hold−down' time (E) are

```
eO = If[# === "−",−1,+1]& /@ o;
eD = If[# === "−",−1,+1]& /@ d;
eC = If[# === "−",−1,+1]& /@ c;
eB = If[# === "−",−1,+1]& /@ b;
eE = If[# === "−",−1,+1]& /@ e;

TableForm[Transpose[{eO,eD,eC,eB,eE}],
 TableHeadings −> {Automatic,{"O","D","C","B","E"}},TableAlignments −> Right]
```

	O	D	C	B	E
1	−1	−1	−1	−1	−1
2	−1	−1	−1	1	1
3	−1	−1	1	−1	1
4	−1	−1	1	1	−1
5	−1	1	−1	−1	1
6	−1	1	−1	1	−1
7	−1	1	1	−1	−1
8	−1	1	1	1	1
9	1	−1	−1	−1	−1
10	1	−1	−1	1	1
11	1	−1	1	−1	1
12	1	−1	1	1	−1
13	1	1	−1	−1	1
14	1	1	−1	1	−1
15	1	1	1	−1	−1
16	1	1	1	1	1

The design matrices for mean and standard deviation parameters are defined as follows

```
X = Transpose[{Table[1,{n}],eO,eD,eC,eO eD,eO eC,eD eC,eO eD eC}];
TableForm[X,
 TableHeadings −> {Automatic,
        {"Constant","O","D","C","O.D","O.C","D.C","O.D.C"}},
 TableAlignments −> Right]
```

	Constant	O	D	C	O.D	O.C	D.C	O.D.C
1	1	−1	−1	−1	1	1	1	−1
2	1	−1	−1	−1	1	1	1	−1
3	1	−1	−1	1	1	−1	−1	1
4	1	−1	−1	1	1	−1	−1	1
5	1	−1	1	−1	−1	1	−1	1
6	1	−1	1	−1	−1	1	−1	1
7	1	−1	1	1	−1	−1	1	−1
8	1	−1	1	1	−1	−1	1	−1
9	1	1	−1	−1	−1	−1	1	1
10	1	1	−1	−1	−1	−1	1	1
11	1	1	−1	1	−1	1	−1	−1
12	1	1	−1	1	−1	1	−1	−1
13	1	1	1	−1	1	−1	−1	−1
14	1	1	1	−1	1	−1	−1	−1
15	1	1	1	1	1	1	1	1
16	1	1	1	1	1	1	1	1

```
Z = Transpose[{Table[1,{n}],eC,eB,eE,eC eB,eC eE,eB eE,eC eB eE}];
TableForm[Z,
 TableHeadings -> {Automatic,
       {"Constant","C","B","E","C.B","C.E","B.E","C.B.E"}},
 TableAlignments -> Right]
```

	Constant	C	B	E	C.B	C.E	B.E	C.B.E
1	1	−1	−1	−1	1	1	1	−1
2	1	−1	1	1	−1	−1	1	−1
3	1	1	−1	1	−1	1	−1	−1
4	1	1	1	−1	1	−1	−1	−1
5	1	−1	−1	1	1	−1	−1	1
6	1	−1	1	−1	−1	1	−1	1
7	1	1	−1	−1	−1	−1	1	1
8	1	1	1	1	1	1	1	1
9	1	−1	−1	−1	1	1	1	−1
10	1	−1	1	1	−1	−1	1	−1
11	1	1	−1	1	−1	1	−1	−1
12	1	1	1	−1	1	−1	−1	−1
13	1	−1	−1	1	1	−1	−1	1
14	1	−1	1	−1	−1	1	−1	1
15	1	1	−1	−1	−1	−1	1	1
16	1	1	1	1	1	1	1	1

The statistical model is defined by

```
ℳ = RegressionModel[{X, Z},
      Distribution → SamplingModel[NormalModel[μ, σ], 3],
      InverseLink → {Function[z, z], Function[z, ℯ^z]}]
```

```
--- RegressionModel ---
```

□ General hidden Markov models

Statistical inference Package SIP contains a function for defining *hidden Markov models*, whose *emission distributions can be any statistical distributions or models*. In addition one can define *hidden Markov regression models* where the emission distributions or models depend on time−varying explanatory variables.

