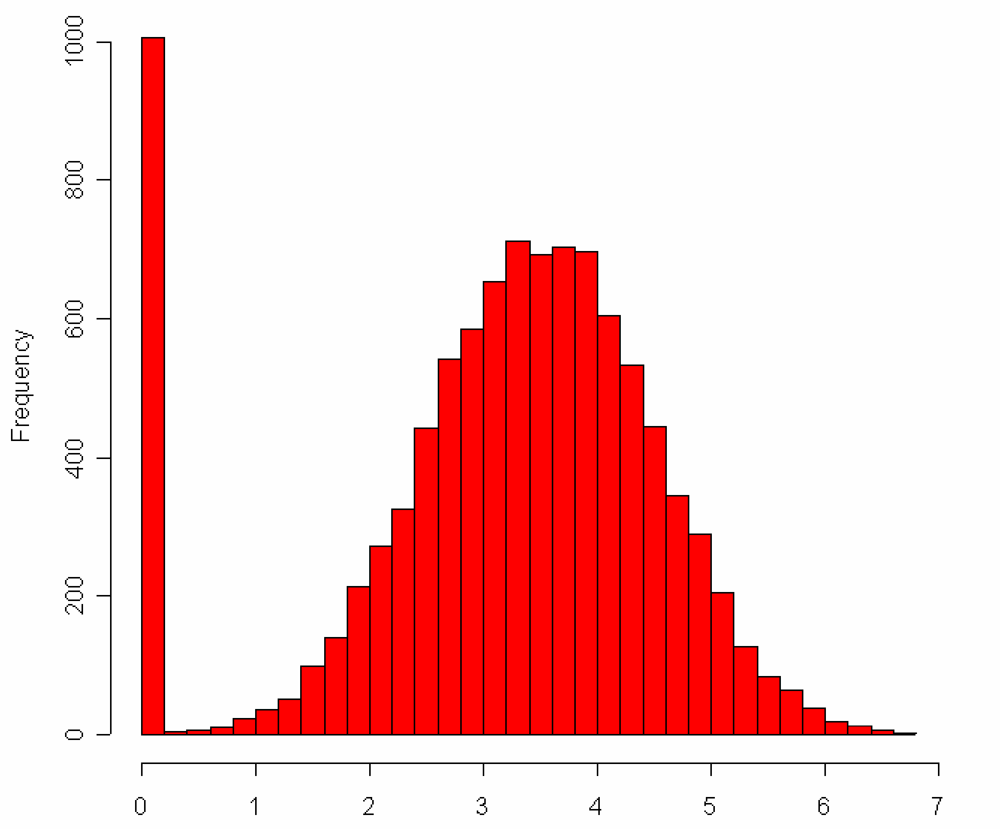


Distributional assumptions in the representation of random taste heterogeneity

Stephane Hess, Imperial College London
Kay W. Axhausen, ETH Zurich

Conference paper STRC 2005

STRC | 5th Swiss Transport Research Conference
Monte Verità / Ascona, March 9-11, 2005

Distributional assumptions in the representation of random taste heterogeneity

Stephane Hess
Centre for Transport Studies
Imperial College London
Exhibition Road
London SW7 2AZ
United Kingdom
Tel. 44 20 7594 6105
Fax 44 20 7594 6102
E-Mail: stephane.hess@imperial.ac.uk

Kay W. Axhausen
Institute for Transport
Planning and Systems
ETH Zürich
CH – 8093 Zürich
Switzerland
Tel. 41 1 633 3943
Fax. 41 1 633 1057
axhausen@ivt.baug.ethz.ch

March 2005

Abstract

This paper examines how well a wide range of parametric distributions can reproduce given target distributions, which are constructed to reflect common assumptions about taste variation in transport models. Using ExpertFit to fit the distributions and to measure the differences between the fitted and the target distribution, a large data set of performance measures is constructed, revealing some systematic patterns of bias. Mixed results are obtained for the given target distributions (Normal censored/uncensored and with/without mass points, Johnson Sb and Lognormal). They show that flexible distributions, such as the Johnson S_U, Gamma, Beta, Erlang, Laplace and Logistic have certain advantages over inflexible distributions, but these advantages disappear for specific constellations. The analysis also shows the usual problems with the heavy tails of the Lognormal distribution, but worryingly also suggests that the recently much-heralded Johnson-S_B distribution does not perform well in all scenarios. The analysis shows that the use of symmetrical distributions, such as the Normal, can lead to problems with the tails, such that, if the Normal is used, no information should be inferred on the basis of its tail behaviour, though it can offer valid estimates of the mean and variance. Overall, these results suggest that, in the absence of software allowing for the use of empirical distributions, or mixtures between continuous and discrete distributions, the most flexible distribution should be used to minimise the risk of biased results.

Keywords

Distributional assumptions – Mixed Logit – Random taste heterogeneity

1. Introduction

Discrete choice models have become the preferred tool for the analysis of individual choices in transportation research. While the majority of applications still make use of closed-form models such as Multinomial Logit (MNL) and Nested Logit (NL), researchers (and to a lesser degree practitioners) are increasingly turning their attention to the Mixed Multinomial Logit (MMNL) model. The MMNL model has the advantage of being able to accommodate random variations in tastes across the population, i.e. variations in the sensitivity to explanatory factors such as travel-time and travel-cost.

The MMNL model uses integration of MNL choice probabilities over the assumed distribution of the taste coefficients, where almost exclusively, continuous statistical distributions are used as the mixing distributions. A major issue in the specification of an MMNL model is the choice of an appropriate mixing distribution, especially so in the absence of *a priori* information about the true shape of the distribution for a given coefficient, or in a more general sense, the sign of the coefficient.

A misguided choice of mixing distribution can lead to poor model performance, and misleading conclusions. A large number of different distributions are available, including unbounded distributions such as the *Normal*, distributions bounded on one side, such as the *Lognormal*, and distributions bounded on both sides, such as the *Triangular and* S_B distributions. For the bounded distributions, a further important distinction needs to made on the basis of whether the bounds are preset (e.g. *Lognormal*), or estimated (e.g. *Triangular* and S_B). This issue is discussed in detail by Hess et al. (2005a), who note that the latter type of distribution has the advantage of still being able to signal problems with data impurities or model misspecifications, which could manifest themselves in the form of a large share for counter-intuitively signed coefficient values, something that is not possible with strictly bounded distributions.

In the case where an *a priori* assumption exists about the true shape of the distribution, or about its nature in terms of bounds, an appropriate choice of mixing distribution can reduce the risk of bias. Nevertheless, it should be recognised that, even in this case, theoretical distributions will not generally be able to offer a perfect approximation to the true distribution. The fact that researchers are generally limited in their choice of distribution by the software package they are using further increases the risk of a significant bias in the approximation. Indeed, the most widely used estimation tools are limited to the Normal and Uniform distribution, and direct transformations thereof, such as the Lognormal and S_B in the case of the Normal, and the Triangular in the case of the Uniform distribution.

The above discussion has shown that, given the risk of bias in the approximation to the true distribution, it is of interest to use as flexible a distribution as possible. In this paper, we discuss specifically the ability of several continuous distributions, including the most heavily used ones in MMNL modelling, to approximate the shape of a range of hypothetical *true* distributions. The analysis also discusses the potential of several distributions that have, to the authors' knowledge, thus far not been used in MMNL modelling. Furthermore, the paper discusses the potential of adapting continuous distributions such as to allow for a heightened mass at a given value, which can for example be useful in the analysis of the prevalence of travellers with a zero value-of-time.

Rather than setting this comparison inside the framework of an actual MMNL analysis, we conduct a purely theoretical analysis. However, the paper offers guidance on how to estimate

MMNL models specified with distributions that are not commonly used in the representation of taste heterogeneity.

The remainder of this paper is organised as follows. In the next section, we briefly review the theory on the Mixed Logit model and discuss the issue of the choice of distribution. Section 3 discusses the empirical framework, while section 4 presents the results. Finally, section 5 summarises the findings of the research and makes some recommendations for improved practice.

2. Distributional Assumptions in the MMNL model

Over recent years, the MMNL model (c.f. McFadden & Train, 2000) has seen increased use in the area of transportation research. The MMNL model has the potential to address the shortcoming of the Multinomial Logit (MNL) model (McFadden, 1974) by allowing for random taste heterogeneity across decision-makers and variable inter-alternative correlation in the unobserved parts of utility. Furthermore, the MMNL model can be specified so as to explicitly account for the repeated choice nature that is a characteristic of many of the more complex datasets used in transportation. Finally, the model can also allow for heteroscedasticity in the error-terms.

In this paper, we will concentrate on the random-coefficients formulation of the MMNL model. In this notation, the MMNL choice probabilities are expressed by means of the integral of MNL choice probabilities over the assumed distribution of the random taste-coefficients:

$$P_{ni} = \int_{\beta} P_{ni}(\beta) f(\beta|\Omega) d\beta \ , \qquad \ldots[1]$$

where $P_{ni}(\beta)$ is the MNL choice probability for alternative i and decision-maker n, given by:

$$P_{ni}(\beta) = \frac{e^{V_{ni}}}{\sum_{j=1}^{I} e^{V_{nj}}} \ , \qquad \ldots[2]$$

where V_{ni} is the observed part of utility, which can be rewritten as $f(\beta, x_{ni})$, where β is a vector of taste-coefficients and x_{ni} is a vector of the attributes of alternative i, as faced by decision-maker n (possibly with interactions with individual-specific socio-demographic attributes). The function $f(\beta|\Omega)$ is the density of the distribution function of the tastes across decision-makers, where the vector Ω gives the attributes of this distribution across the population, typically the mean and standard deviation. In model calibration, the likelihood of the model is maximised with regards to Ω, in order to obtain the most likely parameters of the distribution of β across the sample population.

Except in the case of a trivial distribution function for β, the choice probabilities in equation [1] do not have a closed-form solution, and numerical techniques, typically simulation, need to be used in the estimation and application of the MMNL model. For a discussion of the issues involved with simulation, see for example Hess et al. (2005b).

The choice of distribution for the randomly distributed coefficients (i.e. $f(\beta|\Omega)$) is one of the central issues in the specification of an MMNL model, especially in the case where an *a priori* assumption exists about the sign of a given coefficient, such as for example a travel-time coefficient. For recent discussions of this issue, see for example Train (2003), Hensher &

Greene (2003) and Hess et al. (2005a). By using an unbounded distribution, such as the *Normal*, researchers in effect make an a priori assumption that positive as well as negative values for the given coefficient may exist in the population. The problem is that, due to the shape of the *Normal* distribution (notably because it is symmetrical), it is not clear *a priori* whether results indicating a non-zero probability of a positive travel-time coefficient do in effect reveal the presence of such values, or whether these results are simply an effect of the nature of the *Normal* distribution. This thus clearly constitutes a major drawback of the Normal distribution. On the other hand, the use of the classical choice of bounded distribution, the *Lognormal*, makes an a priori assumption that the sign of the specific coefficient stays constant across individuals. For this reason, Hess et al. (2005a) recommend the use of distributions bounded on either side, where the bounds are estimated from the data, during model calibration. While results indicating a positive probability of a non-negative travel-time coefficient can in this case still be seen as an artefact of the model specification (due to the use an incomplete specification of utility), the risk of bias due to distributional assumptions decreases significantly. Examples of such distributions include the S_B distribution and the *Triangular* distribution. For more details on possible distributions, see for example Hensher & Greene (2003).

While a large number of different available distributions have been used in MMNL modelling, there is as yet a distinct lack of empirical evidence aimed at showing which distribution might be most appropriate in a given scenario. The analysis of the ability of the different distributions to recover the nature of an assumed *true* distribution is the main aim of the analysis described in this paper.

3. Empirical framework

An empirical framework was set up to assess the ability of different distributions to recover the behaviour of an underlying "true" distribution of the marginal utility of travel-time. A set of true distributions were used to generate a number of datasets, to which certain distributions were then fitted, before analysing the results in terms of various criteria. The following three sections describe the separate steps in this process.

3.1 Generation of datasets

Three groups of datasets were generated for the analysis; one based on the Normal distribution (including adapted versions), one based on the S_B distribution, and one based on the Lognormal distribution.

