Percentile estimation for the ability distribution in item response models

Final report on Project 1 of 1.3.301.2, 2004/5

Prepared by:

Murray Aitkin and Irit Aitkin
Department of Psychology
University of Melbourne, Australia

1 Summary

The ability distribution is a fundamental concept in item response models, and plays an important role in current NAEP reports: in particular, the percentiles of this distribution are important in describing changes in this distribution over time.

The information presented in NAEP reports about the percentiles is based on the assumption that these percentiles are identifiable. The results of previous research by the authors under project 1.3.301.2 of 2003/2004 (Final report on Identification of Ability Distributions in IRT models for NAEP items) raised questions about the identifiability of the ability distribution, and therefore about the identifiability of its percentiles. In the conclusions to that report, one point made was:

- Percentiles of the ability distribution are identifiable with any accuracy only if the distribution is modelled *parametrically* – nonparametric estimation does not provide any real accuracy in percentile estimation.

This report follows up that point by examining several parametric models for the ability distribution. The conclusions of the present report are:

1. In the data simulated, with 10 2PL items and 1000 subjects, normal, extreme value and reversed extreme value ability distributions could

be discriminated when the true ability distribution was one of these three.

2. Since the percentiles of these distributions are quite different, the reporting of percentiles from an assumed normal ability distribution is unsound without an examination of the nature of this distribution.

3. This examination can be carried out by methods very similar to Gaussian quadrature which can be readily implemented in available software.

4. A generalization of the normal distribution allows the representation of skewed distributions with varying amounts of skew, and the normal distribution, in a location- and scale-parameter distribution (the *log-cubic distribution*) with a third skewness parameter. This distribution can be fitted with small modifications to available software.

5. In the simulated data with 10,000 subjects, slightly skewed distributions could be distinguished from normal, and approximate confidence limits could be placed on the degree of skewness. Corresponding confidence limits could be placed on the percentiles of the skewed distribution.

6. In the simulated data with 1,000 subjects, a quite wide range of distributions with moderate left and right skew could *not* be discriminated – they were all consistent with the observed test scores.

7. The results of this percentile estimation approach applied to real NAEP data will be reported in the second study.

2 Theoretical framework

Our approach to this project was based on the results from the previous project, which showed that skewed ability distributions could be identified, at least to some extent, by *fully non parametric estimation* of the ability distribution. However the estimation of the *percentiles* of this distribution could not easily be made from the discrete mass-point distribution resulting from the non parametric estimate, nor could the imprecision in the locations and masses be easily allowed for, so the aim of this project was to assess the extent to which a parametric distribution model could be identified, in which the percentiles had an explicit form.

This distribution needed to be sufficiently flexible to allow positive and negative skewness as well as symmetry, and to have easily calculated percentiles. The initial choice for this distribution was the Box-Cox transformed normal distribution, a three-parameter distribution in which the third transformation parameter determines the magnitude and nature of the skewness.

The general form and properties of this distribution were given by Box and Cox (1964). A short description, with details of maximum likelihood model fitting, can be found in Aitkin, Anderson, Francis and Hinde (1989 pp. 119-122). The essential point is that the response variable Y can be transformed by the power λ to normality:

$$Y^{\lambda} \sim N(\mu, \sigma^2)$$

for some unknown λ, μ and σ. The data analysis treats λ as a parameter to be estimated by maximum likelihood. Specific values of λ give well-known transformations: $\lambda = 1$ is the no-transform case, $\lambda = 0$ is the log transformation, $\lambda = 1/2$ is the square-root transformation, $\lambda = -1$ is the reciprocal transformation. (Strictly speaking the transformation is

$$Y(\lambda) = \frac{Y^{\lambda} - 1}{\lambda}$$

which approaches $\log Y$ as $\lambda \to 0$; for other values of λ the subtraction of 1 and division by λ have no effect on the model.) This distribution has been widely used in analysing skewed data; for $\lambda > 1$ the distribution is left-skewed, for $\lambda < 1$ it is right-skewed. The percentiles of the Box-Cox transformed normal distribution are very easily calculated: if $Y^{\lambda} \sim N(\mu, \sigma^2)$, the 100α percentile for Y^{λ} is $\mu + \lambda_{\alpha}\sigma$, and hence that of Y is $[\mu + \lambda_{\alpha}\sigma]^{1/\lambda}$. Standard errors of percentiles can then be obtained from that of $\hat{\lambda}$ by the delta method.

A particular difficulty which we encountered in trying to use this distribution model is that it is not a location- and scale-parameter distribution, although the normal distribution (the case $\lambda = 1$) *is* such a distribution. This is important because of the *latency* of the ability distribution: in the 2PL model with a normal ability distribution the location and scale parameters μ and σ of the normal distribution are confounded with the intercept and slope parameters of the logistic regression and so cannot be estimated. This does not matter for the normal distribution as the shape of the distribution is not determined by these parameters. However for the general Box-Cox distribution when λ is not 1, the shape of the distribution is determined by the values of μ and σ as well as the value of λ.