As an example consider the following counts of epileptic seizures.

```
dataFile = ToFileName[Directory[],"Seizures.dat"];
dataOnEpilepticSeizures = Import[dataFile,"List"]
```

```
{0, 3, 0, 0, 0, 0, 1, 1, 0, 2, 2, 1, 2, 0, 0, 1, 2, 1, 3, 1, 3,
 0, 4, 2, 0, 1, 1, 2, 1, 2, 1, 1, 1, 0, 1, 0, 2, 2, 1, 2, 1, 0,
 0, 0, 2, 1, 2, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0,
 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0,
 0, 2, 1, 0, 1, 1, 0, 0, 0, 2, 2, 0, 1, 1, 3, 1, 1, 2, 1, 0,
 3, 6, 1, 3, 1, 2, 2, 1, 0, 1, 2, 1, 0, 1, 2, 0, 0, 2, 2, 1,
 0, 1, 0, 0, 2, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0,
 0, 1, 3, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0,
 1, 2, 1, 0, 0, 0, 0, 0, 0, 1, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}
```

Assume that the counts can be modelled as generated by a stationary two state hidden Markov model with counts having Poisson distribution in both states.

```
ℳ = HiddenMarkovModel[PoissonModel[μ],
  2, Length[resp], Stationary -> True]
```

```
--- StationaryHiddenMarkovModel ---
```

□ Automatic handling of censored data

Statistical inference Package SIP handles automatically interval censored data. Observations can censored from below, from above, or more generally belong to *any finite union of intervals*.

As an example consider two groups of rats which were exposed to carcinogen DBMA, and the number of days to death due to cancer was recorded ([5]). The columns of the table contain values of group, number of days, and censoring status indicator with 0 denoting uncensored and 1 censored observations.

```
dataFile = ToFileName[Directory[],"Carcinogen.dat"];
dataOnCarcinogenDeaths = Import[dataFile];
```

This shows twenty first units in the dataset.

```
TableForm[Take[dataOnCarcinogenDeaths,20],TableAlignments->Right]
```

```
Group_1     143     0
Group_1     164     0
Group_1     188     0
Group_1     188     0
Group_1     190     0
Group_1     192     0
Group_1     206     0
Group_1     209     0
Group_1     213     0
Group_1     216     0
Group_1     220     0
Group_1     227     0
Group_1     230     0
Group_1     234     0
Group_1     246     0
Group_1     265     0
Group_1     304     0
Group_1     216     1
Group_1     244     1
Group_2     142     0
```

```
group      = dataOnCarcinogenDeaths[[All,1]];
daysToDeath = dataOnCarcinogenDeaths[[All,2]];
status     = dataOnCarcinogenDeaths[[All,3]];
```

The values of days to death form the response

```
resp = MapThread[If[#1 == 1,Interval[{#2,∞}],#2]&, {status, daysToDeath}];
n = Length[resp];
```

and the design matrix for the treatment is

```
X = Transpose[{Table[1,{n}],If[# === "Group_1",1,0]& /@ group}];
```

The statistical model is defined by

```
m = RegressionModel[X, Distribution → LogNormalModel[μ, σ]]
```

```
--- RegressionModel ---
```

LogLikelihoodFunction[$\mathcal{M}$, resp, {β1, β2, σ}]

