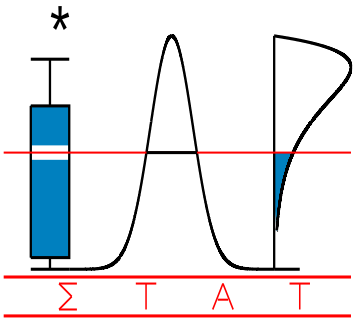
T E C H N I C A L

R E P O R T

0448

THE LOGISTIC-TRANSFORM FOR BOUNDED OUTCOME SCORES

LESAFFRE, E., RIZOPOULOS, D. and S. TSONAKA



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

The logistic-transform for bounded outcome scores

Emmanuel Lesaffre[†*], Dimitris Rizopoulos[†], Spyridoula Tsonaka[†]

Abstract

The logistic transform, originally suggested by Johnson [9], is applied to analyze responses which are restricted to a finite interval, so-called bounded outcome scores. Examples of bounded outcome scores are often standardized to lie in the interval [0,1] and can be found in many research areas. Here, we look at a popular measure in drug compliance research, i.e., the proportion of days the patient has correctly taken a drug and an ADL score, the Barthel index, often used in stroke trials. However, the fact that the score can be 0 or 1 complicates matters. Therefore, a latent score is assumed on (0,1) with a logistic-normal distribution. This approach is examined in the case when the bounded outcome score is a proportion and when it can be considered as a coarse version of a score on (0,1). The usefulness of our approach for clinical trials will be shown for the two-group comparison. A simulation study evaluates the performance (Pr(type I error) and power) of our approach for various distributions on [0,1] in comparison with the two-sample Wilcoxon test. Finally, our approach is illustrated on data from a recent compliance-enhancing clinical trial (THAMES study) conducted in Belgium and on the outcome (Barthel index) of a stroke trial (ECASS-1 study) comparing the effect of placebo and a thrombolytic on patients with an acute ischemic stroke.

Key words: Bounded outcome scores, logistic-transform, compliance, Barthel index.

1 Introduction

Bounded outcome scores are measurements that are restricted to a finite interval, which can be closed, open or half-closed. Examples of bounded outcome scores can be found in many medical disciplines. For instance, in compliance research one measures the proportion of days that a patient correctly takes his/her drug, hereafter denoted as *pdays*. Another example is

[†]Biostatistical Centre, School of Public Health, Catholic University of Leuven, Belgium.

[*]Address for Correspondence: Biostatistical Centre, U.Z. St. Rafaël, Kapucijnenvoer 35, B-3000 Leuven, Belgium. e-mail: emmanuel.lesaffre@med.kuleuven.ac.be

the Barthel-index which is an Activity on Daily Living (ADL) scale which (in one version) has a minimal value of 0 (death or completely immobilised) to 20 (able to perform all daily activities independent) and which jumps with steps of 1. This scale is often used in stroke trials to measure the recovery of a patient in practical terms after an acute stroke.

Typically, bounded outcome scores show a variety of distributions, from unimodal to J- and U-shaped. This peculiar shape of the distribution often necessitates the use of non-parametric methods, like the Wilcoxon test (see Lesaffre *et al.* [15]). Alternatively, one may consider a discretized version of the score and use statistical methods for categorical data, see e.g., Whitehead [23]. While non-parametric tests are often as powerful as parametric tests, possibilities to do on statistical modelling, e.g., when covariate-adjustment is envisaged, are limited. On the other hand, reducing the score to a binary variable invariably reduces the efficiency of the comparison. Hence, a parametric method using the original scores would be welcome.

Here, we explore the use of the logistic-transform first suggested by Johnson [9] to model the distribution of bounded outcome scores. One problem, however, is that most outcome scores are defined on a closed interval while the logistic transform assumes an open interval. This problem could be resolved by assuming a latent variable on (0,1) which gives rise to an observed score taking values on [0,1].

In Section 2 we indicate the usefulness of the logistic transformation for clinical research with bounded outcome scores. In the next section we present methods for fitting distributions on [0,1] assuming latent scores which have a logistic-normal distribution on (0,1). In Section 4 we will look at other distributions than the logistic-normal distribution. In Section 5 a simulation study evaluates the performance (Pr(type I error) and power) of our approach for various distributions on [0,1] in comparison with the two-sample Wilcoxon test. In Section 6 we illustrate our method first on *pdays*, the primary endpoint of the THAMES study, a recent compliance-enhancing intervention study performed in Belgium. In this case the bounded outcome score is a proportion. Secondly we re-analyze the primary endpoint (Barthel index) of the ECASS-1 study, an early placebo-controlled randomized clinical trial evaluating the effect of a thrombolytic drug on patients with an acute ischemic stroke. In this case we assume that the bounded outcome score is a coarsened version of a latent score on (0,1). Finally, in Section 7 we summarize our results and make some suggestions for further research in this area.

2 The logistic transformation and its application to clinical research

Johnson [9] suggested the logistic transformation $Z = \alpha + \beta \log\left(\frac{U-a}{b-U}\right)$ where U is a score on the interval (a, b). The aim of Johnson was to achieve standard normality. In the case of proportions $a = 0$ and $b = 1$. Here we take $\alpha = 0$ and $\beta = 1$ and assume that the logistic transformation achieves a normal density $N(\mu, \sigma^2)$. In general, when Z has density $f(z) \equiv f(z; \boldsymbol{\theta})$ then U has as density $g(u) \equiv g(u; \boldsymbol{\theta}) = f(\text{logit}(u))\frac{1}{u(1-u)}$, where $\text{logit}(u) = \log\left(\frac{u}{1-u}\right)$. When $Z \sim N(\mu, \sigma^2)$, then U has a *logistic-normal* distribution, denoted by $LN(\mu, \sigma^2)$ and $\boldsymbol{\theta}^T = (\mu, \sigma^2)$. While the normal density is certainly most important to us, transformations to other distributions like the t-distribution and the logistic distribution are also of interest here.

The logistic-normal distribution can have very different shapes depending on the choice of μ and σ^2. In Figure 1, we show the different shapes of the logistic-normal distribution, from a unimodal distribution to a J- and U-shaped distribution. Of course similar shapes appear when transformation to a t- or logistic distribution is envisaged. Hence, the logistic transformation is very well suited to model a variety of distributions on (0,1). A similar property holds for the Beta family, but Aitchison and Begg [1] indicate that the logistic-normal distribution is richer and can approximate any Beta density to a sufficient degree.

It is clear that when the bounded outcome scores have a logistic-normal distribution then the analysis can be done on the z-scale using classical statistical analyses assuming a Gaussian distribution. For instance, when comparing the effect of a new treatment versus a standard treatment at the end of the study a simple unpaired t-test can be used on the z-values, instead of a non-parametric test on the u-values. Further, a 95% confidence interval can be obtained for the true difference Δ (> 0) between the two treatments on the z-scale. An interpretation of Δ in a particular case is given in the following paragraph.

The logistic transformation is also a useful tool for sample size calculations in a clinical trial with a bounded outcome score U as primary endpoint. Most sample size calculations assume the location-shift alternative. Namely, it is assumed that under H_0 the density of the primary endpoint is $g_0(u)$ and that under H_a this density is shifted to, say the right with an amount Δ (> 0). But with a bounded outcome score U the location-shift alternative is most often not an option because of the boundaries at 0 and 1. This lead Lesaffre and De Klerk [14] to estimate the necessary sample size by modelling historical data on *pdays*, the proportion of days the patient correctly takes his/her drug, by a mixture of two Beta densities

and assuming a different effect of the intervention treatment on the two subpopulations of patients (represented by the two beta densities). Alternatively, if the distribution of U under the null- and alternative hypothesis is logistic-normal, one could postulate the location-shift alternative on the transformed (e.g., normal) scale. That is, we assume that the distribution of Z under H_0 is $N(\mu, \sigma^2)$, while under the alternative it is $N(\mu + \Delta, \sigma^2)$). If the logistic transformation yields on the z-scale a logistic distribution, then the effect-size $\frac{\Delta}{\sigma}$ can be interpreted as a log-odds ratio for the cumulative distributions on the z-scale, i.e.

$$\frac{\Delta}{\sigma} = \log\left(\frac{F_0(z)/(1-F_0(z))}{F_\Delta(z)/(1-F_\Delta(z))}\right), \tag{1}$$

where $F_0(z)$ is the cumulative logistic distribution under H_0 and $F_\Delta(z)$ the cumulative logistic distribution under H_a. In this respect, we obtain a generalisation of the *proportional odds model* for ordinal data described by McCullagh [16] to continuous data, see also Whitehead [23]. Further, because of the monotone transformation from U to Z, the same property must hold for U. Namely, (1) also holds when replacing $F_0(z)$ and $F_\Delta(z)$ by the corresponding cdf's of U, i.e., $G_0(u)$ and $G_\Delta(u)$, respectively. Unfortunately, this result does not hold for the normal distribution. Nevertheless, for the transformed normal distributions, it is easily seen that the proportion of individuals better off with the new treatment than with the control treatment is equal to $P_\Delta = \Phi\left(\frac{\Delta}{\sigma\sqrt{2}}\right)$ and because the logistic transformation is monotone the same property holds on the original scale. Hence, using the location-shift alternative on the transformed scale allows to rank different alternatives easily because $\Delta_1 < \Delta_2$ implies that on the original scale the distribution under the first alternative hypothesis is stochastically smaller than under the second alternative hypothesis. Hence, power and sample size calculations are relatively easy since they can be done on the transformed scale. Further, a 95% C.I. for P_Δ can be obtained using the Delta Method when estimates for Δ, σ and their covariance matrix are available.