It became clear in our attempts to fit the model that μ and σ had to be specified in some arbitrary way to maximize the likelihood, but this meant that the skewness was determined not just by λ but by the arbitrary choice of the values of μ and σ, which could not be estimated from the test item data. Even with this specification we were unable to find an effective method for estimating the skewness parameter and maximizing the likelihood.

We therefore changed our approach. Instead of using a three-parameter distribution with a single skewness parameter to be estimated which is not a location- and scale-parameter distribution, we considered *two* two-parameter distributions which are location- and scale-parameter distributions with *fixed* positive and negative skewness which does not have to be estimated. We fitted these two distributions by maximum likelihood using numerical quadrature (in a similar way to Gaussian quadrature), and compared them with the normal distribution, fitted by Gaussian quadrature, using generated ability data which came from one of these three distributions.

The skewed distribution we used is the *extreme value distribution* (denoted by *ev*), described on pp. 283-285 of Aitkin et al. The probability density function (with argument y) is

$$f(y \mid \theta, \phi) = \frac{1}{\phi} \exp\left[\frac{y-\theta}{\phi} - \exp\left(\frac{y-\theta}{\phi}\right)\right]$$

where θ is the location parameter and ϕ is the scale parameter. These parameters are not the mean and standard deviation of the distribution, though they are closely related to them; these are

$$\mu = \theta + \phi\psi(1), \ \ \sigma = \phi\sqrt{\psi'(1)},$$

where $\psi(x)$ and $\psi'(x)$ are the digamma and trigamma functions (the first and second derivatives respectively of the log gamma function), with $\psi(1) = -0.5771$ and $\psi'(1) = 1.645$. Thus the mean and standard deviation are

$$\mu = \theta - 0.577\phi, \ \sigma = 1.283\phi.$$

This distribution has fixed negative skew, given by

$$k_3 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\psi''(1)}{[\psi'(1)]^{3/2}} = -1.139,$$

where $\psi''(x)$ is the tetragamma function, the third derivative of the log gamma function.

By changing the sign of the response variable, the extreme value distribution becomes the *reversed extreme value distribution* (denoted by *rev*), described on pp. 285-286 of the same book, which has fixed positive skew of the same magnitude. The *ev* and *rev* distributions with $\theta = 0$ and $\phi = 1$ are called *standardized* for consistency with the normal distribution; their means are -0.577 and 0.577, and their standard deviations are 1.283. These two distributions are shown with the standard normal distribution in Figure 1 (normal – solid curve, *ev* – dotted curve, *rev* – dashed curve).

The percentiles of these distributions are explicit functions of the location and scale parameters, as for the normal distribution: for the *ev* distribution the 100α percentile is $\theta + \phi \log[-\log(1-\alpha)]$, while for the *rev* distribution it is $\theta - \sigma \log[-\log(\alpha)]$.

The aim of our analysis was to assess whether these distributions can be discriminated when the abilities are generated from one of them. Since the percentiles of the three distributions are quite different (see Table 1 below), it is important to know whether the unobserved ability distribution is skewed (right or left) or symmetric. The three distributions cover this range, though they are not members of a single family indexed by one skewness parameter.

Table 1: Table of percentiles for the three distributions
extreme value (ev), normal (N), reversed extreme value (rev)
(location parameter = 0, scale parameter = 1)

%	01	05	10	25	50	75	90	95	99
ev	-4.60	-2.97	-2.25	-1.25	-0.37	0.33	0.83	1.10	1.53
N	-2.33	-1.65	-1.28	-0.68	0.00	0.68	1.28	1.65	2.33
rev	-1.53	-1.10	-0.83	-0.33	0.37	1.25	2.25	2.97	4.60

If the maximized likelihood is not much affected by the different ability model assumptions, that is the three distributions fit about equally well, then the percentiles of the ability distribution are essentially not estimable, since all three distributions *could* be the true ability distribution, but with quite different percentiles. However if we *can* discriminate between the distributions, we will need to use percentiles for the *appropriate distribution* – the use of the normal percentiles will not be adequate if the ability distribution can be clearly identified as skewed.