Aside from the simple Normal distribution, three adapted versions were used;

- a Normal distribution with a specific mass at a given point
- a truncated Normal distribution
- a censored Normal distribution, with any censored mass assigned to the lower endpoint

For the basic Normal distribution, the following four settings were used:

- $\mu=3$, $\sigma=1$
- $\mu=6$, $\sigma=1$
- $\mu=2$, $\sigma=3$
- $\mu=1$, $\sigma=6$

These represent cases with a mean larger than the standard deviation, and cases with the standard deviation larger than the mean, as well as cases with an essentially exclusively positive domain, and cases with a spread over the positive as well as negative part of the space of real numbers. There is no need to additionally use a case with a *purely* negative domain; a simple sign change could be used in the model fitting exercise to revert to the *purely* positive case. This set of experiments thus allows for the representation of cases of attributes with purely signed effects, as well as attributes that some individuals value positively, while some others value them negatively.

The next set of experiments uses a Normal distribution with a mass at a given point, allowing for example for the representation of special cases in value-of-time modelling, such as a zero value-of-time for part of the population (c.f. Cirillo & Axhausen, 2004).

The settings used were as follows:

- μ=0.6, σ=2, mass of 0.1 at 0
- μ=0.6, σ=0.15, mass of 0.1 at 0
- μ=0.6, σ=0.15, mass of 0.2 at 0
- μ=0, σ=1, mass of 0.1 at 0

This includes cases of a centrally-located mass-point, a mass below the lower 95% confidence limit, as well as a mass at zero for attributes that some individuals care about positively, some negatively, and some don't at all. There is no interest in further using a *purely* positive case with a mass somewhere in the middle, or a case with a mass at the upper end-point, as both can be obtained as minor modifications of the above cases.

In the truncated case, with mean μ and standard deviation σ, and lower truncation point α, the truncated Normal draw produced on the basis of a uniform draw *r* is given by:

$$\mu + \sigma \cdot \Psi\left[(1-r)\cdot \varphi(\alpha,\mu,\sigma) + r\right], \qquad \ldots[3]$$

where Ψ is in the inverse standard cumulative Normal distribution, and φ is the cumulative Normal distribution. The parameter values used with this distribution were:

- μ=0.6, σ=2, α=0.2
- μ=0.6, σ=2, α=0
- μ=0.4, σ=0.2, α=0
- μ=0.4, σ=0.6, α=0.4

This again covers a selection of cases, and there is no need to use examples with upper truncation, as it can be obtained by simple reflection.

The final type of distribution is based on producing a set of draws from a Normal distribution, and replacing all values below the lower censor by the value of that censor. The parameter values used with this distribution were:

- μ=0.6, σ=0.6, α=0
- μ=0.6, σ=0.3, α=0
- μ=0.6, σ=0.9, α=0
- μ=0.6, σ=0.3, α=0.3

Again, the case of an upper censor can be obtained by simple reflection.

For the Lognormal distribution, the following settings were used:

- μ =-3, σ =2
- μ =-2, σ =1
- μ =0, σ =0.5
- μ =-2, σ =0.5

This includes varying cases with means closer or further away from zero, and variations in the weight of the tail of the distribution. Cases with a sign-change are not explored in the present analysis; given the exact symmetry, the results of the approximations would be identical, as a sign-change can be used in conjunction with any distribution.

Finally, for the S_B distribution, the following settings were used:

- mu=0, s=1, a=0.1, b=0.3
- mu=-2, s=1, a=0, b=0.5
- mu=2, s=1, a=0.5, b=0.7
- mu=1, s=9, a=0.1, b=0.5

In addition to the parameters of the underlying Normal distribution, the specification has two additional parameters, **a** and **b**, which define the bounds of the distribution (transformation from the initial 0-1 domain). The above four cases include a symmetrical example (with a flat plateau), a left-skewed example, a right-skewed example, and a bi-modal example.

For every choice of distribution, a dataset of 8,000 independent observations was generated using the statistical package *R*, resulting in 24 separate datasets.

3.2 Fitting of distributions

In the analysis of the approximation performance of different distributions, the distribution fitting software ExpertFit (EF) was used. This optimises the parameters of a target distribution so as to most closely replicate the behaviour of the true distribution.

A total of thirty-two continuous distributions were available for the analysis. While some of these distributions have been used previously in MMNL modelling, a large number have not. To the authors' knowledge, this for example includes the Beta distribution and the Johnson SU distribution. At this point, it should be noted that not all of the distributions given below could be used in each of the experiments, for model fitting reasons. Additionally, it should be noted that empirical distributions were not included at this stage of the analysis.

The thirty-two distributions can be split into three separate categories, as follows.

3.2.1 Unbounded distributions

These ten distributions are unbounded on either side. As such, they are occasionally easier to estimate, but can produce problems in terms of producing high probabilities of counter-intuitively signed coefficients (c.f. Hess et al., 2005a)

- Cauchy
  - Symmetrical, mean and variance do not exist
- Error
  - Symmetrical, μ=0
- Exponential power
  - Symmetrical

- Extreme value type A
    - Right-skewed
- Extreme value type B
    - Left-skewed
- Johnson SU
    - Can be symmetrical & asymmetrical
- Laplace
    - Symmetrical
- Logistic
    - Symmetrical
- Normal
    - Symmetrical
- Student's t,
    - Symmetrical, μ=0

3.2.2 Left-bounded distributions

These seventeen distributions are bounded on the left-hand side. For some distributions, the lower bound is fixed at zero, while for others, the distribution has an additional location parameter, allowing for a variable lower bound. In some cases, this bound is however fixed to be strictly positive. In all cases, an external shift of the final draws can however be used to transform the lower bound to essentially any desired real value.

We now look at the individual distributions in more detail, each time giving the initial domain.

- Chi-square
    - Domain = $(0,\infty)$, variance proportional to mean ($\sigma^2=2\mu$).
- Erlang
    - Domain = (γ,∞), no constraints on γ
- Exponential
    - Domain = (γ,∞), no constraints on γ
- F
    - Domain = $(0,\infty)$
- Gamma
    - Domain = (γ,∞), no constraints on γ
- Inverse Gaussian
    - Domain = (γ,∞), no constraints on γ
- Inverted Weibull
    - Domain = (γ,∞), no constraints on γ
- Log-Laplace
    - Domain = (γ,∞), no constraints on γ
- Log-logistic
    - Domain = (γ,∞), no constraints on γ
- Lognormal
    - Domain = (γ,∞), no constraints on γ, known for heavy tail
- Pareto
    - Domain = (γ,∞), $\gamma>0$
- Pearson type V
    - Domain = (γ,∞), no constraints on γ

- Pearson type VI
    - Domain = (γ,∞), no constraints on γ
- Random walk
    - Domain = (γ,∞), no constraints on γ
- Rayleigh
    - Domain = (γ,∞), no constraints on γ
- Wald
    - Domain = (γ,∞), no constraints on γ
- Weibull,
    - Domain = (γ,∞), no constraints on γ

3.2.3 Distributions bounded on both sides

The final five distributions available are bounded on both sides, with bounds estimated from the data. This minimises the risk of results wrongly indicating a positive probability of a counter-intuitively signed coefficient, while still allowing for such values to manifest themselves, for example due to data or model irregularities (c.f. Hess et al., 2005a).

- Beta
    - Can be symmetrical or asymmetrical, in which case it can be left-skewed or right-skewed, can be bi-modal
- Johnson SB
    - Can be symmetrical or asymmetrical, in which case it can be left-skewed or right-skewed, can be bi-modal, moments extremely complicated, generally simulated
- Power function
    - Can be flat, linear, or left or right-skewed
- Triangular
    - Can be symmetrical or asymmetrical, in which case it can be left-skewed or right-skewed
- Uniform
    - Equal probability for all values in domain

4. Results

After fitting a distribution to a dataset, EF provides statistics on quality of fit, including the mean difference in the density function between the true and target distribution, the maximum difference between the density functions, and the likelihood-function value. This latter measure gives the probability of obtaining the data from the given distribution, where with density f(x), and observed points $x_1,\ldots,x_N$, the likelihood is given by $f(x_1)\cdot\ldots\cdot f(x_N)$. The software additionally provides the moments of the sample and the target distribution (where possible to compute), and the percentile tables.

Finally, EF computes three goodness-of-fits tests on the basis of the true and estimated density function, namely the Anderson-Darling Test, the Kolmogorov-Smirnov Test, and the Equal-Probable Chi-Square Test. The former two tests have the problem that they cannot be computed for all the distributions used in the present analysis. Although hampered by other shortcomings, the Equal-Probable Chi-Square test faces no restrictions in terms of applicable distributions, and as such, was used in the present analysis.

Aside from the retrieval of the mean and variance of the distribution, the difference in the density function, and the overall likelihood function, it is of interest to look at the performance of the different distributions in the tails of the distribution. This applies specifically to the share of observations below a given value, for example observations that are not of the expected sign. EF is unable to produce these directly, so that a substitute was used, in the form of the difference between the true and fitted distribution in the values for the lower respectively upper octiles of the distribution.

Additionally, two non-linear measures were defined, emphasising the tails of the distribution. These measures use a weighted sum between the differences of different percentiles. In the first case, the measure is defined as:

$$W_A = \sum_{i=1,..7} \left| f(p_{i,t}) - f(p_{i,f}) \right| (1 + \frac{|50 - p_i|}{100})^2 , \qquad \ldots[4]$$

where

pi = Percentile

with:

- i=1: minimum
- i=2: lower octile
- i=3: lower quartile
- i=4: median
- i=5: upper quartile
- i=6: upper octile
- i=7: maximum

f(pi,t) = Value of the target distribution percentile pi

f(pi,f) = Value of the fitted distribution percentile pi

The inclusion of the extremes in this measure can cause problems in the presence of large outliers, such that an additional version is defined without the extreme values, as:

$$W_B = \sum_{i=2,..6} \left| f(p_{i,t}) - f(p_{i,f}) \right| (1 + \frac{|50 - p_i|}{100})^2 \qquad \ldots[5]$$

We now look at the actual results obtained in the fitting exercises. The first group of tests includes the full set of distributions, while in the second group, which additionally uses segmentation by true distributions, only the best-performing target distributions are included.

4.1 All distributions

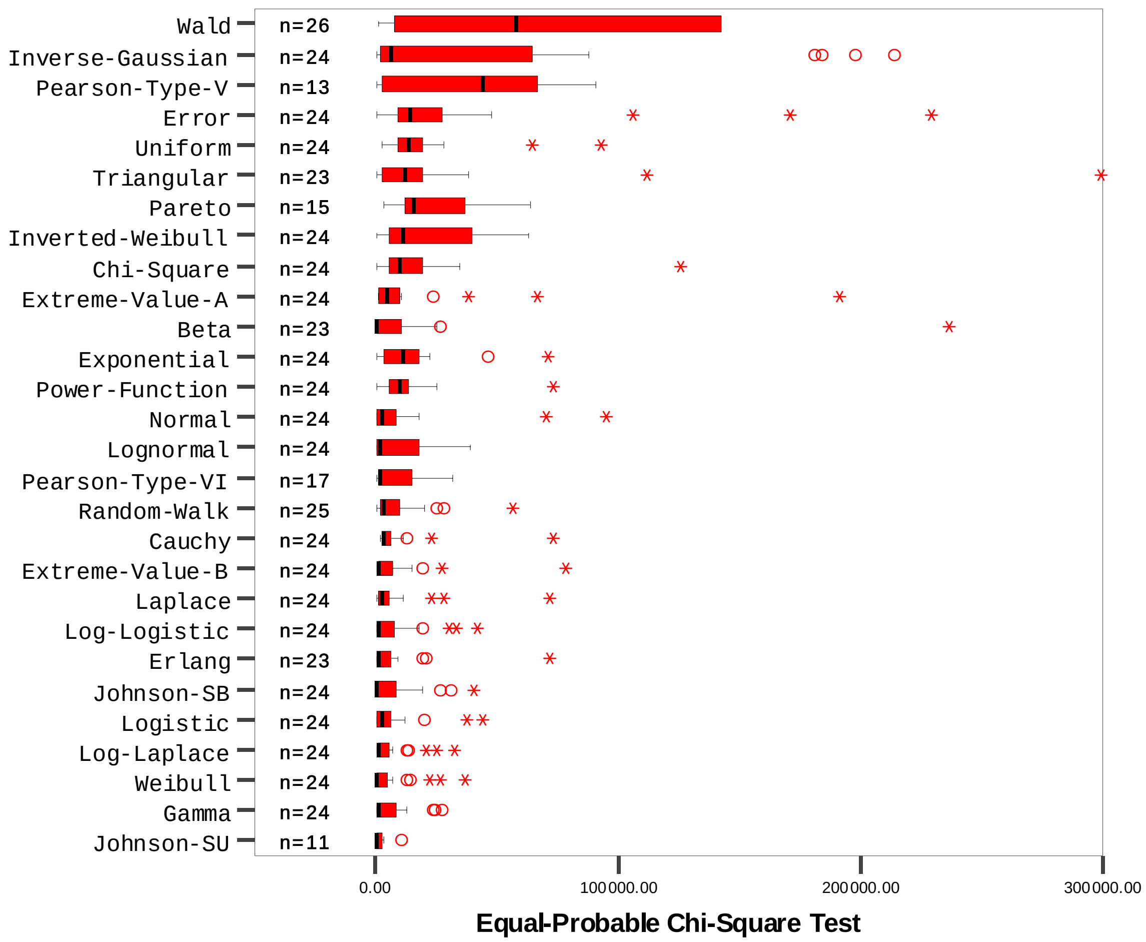
The sets of comparisons presented here use the full group of target distributions, and the various measures are calculated across all experiments in which the specific distributions were used. For most distributions, this number was equal to 24, i.e. the distributions were used in all experiments. However, it was occasionally not possible to fit a given distribution, such that the number varies. This is especially the case for the Johnson S_U distribution, which was used in only 11 experiments, as well as the two Pearson distributions, and the Pareto distribution. As such, the reliability of the results for these distributions is somewhat lower, especially for the S_U distribution. Additionally, in some of the comparisons, the number is lower as the specific measure of performance could not be evaluated. Finally, the number of experiments for the Wald distribution is 26, as two different fitting approaches were used in two experiments. Box-plots are used for the comparisons, to give an indication of the spread of performance across experiments as well as across distributions.