$$-\text{Log}[142] - \frac{(-\beta 1 + \text{Log}[142])^2}{2\,\sigma^2} - \text{Log}[143] - \frac{(-\beta 1 - \beta 2 + \text{Log}[143])^2}{2\,\sigma^2} -$$
$$\text{Log}[156] - \frac{(-\beta 1 + \text{Log}[156])^2}{2\,\sigma^2} - \text{Log}[163] - \frac{(-\beta 1 + \text{Log}[163])^2}{2\,\sigma^2} -$$
$$\text{Log}[164] - \frac{(-\beta 1 - \beta 2 + \text{Log}[164])^2}{2\,\sigma^2} - 2\,\text{Log}[188] -$$
$$\frac{(-\beta 1 - \beta 2 + \text{Log}[188])^2}{\sigma^2} - \text{Log}[190] - \frac{(-\beta 1 - \beta 2 + \text{Log}[190])^2}{2\,\sigma^2} -$$
$$\text{Log}[192] - \frac{(-\beta 1 - \beta 2 + \text{Log}[192])^2}{2\,\sigma^2} - \text{Log}[198] - \frac{(-\beta 1 + \text{Log}[198])^2}{2\,\sigma^2} -$$
$$\text{Log}[205] - \frac{(-\beta 1 + \text{Log}[205])^2}{2\,\sigma^2} - \text{Log}[206] - \frac{(-\beta 1 - \beta 2 + \text{Log}[206])^2}{2\,\sigma^2} -$$
$$\text{Log}[209] - \frac{(-\beta 1 - \beta 2 + \text{Log}[209])^2}{2\,\sigma^2} - \text{Log}[213] -$$
$$\frac{(-\beta 1 - \beta 2 + \text{Log}[213])^2}{2\,\sigma^2} - \text{Log}[216] - \frac{(-\beta 1 - \beta 2 + \text{Log}[216])^2}{2\,\sigma^2} -$$
$$\text{Log}[220] - \frac{(-\beta 1 - \beta 2 + \text{Log}[220])^2}{2\,\sigma^2} - \text{Log}[227] -$$
$$\frac{(-\beta 1 - \beta 2 + \text{Log}[227])^2}{2\,\sigma^2} - \text{Log}[230] - \frac{(-\beta 1 - \beta 2 + \text{Log}[230])^2}{2\,\sigma^2} -$$
$$2\,\text{Log}[232] - \frac{(-\beta 1 + \text{Log}[232])^2}{\sigma^2} - 4\,\text{Log}[233] - \frac{2\,(-\beta 1 + \text{Log}[233])^2}{\sigma^2} -$$
$$\text{Log}[234] - \frac{(-\beta 1 - \beta 2 + \text{Log}[234])^2}{2\,\sigma^2} - \text{Log}[239] - \frac{(-\beta 1 + \text{Log}[239])^2}{2\,\sigma^2} -$$
$$\text{Log}[240] - \frac{(-\beta 1 + \text{Log}[240])^2}{2\,\sigma^2} - \text{Log}[246] - \frac{(-\beta 1 - \beta 2 + \text{Log}[246])^2}{2\,\sigma^2} -$$
$$\text{Log}[264] - \frac{(-\beta 1 + \text{Log}[264])^2}{2\,\sigma^2} - \text{Log}[265] - \frac{(-\beta 1 - \beta 2 + \text{Log}[265])^2}{2\,\sigma^2} -$$
$$2\,\text{Log}[280] - \frac{(-\beta 1 + \text{Log}[280])^2}{\sigma^2} - 2\,\text{Log}[296] - \frac{(-\beta 1 + \text{Log}[296])^2}{\sigma^2} -$$
$$\text{Log}[304] - \frac{(-\beta 1 - \beta 2 + \text{Log}[304])^2}{2\,\sigma^2} - \text{Log}[323] - \frac{(-\beta 1 + \text{Log}[323])^2}{2\,\sigma^2} -$$
$$18\,\text{Log}[2\,\pi] - 36\,\text{Log}[\sigma] + \text{Log}\Big[1 + \frac{1}{2}\Big(-1 - \text{Erf}\Big[\frac{-\beta 1 + \text{Log}[204]}{\sqrt{2}\,\sigma}\Big]\Big)\Big] +$$
$$\text{Log}\Big[1 + \frac{1}{2}\Big(-1 - \text{Erf}\Big[\frac{-\beta 1 - \beta 2 + \text{Log}[216]}{\sqrt{2}\,\sigma}\Big]\Big)\Big] +$$
$$\text{Log}\Big[1 + \frac{1}{2}\Big(-1 - \text{Erf}\Big[\frac{-\beta 1 - \beta 2 + \text{Log}[244]}{\sqrt{2}\,\sigma}\Big]\Big)\Big] +$$
$$\text{Log}\Big[1 + \frac{1}{2}\Big(-1 - \text{Erf}\Big[\frac{-\beta 1 + \text{Log}[344]}{\sqrt{2}\,\sigma}\Big]\Big)\Big]$$

□ Profile likelihood confidence intervals for more general interest functions

In *Statistical inference Package SIP* profile likelihood based confidence interval can be calculated for *any linear or smooth nonlinear interest function* of parameters.