3 Fitting the *ev* and *rev* distributions

We fitted the two distributions by a slight modification of Gaussian quadrature (GQ). In GQ, we replace the integral over the continuous normal distribution by a finite sum over a discrete approximation to the normal distribution. Tables of the discrete approximation are available, optimized for closeness of agreement with the continuous normal distribution. However, discrete integral approximations can be computed without any special tables. The simplest such approximation, which can be computed for any distribution $f(y)$, uses an equally spaced grid of K values y_k, $k = 1, \ldots, K$ of y, calculates the density ordinates $f_k = f(y_k)$, and normalizes these to probabilities $p_k = f_k / \sum_{k=1}^{K} f_k$. The probabilities p_k and locations y_k are then used in exactly the same way as in Gaussian quadrature, replacing the Gaussian masses and mass-points. We followed this approach for the *ev* and *rev* distributions, using a $K = 14$ point grid. This approach is used in the current NAEP analysis for the normal distribution, with 40 mass-points at steps of 0.25 from -5 to 5, for the initial fitting of the large conditioning model.

4 Results

We fitted each of the three distributions to samples drawn from the standard normal distribution, and then repeated this process for samples from the *ev* and *rev* distributions. We report here the results from 50 samples from each distribution. The model fitted is the 10-item 2PL model with 1000 subjects and no person-level variables, used in previous simulation studies.

We report the average values of -2 log (maximized likelihood) under each assumed fitted distribution, when the sample is drawn from one of the three distributions. These values are reported as deviance *differences* relative to the value of -2 log (maximized likelihood) for the true ability distribution, and are denoted by "dd" in the table below. Large values of dd represent clear discrimination in favour of the correct distribution. We also report the proportion (prop < 0) of samples in which the deviance difference was *negative*, indicating a better fit of the wrong distribution, and the proportion (prop > 5) of samples with deviance differences greater than 5.0, representing strong evidence (a ratio of maximized likelihoods greater than 12) in favour of the *correct* distribution.

Table of -2 log(maximized likelihood)

			True distribution		
			ev	normal	rev
	ev	mean dd	0	14.20	58.56
		sd dd		8.00	15.69
		prop < 0		0.02	0
		prop > 5		0.90	1
Fitted distribution	normal	mean dd	15.34	0	14.90
		sd dd	8.57		7.81
		prop < 0	0.06		0.02
		prop > 5	0.88		0.90
	rev	mean dd	58.61	16.21	0
		sd dd	15.95	8.28	
		prop < 0	0	0	
		prop > 5	1	0.94	

The *ev* and *rev* distributions could always be distinguished from each other, with deviance differences of 20 or more, and the intermediate normal could be distinguished with high probability (.88-.94) from the *ev* or the *rev*.

5 Conclusions

It is possible to clearly discriminate the skewed and the normal distributions from each other, in samples of size 1,000 with the item parameters used. This implies that normal percentiles should not be routinely used in reporting without an assessment of the form of the ability distribution: in NAEP-size samples it should be very clear whether a normal assumption is reasonable, if the true distribution has skewness of the order of *ev* or *rev*.

An important question is whether a distribution could be found which is of location -scale form and which has a single parameter controlling the skewness which could be estimated, allowing confident reporting of percentiles of the ability distribution. We now propose such a distribution.

6 The log-cubic distribution

Consider the form of the extreme value density:

$$f(y \mid \theta, \phi) = \frac{1}{\phi} \exp \left[\frac{y-\theta}{\phi} - \exp\left(\frac{y-\theta}{\phi} \right) \right].$$

Taking logs, and expanding the second exponential term, we have

$$\begin{aligned} \log f(y \mid \theta, \phi) &= -\log \phi + \frac{y-\theta}{\phi} - \left[1 + \frac{y-\theta}{\phi} + \frac{(y-\theta)^2}{2!\phi^2} + \frac{(y-\theta)^3}{3!\phi^3} + \ldots \right] \\ &= -\log \phi - 1 - \frac{(y-\theta)^2}{2!\phi^2} - \frac{(y-\theta)^3}{3!\phi^3} + \ldots . \end{aligned}$$

If we *truncate* the exponential series at the cubic term, and introduce a parameter γ into the cubic term to represent the total effects of terms above the cubic, we obtain the *log-cubic distribution*, which has the probability density function

$$f(y \mid \theta, \phi, \gamma) = c \cdot \exp \left[-\frac{1}{2} \left(\frac{y-\theta}{\phi} \right)^2 + \gamma \left(\frac{y-\theta}{\phi} \right)^3 \right],$$

where the integrating constant c is a function of ϕ and γ. This distribution was proposed and used by Holland and Thayer (1986, 2000) together with the more general *log-quartic distribution* which adds the fourth-degree term. The log-cubic and log-quartic distributions are location- and scale-parameter distributions like the normal and extreme value distributions. They have the attractive feature that the first three (for the log-cubic) or first four (for the log-quartic) sample moments are sufficient statistics for the parameters, and a wide variety of distributional shapes can be fitted with these distributions. They are not widely known or used because they suffer from a serious theoretical difficulty: the integrating constants which make the distributions integrate to 1 are not known, and hence maximum likelihood fitting of the distributions to *observable* data cannot be carried out, because the way in which the parameters ϕ and γ (and the coefficient δ of the fourth power in the log-quartic) appear in the likelihood through c is unknown.