4.1.1 Equal chi-squared test

The first measure used in the comparison of the different target distributions, across all 24 experiments, is the equal chi-squared test. Despite restrictions with this method (given the missing guidance on optimal interval size), the measure gives an indication of the performance of the different distributions.

The results are summarised in Figure 1, with distributions ranked according to the mean test-value across experiments. The analysis shows some variation in performance, with very poor performance by a set of distributions. Two commonly used distributions, the Triangular and the Lognormal, show great variation in performance, although the mean performance of the Lognormal distribution is better than that of the Normal distribution. Good performance is obtained by a set of distributions (essentially everything below Cauchy), including the two Johnson distributions, the Gamma, and the Weibull.

Figure 1 Equal-Probable Chi-Square test, all experiments and all distributions

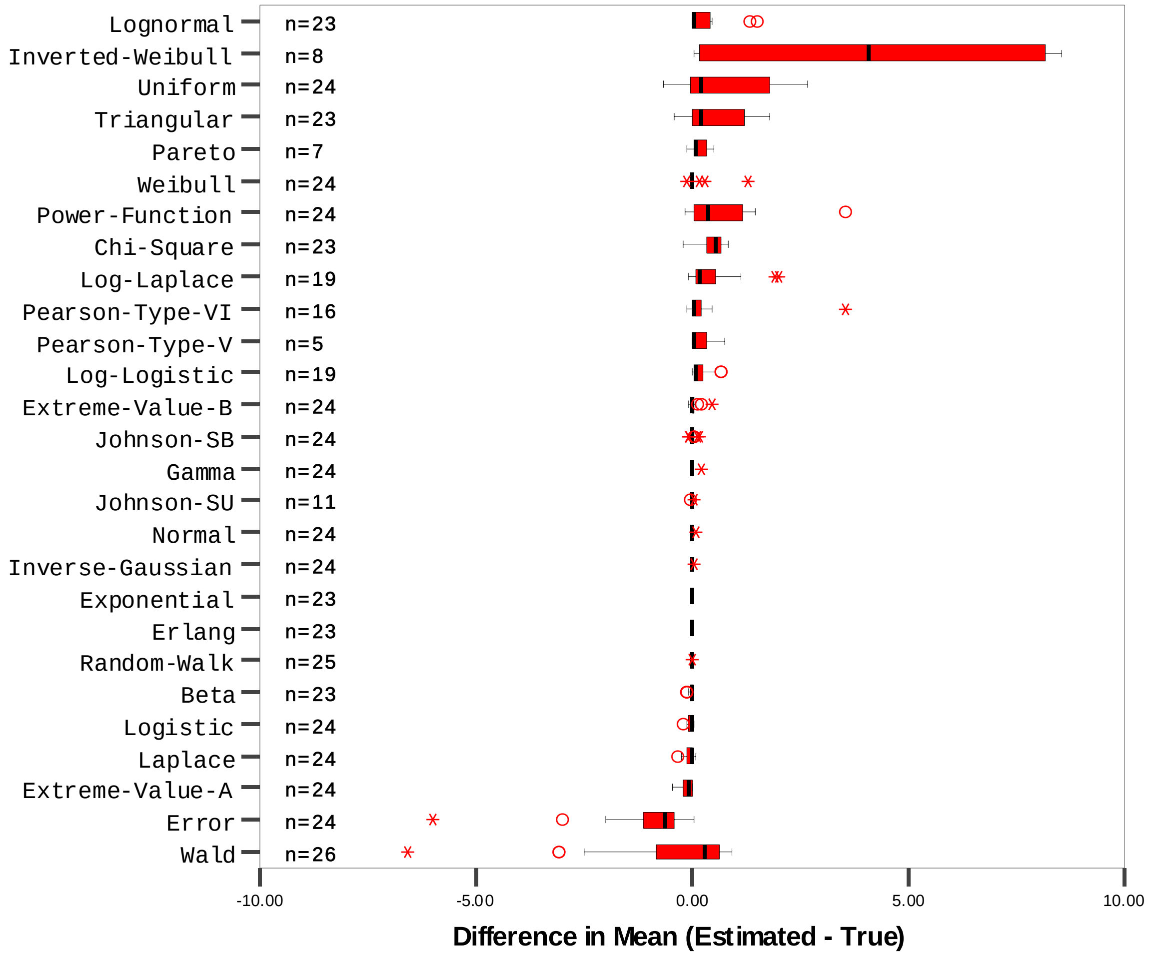


4.1.2 Recovery of true mean

The recovery of the mean of the true distribution is of crucial importance in Mixed Logit analysis. As such, the comparison between the mean values produced by true and target distributions is an important measure of performance. The Cauchy distribution was not included in this comparison (moments not available), while in addition to the distributions listed in the introduction to section 4.1., only a limited number of comparisons were possible for the Inverted Weibull distribution.

The results are summarised in Figure 2, with distributions sorted by the mean difference between the means of the true and target distributions. The analysis shows good overall performance by the distributions, with some exceptions, with distributions leading to over-, respectively underestimated means. Exceptions in the first group include the Lognormal, which, although offering good median performance, is ranked poorly because of some huge outliers. Other examples in this group include the Inverted Weibull, Uniform, and Triangular distributions, with the Error and Wald distributions leading to bias to the left of the true mean value.

Figure 2 Recovery of mean values, all experiments and all distributions

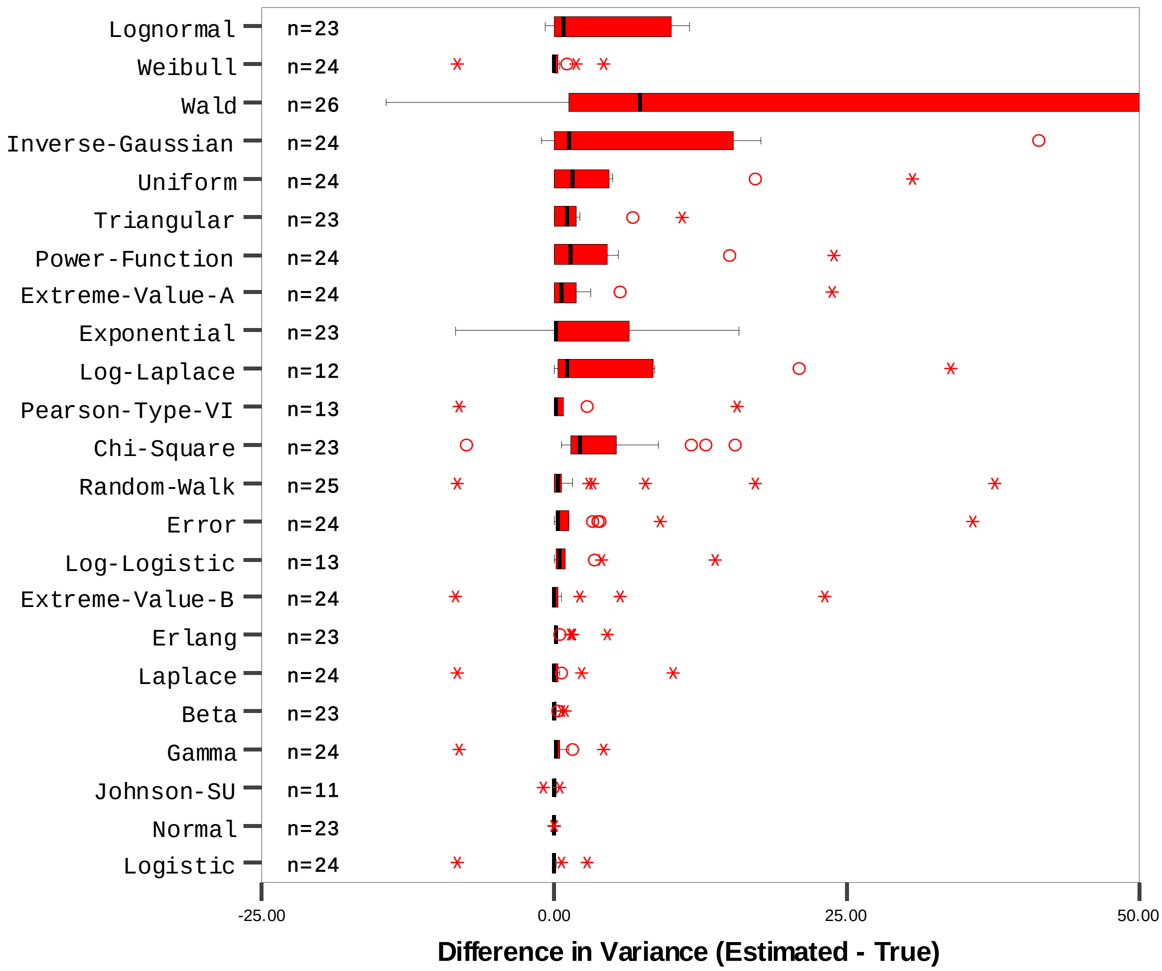


4.1.3 Recovery of true variance

In the analysis of the recovery of the variance of the true distribution, five distributions had to be excluded from the analysis, as the variance was not available; these are the Pearson Type V, Pareto, Johnson S_B, Inverted Weibull, and Cauchy distributions. In addition to the distributions listed in the introduction to section 4.1., only a limited number of comparisons could be made for the Log-Logistic and Log-Laplace distributions.

The results are summarised in Figure 3, ranked by the mean difference in the variance. Most distributions overestimate the true variance, but while this overestimation is acceptable for about half of the distributions, problems exist for example in the case of the Lognormal (acceptable median performance, but some big outliers) and Triangular distributions.