The calculation of profile likelihood based confidence intervals for various functions of parameters in the case of waiting times of eruptions of Old Faithful provides examples of this.

□ More general hypotheses

Statistical inference Package SIP can handle statistical hypotheses corresponding restricted statistical models defined by *any linear or smooth nonlinear functions* of the parameters.

The calculation of likelihood ratio tests for various hypotheses on parameters in the case of waiting times of eruptions of Old Faithful provides examples of this.

□ Random observation from any statistical distribution or model

The uncertainties of statements made in the statistical inference are derived from the probability distributions of the statistical model. In case of *likelihood based inference*, on which *Statistical inference Package SIP* relies, these uncertainties come from certain asymptotic distributional results of very general nature. From this generality follows that the same method for deriving inferential statements and their uncertainties can be applied in a very large collection of statistical models.

□ Symbolic computation of properties of statistical models

Statistical inference Package SIP contains *32* functions for calculating *properties of statistical distributions and models*. In addition to numerical arguments and results, almost all of these functions accept *symbolic arguments and give symbolic results*. The package has been designed so that the results of various functions in it can easily be given as input to other *MATHEMATICA* functions.

■ Stages of statistical inference in *Statistical inference Package SIP*

□ Data

Statistical inference starts with *empirical data*, which consists of *units* that have been selected into the empirical study. The data set contains one or more *statistical variables* which give for statistical units values of some properties of the units. Statistical units are often called *subjects*, *cases*, *runs*, etc.

□ Response

The first task in the analysis is to decide which statistical variable or variables forms *response* or responses, respectively. The response is the statistical variable whose *distribution* or *properties* of the distribution are of interest in the empirical study. One might be interested in the form of the distribution or more generally how the the distribution or some property of it depends on other statistical variables observed in the study. Responses are often called also *'dependent'* variables or *criterion* variables.

□ Statistical model

After response has been selected one has to decide which kind of *statistical model* is used to analyze the data. The first aspect to attend to in deciding about the statistical model is *dependence* between the statistical units.

If the observations from different statistical units are *independent*, that is, if the value of the response for some statistical unit in no way affects it's values for other statistical units, one

has to select a statistical model for independent observations. In *Statistical inference Package SIP* there is a large collection of such models: SamplingModel, IndependenceModel, LinearRegressionModel, LogisticRegressionModel, etc. SamplingModel is used when in the data there is only one group of units and one is interested in the distribution of the response in this one group. IndependenceModel is used when the data set contains finite number of groups and one is interested in the differences of the distributions of the response among these groups. Finally, various regression models are used when the data set contains other statistical variables whose affect on the distribution of the response is of interest. These other statistical variables are often called *factors*, *explanatory* variables, '*independent*' variables etc.

If the observations from different statistical units are *dependent*, that is, if the value of a response variable for some statistical unit affects it's values for other statistical units, one has to select a statistical model for dependent observations. In *Statistical inference Package SIP* there are three such models: MarkovChainModel, MarkovProcessModel, and StochasticProcessModel. In all these units are assumed to be arranged in fixed order. Often the order is the time order of the observations, but it need not be. The models differ with respect to the nature of the response and to the nature of the dependence of an observation on the previous observations.

□ Interest function

The statistical models in *Statistical inference Package SIP* are so called *parametric* statistical models. This means that the distribution of the response is assumed to have some known functional form with finite number of real valued unknown quantities called *parameters*. Statistical inference then concerns the unknown values of the parameters or more generally the unknown values of some real valued functions of the parameters. These functions are called *interest functions*.