However, for our purposes for the log-cubic as an ability distribution, this defect is irrelevant, because we integrate over this distribution numerically by quadrature, and in the discrete form of the distribution the integrating constant disappears when we divide each mass point ordinate by the sum of the ordinates. The use of the distribution by Holland and Thayer also avoided this problem in a similar way, by embedding it in a log-linear model framework where the integrating constant did not need to be known. Applications of the same approach can be found in Lindsey and Mersch (1992) and Aitkin (1995).

It is immediately clear from the form of the density that γ is a skewness parameter: if it is zero, the distribution reduces to the normal. If it is negative, the density is increased for $y < \theta$ and decreased for $y > \theta$, giving left skew, and conversely if γ is positive, giving right skew.

7 Percentiles of the log-cubic distribution

The log-cubic distribution does not have an analytic cumulative distribution function, and so there is no immediate calculation of the percentiles of the distribution. However for any given γ, we may again compute the density ordinates, this time over a very fine mesh, and normalize them to sum to 1.0. They can then be cumulated to give a very fine mesh approximation to the continuous cdf, sufficiently accurately to give any required percentiles precisely.

We give below a table of moments and percentiles, computed over a 101-point grid from $y = -4(0.01)6$ for $\gamma = 0(0.01)0.06$, for the standardized distributions with $\theta = 0$ and $\phi = 1$, and include the *rev* distribution from the table above. The positive range of γ is limited to 0.06; for greater values of γ the cubic term in the density increases so rapidly for large y that it diverges to infinity. The range of y covers the effective range of the right-skewed distributions for positive γ. All these distributions have positive skew; reversing the sign of γ reverses the skewness to negative.

Table 2: Moments and percentiles of the log-cubic and *rev* distributions

γ	0.00	0.01	0.02	0.03	0.04	0.05	0.06	*rev*
μ	0.000	0.030	0.061	0.094	0.131	0.173	0.240	0.577
σ	1.000	1.002	1.007	1.018	1.035	1.061	1.136	1.283
skew	0.000	0.061	0.125	0.198	0.291	0.437	0.736	1.139
01	-2.33	-2.30	-2.25	-2.19	-2.14	-2.10	-2.05	-1.53
05	-1.65	-1.65	-1.61	-1.58	-1.54	-1.50	-1.45	-1.10
10	-1.28	-1.30	-1.27	-1.23	-1.20	-1.18	-1.15	-0.83
25	-0.68	-0.70	-0.68	-0.66	-0.63	-0.61	-0.58	-0.33
50	0.00	-0.03	-0.01	0.01	0.03	0.06	0.10	0.37
75	0.68	0.65	0.67	0.71	0.74	0.78	0.84	1.25
90	1.28	1.27	1.31	1.37	1.42	1.50	1.62	2.25
95	1.65	1.65	1.71	1.77	1.86	1.96	2.15	2.97
99	2.33	2.35	2.45	2.57	2.72	2.97	3.51	4.60

We show in Figure 2 the normal and reversed extreme value distributions and the log-cubic distributions for $\gamma = 0.02, 0.04$ and 0.06 (normal – solid symmetric, $\gamma = 0.02$ – dotted, $\gamma = 0.04$ – dot-dash, $\gamma = 0.06$ – dashed, rev – solid asymmetric).

The skew of the log-cubic distributions is mild compared to that of the *rev* distribution. The median and lower percentiles vary less with γ than the upper percentiles.

8 Fitting the log-cubic distribution, and estimating γ from sample data

To use this distribution for model fitting by quadrature, we need to specify the value of γ. For this fixed γ, we proceed as for the extreme value distributions: we define an equally spaced grid of $K = 27$ values y_k, $k = 1, \ldots, K$ of y, calculate the log-cubic density ordinates $f_k = f(y_k \mid \theta, \phi, \gamma)$, and normalize these to probabilities $p_k = f_k / \sum_{k=1}^{K} f_k$. The probabilities p_k and locations y_k are then used in exactly the same way as in Gaussian quadrature, replacing the Gaussian masses and mass-points.

By defining a grid of values of γ and repeating the quadrature over the grid, the set of values of the maximized likelihoods for each γ gives the *profile likelihood* in γ; the grid value of γ with the largest likelihood is an approximation to the MLE $\hat{\gamma}$ of γ. The approximation can be refined by refining the grid around the approximate MLE. An approximate 95% confidence interval for the true γ is the set of values of γ whose values of the deviance $-2 \log L_{max}$ are within 3.84 of the value at $\hat{\gamma}$. If this interval contains the value 0, the normal ability distribution is consistent with the observed test data – it is a possible representation of the ability distribution.