Figure 3 Reproduction of variance, all experiments and all distributions

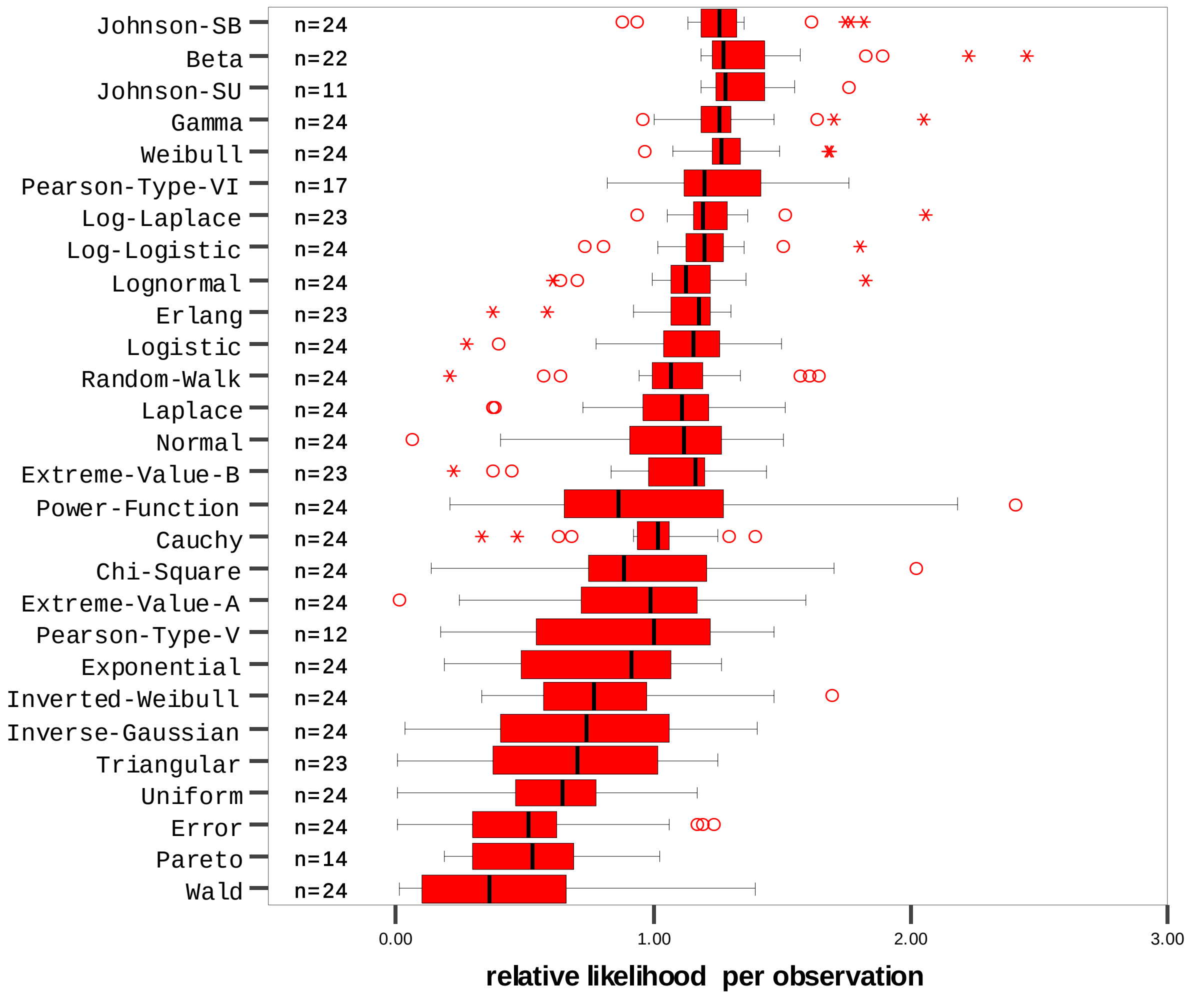


4.1.4 Likelihood

In the analysis of the performance in terms of likelihood, the relative likelihood was used. As such, for each experiment, the mean likelihood across distributions was calculated in each experiment, and the relative likelihood for each distribution (in relation to the mean likelihood) was used as the measure of performance. As such, a value larger than 1 indicates above average performance for that distribution.

The results are summarised in Figure 4. They show very good performance by the two Johnson distributions, the Beta and Gamma distributions and the Weibull distributions. For other distributions, such as the Pearson distributions, the performance is highly variable. This is also the case for the commonly used Lognormal and Normal distributions, where especially the latter sometimes offers better-than-average, and sometimes lower-than-average performance.

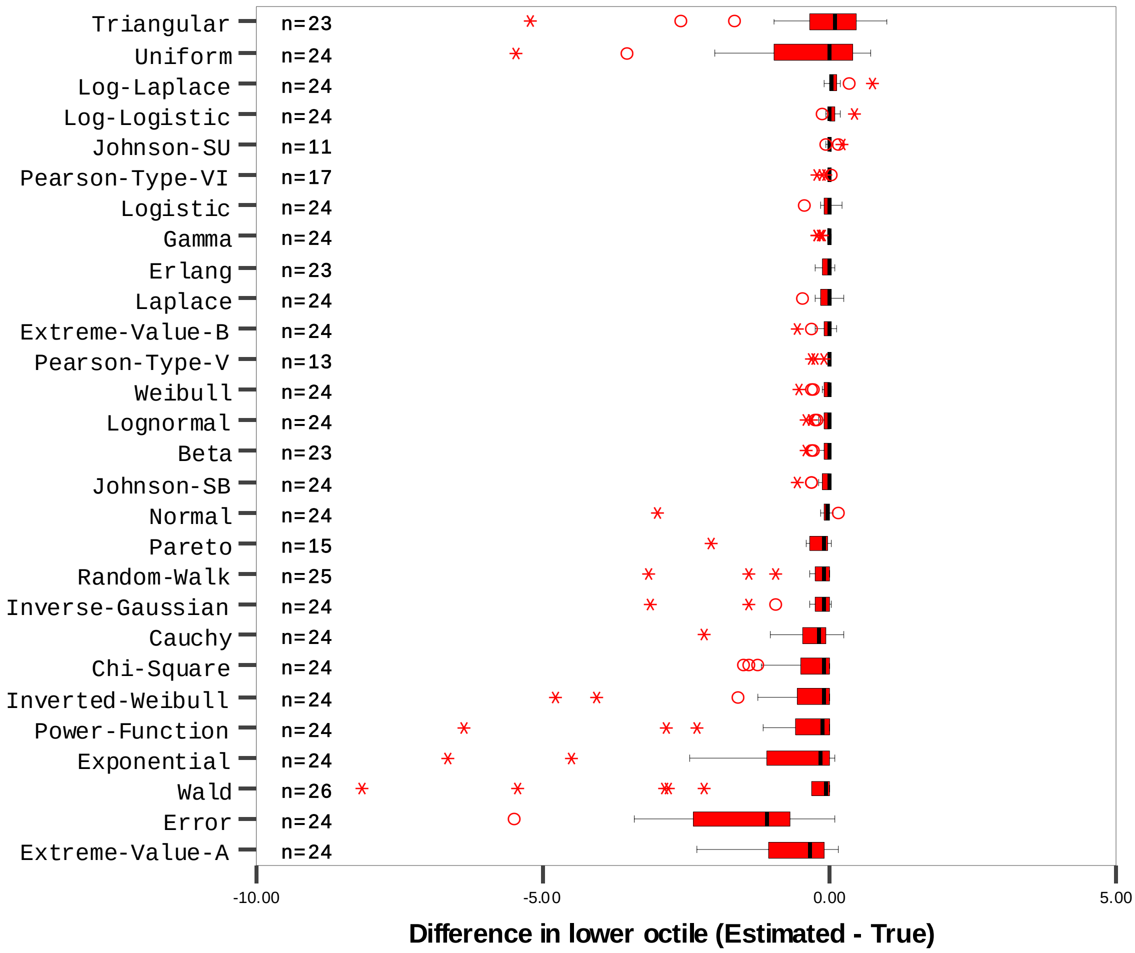
Figure 4 Relative likelihood performance, all experiments and all distributions



4.1.5 Lower Octile

The next part of the analysis looks at the recovery of the lower octile, with results shown in Figure 5. Again, some performance perform very well, notably the two Johnson distribution, and the Beta and Gamma, while others, namely the Triangular and Uniform, overestimate the lower octile, and others, most notably the Extreme Value A and Error distributions, underestimate the lower octile. The mean performance of the Normal is acceptable, though there are some outliers, with underestimated lower octiles. This is caused mainly by the experiments using true distributions bounded to the left, where the symmetric nature of the Normal can lead to problems with the lower limits, as stressed by Hess et al. (2005a).

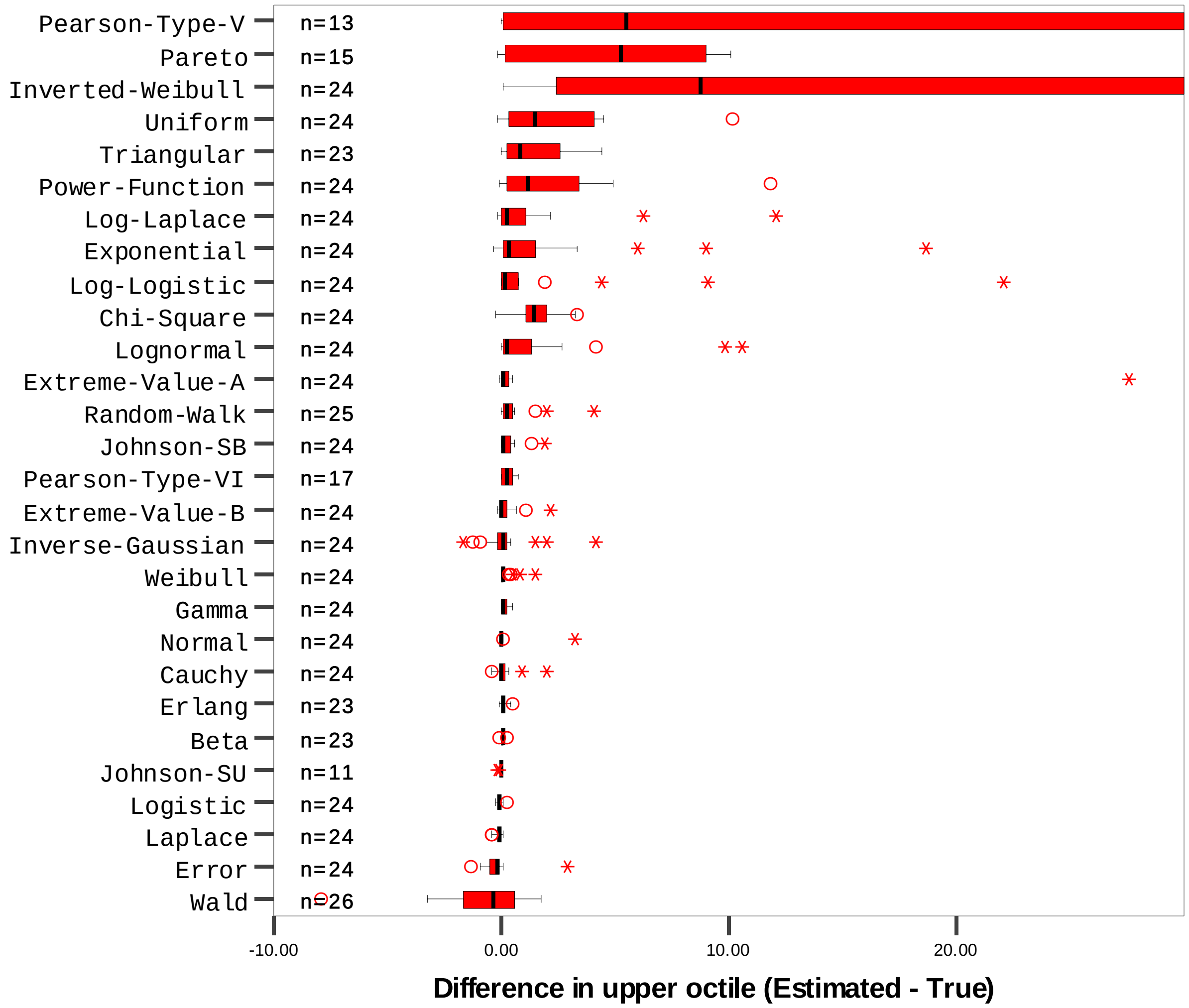
Figure 5 Recovery of lower octile, all experiments and all distributions



4.1.6 Upper Octile

A clearer picture emerges for the analysis using the recovery of the upper octile as the performance measure, as shown in Figure 6. Here, significant problems with overestimation are exhibited by the Pearson Type V, Pareto and Inverted Weibull distributions, while some problems with overestimated upper octiles (in some of the experiments), exist for Uniform, Triangular, and Lognormal distributions, amongst others. The Normal performs very well in these experiments, which may be a reflection of the use of true distributions with a longer tail to the right.