The logistic transformation is also useful in statistical modelling of bounded outcome scores on (0,1). Indeed, the logit regression model

$$\log\left(\frac{U}{1-U}\right) = \boldsymbol{x}^T\boldsymbol{\beta} + \sigma Z, \tag{2}$$

with $Z \sim N(0,1)$ has been used in various applications (see e.g., [10]). We would argue that this feature is one of the major advantages of our approach, especially in clinical trial applications when baseline covariates need to be taken into account.

Despite the attractiveness of working on the transformed scale, in practice bounded outcome scores often lie in the closed interval [0,1] implying that the logistic transform is not possible. A way out to this problem is to assume a latent bounded score on (0,1) which is

coarsely measured giving rise to the observed bounded outcome scores on [0,1]. In the next section we will look at some possible approaches to model scores on [0,1], but we consider only the logistic-normal case.

3 Modelling bounded outcome scores on [0,1]

Often bounded outcome scores are proportions or fractions. Aitchison and Shen [3] give examples in different research areas, e.g., as a medical application they mention the proportion of serum proteins in blood. Compliance research, as seen above, provides other examples. Inevitably, the true proportion for an individual is almost never known because it either involves destructive and/or exhaustive measuring.

In this section we will denote a score on [0,1] by Y_i to distinguish it from a score $U_i \in (0,1)$. We consider three major cases. In the first case, the bounded outcome scores Y_i are proportions equal to r_i/N_i $(i = 1, \ldots, n)$ whereby r_i is the ith count out of N_i units. For instance, in compliance research, $pdays_i = r_i/N_i$, whereby r_i is the number of days the ith patient has correctly taken the drug in N_i days. In the second case, the Y_is are discrete random variables on [0,1]. So we will assume a latent structure on (0,1) behind the bounded outcome score. An example of this type is the Barthel index (see e.g., Lesaffre *et al.* [15]). In the third case, the Y_is are fractions, e.g., the fraction of a substance in some material (e.g., the fraction of serum proteins in blood).

3.1 Modelling proportions on [0,1]

When, given the individual, the bounded outcome score is a proportion derived from a series of independent Bernoulli experiments, then an obvious choice is to work with a binomial distribution. Namely, we could assume that,

$$r_i \sim Bin(U_i, N_i) \qquad (i = 1, \ldots, n) \tag{3}$$

with $U_i \sim LN(\mu, \sigma^2)$. Thus, we assume that given U_i, r_i has a binomial distribution and that U_i is a continuous latent random variable which has a logistic-normal distribution. For each value of U_i one observes N_i binary outcomes W_{ij} $(j = 1, \ldots, N_i)$ summing up to r_i. For the compliance example, U_i could be interpreted as the (true) latent adherence of the ith patient to the drug. The observed adherence $pdays_i = r_i/N_i$ is a coarse measure for the latent score

depending on the operational definition of when the drug is taken correctly and on the length of the period that the patient is observed. Observe that this model is actually a classical measurement error model [4], specifying the distribution $f(Y|U)$.

Model (3) can be extended by replacing μ by $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and hence becomes an example of a *generalized mixed effects model*, whereby conditional on U_i the W_{ij} are assumed to be independent. Fitting such a model can be done with e.g., the SAS procedure NLMIXED using Maximum Likelihood or the function `glmmPQL` in package `MASS` [20] in R [18] using Penalized Quasi-Likelihood. Hence, actually we are modelling overdispersion. An alternative way to deal with the overdispersion problem is to use quasi-likelihood. In that case, we assume that the variance function of our model is of the form $V(\phi, p_i) = \phi N_i p_i(1-p_i)$ instead of the usual $V(p_i) = N_i p_i(1-p_i)$ which is assumed by the binomial model. Although, there is no statistical model leading to the variance function $V(\phi, p_i)$, the parameters can be estimated and they also have the usual asymptotic properties [17]. Nowadays, standard software is available for fitting quasi-likelihood models, e.g., function `glm` in R or the SAS procedure GENMOD.

3.2 Modelling discrete bounded outcome scores on [0,1] which are not proportions

When Y_i is a discrete random variable on [0,1] (e.g., the Barthel index), but not a proportion then it is not immediately obvious what model to choose for the scores. But, we could assume that the observed score is a coarse measure of the true latent score which lies in the interval (0,1). This assumption can often be made, but not always as discussed in Section 6.2.

With a latent score on (0,1) there could be several mechanisms that generate observed scores on [0,1]. One mechanism is *rounding-off*. Namely, the latent score U_i is rounded off to Y_i when $Y_i - \epsilon \leq U_i < Y_i + \epsilon$, where ϵ is the smallest rounding-off error. *Digit preference* and *truncation* are other examples of coarsened recording of the latent score. The last type of coarsening could happen when the observed score Y_i can only take a finite number of different values and that, say, $y_i = k/m$ when $k/m \leq U_i < (k+1)/m$, where $k = 1, \ldots, m$ and m is the maximum score in the original scale.

The framework of our approach is that of coarsened data, as formalized by Heitjan and Rubin [8] and Heitjan [7]. They considered deterministic or stochastic coarsening. In particular, when there is more than one possible procedures to generate the Y_i from the U_i, then

one speaks of a stochastic coarsening.

In general, we assume $\frac{m}{s}+1$ disjoint intervals $[(a_l, b_l)]$ (a_l, b_l are real numbers in $[0,1]$ and $[(.,.)]$ denotes the four possible kinds of intervals) for which $\bigcup_{l=1}^{\frac{m}{s}+1}[(a_l, b_l)] = [0,1]$ and each one contains only one $y_i = k_i/m$, $(k = 0, s, \ldots, m)$, then

$$y_i(u_i) = \left\{ \frac{k}{m} : u_i \in [(a_i, b_i)], k = 0, s, \ldots, m \right\}, \tag{4}$$

where s denotes the step length. For different procedures generating a_l and b_l, we obtain different coarsening mechanisms, i.e., rounding-off, digit preference, etc. In case of stochastic coarsening for different individuals different coarsening mechanisms apply, hence different sets of $[(a_l, b_l)]$ apply. For deterministic coarsening and stochastic coarsening when the observed y_i are coarsened at random (CAR), the likelihood becomes

$$L(\boldsymbol{\theta}; y) = \prod_{i=1}^{n} \int_{a_i}^{b_i} g\left(u_i; \boldsymbol{\theta}\right) du, \tag{5}$$

where g is the probability density function of the logistic-normal distribution.

This integral can be calculated by means of the transformation $z = \text{logit}\,(u)$, and so we get that the likelihood of the sample becomes the likelihood of interval-censored normally distributed random variables,

$$L(\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma)^T; y) = \prod_{i=1}^{n} \left[\Phi\left(\frac{z_i^{(u)} - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} \right) - \Phi\left(\frac{z_i^{(l)} - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} \right) \right], \tag{6}$$

where $z_i^{(l)} = \text{logit}(a_i)$, $z_i^{(u)} = \text{logit}(b_i)$ and $\Phi(.)$ is the distribution function of the standard normal distribution.

The Maximum Likelihood estimates for this model can easily be obtained using standard numerical procedures like the Newton-Raphson or the BFGS (Broyden, Fletcher, Goldfarb and Shanno) algorithm. Alternatively, an EM algorithm can be written involving the calculation of means of truncated normal distributions.

Observe that while in the previous subsection we specified the model $f(Y|U)$ now we have specified $f(U|Y)$, although in a rather simple way. One could specify other models for $f(U|Y)$, for instance a normal kernel $N(0, \sigma_\delta^2)$ truncated at the boundaries. This corresponds to the Berkson measurement error model. However, it is not clear to what physical phenomenon, like rounding-off, this model corresponds to.