□ Confidence interval

Given a real valued interest function of the parameters the problem is to find those values which are *supported* by the *statistical evidence*, that is, by observed values of response and it's statistical model.

The *Statistical inference Package SIP* function ProfileInterval calculates the so called profile likelihood based confidence interval for any smooth interest function and statistical model. On default the result is an (approximate) 95%−level confidence interval for the unknown value of the interest function.

The intepretation of the interval is such that the statistical evidence supports every value inside the interval more than any value outside the interval. The confidence level is a measure of reliability of the statement that the unknown value of the interest function belongs to the actual computed interval or in other words one minus the confidence level measures the risk involved in making the statement that the unknown value of the interest function belongs to the actual computed interval.

□ Significance test

In *Statistical inference Package SIP* the function LRTest is used to perform significance tests. The result of the function is the *observed significance level* that measures the risk in making the statement that the statistical hypothesis is 'false', that is, making the statement that the belief of the researcher is 'true'. The observed significance level thus tells whether there is in the data and model enough evidence to support the claim of the researcher.

Although significance tests are widely used one should, however, be aware of their limited usefulness. First, if the result is *not significant*, that is, if there is not enough evidence to make the statement that the statistical hypothesis is 'false', it does not follow that the statistical hypothesis is 'true'. It merely means that there is because of various possible reasons not enough information in the statistical evidence to refute the statistical hypothesis, even if it is 'false', or that the statistical hypothesis might actually be 'true'. Secondly, in case that the result of the test is *significant*, the significance test usually does not tell how much and in which way the considered interest functions deviate from values they have under the statistical hypothesis. This all means that (profile likelihood based) confidence intervals provide a much more informative way of statistical inference.

□ Additional remarks

Statistical inference Package SIP has been designed so that in constructing more complicated statistical models from statistical distributions and other statistical models the component statistical distributions and models can be any distributions and models. For example, one can construct a mixture model of logistic regression models or nonlinear regression model of a Markov process model. Thus in fact the response can consist of independent vectors whose components are dependent.

Because profile likelihood based confidence intervals and likelihood ratio tests require the calculation of the (constrained) maximum likelihood estimates there is in *Statistical inference Package SIP* a function (MLEFit) for this. Statistical inference, however, should never consist of plain maximum likelihood estimation, but instead of calculation of confidence intervals and/or significance tests, which are proper forms of statistical inference containing assessments of the reliabilities of the statements made.

Various properties of statistical models can be calculated symbolically or numerically.

■ Example

This loads the package.

```
<<StatisticalInference`
```

□ Data

In an experiment testing the effectiveness of a pesticide, two groups of flies were exposed to the pesticide for 30 and 60 seconds, respectively. The quantity measured was the time elapsed from the instant the fly was exposed to the pesticide to the time of reaction (Denker et al 1998, p. 400). The columns of the following table contain values of exposure time and reaction time, respectively.

```
dataFile = ToFileName[Directory[],"Flies.dat"];
(dataOnFlies = Import[dataFile])//TableForm
```

```
30      3.
30      5.
30      5.
30      7.
30      9.
30      9.
30      10.
30      12.
30      20.
30      24.
30      24.
30      34.
30      43.1
30      46.
30      57.9
30      140.
60      2.
60      5.
60      5.
60      7.
60      8.
60      9.
60      14.
60      18.
60      24.
60      26.
60      26.
60      34.
60      37.1
60      42.
60      89.9
```

These expressions extract the variables from the table.

```
exposureTimes = dataOnFlies[[All,1]];
reactionTimes = dataOnFlies[[All,2]];
```

□ Response

Response consists of two vectors of reaction times with sizes 16 and 15, respectively.

```
resp={Extract[reactionTimes,Position[exposureTimes,30]],
      Extract[reactionTimes,Position[exposureTimes,60]]};
```

□ Statistical model

Assume that observations can be considered as independent samples from some possibly different gamma−models.

```
M=IndependenceModel[
    {SamplingModel[GammaModel[ν,μ],Length[First[resp]]],
     SamplingModel[GammaModel[ν,μ],Length[Last[resp]]]}]
```

```
--- IndependenceModel ---
```