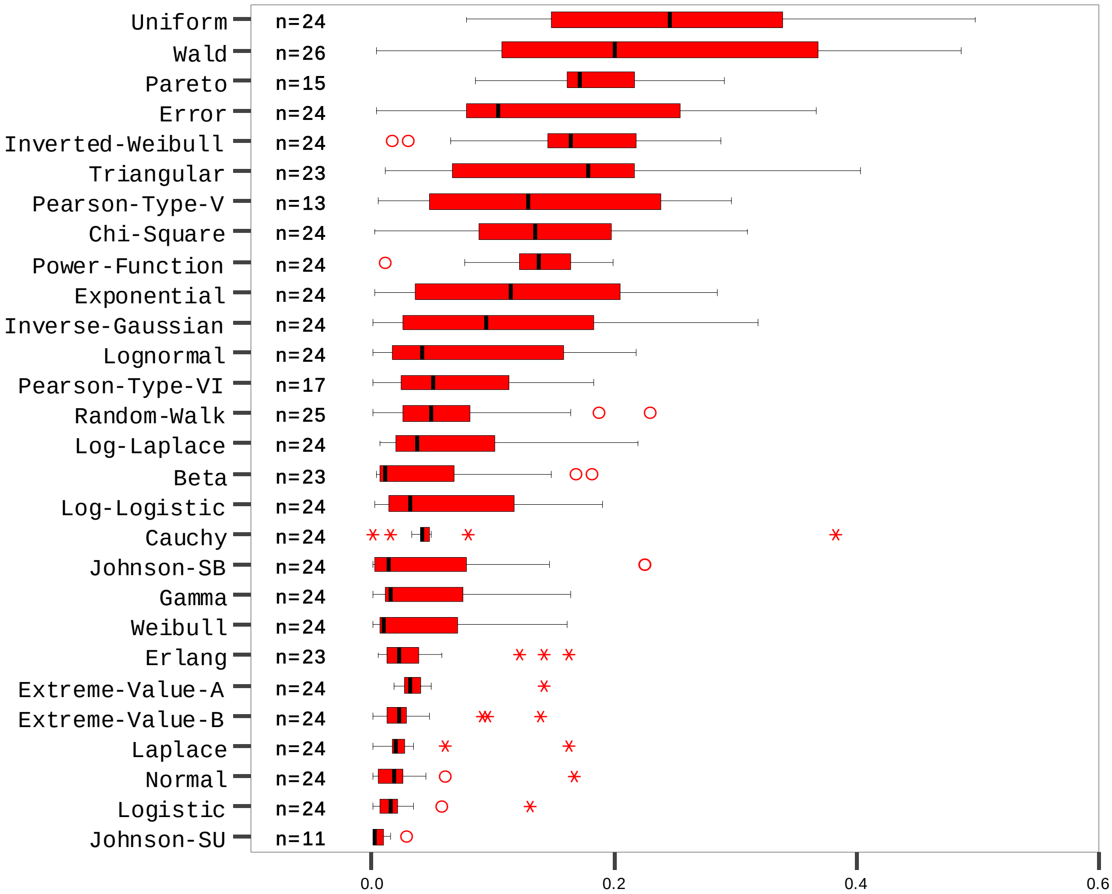
Figure 6 Recovery of upper octile, all experiments and all distributions



4.1.7 DF-mean

The results of the comparison using the mean difference between the true and target density function are summarised in Figure 7. They show good performance for a set of distributions, including the Johnson S_U, Logistic and Normal, with acceptable performance by the Gamma, Johnson S_B and Beta, amongst others.

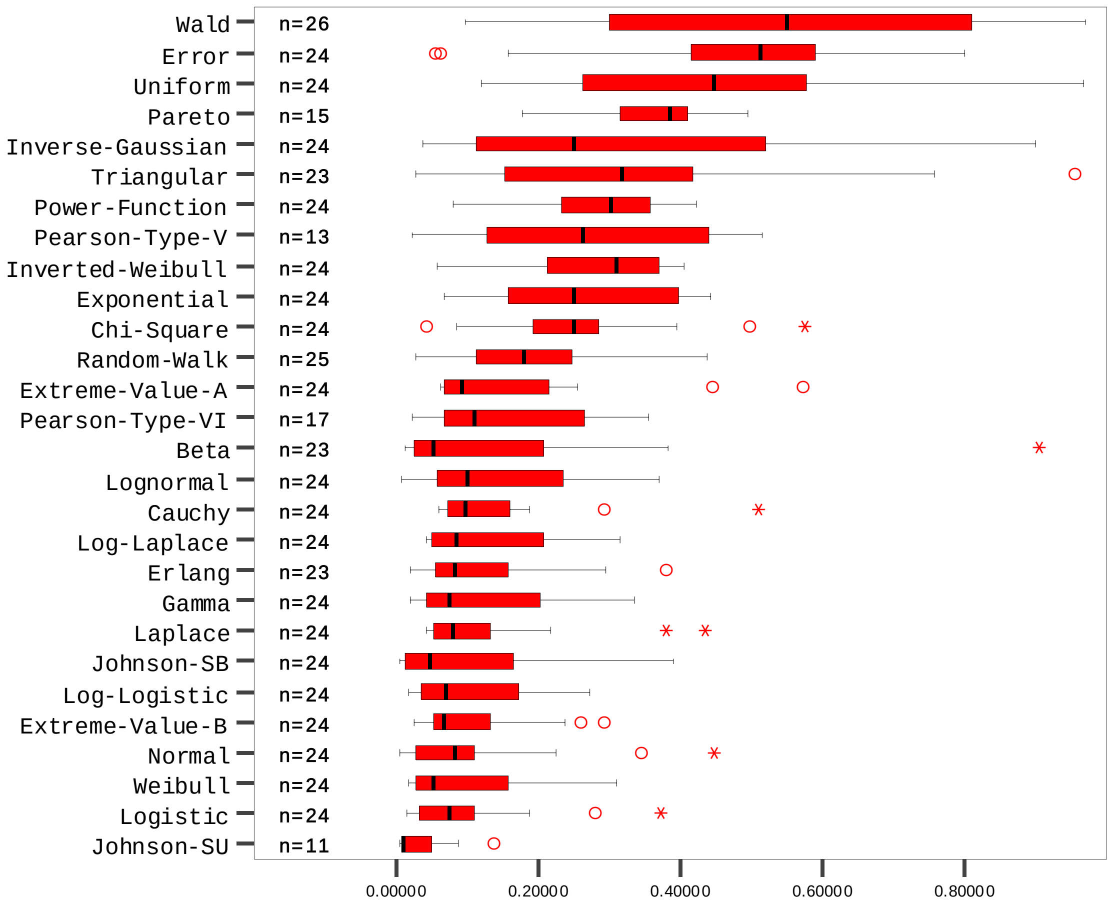
Figure 7 Mean difference in density function, all experiments and all distributions



4.1.8 DF-max

The results using the maximum difference between the true and target density functions are shown in Figure 8; they are largely consistent with those in section 4.1.7., and 4.1.1.

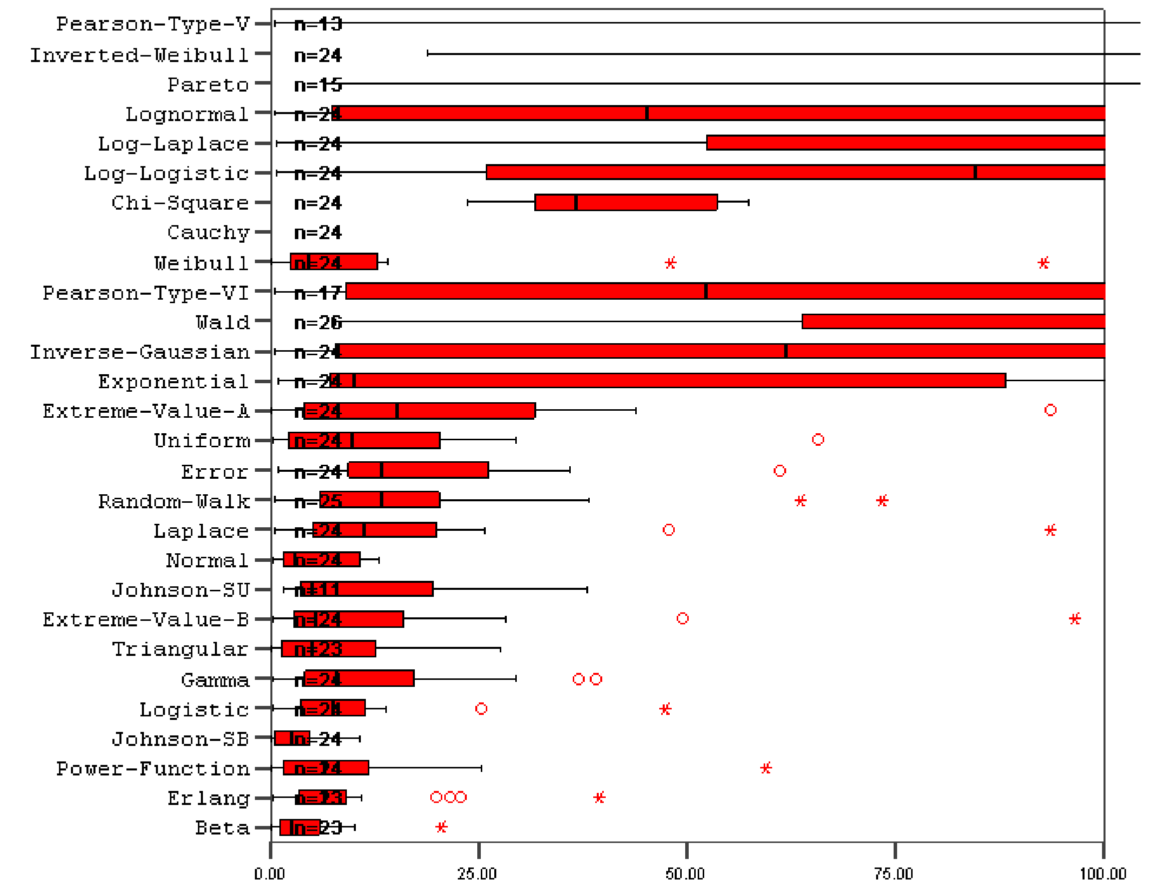
Figure 8 Maximum difference in density function, all experiments and all distributions



4.1.9 Weighted difference approach A

As expected, the use of the first weighted difference approach, given by equation [4] yields very high values for some of the distributions, showing problems with the extremes of the distributions. As shown in Figure 9, this is for example the case with the Lognormal distribution, highlighting the well-documented problems with the heavy tail for this distribution. With this measure, the best performance is obtained by the Beta distribution.

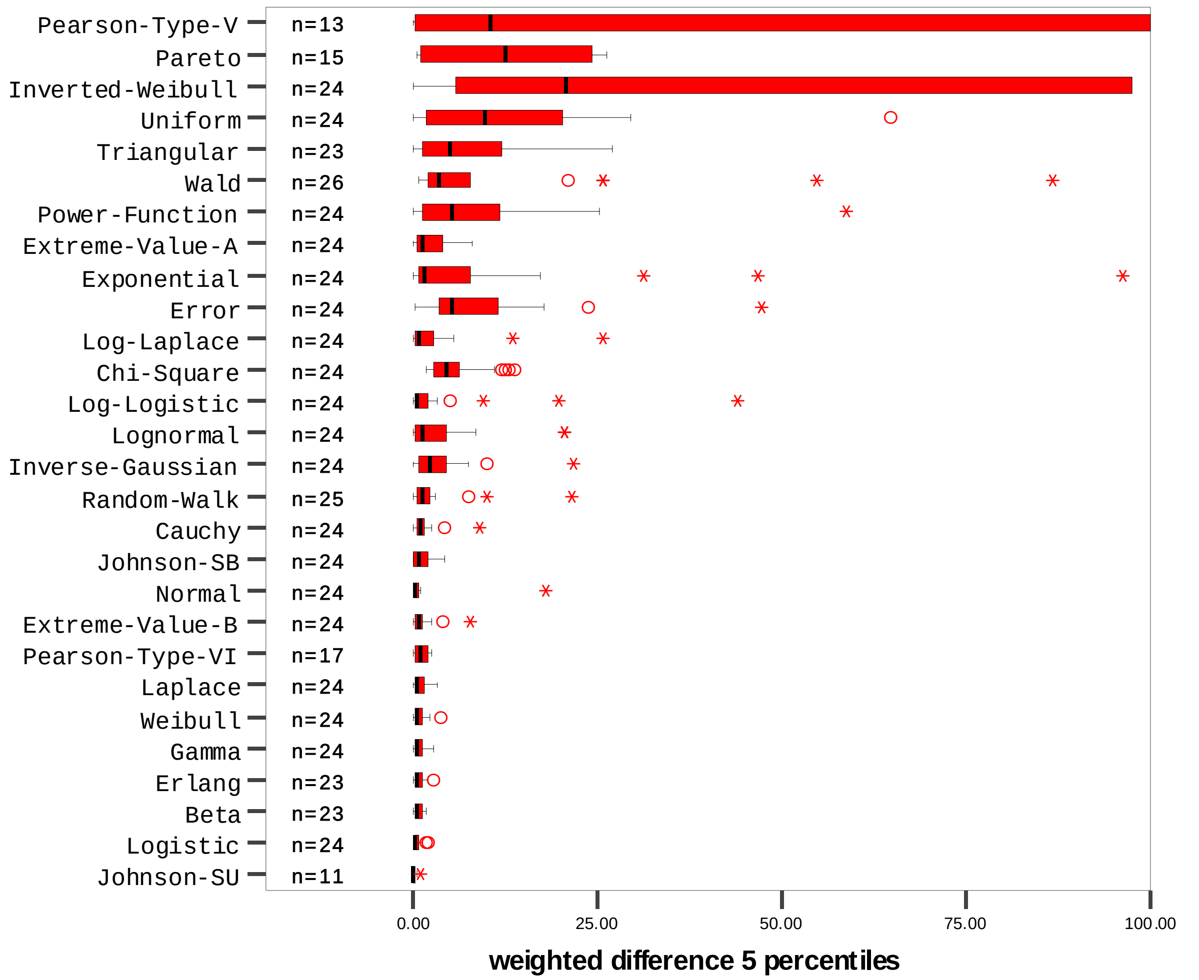
Figure 9 Weighted difference approach A, all experiments and all distributions



4.1.10 Weighted difference approach B

The results using the approach described by equation [5] are comparable to those obtained with equation [4], but with fewer extreme values, as highlighted in Figure 10. This is notably the case for the Lognormal distribution, which now offers better performance; this suggests that the problems with the heavy tails are principally due to the upper few percentiles, towards the upper end of the upper octile. The best performance is now offered by the Johnson S_U distribution.