The previous mechanisms (rounding-off, digit preference, etc.) assume only coarsened recording of the true latent scores. Of course even without coarsening, the observed scores

may deviate from the latent scores purely by random variation (as in the previous subsection). In other words, $Y_i = U_i + \delta_i$, where δ_i is the measurement error for the ith subject. When measurement error happens on the transformed scale, i.e., $Y_i = \text{expit}(Z_i + \delta_i)$, where expit is the inverse of logit and $\delta_i \sim N(0, \sigma_\delta^2)$ then $Y_i \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2 + \sigma_\delta^2)$ if $Z_i \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2)$. If there is no information on σ_δ there is no way of retrieving the distribution of the true latent score, and actually in many applications there is no need to do this. Hence, if measurement error is present (which is often the case) we will assume here that the observed score Y_i is obtained from coarsened recording of $U_i + \delta_i$, which has a logistic-normal distribution. But, for simplicity reasons we denote the variance on the transformed scale again σ^2.

3.3 Modelling fractions on [0,1]

Fractions typically have a value in $[0, 1]$. The values in (0,1) can be assumed realizations of a continuous random variable. For instance, the proportion of mercury in amalgam is the ratio of 2 continuous random variables with numerator the weight of mercury and denominator the total weight. But, because some alloys do not contain the specific component this ratio can be zero. The other extreme is that the alloy is not an alloy but a pure substance. Thus, although a fraction is a proportion the method explained in Section 3.1 cannot be used here. On the other hand, a mixture distribution could describe the distribution of fractions by adding to a classical distribution on $(0, 1)$ two point-mass distributions at the boundaries, i.e., at 0 and at 1. This is similar to the models considered by Zhou and Tu [24] and Lachenbruch [11, 12] who examined two-part models for positive random variables with a zero-part and hence we call it a three-part model. Formally, the three-part model is

$$g(Y_i) \equiv g(Y_i, \boldsymbol{d}_i) = \prod_{j=1}^{3} \pi_{ij}^{d_{ij}}, \tag{7}$$

where

$$\pi_{ij} = \begin{cases} p_1 I(Y_i = 0) & \text{if } \boldsymbol{d}_i = (1,0,0) \\ p_2 I(Y_i = 1) & \text{if } \boldsymbol{d}_i = (0,1,0) \\ (1 - p_1 - p_2) g(u_i|\boldsymbol{x}_i; \boldsymbol{\theta}) & \text{if } \boldsymbol{d}_i = (0,0,1) \end{cases}$$

where $\boldsymbol{d}_i$ is the indicator vector specifying when $U \in (0, 1)$ is observed, $I(y = a)$ is the Dirac function being 1 for $y = a$ and zero elsewhere and $p_1 > 0$, $p_2 > 0$, $p_1 + p_2 < 1$. Clearly, the proportions p_1, p_2 represent the proportions of subjects who have a score 0, 1 respectively and need to be estimated from the data. A useful extension could be to include covariates in the 0- and 1-parts, as well. Namely, the probability of belonging to one of the three parts

can be given by the regression model:

$$h(p_{ij}) = \boldsymbol{w}_i^T \boldsymbol{\alpha}_j \ (j = 1, 2)$$

where e.g., h is the logit function, $\boldsymbol{\alpha}_j$ $(j = 1, 2)$ are two parameter vectors and $\boldsymbol{w}_i$ is a set of covariates not necessarily the same as $\boldsymbol{x}_i$ (i.e., the covariates in the continuous part).

The three-part model could also be considered in the two previous sections, in case the logistic transform does not provide a good fit to the data or its use is not supported by the problem's nature as will be seen in Section 6.2.

4 Alternatives to the logistic-normal distribution

In Section 3 we emphasized the logistic-normal distribution, but it is not the only suitable distribution resulting from the logistic transformation and not necessarily the best one either. Indeed, statistical inference based on the normal distribution is known to be vulnerable to outliers. A more robust approach is to use the logistic-t distribution instead of the logistic-normal [13]. Of course, for scores on (0,1) a part of the attractiveness of the approach would be lost because we would need to replace the classical techniques, like the t-test, by more sophisticated ones. The same remark applies for a proportion on [0,1] with a latent score on (0,1), since the current software (e.g., SAS proc NLMIXED) assumes normality for the latent scores. For the rounding-off mechanism changing from a logistic-normal to a logistic-t distribution is relatively straightforward since it only involves replacing the normal integrals by t-integrals. Moreover, using the BFGS algorithm as optimization procedure requires only the first-order derivatives to be adapted which can easily be computed.

Another extension to the logistic-normal distribution is obtained by changing the logistic transformation. Aitchison and Lauder [2] suggested the Box-Cox transformation family

$$z_i = \frac{(u_i/(1-u_i))^\lambda - 1}{\lambda},$$

for which when $\lambda \to 0$ corresponds to the logistic transformation. However, the authors did not document the usefulness of their extension. For the analysis of clinical trial data, we suspect that the extra flexibility has little to offer.

5 Simulation study

The purpose of this simulation study is to evaluate our proposals for analyzing bounded outcome scores on [0,1] in the two-group situation and under the shift-alternative on the transformed scale. More specifically we focus on the cases described in Sections 3.1 and 3.2 and are interested in the probability of the Type I error and the power for different alternatives. Further, we assume below that the first treatment group pertains to the control treatment and the second treatment group to the experimental treatment.

5.1 Set up of simulation study

All simulations involve two-group comparisons. We first look at scores on (0,1). The mean and variance from the control and experimental treatments are specified on the transformed scale and vary from 0 to 3 and from 1 to 4, respectively, to ensure different shapes of the distribution on the original scale. In particular, to achieve U-shaped and Unimodal distributions we used $\mu = 0$ and σ equal to 4 and 1, respectively. For the J-shaped we used $\mu = 3$ and $\sigma = 1$. Further, we vary Δ/σ from 0 (H_0) to 0.2, 0.5 and 1, corresponding to no, small, moderate and large treatment effects, respectively. In all cases the mean of the model is $\mu + \Delta \boldsymbol{x}_\Delta$, where $\boldsymbol{x}_\Delta$ is the treatment indicator. The sample sizes for the treatment groups vary and are taken as $n_1 = n_2 = 20$, 50 and 100. We sampled from a logistic-normal, a logistic-t and a logistic-logistic distribution. Each situation was sampled 1000 times. For all sampled situations we apply the two-sample t-test on the transformed scale and the Wilcoxon test (on the original scale). While these simulations actually represent a power comparison between a two-sample t-test and a Wilcoxon test, they are not of primary interest to us. However, they are useful to serve as a benchmark to the simulations on [0,1].

To simulate scores y_i on [0,1] we use the simulated scores u_i on (0,1). Firstly, we create proportions (like for the compliance measure) by simulating from the Binomial distribution with probability u_i and N_i fixed first at 30 and then varying stochastically from 10 to 200. Then we compared the Wilcoxon test on the original proportions with an analysis using a generalized mixed-effects model (i.e., SAS proc NLMIXED). Secondly, the y_is are obtained using the coarsening mechanism that $y_i = k/m$ when $(k - 0.5s)/m \leq U_i < (k + 0.5s)/m$, where $(k = 1, \ldots, m)$ with $s = 1$ and $m = 10$, 30 and 50. In the boundaries, we have $y_i = 0$ when $0 < U_i < 0.5s/m$ and $y_i = 1$ when $(m - 0.5s)/m < U_i < 1$. In this case we compare the Wilcoxon test with the model of Section (3.2) assuming as reference distribution for the latent variable the logistic-normal. For this model an S/R function was written by the authors

which is available upon request.

In a second step the effect of including a baseline covariate into our models is examined. The purpose is to show the advantage of our approach above versus a non-parametric approach by allowing for covariate adjustment thereby increasing the power of the statistical comparison between the treatment groups. This simulation was inspired by the results obtained from the THAMES and from the survival part of the ECASS-1 study. More specifically we have taken the value of Δ from these studies. Further, we have taken for $\mu = 0, 1.5$ and 3 and for $\sigma = 1, 2, 3$. Moreover, we included as baseline value realizations of the variable $\text{logit}(u_0)$ where $U_0 \sim LN(\mu, \sigma^2)$ and $\rho = cor(\text{logit}(U), \text{logit}(U_0)) = 0.3$ or 0.7. For each case the power for the Binomial and the Coarsening model was computed and compared with the one of the Wilcoxon test. These results are based on 1000 replications, sample size n equal to 200 and $N_i = m = 30$.

5.2 Simulation results

5.2.1 Results on (0,1)

First we compared (results not shown) the achieved significance level and power of the t-test on the logistically transformed score and the Wilcoxon test on the original score are compared on random variables with a logistic-normal, a logistic-t and a logistic-logistic distribution, respectively. This is actually a standard comparison of the test characteristics of the t-test and the Wilcoxon test. The results are not particularly of interest here but were useful as a benchmark for the results in the following subsections. The simulation results indicated (as expected) that the t-test and the Wilcoxon test have similar characteristics under the null- and alternative hypotheses, with the Wilcoxon test having a slightly higher power for the logistic-t and logistic-logistic distributions.