If the confidence interval excludes 0, a skewed distribution is needed, though the degree of skew may be small if the confidence interval nearly includes zero.

This approach was implemented in GLLAMM and results are reported below for a sample drawn from the distribution with $\gamma = 0.03$, giving a skewness coefficent of 0.2. As can be seen from Figure 2, the distributions with $\gamma \leq 0.4$ are hard to distinguish from normal.

The log-cubic distribution is fitted over a grid of values $-0.06(0.01)0.06$, and the *ev* and *rev* distributions are also fitted. The values of the deviance (dev $= -2 \log L_{max}$) are given below (with 12,300 subtracted), and are graphed in Figure 3, excluding the *ev* and *rev* values.

γ	*ev*	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	0.00
dev	105.58	94.16	91.14	90.00	89.32	88.86	88.56	88.36
γ	0.01	0.02	0.03	0.04	0.05	0.06	*rev*	
dev	88.26	88.27	88.41	88.74	89.48	92.02	101.12	

Taking 88.26 as the minimum deviance, the interval ± 3.84 around the minimum (shown by the straight line in Figure 3) includes all values of γ in the range $(-0.053, 0.060)$; the *ev* and *rev* distributions are both excluded, but the normal is not. This range of possible values of γ does not provide any great precision in the extreme percentiles, though the median is better: for the median the corresponding interval is $(-0.07, 0.10)$ (the true value is 0.01), while for the 75-th percentile it is $(0.60, 0.84)$ (true value 0.71) and for the 90-th percentile it is $(1.17, 1.62)$ (true value 1.37). For the normal distribution these percentiles are 0, 0.68 and 1.28 respectively.

Since NAEP sample sizes are typically much larger than 1000, these results do not rule out the possibility of more precise estimation of ability distribution percentiles in NAEP studies. To assess this we repeated the simulation above, with a subject sample size of 10,000 instead of 1000. The deviance table is given below, after subtracting 127,600, and the values are graphed in Figure 4, excluding the *ev* and *rev* values.

γ	ev	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	0.00
dev	312.66	168.32	132.16	116.34	105.68	97.56	91.08	85.78
γ	0.01	0.02	0.03	0.04	0.05	0.06	rev	
dev	81.50	78.12	75.78	74.78	76.44	90.82	158.18	

The minimum deviance occurs close to $\gamma = 0.04$, and the approximate 95% confidence interval shrinks considerably, to $(0.019, 0.055)$. This *excludes* the normal distribution. The corresponding 95% confidence intervals for the median, 75th and 90th percentiles are $(-0.01, 0.08)$, $(0.67, 0.81)$ and $(1.31, 1.56)$.

9 Discussion

This project has established that ability distributions with substantially different skew can be discriminated, in terms of their skewness, from modest sample sizes (1000 here) with relatively small numbers of items (10 here). Small to moderate differences in skew need larger sample sizes to identify; with 10,000 subjects quite small differences in skew could be identified with 10 items.

These results will be evaluated on real NAEP data, and are important for the reporting of ability distribution percentiles; these are currently based on the empirical distribution of plausible values of ability generated from the posterior distribution of individual ability assuming a *normal* distribution of ability. The model-based approach proposed here replaces the plausible value generation by direct fitting of a model which allows ability to be skewed; this is less general than non parametric estimation, but more general than the single normal distribution currently used, and provides direct estimates of percentiles, and the precisions of the estimates, from the estimated log-cubic distribution (which the non parametric approach does not). This distribution has the same scaling features (location and scale parameters) as the normal, and allows the reporting of percentiles which do represent, through the distribution family, the degreee of skewness of the underlying abilities.

The information in actual NAEP data about the ability distribution remains to be determined. This will be investigated in Project 2 of this work.

10 References

Aitkin, M. (1995) Probability model choice in single samples from exponential family models using Poisson log-linear modelling, and model comparison using Bayes and posterior Bayes factors. *Statistics and Computing* **5**, 113-120.

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989) *Statistical Modelling in GLIM.* Clarendon Press, Oxford.

Box, G.E.P. and Cox, D.R. (1964) An analysis of transformations (with Discussion). *Jour. Roy. Statist. Soc. B*, **26**, 211-252.

Holland, P.W. Thayer, D.T. (1986) *Notes on the use of log-linear models for fitting discrete probability distributions*, ETS Research Report 87-31, Educational Testing Service, Princeton N.J.

Holland, P.W. Thayer, D.T. (2000) Univariate and bivariate loglinear models for discrete test score distributions. *J. Educational and Behavioral Statistics*, **25**, 133-183.