Figure 10 Weighted difference approach B, all experiments and all distributions



4.1.11 Summary for first set of comparisons

The results of this first set of comparisons have shown some differences depending on the measure used. Nevertheless, some consistent results are obtained. These notably show good performance for the two Johnson distributions, and, depending on the measure used, also the Gamma, Beta and Erlang distributions. The Normal offers acceptable performance overall, while the well-documented problems with the upper tail of the Lognormal are also highlighted by some of the measures.

4.2 Best fitting distributions

The second part of the analysis uses only the best-fitting distributions, but additionally uses segmentation by dataset, thus showing the performance of the different target distributions on the separate true distributions.

We first summarise the results from section 1, by ranking the distributions according to the different criteria, as shown in Table 1. In this comparison, the absolute values were used for non-strictly signed measures, to avoid cancelling-out effects. In this way, left-ward bias is judged in the same way as right-ward bias, at least in this ranking.

Table 1: Overall ranking of distributions by different criteria

absolute difference in mean	absolute difference in variance	Log-Likelihood Function	absolute difference in lower octile	absolute difference in upper octile	DF-Mean	DF-Max	weighted difference 7 percentiles	weighted difference 5 percentiles
Erlang	Normal	Pearson-V	Pearson-VI	Joh.-SU	Joh.-SU	Joh.-SU	Beta	Joh.-SU
Exponential	Pearson-V	Pearson-VI	Gamma	Logistic	Logistic	Logistic	Erlang	Logistic
R-Walk	Beta	Joh.-SB	Pearson-V	Beta	Normal	Weibull	P.-Funct.	Beta
Inv.-Gau.	Joh.-SU	Weibull	Joh.-SU	Laplace	Laplace	Normal	Joh.-SB	Erlang
Normal	Erlang	Gamma	Log-Logis.	Erlang	E-Val.-B	E-Val.-B	Logistic	Gamma
Gamma	Logistic	Log-Lapl.	Lognormal	Gamma	E-Val.-A	Log-Logis.	Gamma	Weibull
Joh.-SU	Gamma	Log-Logis.	Erlang	Normal	Erlang	Joh.-SB	Triangular	Laplace
Joh.-SB	Laplace	Beta	Logistic	Weibull	Weibull	Laplace	E-Val.-B	Pearson-VI
Beta	E-Val.-B	R-Walk	Beta	Cauchy	Gamma	Gamma	Joh.-SU	E-Val.-B
Logistic	Log-Logis.	Lognormal	Weibull	E-Val.-B	Joh.-SB	Erlang	Normal	Normal
E-Val.-B	Error	Logistic	Joh.-SB	Pearson-VI	Cauchy	Log-Lapl.	Laplace	Joh.-SB
Laplace	R-Walk	Erlang	E-Val.-B	Joh.-SB	Log-Logis.	Cauchy	R-Walk	Cauchy
Log-Logis.	Chi-Square	Laplace	Log-Lapl.	Error	Beta	Lognormal	Error	R-Walk
Pearson-V	Log-Lapl.	E-Val.-B	Laplace	R-Walk	Log-Lapl.	Beta	Uniform	Inv.-Gau.
Pearson-VI	Pearson-VI	Cauchy	Normal	Inv.-Gau.	R-Walk	Pearson-VI	E-Val.-A	Lognormal
Log-Lapl.	Exponential	Normal	Pareto	E-Val.-A	Pearson-VI	E-Val.-A	Exponential	Log-Logis.
Chi-Square	E-Val.-A	Pareto	R-Walk	Lognormal	Lognormal	R-Walk	Inv.-Gau.	Chi-Square
E-Val.-A	P.-Funct.	P.-Funct.	Inv.-Gau.	Chi-Square	Inv.-Gau.	Chi-Square	Wald	Log-Lapl.
Error	Triangular	Chi-Square	Chi-Square	Log-Logis.	Exponential	Exponential	Pearson-VI	Error
P.-Funct.	Uniform	Joh.-SU	Cauchy	Exponential	P.-Funct.	Inv.-Wei.	Weibull	Exponential
Weibull	Inv.-Gau.	Exponential	Inv.-Wei.	Wald	Chi-Square	Pearson-V	Cauchy	E-Val.-A
Pareto	Wald	Inv.-Wei.	P.-Funct.	Log-Lapl.	Pearson-V	P.-Funct.	Chi-Square	P.-Funct.
Wald	Weibull	E-Val.-A	Triangular	P.-Funct.	Triangular	Triangular	Log-Logis.	Wald
Triangular	Lognormal	Inv.-Gau.	Exponential	Triangular	Inv.-Wei.	Inv.-Gau.	Log-Lapl.	Triangular
Uniform		Uniform	Wald	Uniform	Error	Pareto	Lognormal	Uniform
Inv.-Wei.		Error	Error	Inv.-Wei.	Pareto	Uniform	Pareto	Inv.-Wei.
Lognormal		Triangular	Uniform	Pareto	Wald	Error	Inv.-Wei.	Pareto
		Wald	E-Val.-A	Pearson-V	Uniform	Wald	Pearson-V	Pearson-V

Although there are some differences across criteria, some overall conclusions are possible, as already indicated in section 4.1.11. It seems like the Triangular is lacking flexibility, that the Lognormal offers terrible performance for mean and variance, as expected, and that the Normal does very well for mean and variance, while it performs poorly for the lower octile, especially in the presence of left-bounded distributions, as we will see later.

From the original set of distributions, the following thirteen were retained. They include the best-fitting distributions, along with some of the most-commonly used ones. Additionally, they give a fairly representative selection from the three groups of distributions, although the balance has somewhat shifted towards unbounded distributions, when compared to the frequencies in the list set out in sections 3.2.1. and 3.2.2.

- ***Retained unbounded distributions***
  - Extreme Value Type B
  - Johnson S_U
  - Laplace
  - Logistic
  - Normal
- ***Retained left-bounded distributions***
  - Erlang
  - Gamma
  - Log-Logistic
  - Lognormal
  - Pearson Type VI
  - Weibull
- ***Retained distributions bounded on both sides***
  - Beta
  - Johnson S_B

It is of interest to look at the frequencies for the different distributions in the specific experiments, as summarised in Table 3. They show low representation for the Johnson SU and also for the Pearson Type VI, although the latter is used in at least one experiment in each group.

Table 2 Experiments by distribution and data-set

Target distribution	Dataset: Normal	Normal truncated	Normal censored	Normal with mass	S_B	Lognormal	**Total**
Beta	4	4	4	4	4	3	**23**
Erlang	4	4	4	4	4	3	**23**
Extreme-Value-B	4	4	4	4	4	4	**24**
Gamma	4	4	4	4	4	4	**24**
Johnson-SB	4	4	4	4	4	4	**24**
Johnson-SU	3	0	1	4	0	3	**11**
Laplace	4	4	4	4	4	4	**24**
Log-Logistic	4	4	4	4	4	4	**24**
Logistic	4	4	4	4	4	4	**24**
Lognormal	4	4	4	4	4	4	**24**
Normal	4	4	4	4	4	4	**24**
Pearson-Type-VI	1	4	4	2	2	4	**17**
Weibull	4	4	4	4	4	4	**24**
Total	**48**	**48**	**49**	**50**	**46**	**49**	**290**

Given the experiences from section 3, only three measures of performance were retained for the analysis using the more limited set of distributions, namely the mean and maximum difference in the density function, and the weighted approach using 5 percentile points (equation [5]).

With a few exceptions (notably the Lognormal), the distributions used in this part of the analysis give relatively stable performance, such that only the mean values were used in the comparison, where an additional segmentation by dataset was however used (reducing the number of values to a maximum of 4 in each average). In each case, the results also include the mean performance across datasets for each distribution, taking into account the number of experiments in each group.

4.2.1 DF-mean

The first analysis in this part of the paper looks at the mean difference between the true and target density functions, with results summarised in Table 3 and Figure 11.

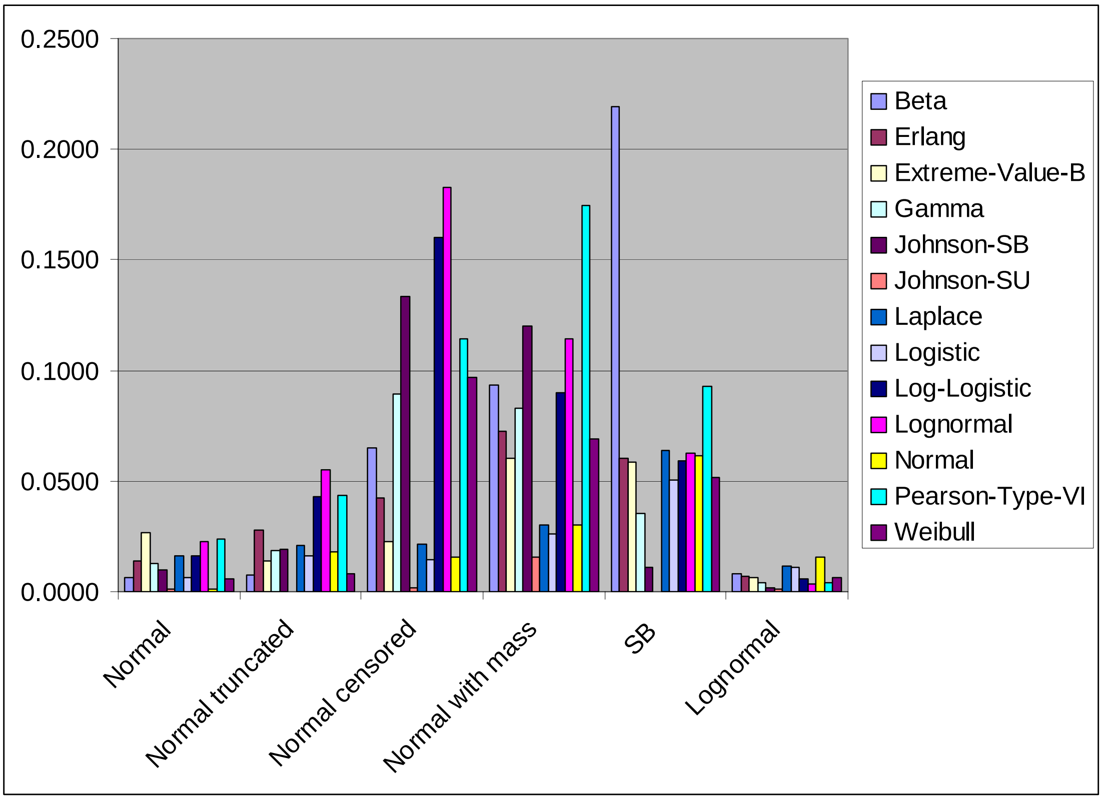
Only a very limited number of comparisons were available for the Johnson S_U distribution, but those available show very good performance, with by far the best performance on average. Surprisingly, it also seems to outperform the Lognormal distribution on the actual Lognormal data (along with the Johnson S_B), which could be an indication of high sampling error in the data.