5.2.2 Results for the Binomial model

In Table 1 the comparison between the Wilcoxon test and the Binomial model with a normal random-effect and $N_i = 30$ is presented. In the majority of the cases we observe that the performance of the Wilcoxon test is nearly identical with the one of the Binomial model. The Type I error is close to 0.05 for all the cases even for the small sample size $n = 40$. Further, we observe the expected positive association of the power with the sample size and the effect size

Δ/σ. However, the power does also depend on the shape of the distribution, in contrast to the results of the previous subsection. In particular, we observe that more powerful is the U-shaped, followed by the Unimodal and the J-shaped, especially for small N_i. An explanation of this phenomenon lies in the fact that the latent score U_i (true probability of success) is not known but only the observed proportion is. When all the true proportions are relatively close to 0 or to 1 the observed proportions will be relatively close to each other, especially when N_i are small. A proof of this is seen in the power of the Wilcoxon test which shows a similar behavior. When N_i varies from 10 to 200 (results not shown), the power increases and the dependence of the power on the shape of the distribution gradually decreases.

5.2.3 Results for the Coarsening model

In Table 2 the comparison between the Wilcoxon test and the model based on the coarsening mechanism of Section 3.2 with $m = 10$ is presented. Basically the same phenomenon is seen as for the previous case. Further, a similar explanation of why the J-shaped model performs worse can be given. However, as m and n increase, the distribution shape dependency of the power gradually disappears.

5.2.4 Including covariates

The results for the power calculations when taking the response measured at baseline into account are presented in Tables 3 and 4. First of all, we observe that now the Wilcoxon can have a much lower power than the parametric models taking the baseline covariate into account. Further, as σ and ρ increase the difference in power between the parametric models and the Wilcoxon test increases. Finally, we observe that this difference is more obvious for the THAMES study since in this case the estimated Δ is lower and thus more difficult to detect by the nonparametric test.

6 Applications

6.1 A compliance-enhancing intervention study: THAMES study

Recently, an open-label, multicenter compliance-enhancing intervention (THAMES) study was completed in Belgium to measure the effect of a program of pharmaceutical care, designed to enhance adherence to atorvastatin treatment. All patients, aged 18 years or above, who had been taking atorvastatin for at least 3 months were included in the study provided they usually got their medication in one of the pharmacies participating in the study. Four well-defined districts were identified, two in Flanders (North of Belgium) and two in Wallonia (South of the country). In both Flanders and Wallonia, all pharmacists in one of the districts were to apply measures to improve compliance and enhance persistence, whereas in the second district no such measures were taken. All pharmacists were equipped with the MEMS system, an electronically monitored pharmaceutical package designed to compile the dosing histories of ambulatory patients taking oral medications (see [19]). At the first visit, the pharmacist instructed the patient how to use the monitored packages. Depending on the number of tablets prescribed and posology, the patient was instructed to return to the pharmacy every month or every 3 months. The total study duration was 12 months. Therefore, the number of visits to the pharmacy ranged from 5 to 13. At each visit, the patient's dosing history was checked by means of the electronic monitoring system. The period between the first and second visit was considered to be the baseline period. In addition to the compilation of dosing histories as described above, pharmaceutical care measures to enhance compliance/persistence with the prescribed dosing regime were applied from the second pharmacy visit onwards, i.e. after the baseline period. The pharmacist used several tools to improve compliance, more details can be found in [21]. In case the patient failed to appear for the last visit, contact was made with the patient by telephone or postcard in order to obtain the reasons for drop-out.

The primary efficacy parameter of the THAMES study was adherence to prescribed therapy in the post-baseline period whereby adherence was defined for each patient as the proportion of days during which the MEMS record showed that the patient had opened the pill container. This variable was also estimated at baseline (baseline adherence) between the first intake of drug and the time of the second visit to the pharmacy. The second pharmacy visit was the first time the pharmacist was to counsel with the patient based on observed dosing history data. Finally for the calculation of the "post-baseline adherence" the post-baseline period was arbitrarily cut-off at day 300.

6.1.1 Baseline covariates

The THAMES study could not be randomized due to practical difficulties. Therefore, we need to compare the baseline covariates of the intervention and control groups. In Table 5 we compared: gender, age, weight, work status (unemployed versus employed), a cardiovascular risk score (see Vrijens *et al.* [21]), family history of CHD and the *pdays* at baseline with the appropriate statistical techniques. Especially, the adherence at baseline is of interest, the Wilcoxon test gives a significant value (p-value = 0.011). The reasons for this significant difference at the start are not clear, but it requires that the imbalance at the start needs to be taken into account. We were aware of the potential dangers of correcting for baseline covariates in the presence of imbalance at baseline (see e.g., Wainer and Brown [22]). However, for illustrative purposes we will still perform an ANCOVA type of analysis in the next subsection, for instance to show that our approach easily incorporates baseline covariates.

6.1.2 Efficacy comparison

Initially, we checked for a treatment effect without correcting for any baseline covariates. Both the Wilcoxon test and the Binomial model give a significant intervention effect with p-value < 0.001. In Figure 3 the histograms of the two treatment groups with superimposed kernel estimates and fitted distributions are presented. We observe that the estimated logistic-normal distributions provide a very good fit to the observed data. Finally, $\widehat{P_\Delta} = 0.657$ with corresponding 95% C.I. using the Delta method $(0.601, 0.713)$, which also supports the intervention effect.

In the second part of the analysis we take into account the baseline covariates mentioned in Section 6.1.1. Note that the logit of baseline adherence was taken as covariate. In Table 6 the results of the Binomial model with a logistic-normal random effect are shown. A highly significant intervention effect (p-value < 0.001) is seen. Also, most of the covariates do not seem to have a strong impact on the response. As mentioned above our analysis uses a generalized linear mixed model to estimate the intervention effect. A competitor is a marginal GEE model. In Table 7 the results are shown from an analysis using R function `glm` (see also Section 3.1). As expected the intervention effect is lower now. On the other hand the regression coefficients of the other regressors are larger in absolute value than for the previous analysis indicating that in the generalized linear mixed approach the random variable Z captures much of the extra-Binomial variation. While we are aware of the dispute between statisticians regarding the use of a random effects model versus a marginal model,

we believe that here the generalized mixed effects model is to be preferred since the estimated intervention effect can clearly be related to the obtained effect if the true compliance measures were known.

6.2 A stroke trial: ECASS-1 study

The ECASS stroke study is a double-blind randomized parallel study designed to compare the effect of alteplase (a thrombolytic drug) and placebo in patients with an acute ischaemic stroke. The primary outcome is patient status at 3 months assesses by the Barthel index, a measure of quality-of-life assessing the ability of patients to perform daily activities. For more details we refer to [6] and [5]. Several versions of the Barthel index exist. In the ECASS-1 study the one used has scale from 0 to 100 with steps of 5. This version can be standardized such that the values lie in [0,1]. The higher the score the better the patient can live an independent life with 1 as the maximal value when the patient can live an independent life without the help of others. The score could also be considered as a reflection of a latent scale which measures the ability to cope with a handicap resulting from, say, a cerebrovascular stroke. In this way it is plausible that the latent score for most if not all patients is less than 1. This is medically supported since an observed score of 1 does not necessarily imply complete neurologic recovery of the patient. Further it is also true that for a patient who survived a stroke with a (observed) score of zero his/her true latent score is close to zero but not really zero. But, clearly this is different from a patient who has an observed score of zero because he did not survive the stroke. For this patient our assumption of a latent score on (0,1) is not tenable. We could overcome this conflict by using a two-part model with one point-mass distribution at 0 describing the patients who did not survive the stroke and a continuous latent variable in (0,1) reflecting the ability for the remaining patients. Observe that this model is not exactly the same as the one described in Section 3.3. Indeed, here the mixture is based on the extra information we have on the patient (survived versus deceased) and not in the first place on the observed value of zero for the bounded outcome score.