Lindsey, J.K. and Mersch, G. (1992) Fitting and comparing probability distributions with log linear models. *Computational Statistics and Data Analysis* **13**, 373-384.

11 Figures

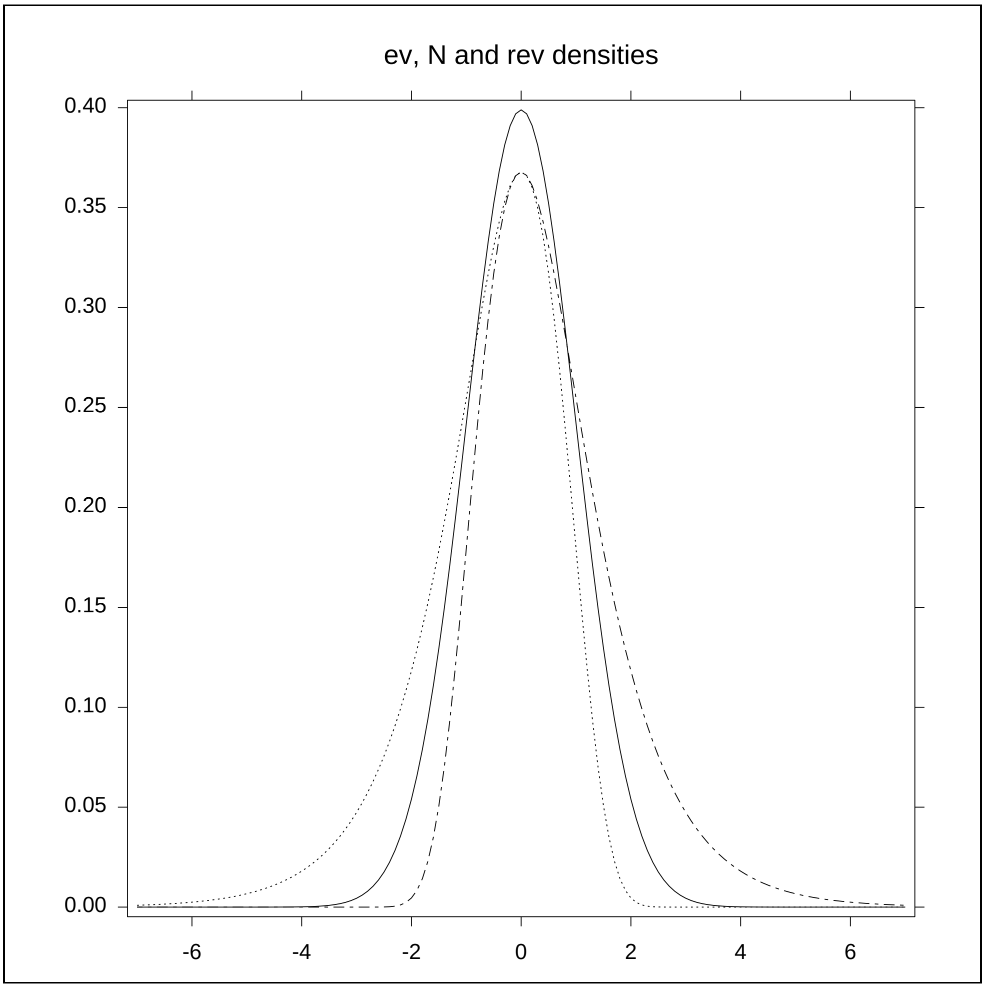

Figure 1: Normal, *ev* and *rev* densities

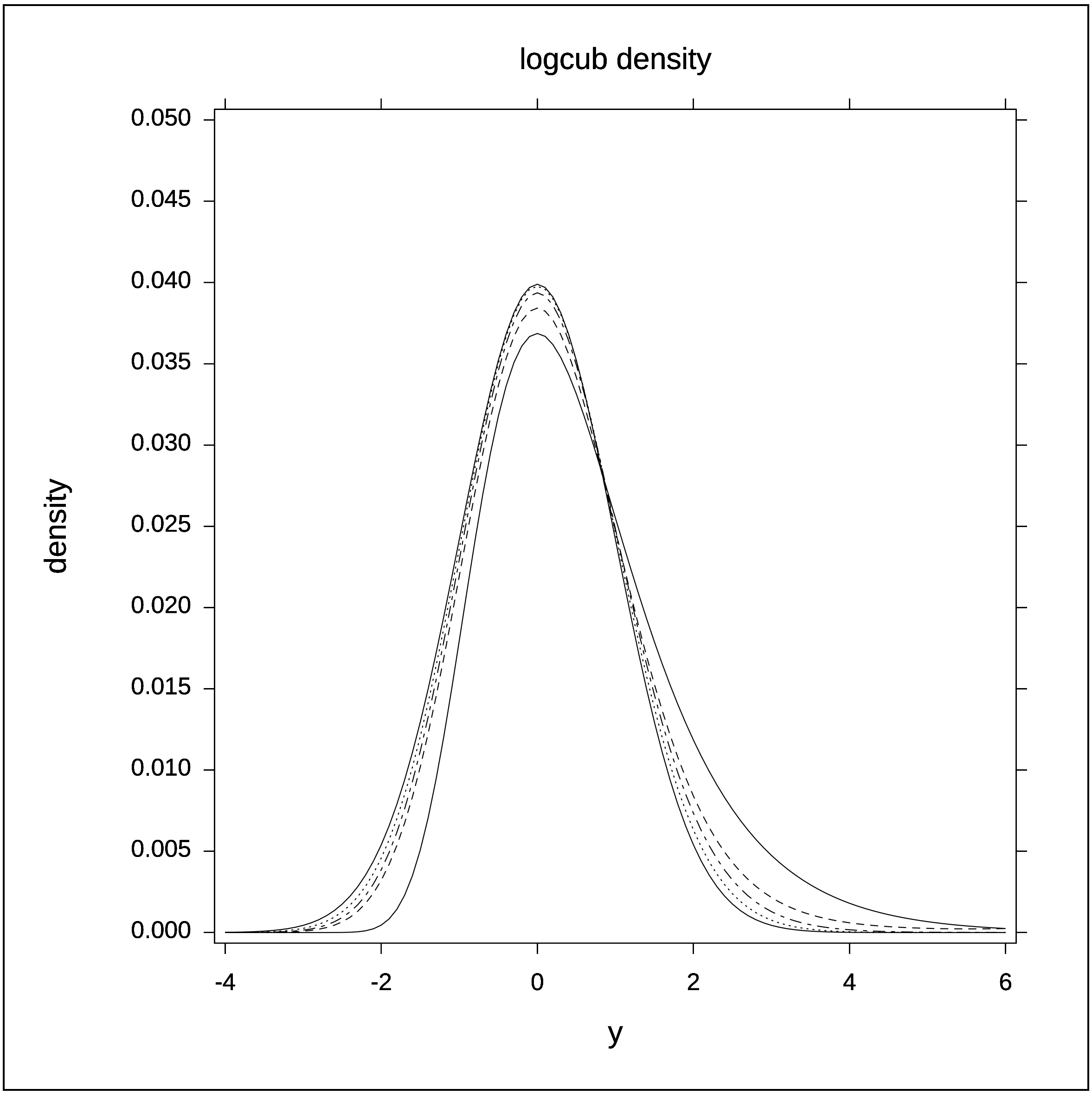

Figure 2: Normal, *rev* and log-cubic densities