Very good performance is also obtained with the Logistic distribution, and to a lesser extent the Laplace distribution. The performance of the Normal distribution is also generally good, though it struggles somewhat when the true distribution is S_B or Lognormal, as do the Logistic and Laplace. Finally, while the Beta works well in some cases, it seems to have major problems when the true distribution is S_B. Furthermore, all distributions have problems with true distributions with a mass at a point (either as an endpoint or in between), though some are able to cope with it better than others, above all the Logistic and Laplace, and, where available, the Johson S_U.

Table 3 Performance for best-fitting distributions, mean difference in density function

Target distribution	Dataset: Normal	Normal truncated	Normal censored	Normal with mass	S_B	Lognormal	Weighted Mean
Beta	0.0061	0.0075	0.0651	0.0932	0.2190	0.0084	**0.0691**
Erlang	0.0138	0.0278	0.0422	0.0723	0.0605	0.0072	**0.0386**
Extreme-Value-B	0.0270	0.0140	0.0226	0.0603	0.0585	0.0064	**0.0326**
Gamma	0.0130	0.0187	0.0891	0.0830	0.0353	0.0043	**0.0421**
Johnson-SB	0.0099	0.0190	0.1333	0.1199	0.0108	0.0018	**0.0512**
Johnson-SU	0.0014		0.0019	0.0156		0.0009	**0.0034**
Laplace	0.0165	0.0210	0.0217	0.0303	0.0637	0.0113	**0.0281**
Logistic	0.0062	0.0160	0.0147	0.0262	0.0506	0.0109	**0.0212**
Log-Logistic	0.0162	0.0432	0.1601	0.0899	0.0593	0.0060	**0.0649**
Lognormal	0.0227	0.0550	0.1827	0.1145	0.0629	0.0038	**0.0766**
Normal	0.0012	0.0181	0.0154	0.0301	0.0614	0.0159	**0.0240**
Pearson-Type-VI	0.0239	0.0436	0.1142	0.1745	0.0925	0.0042	**0.0786**
Weibull	0.0057	0.0084	0.0969	0.0691	0.0519	0.0066	**0.0412**

Figure 11 Performance for best-fitting distributions, mean difference in density function



4.2.2 DF-max

The results for the analysis using the maximum difference in the density function as the measure of performance are summarised in Table 4 and Figure 12. It is of interest to look at the correlation between these results and those obtained for the mean differences in the density function. The overall correlation (looking at the weighted mean performance) is quite high, at 0.77. When looking at the correlation in separate groups, i.e. different true distributions, as in Table 5, it becomes visible that this applies only to a lesser degree for the Lognormal and censored Normal datasets. The distribution-specific correlation-levels also show high correlation in most cases. The correlation is lower for the Normal distribution, Laplace, Logistic and Extreme-Value-B, while it is surprisingly negative for the Logistic, suggesting that good mean performance is linked to poor maximum error in this case, and vice-versa. Overall, for stability reasons, distributions with high levels of correlation between the two measures are preferable, which again speaks in favour of the Johnson distributions, and the Gamma and Beta distributions, where the problems of the latter with some target distributions however causes concerns. The problems of the Laplace, Logistic and Normal distributions with the S_B and Lognormal distributions persist.

Table 4 Performance for best-fitting distributions, maximum difference in density function

Target distribution	Dataset: Normal	Normal truncated	Normal censored	Normal with mass	S_B	Lognormal	Weighted Mean
Beta	0.0216	0.0270	0.2067	0.2139	0.3030	0.0604	**0.1422**
Erlang	0.0448	0.0828	0.1768	0.1797	0.1624	0.0704	**0.1216**
Extreme-Value-B	0.0617	0.0457	0.0929	0.1373	0.1567	0.1180	**0.1014**
Gamma	0.0431	0.0523	0.2325	0.1986	0.1149	0.0702	**0.1207**
Johnson-SB	0.0305	0.0549	0.2705	0.2232	0.0325	0.0139	**0.1082**
Johnson-SU	0.0080		0.0064	0.0819		0.0144	**0.0186**
Laplace	0.0463	0.0967	0.0645	0.0917	0.1849	0.2099	**0.1116**
Logistic	0.0192	0.0772	0.0549	0.0895	0.1403	0.1773	**0.0894**
Log-Logistic	0.0376	0.0700	0.2198	0.1614	0.1155	0.0184	**0.1075**
Lognormal	0.0637	0.1054	0.2971	0.2283	0.1214	0.0201	**0.1445**
Normal	0.0073	0.0703	0.0605	0.0909	0.1559	0.2183	**0.0954**
Pearson-Type-VI	0.0675	0.0962	0.2594	0.3398	0.1268	0.0483	**0.1610**
Weibull	0.0208	0.0271	0.2093	0.1649	0.0886	0.0544	**0.0959**

Figure 12 Performance for best-fitting distributions, maximum difference in density function

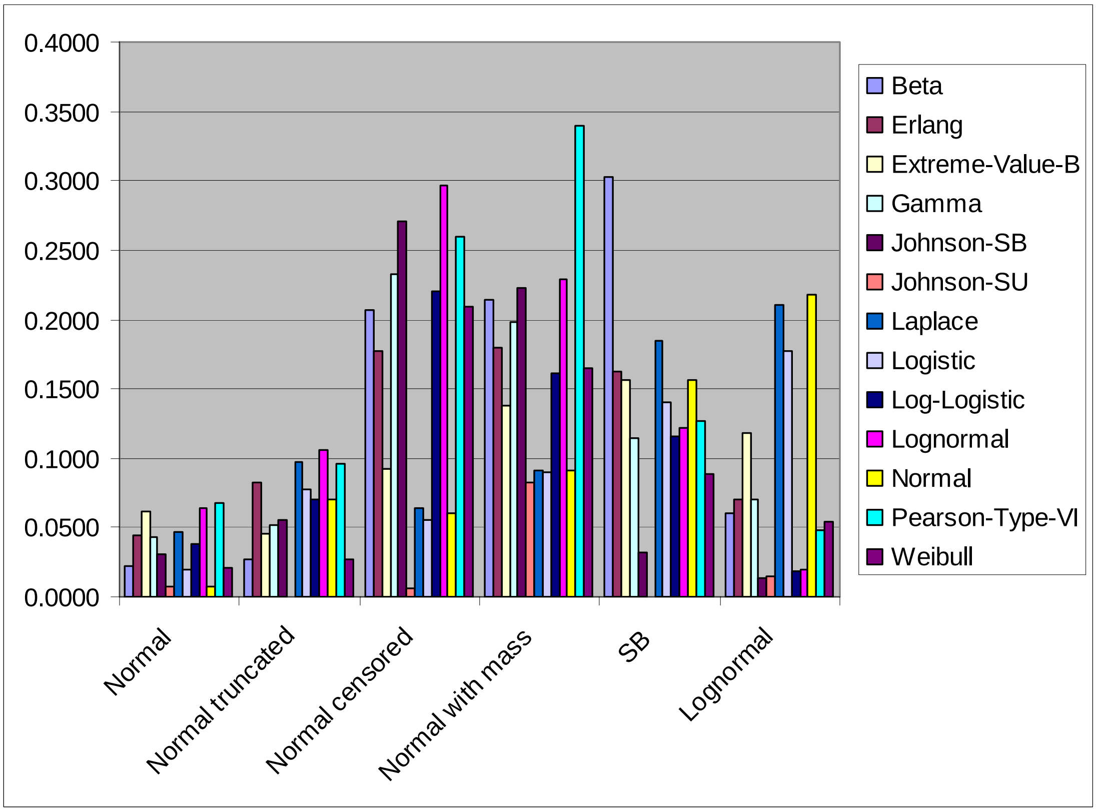


Table 5 Correlation between mean and maximum difference in density function, segmentation by true distribution

Normal	Normal truncated	Normal censored and mass point	Normal with mass point	Sb	Lognormal
0.887134	0.807444	0.563389	0.86657	0.885776	0.679837

Table 6 Correlation between mean and maximum difference in density function, segmentation by target distribution

Johnson-SB	0.99759
Lognormal	0.99142
Johnson-SU	0.98877
Gamma	0.97176
Weibull	0.97028
Pearson-Type-VI	0.96140
Beta	0.92244
Erlang	0.89678
Extreme-Value-B	0.67717
Normal	0.47738
Laplace	0.33554
Log-Logistic	0.31148
Logistic	-0.30801

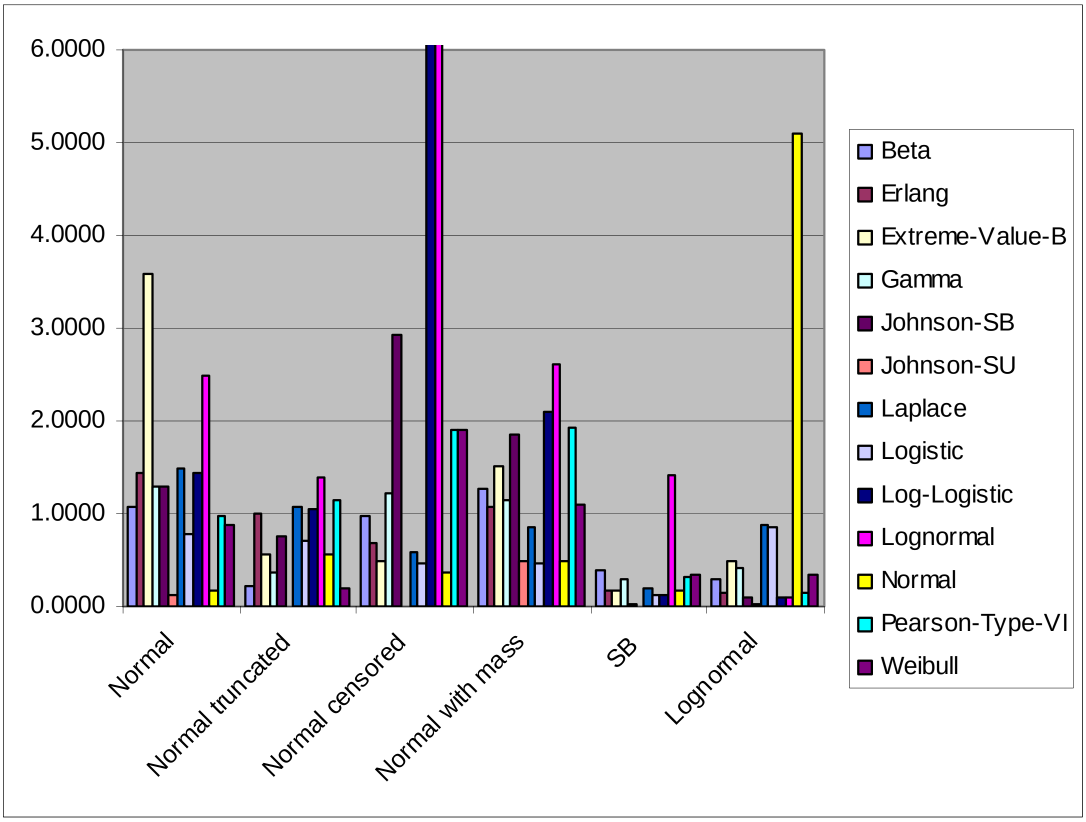
4.2.3 Weighted difference, 5-points

The final part of the analysis looks at the performance obtained when using the weighted difference defined by equation [5], with results summarised in Table 7, and Figure 13, where the Figure had to be censored above, due to the very poor performance of the Log-Logistic and Lognormal on the Normal dataset. The results repeat the problems that the Normal has in the case of the Lognormal dataset. A closer analysis shows this to be due largely to the poor performance for the lower extreme value, the lower octile, and to lesser degree the lower quantile. This is especially the case in the example using a mean close to zero, with a high standard deviation. This validates claims and empirical results by Hess et al. (2005a), who note that the Normal may in this case falsely indicate a large negative share for the distribution.