In particular, combining models (5) and (7), the model for the Barthel index which takes into account patient mortality has the form,

$$g(y_i, d_i; \boldsymbol{\alpha}, \boldsymbol{\theta}) = p_i^{d_i} \left[(1 - p_i) g(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})\right]^{1-d_i}$$

where d_i is the indicator of death, $\text{logit}(p_i) = \boldsymbol{w}_i^T \boldsymbol{\alpha}$ and $g(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}) = \int_{a_i}^{b_i} g\left(u_i|\boldsymbol{x}_i; \boldsymbol{\theta}\right) du$. The

corresponding log-likelihood is given by,

$$\ell\left(\boldsymbol{\theta},\boldsymbol{\alpha};\boldsymbol{y},\boldsymbol{d}\right)=\sum_{i=1}^{n}\left\{d_i\left(\boldsymbol{w}_i^T\boldsymbol{\alpha}-\mathcal{C}_i\right)+\left(1-d_i\right)\left[\log\left(\Phi\left(\frac{z_i^{(u)}-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)-\Phi\left(\frac{z_i^{(l)}-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)\right)-\mathcal{C}_i\right]\right\} \quad (8)$$

where $\boldsymbol{\theta}=(\boldsymbol{\beta}^T,\sigma)^T$, $\mathcal{C}_i=\log\left(1+\exp\left(\boldsymbol{w}_i^T\boldsymbol{\alpha}\right)\right)$ and $z_i^{(l)},z_i^{(u)}$ are given in Section 3.2. Since the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ have distinct parameter spaces, model (8) splits up in a logistic regression for the mortality rate and the coarsening model for the latent ability only for the survivors.

According to our remarks in the beginning of this section, a reasonable coarsening mechanism in the case of Barthel index could be,

$$y_i(u_i)=\left\{\frac{k}{m}:u_i\in\left(\frac{k-s/2}{m},\frac{k+s/2}{m}\right),k=0,s,\ldots,m\right\},$$

where $m=100$ and $s=5$ in our case and when $\frac{k-s/2}{m}<0$ or $\frac{k+s/2}{m}>1$ then $y_i(u_i)$ equals 0 or 1, respectively. Before adopting this mechanism and use model (8) for our analysis we should consider any discrepancies from the CAR assumption. The use of this mechanism assumes two different lengths for the intervals in which u_i might lie, namely $s/2m$ if $u_i\in(0,s/2m)$ or $u_i\in(0,1-s/2m)$ and s/m in any other case. However, based on Corollary 1 of Heitjan and Rubin [8], we could assume CAR here because the type of coarsening used is known from y_i and moreover for all the values of $u_i\in y_i$ we obtain the same coarsening mechanism.

In accordance with [15] we first perform a Wilcoxon test on all patients ignoring their mortality status. This gives a non significant treatment effect with p-value $=0.104$. Observe that this analysis mainly addresses the more practical question of how beneficial the thrombolytic treatment is overall (combining ability with mortality).

Using model (8) we first divided the total study sample into survivors and deaths. The model uses a logistic regression to predict mortality. The second part of the model only involves survivors and on this part the model introduced in Section 3.2 is used. As mentioned above, when the parameters of the two parts of the model are different, the analysis splits up into two independent parts. The results for the effect of treatment and baseline covariates on mortality are presented in Table 8. In particular, we observe no treatment effect (p-value $=0.080$) but only a significant age effect (p-value <0.001).

With regard to the second part of the model, we started our analysis without using any baseline covariate. Both the Wilcoxon test and the coarsening model give a significant treatment effect with corresponding p-values equal to 0.002 and 0.003, respectively. In Figure 4

the histograms of the two treatment groups with the superimposed kernel estimates and the fitted distributions are presented. We observe also here that the estimated logistic-normal distributions provide a very good fit to the observed data. The estimated P_Δ equals 0.583 with corresponding 95% C.I. using the Delta method $(0.530, 0.637)$, showing the treatment effect. In a second step the baseline covariates are introduced into the model. Table 9 presents the results of the Coarsening model. A more significant treatment effect (p-value = 0.001) is obtained by introducing the covariates and again a significant age effect (p-value < 0.001) but of course in the other direction now.

7 Conclusion

The assumption of a latent variable on (0,1) together with the logistic transform facilitates the modelling of the distribution of the bounded outcome score on [0,1] and allows baseline covariate adjustment. We believe that this is the main advantage of our approach over other approaches.

Despite the attractiveness of our approach, some critical remarks are needed. For the Binomial model, we assumed that a random intercept model is adequate. For the compliance example, this assumption is probably justified when the period upon which *pdays* is calculated is relatively short. However, for longer periods it is likely that a random slope or serial correlation is needed. On the other hand for the Coarsening model, a particular coarsening mechanism needs to be chosen in order to calculate the integrals in (5). However, it is not always clear which coarsening mechanism is underlying a discrete outcome score. For instance, the coarsening mechanism chosen for the Barthel index was driven by logical reasoning but there is no way of verifying whether this coarsening mechanism actually applies here.

We believe that future research is needed to render our approach more useful in practice. For instance, it would be helpful to have analytical expressions which allow us to calculate the necessary sample size when planning a clinical trial with bounded outcome scores given μ, Δ, σ and (a) N for the Binomial model or (b) m for the Coarsening model. Further, repeated models for bounded outcome scores using either a multivariate approach as suggested by Aitchison and Shen [3] or incorporating random effects could be useful. Finally, a closer look is needed on how to include bounded outcome scores as covariates in a regression model.

Acknowledgements

All authors acknowledge support from the Interuniversity Attraction Poles Program P5/24 − Belgian State − Federal Office for Scientific, Technical and Cultural Affairs.
The authors also thank Pfizer Belgium for the permission to use the data of the THAMES study and Boehringer Ingelheim for the permission to use the ECASS-1 data.

References

[1] J. Aitchison and C. Begg, *Statistical diagnosis when cases are not classified with certainty*, Biometrika **63** (1976), 1–12.

[2] J. Aitchison and I. J. Lauder, *Kernel density estimation for compositional data*, Applied Statistics **34** (1985), 129–137.

[3] J. Aitchison and S. Shen, *Logistic-normal distributions: Some properties and uses*, Biometrika **67** (1980), 261–272.

[4] R. Carroll, D. Ruppert, and L. Stefanski, *Nonlinear Measurement Error Models*, Monographs on Statistics and Applied Probability, vol. 63, Chapman and Hall, New York, 1995.

[5] C. Collin, D. Wade, S. Davies, and V. Horne, *The Barthel ADL index: a reliability study*, International Disability Studies **10** (1988), 61–63.

[6] C. Granger, G. Albrecht, and B. Hamilton, *Outcome of comprehensive medical rehabilitation: measurements of PULSES profile and the Barthel index*, Archives of Physical and Medical Rehabilitation **60** (1979), 145–154.

[7] D. Heitjan, *Ignorability and coarse data: Some biomedical examples*, Biometrics **49** (1993), 1099–1109.

[8] D. Heitjan and D. Rubin, *Ignorability and coarse data*, Annals of Statistics **19** (1991), 2244–2253.

[9] N. Johnson, *Systems of frequency curves generated by methods of translation*, Biometrika **36** (1949), 149–176.

[10] R. Kieschnick and B. McCullough, *Regression analysis of variates observed on (0,1): percentages, proportions and fractions*, Statistical Modelling **3** (2003), 193–213.

[11] P. Lachenbruch, *Comparisons of two-part models with competitors*, Statistics in Medicine **20** (2001), 1215–1234.

[12] ———, *Analysis of data with excess zeros*, Statistical Methods in Medical Research **11** (2002), 297–302.

[13] K. Lange, R. Little, and J. Taylor, *Robust statistical modeling using the t distribution*, Journal of the American Statistical Association **84** (1989), 881–896.

[14] E. Lesaffre and E. de Klerk, *Estimating the power of compliance - improving methods*, Control Clinical Trials **21** (2000), 540–551.

[15] E. Lesaffre, I. Scheys, J. Fröhlich, and E. Bluhmki, *Calculation of power and sample size with bounded outcome scores*, Statistics in Medicine **12** (1993), 1063–1078.

[16] P. McCullagh, *Regression models for ordinal data*, Journal of the Royal Statistical Society, Series B **42** (1980), 109–142.

[17] P. McCullagh and J. Nelder, *Generalized Linear Models*, 2nd ed., Chapman and Hall, London, 1989.

[18] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2004, ISBN 3-900051-00-3, http://www.r-project.org.

[19] J. Urquhart, *The electronic medication event monitor: Lessons for pharmacotherapy*, Clinical Pharmacokinetic **32** (1997), 345–356.

[20] W. Venables and B. Ripley, *Modern Applied Statistics with S*, 4th ed., Springer-Verlag, New York, 2002.

[21] B. Vrijens, A. Belmans, K. Matthijs, E. De Klerk, and E. Lesaffre, *Effect of patient intervention and compliance-enhancing pharmaceutical care on adherence with Atorvastatin*, Submitted for publication, 2004.

[22] H. Wainer and L. Brown, *Two statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data*, American Statistician **58** (2004), 117–123.

[23] J. Whitehead, *Sample size calculations for ordered categorical data*, Statistics in Medicine **12** (1993), 2257–2271.