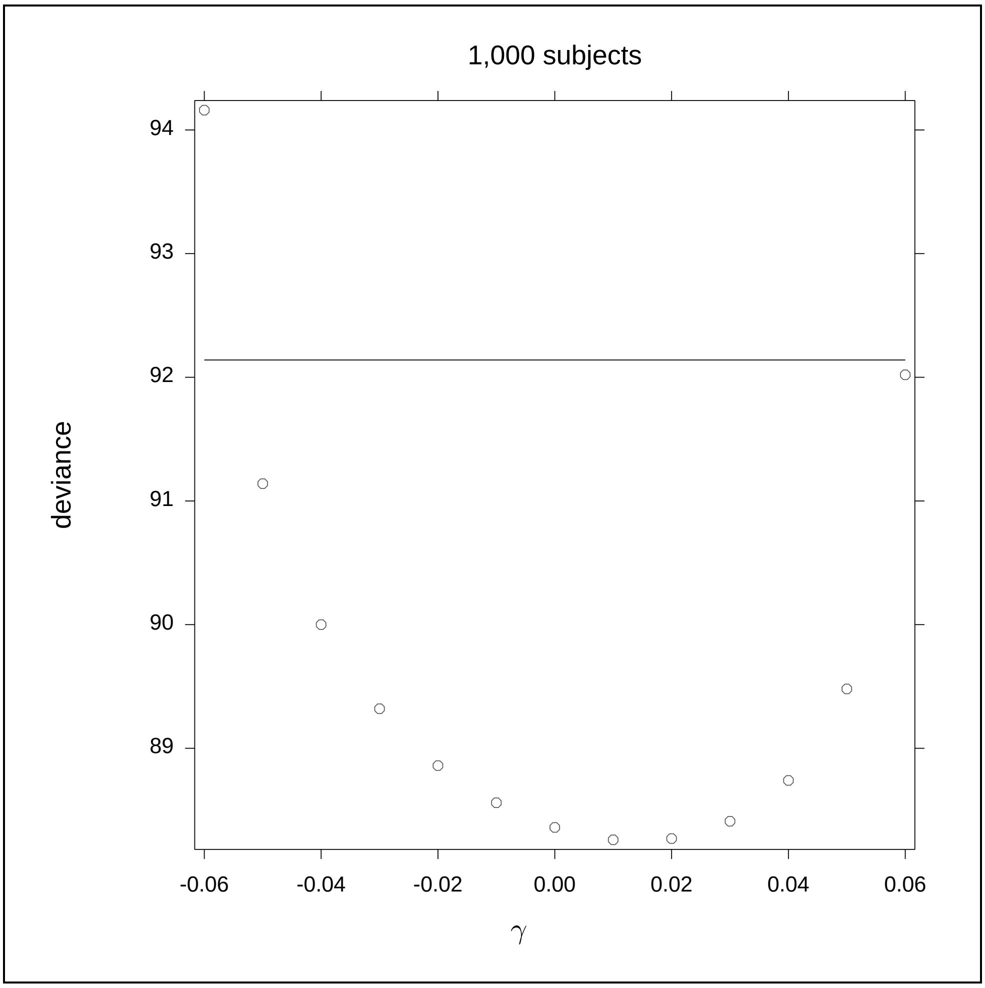

Figure 3: Deviances for sample of 1,000

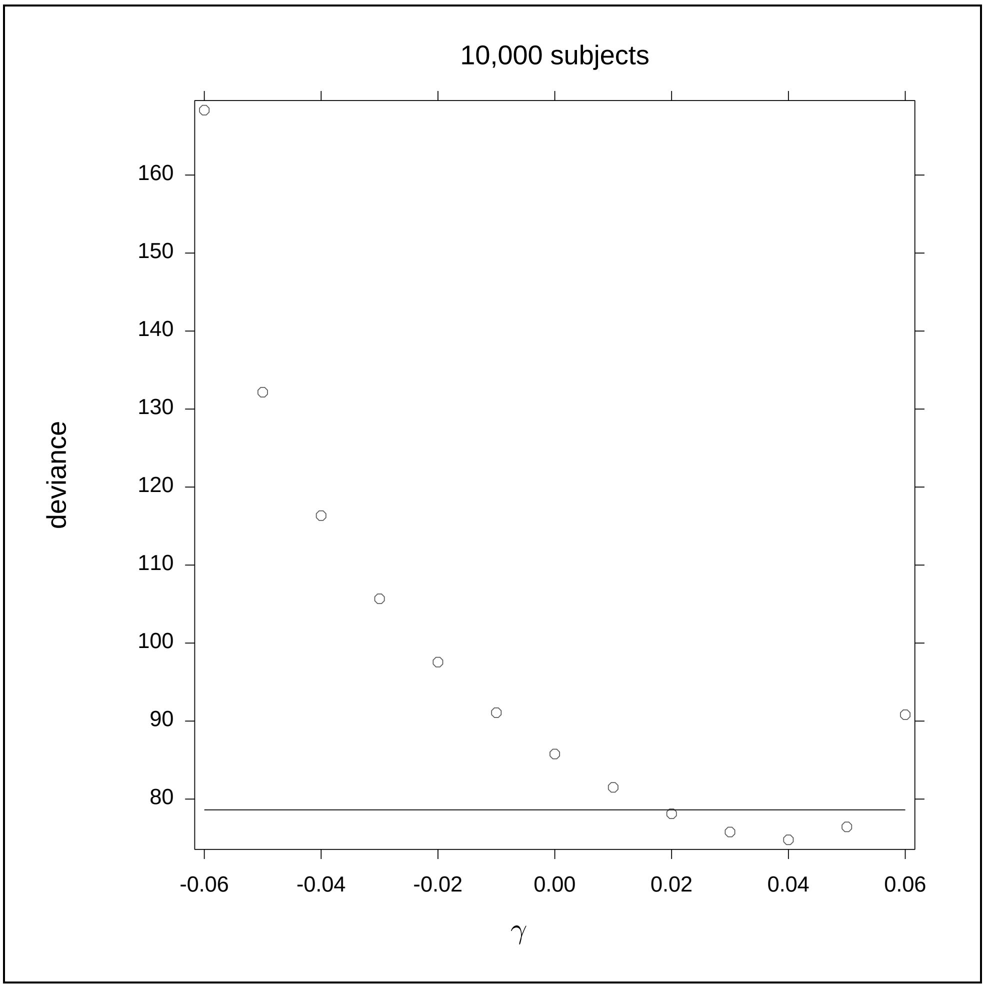

Figure 4: Deviances for sample of 10,000