Table 7 Performance for best-fitting distributions, weighted difference using 5 percentiles

Target distribution	Normal	Normal truncated	Normal censored	Normal with mass	S_B	Lognormal	**Weighted Mean**
	Dataset						
Beta	1.0854	0.2229	0.9773	1.2648	0.3819	0.2910	**0.7218**
Erlang	1.4284	1.0043	0.6793	1.0848	0.1746	0.1522	**0.7801**
Extreme-Value-B	3.5961	0.5488	0.4971	1.5198	0.1732	0.4813	**1.1645**
Gamma	1.2984	0.3683	1.2136	1.1357	0.2812	0.4267	**0.8030**
Johnson-SB	1.2947	0.7608	2.9349	1.8631	0.0214	0.0905	**1.2075**
Johnson-SU	0.1109		0.0000	0.4985		0.0222	**0.1089**
Laplace	1.4838	1.0786	0.5969	0.8624	0.1986	0.8686	**0.8473**
Logistic	0.7745	0.7073	0.4603	0.4596	0.1315	0.8543	**0.5520**
Log-Logistic	1.4371	1.0393	18.7060	2.0883	0.1263	0.1088	**4.0832**
Lognormal	2.4975	1.3993	13.9130	2.6118	1.4114	0.1085	**3.8112**
Normal	0.1818	0.5612	0.3598	0.4852	0.1691	5.0859	**0.9690**
Pearson-Type-VI	0.9872	1.1347	1.9039	1.9373	0.3127	0.1364	**1.1092**
Weibull	0.8736	0.1893	1.9003	1.1051	0.3398	0.3424	**0.8113**

Figure 13 Performance for best-fitting distributions, weighted difference using 5 percentiles



4.2.4 Summary

As a final summary, the measure from section 4.3.3. was used to calculate the ranking of the different distributions in the different experiments, as shown in Table 8. Overall, these are consistent with the results found when using the measures in section 4.3.1. and 4.3.2., though with some differences, notably for the Beta distribution. It should be noted that the seemingly better performance by the Johnson S_U than by the Normal on Normal data is down to lower number of experiments with Johnson S_U, and sampling error. A similar reasoning applies when the true distribution is Lognormal.

Table 8 Ranking of different distributions, using 5-point weighted difference

Dataset

Target distribution	Normal	Normal truncated	Normal censored	Normal with mass	S_B	Lognormal	**Weighted Mean**
Beta	6	2	7	8	11	7	**3**
Erlang	9	8	6	5	6	6	**4**
Extreme-Value-B	13	4	4	9	5	10	**10**
Gamma	8	3	8	7	8	9	**5**
Johnson-SB	7	7	11	10	1	2	**11**
Johnson-SU	1		1	3		1	**1**
Laplace	11	10	5	4	7	12	**7**
Logistic	3	6	3	1	3	11	**2**
Log-Logistic	10	9	13	12	2	4	**13**
Lognormal	12	12	12	13	12	3	**12**
Normal	2	5	2	2	4	13	**8**
Pearson-Type-VI	5	11	10	11	9	5	**9**
Weibull	4	1	9	6	10	8	**6**

In summary, the following observations can be made:

- The best overall performance is offered by the Johnson S_U distribution, though it remains to be seen whether its applicability is similarly limited in practice (i.e. Mixed Logit analysis) as it is in ExpertFit. The performance is also somewhat masked by the low number of comparisons conducted for this distribution.
- Good performance is also in general obtained with the seemingly thus far unused (or at least not commonly used) Beta, Erlang, Gamma, Logistic, and Weibull distributions.
- For the S_B distribution, good performance is only obtained on the S_B and Lognormal datasets, which causes some concerns.
- The Lognormal distribution offers some of the poorest performances, largely due to its tail and strict shape assumptions.
- Good performance is in some cases also obtained by the Extreme-Value Type B distribution.
- It seems that the Normal distribution performs very well in most cases, though the performance on the Lognormal dataset causes a lot of concern. Also, the results from section 4.2.3., and 4.1. in general, highlight the need to be careful not to infer conclusions based on the tail behaviour of this distribution (and other unbounded ones).

It should be noted that none of the distributions performs overly well in the presence of a point with heightened mass, though some are able to deal with it better than others. For this, the most reliable measure seems to be the maximum difference in the density function, which suggests better-than-average performance for the Johnson S_U (where available), the Laplace and Logistic distributions, and surprisingly also the Normal distribution.

5. Summary, Conclusions and Recommendations

In this paper, we have conducted an empirical comparison of various continuous distributions, with the aim of offering guidance for the choice of distribution in Mixed Logit analysis. The comparison included a large number of target distributions, with 6 different types of true distributions being used to generate the 24 separate datasets used in the analysis.

The results from the analysis are mixed. They show that flexible distributions, such as the Johnson S_U, Gamma, Beta, Erlang, Laplace and Logistic have certain advantages over inflexible distributions. However, in some cases, problems with model fitting may arise with the use of these distributions. The results also show that the commonly used Normal distribution performs well in terms of recovering the mean and variance of the true distribution, as well as the overall shape of the density function, but that its symmetrical nature can lead to problems in the tails of the distribution, if the true distribution is asymmetrical. The analysis also shows the usual problems with the heavy tails of the Lognormal distribution, but worryingly also suggests that the recently much-heralded Johnson-S_B distribution does not perform well in all scenarios.

In the context of recent discussions of the distribution of value of travel-time savings (VTTS) (c.f. Cirillo & Axhausen, 2004, Hess et al., 2005a), it is of interest to look at the performance of the different distributions in the presence of a true distribution bounded at one end, and distributions having an inflated mass at one point, with a view to allowing for a zero VTTS.

The analysis shows that the use of symmetrical distributions, such as the Normal, can lead to problems with the tails, such that, if the Normal is used, no information should be inferred on the basis of its tail behaviour, though it can offer valid estimates of the mean and variance. Problems can also be caused by distributions bounded to one side, as their generally long tails to the other side can cause problems. On the other hand, it also seems that distributions bounded on both sides, such as the Johnson S_B, offer suboptimal performance in some cases. At least in the examples presented here, unbounded, but flexible distributions (allowing for asymmetry), such as the Johnson S_U, Laplace and Logistic seem to be preferable. In this case, it is however still important to be careful not to infer too much information on the basis of the extreme values in the tails of the distributions.

In terms of recovering a mass at a given point, none of the distributions seems to be able to cope very well. As such, it is of interest to use mixtures of distributions. In the case where the main distribution is to be continuous, with an inflated mass at one point, only zero should really be seen as a candidate for such a point, given its specific role. As such, the location of the mass need not be estimated; this belongs to the domain of purely discrete mixtures of discrete choice models, as described by Hess et al. (2005c). Fixing the location of the support point also greatly facilitates the estimation of the model.

In this case, the following approach can be used. Let's assume we want to use a Normal distribution, with a mass at zero. In this case, we estimate three parameters; the mean of the

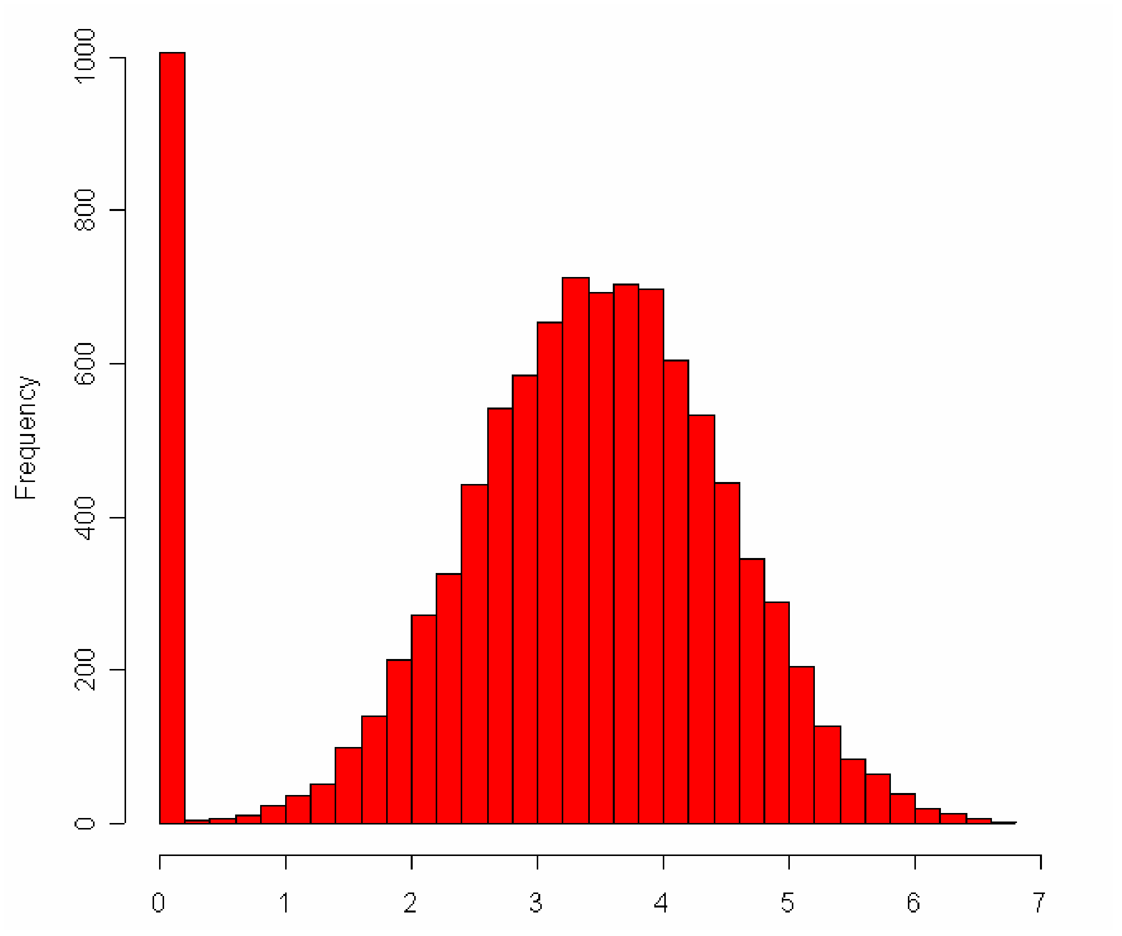
Normal distribution, μ, its standard deviation, σ, and the mass at zero, γ. In the estimation code, simulation over the distribution is used at each iteration of the optimisation algorithm. In each iteration of the simulation process, a random draw (or quasi-random draw), say *r*, is used, where *r* is contained between 0 and 1. With this draw *r*, a draw from the N(μ,σ) distribution with a mass of γ at zero is produced as follows. If the draw *r* is smaller than γ, the value of the draw from the distribution at this iteration of the simulation algorithm is set to 0. Otherwise, it is set to

$$\mu+\sigma\cdot\Phi^{-1}((r-\gamma)/(1-\gamma)), \quad \ldots[6]$$

where $\Phi^{-1}()$ is the Inverse Cumulative Normal distribution.

As an example, with μ=3.5, σ=1, and γ=0.1, the distribution obtained with this approach has the shape shown in Figure 14 (using 10,000 draws). A similar approach can be used with any other underlying continuous distribution, by replacing $\Phi^{-1}()$ by the appropriate inverse transform.

Figure 14 Distribution of N(3.5,1) with 10% mass at 0



6. Acknowlegdments

The authors would like to thank John Polak for helpful comments and suggestions.

7. References

Cirillo, C. and K.W Axhausen (2004), Evidence on the distribution of values of travel time savings from a six-week diary, *Arbeitsbericht Verkehrs- und Raumplanung,* **212**, IVT, ETH Zürich, Zürich.

Hensher, D.A. and W.L. Greene (2003) The mixed logit model: The state of practice, *Transportation*, **30** (2) 133-176.

Hess, S., M. Bierlaire and J.W. Polak (2005a) Estimation of value-of-time using Mixed Logit models, *Transportation Research A*, **39** (3) 221-236.

Hess, S., K.E. Train and J.W. Polak (2005b) On the use of a Modified Latin Hypercube Sampling (MLHS) approach in the estimation of a Mixed Logit model for vehicle choice, Transportation Research B, forthcoming.

Hess, S., M. Bierlaire and J.W. Polak (2005c) Discrete mixtures of GEV models, paper presented at the 5th Swiss Transport Research Conference, Monte Verita, Ascona, March 2005.

McFadden, D. (1974) Conditional logit analysis of qualitative choice behaviour, in P. Zarembka (ed.) *Frontiers in Econometrics*, 105-142, Academic Press, New York.

McFadden, D. and K.E. Train (2000) Mixed MNL models for discrete response, *Journal of Applied Econometrics,* **15**, 447-470.

Train, K.E. (2003) *Discrete choice analysis with simulations*, Cambridge University Press, Cambridge.