[24] X.-H. Zhou and W. Tu, *Comparison of several independent population means when their samples contain log-normal and possibly zero observations*, Biometrics **55** (1999), 645–651.

Effect Size Δ/σ	Sample Size $n = n_1 + n_2$	Dist. Shape	Distributions Logistic-Normal	Logistic-t ($df = 4$)	Logistic-Logistic
0.0	40	Unimodal	0.064/0.056	0.054/0.045	0.044/0.042
		U-Shaped	0.041/0.038	0.050/0.047	0.056/0.048
		J-Shaped	0.055/0.050	0.057/0.058	0.064/0.058
	100	Unimodal	0.059/0.056	0.048/0.044	0.050/0.045
		U-Shaped	0.070/0.068	0.047/0.043	0.039/0.043
		J-Shaped	0.068/0.055	0.054/0.049	0.046/0.045
	200	Unimodal	0.068/0.068	0.044/0.046	0.053/0.053
		U-Shaped	0.045/0.045	0.051/0.051	0.053/0.048
		J-Shaped	0.058/0.052	0.045/0.052	0.060/0.048
0.2	40	Unimodal	0.083/0.069	0.116/0.109	0.097/0.076
		U-Shaped	0.097/0.089	0.127/0.123	0.098/0.097
		J-Shaped	0.062/0.057	0.077/0.073	0.084/0.065
	100	Unimodal	0.149/0.147	0.173/0.176	0.156/0.139
		U-Shaped	0.164/0.168	0.232/0.230	0.180/0.174
		J-Shaped	0.101/0.101	0.109/0.114	0.127/0.127
	200	Unimodal	0.264/0.262	0.303/0.323	0.262/0.269
		U-Shaped	0.255/0.254	0.368/0.380	0.282/0.283
		J-Shaped	0.179/0.175	0.182/0.196	0.182/0.178
0.5	40	Unimodal	0.305/0.279	0.367/0.368	0.331/0.315
		U-Shaped	0.296/0.295	0.416/0.410	0.336/0.335
		J-Shaped	0.213/0.195	0.217/0.200	0.213/0.180
	100	Unimodal	0.630/0.617	0.709/0.734	0.678/0.681
		U-Shaped	0.649/0.645	0.820/0.820	0.707/0.715
		J-Shaped	0.424/0.381	0.444/0.447	0.425/0.422
	200	Unimodal	0.917/0.900	0.949/0.965	0.917/0.913
		U-Shaped	0.919/0.913	0.978/0.981	0.954/0.952
		J-Shaped	0.728/0.689	0.724/0.759	0.707/0.714
1.0	40	Unimodal	0.816/0.783	0.873/0.882	0.847/0.831
		U-Shaped	0.805/0.788	0.928/0.926	0.844/0.861
		J-Shaped	0.581/0.531	0.586/0.583	0.566/0.535
	100	Unimodal	0.996/0.995	0.998/1.000	0.998/0.998
		U-Shaped	0.998/0.998	0.999/1.000	0.997/0.997
		J-Shaped	0.936/0.913	0.913/0.930	0.914/0.901
	200	Unimodal	1.000/1.000	1.000/1.000	1.000/1.000
		U-Shaped	1.000/1.000	1.000/1.000	1.000/1.000
		J-Shaped	0.997/0.997	0.999/0.999	0.996/0.998

Table 1: Simulation results (based on 1000 simulations) for the Binomial model with Normal random-effect versus Wilcoxon test when N_i equals 30. Each entry refers to the proportion of times the null hypothesis is rejected in the Binomial model/Wilcoxon test. In each case the model assumes normal distribution for the latent variable.

Effect Size Δ/σ	Sample Size $n = n_1 + n_2$	Dist. Shape	Distributions		
			Logistic-Normal	Logistic-t ($df = 4$)	Logistic-Logistic
0.0	40	Unimodal	0.031/0.042	0.042/0.048	0.026/0.049
		U-Shaped	0.038/0.057	0.046/0.050	0.047/0.035
		J-Shaped	0.048/0.046	0.039/0.043	0.041/0.043
	100	Unimodal	0.043/0.042	0.026/0.045	0.026/0.053
		U-Shaped	0.051/0.056	0.052/0.053	0.061/0.065
		J-Shaped	0.050/0.052	0.049/0.055	0.054/0.052
	200	Unimodal	0.047/0.047	0.058/0.058	0.056/0.049
		U-Shaped	0.051/0.056	0.048/0.045	0.051/0.046
		J-Shaped	0.047/0.050	0.047/0.052	0.040/0.043
0.2	40	Unimodal	0.080/0.087	0.056/0.102	0.150/0.097
		U-Shaped	0.096/0.098	0.101/0.100	0.085/0.082
		J-Shaped	0.075/0.064	0.081/0.083	0.088/0.079
	100	Unimodal	0.145/0.156	0.127/0.232	0.109/0.175
		U-Shaped	0.137/0.134	0.194/0.204	0.176/0.171
		J-Shaped	0.139/0.137	0.188/0.186	0.162/0.161
	200	Unimodal	0.188/0.275	0.191/0.363	0.174/0.286
		U-Shaped	0.249/0.253	0.374/0.388	0.273/0.282
		J-Shaped	0.209/0.213	0.387/0.314	0.275/0.263
0.5	40	Unimodal	0.273/0.280	0.226/0.397	0.184/0.309
		U-Shaped	0.297/0.295	0.405/0.403	0.322/0.327
		J-Shaped	0.270/0.267	0.385/0.350	0.287/0.297
	100	Unimodal	0.691/0.629	0.794/0.789	0.604/0.675
		U-Shaped	0.627/0.632	0.799/0.810	0.673/0.683
		J-Shaped	0.575/0.541	0.676/0.628	0.604/0.637
	200	Unimodal	0.934/0.929	0.810/0.975	0.895/0.938
		U-Shaped	0.895/0.889	0.978/0.983	0.938/0.941
		J-Shaped	0.874/0.866	0.911/0.943	0.885/0.902
1.0	40	Unimodal	0.851/0.826	0.878/0.901	0.858/0.845
		U-Shaped	0.823/0.806	0.945/0.949	0.857/0.860
		J-Shaped	0.778/0.731	0.788/0.807	0.722/0.759
	100	Unimodal	0.992/0.996	0.993/0.999	0.984/0.998
		U-Shaped	0.997/0.996	1.000/1.000	0.998/0.999
		J-Shaped	0.996/0.979	0.980/0.994	0.988/0.989
	200	Unimodal	1.000/1.000	0.986/1.000	0.995/1.000
		U-Shaped	1.000/1.000	1.000/1.000	1.000/1.000
		J-Shaped	1.000/1.000	1.000/1.000	1.000/1.000

Table 2: Simulation results (based on 1000 simulations) for the Coarsening model versus Wilcoxon test when m equals 10. Each entry refers to the proportion of times the null hypothesis is rejected in the Coarsening model/Wilcoxon test. In each case the model assumes normal distribution for the latent variable.

μ	σ	$\rho = 0.3$		$\rho = 0.7$	
		Coarsening/Wilcoxon	Binomial/Wilcoxon	Coarsening/Wilcoxon	Binomial/Wilcoxon
	1.0	1.000/1.000	1.000/1.000	1.000/1.000	1.000/1.000
0.0	2.0	0.866/0.816	0.806/0.756	0.988/0.798	0.959/0.784
	3.0	0.474/0.434	0.498/0.474	0.786/0.463	0.690/0.432
	1.0	1.000/1.000	0.996/0.996	1.000/1.000	1.000/0.994
1.5	2.0	0.868/0.826	0.794/0.727	0.982/0.792	0.946/0.745
	3.0	0.479/0.442	0.437/0.398	0.702/0.450	0.657/0.428
	1.0	1.000/1.000	0.984/0.972	1.000/1.000	1.000/0.980
3.0	2.0	0.790/0.754	0.705/0.638	0.960/0.789	0.892/0.662
	3.0	0.496/0.452	0.441/0.380	0.666/0.430	0.589/0.389

Table 3: Simulation results (based on 1000 simulations) for the Coarsening model and the Binomial model, respectively versus the Wilcoxon test using the treatment effect Δ obtained from the THAMES study. Each entry refers to the proportion of times the null hypothesis is rejected in the Coarsening model/Binomial model/Wilcoxon test.