This gives the observed log−likelihood function of the model

LogLikelihoodFunction[$\mathcal{M}$,resp,{{ν30,μ30},{ν60,μ60}}]

$$44.9161\,(-1+\nu30)-\frac{449.\,\nu30}{\mu30}+40.7346\,(-1+\nu60)-\frac{347.\,\nu60}{\mu60}+16\,\nu30\,\text{Log}\left[\frac{\nu30}{\mu30}\right]+15\,\nu60\,\text{Log}\left[\frac{\nu60}{\mu60}\right]-16\,\text{Log}[\text{Gamma}[\nu30]]-15\,\text{Log}[\text{Gamma}[\nu60]]$$

and this the observed score function.

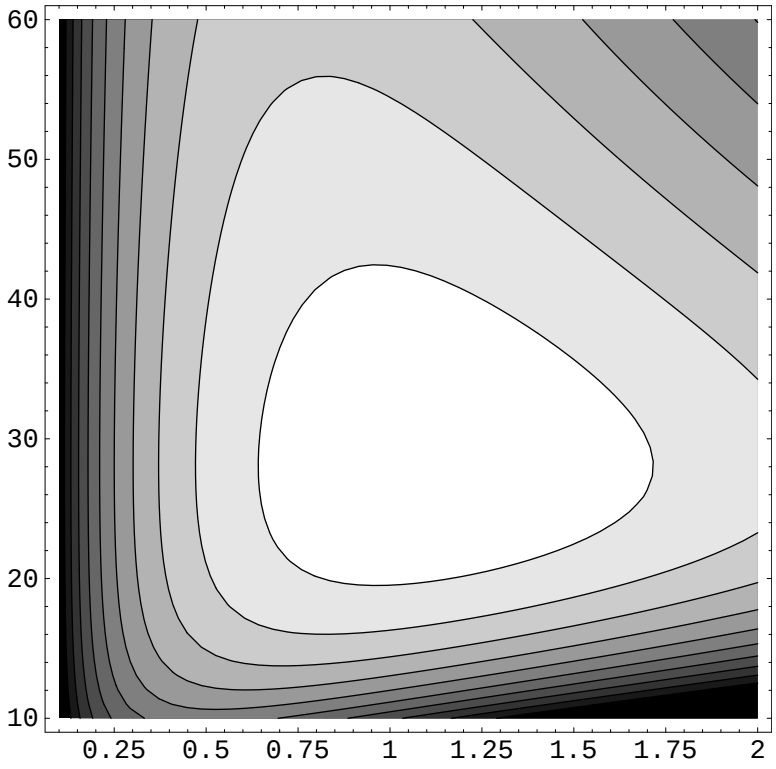
ScoreFunction[$\mathcal{M}$,resp,{{ν30,μ30},{ν60,μ60}}]//MatrixForm

$$\begin{pmatrix} 60.9161-\frac{449.}{\mu30}+16\,\text{Log}\left[\frac{\nu30}{\mu30}\right]-16\,\text{PolyGamma}[0,\nu30] \\ \frac{449.\,\nu30}{\mu30^2}-\frac{16\,\nu30}{\mu30} \\ 55.7346-\frac{347.}{\mu60}+15\,\text{Log}\left[\frac{\nu60}{\mu60}\right]-15\,\text{PolyGamma}[0,\nu60] \\ \frac{347.\,\nu60}{\mu60^2}-\frac{15\,\nu60}{\mu60} \end{pmatrix}$$

□ Fitted model

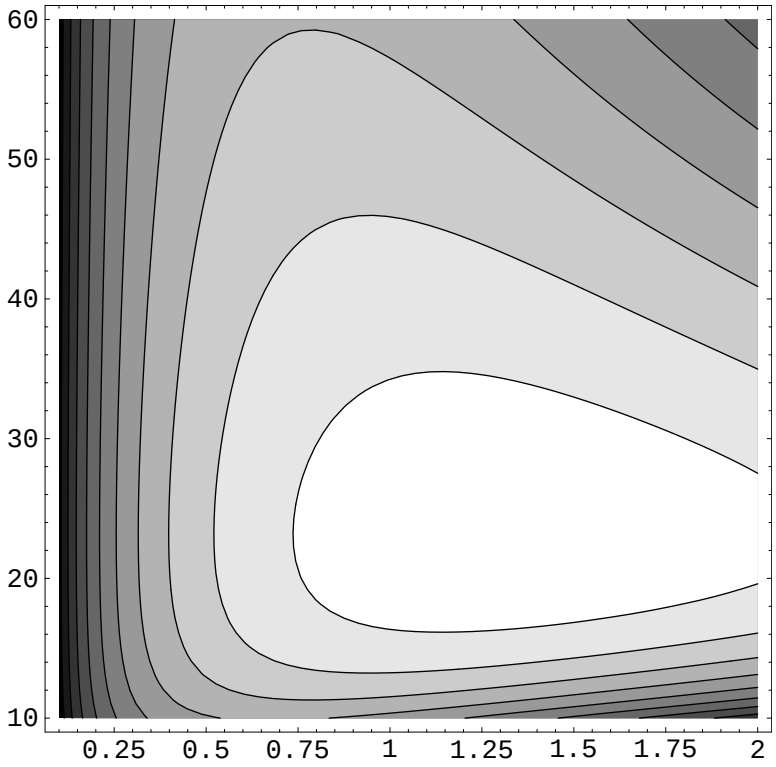
This is the contour plot of the observed log−likelihood function for the 30 seconds exposure time group

ContourPlot[Evaluate[LogLikelihoodFunction[SamplingModel[GammaModel[ν,μ],Count[exposureTimes,



and this the same for the 60 seconds exposure time group.

```
ContourPlot[Evaluate[LogLikelihoodFunction[SamplingModel[GammaModel[ν,μ],Count[exposureTimes,
```



This gives the fitted model

```
ℱ=MLEFit[ℳ,resp]
```

```
--- FittedModel ---    Convergence: True
```

and the maximum likelihood estimate is

```
MLParameterEstimate[ℱ]//N
```

```
{{1.08475, 28.0625}, {1.31645, 23.1333}}
```

□ Profile interval for difference between the means

This calculates (approximate) 95% confidence interval to the difference of the means

```
ℐ = ProfileInterval[ℱ, μ30 - μ60, {{ν30, μ30}, {ν60, μ60}}]
```

```
--- ProfileIntervalModel ---    Convergence: True
```

and the confidence interval is

```
ConfidenceLimits[ℐ]
```

```
{-13.1179, 26.3399}
```

□ Likelihood ratio test for both samples coming from exponential model

This calculates the likelihood ratio test for the statistical hypothesis that both shapes are equal to one, i.e., both samples are samples from exponential models.

```
𝒯=LRTest[ℱ,{μ30,μ60},{{1,μ30},{1,μ60}}]
```

```
--- LRTestModel ---
```

Observed significance level of the test is

```
ObservedSignificanceLevel[𝒯]
```

```
0.699753
```

■ References

[1] A. O'Hagan, *Kendall's Advanced Theory of Statistics Volume 2B, Bayesian Inference,* London: Edward Arnold, 1994.

[2] Y. Pawitan, *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford: Clarendon Press, 2001.

[3] A. Azzalini, *Statistical Inference: Based on the Likelihood.* London: Chapman and Hall, 1996.

[4] J. J. Pignatiello and J. S. Ramberg, "Contribution to discussion of offline quality control, parameter design and the Taguchi method," *Journal of Quality Technology*, **17**, 1985 pp. 198–206.

[5] M. E. Stokes, C. S. Davis, and G. G. Koch, *Categorical Data Analysis: Using the SAS System,* Cary, NC: SAS Institute Inc., 1995.

[6] J. D. Kalbfleisch and R. I. Prentice, *Statistical Analysis of Failure Time Data*. New York: Wiley, 1980.