μ	σ	$\rho = 0.3$		$\rho = 0.7$	
		Coarsening/Wilcoxon	Binomial/Wilcoxon	Coarsening/Wilcoxon	Binomial/Wilcoxon
	1.0	1.000/1.000	1.000/1.000	1.000/1.000	1.000/1.000
0.0	2.0	0.970/0.946	0.956/0.936	0.998/0.956	1.000/0.942
	3.0	0.708/0.659	0.682/0.658	0.929/0.671	0.886/0.644
	1.0	1.000/1.000	1.000/1.000	1.000/1.000	1.000/1.000
1.5	2.0	0.976/0.962	0.948/0.921	1.000/0.948	1.000/0.956
	3.0	0.695/0.669	0.695/0.648	0.905/0.688	0.864/0.636
	1.0	1.000/1.000	1.000/1.000	1.000/1.000	1.000/0.999
3.0	2.0	0.950/0.913	0.914/0.878	0.999/0.948	0.985/0.869
	3.0	0.697/0.645	0.622/0.569	0.876/0.639	0.824/0.601

Table 4: Simulation results (based on 1000 simulations) for the Coarsening model and the Binomial model, respectively versus the Wilcoxon test using the treatment effect Δ obtained from the survivor's part of the ECASS-1 study. Each entry refers to the proportion of times the null hypothesis is rejected in the Coarsening model/Binomial model/Wilcoxon test.

Variable	Levels	Intervention	Control	p-value
Adherence	–	0.926	0.895	0.011
Gender	Males	102(54.84%)	85(46.70%)	
	Females	84(45.16%)	97(53.30%)	0.116
Age	–	61.96	60.71	0.231
Weight	–	77.22	77.83	0.735
Work	Unemployed	46(24.73%)	64(35.16%)	
	Employed	140(75.27%)	118(64.84%)	0.029
Card. Risk	–	12.74	10.52	0.013
Fam. Hist.	No	145(77.54%)	142(78.02%)	
	Yes	42(22.46%)	40(21.98%)	0.911

Table 5: THAMES study: Comparison of baseline covariates for difference in the two groups. For categorical variables frequencies (percentages) and for continuous variables the mean, are reported.

Variable	Estimate	Std. Error	p-value
Intercept	0.545	0.249	0.029
Baseline	0.648	0.051	<0.001
Intervention	0.818	0.161	<0.001
Gender	−0.171	0.199	0.391
Age	0.009	0.011	0.403
Weight	0.005	0.006	0.429
Work	−0.240	0.233	0.302
Card. Risk	−0.021	0.012	0.064
History	−0.444	0.191	0.021
σ	1.433	0.062	

Table 6: THAMES study: Parameter estimates, standard errors and p-values for the Binomial model with normal random-effect for the Compliance data.

Variable	Estimate	Std. Error	p-value
Intercept	0.814	0.242	0.001
Baseline	0.545	0.055	<0.001
Intervention	0.706	0.163	<0.001
Gender	−0.142	0.193	0.465
Age	0.029	0.011	0.012
Weight	0.010	0.006	0.088
Work	−0.748	0.243	0.002
Card. Risk	−0.027	0.011	0.013
History	−0.609	0.175	<0.001
ϕ	46.623		

Table 7: THAMES study: Parameter estimates, standard errors and p-values for the Binomial model assuming overdispersion for the Compliance data.

Variable	Estimate	Std. Error	p-value
Intercept	−2.574	0.263	<0.001
Treatment	0.516	0.294	0.080
Gender	−0.340	0.307	0.268
Age	0.937	0.228	<0.001

Table 8: ECASS-1 study: Parameter estimates, standard errors and p-values for the logistic regression for the mortality rate.

Variable	Estimate	Std. Error	p-value
Intercept	2.655	0.258	<0.001
Treatment	1.059	0.331	0.001
Gender	−0.467	0.339	0.169
Age	−0.793	0.171	<0.001
σ	3.317	0.165	

Table 9: ECASS-1 study: Parameter estimates, standard errors and p-values for the Coarsening model for the Barthel index.

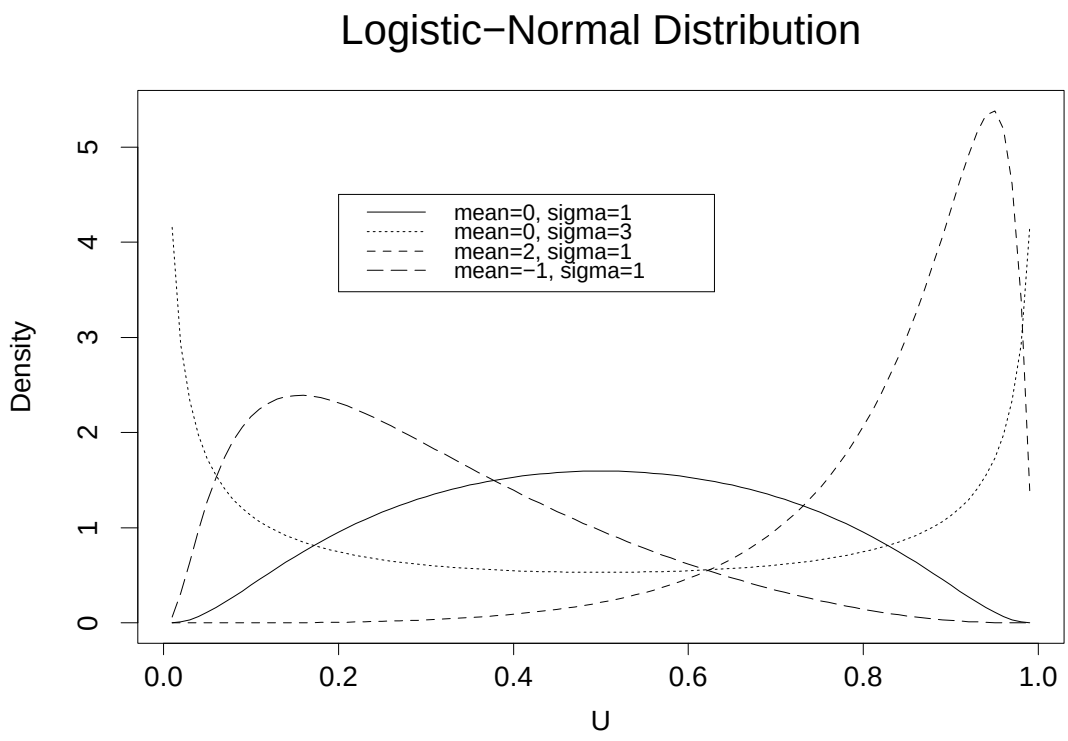


Figure 1: Different logistic-normal distributions

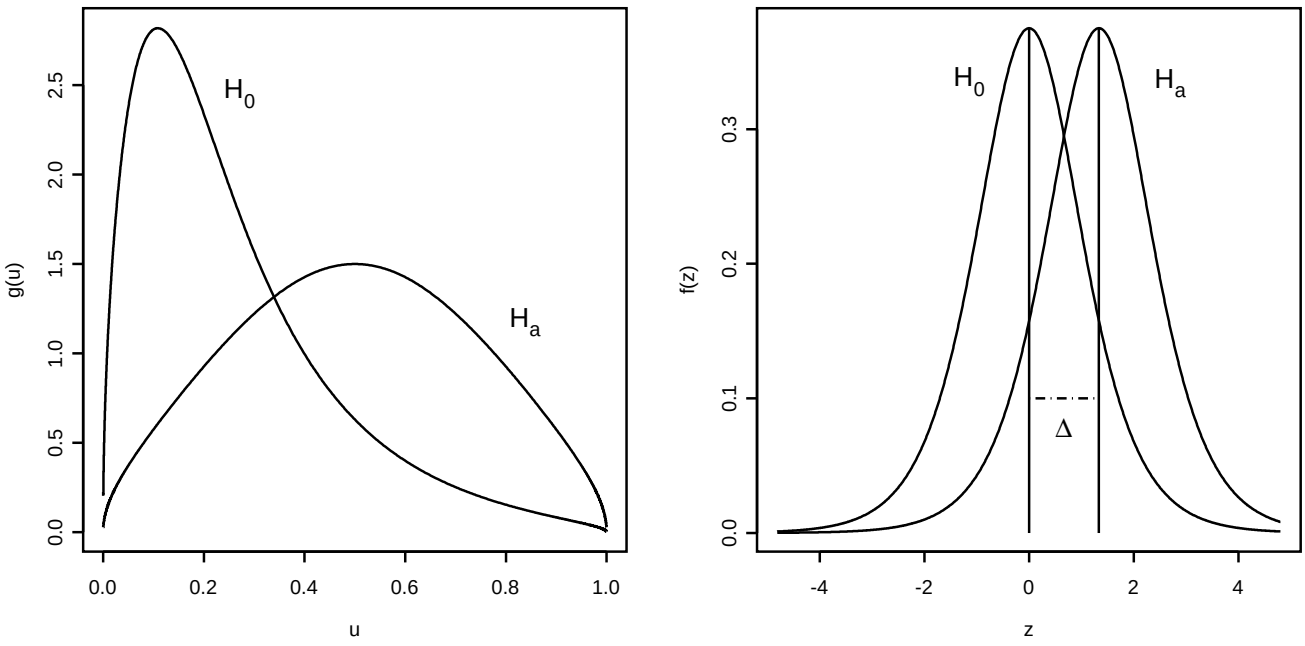
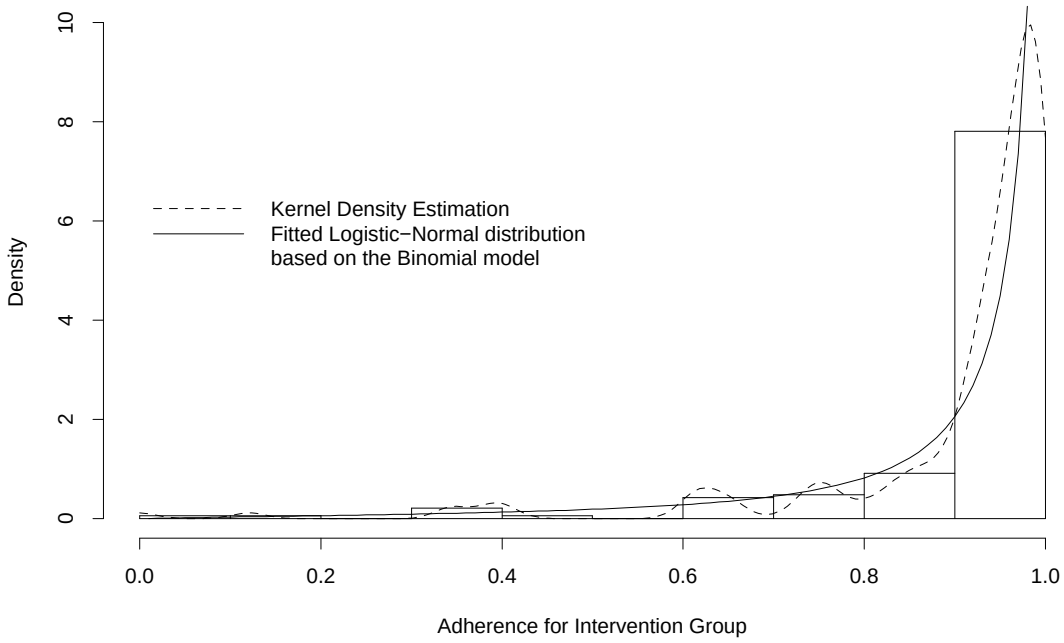


Figure 2: Correspondence of the location-shift alternative between the original and transformed scale



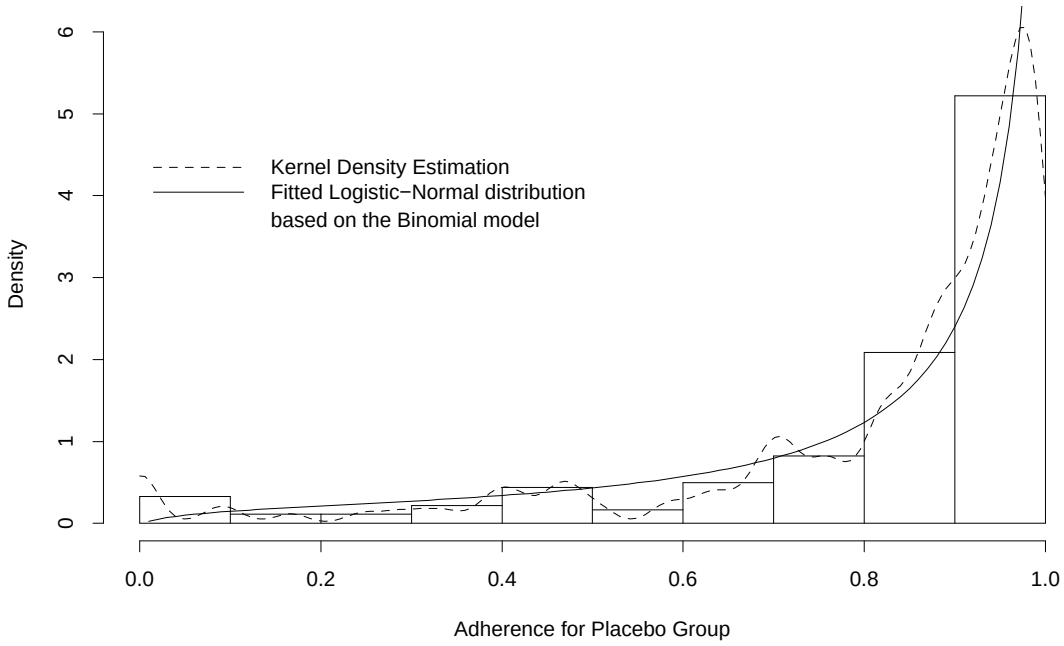


Figure 3: THAMES study: Proportion of correct dosing days

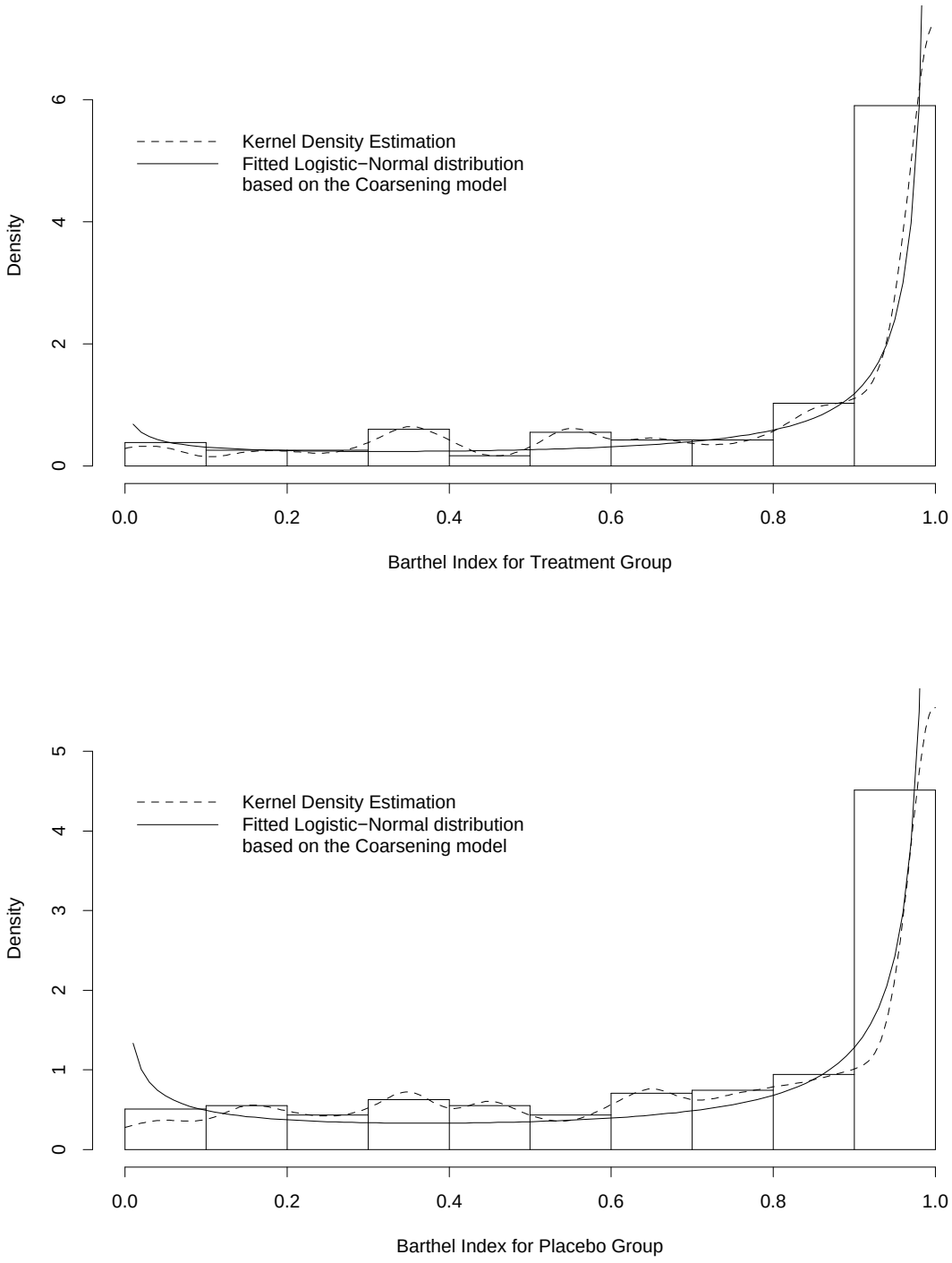


Figure 4: ECASS-1 study: Barthel Index for survivors