A note on the sample size calculation for a longitudinal Bernoulli distributed (binary) endpoint

Project Multivariate and Hierarchical Data – Discovering Associations

Steven Abrams, Luc Bijnens, Yannick Vandendijck, Johan Verbeeck

Objective

In this technical note we provide additional details concerning the sample size calculation steps required to determine the optimal sample size in case of a Bernoulli distributed endpoint (i.e., a binary variable), measured repeatedly over time (referred to as longitudinal data). In contrast to the Poisson distributed outcome to quantify the freshness of roses (i.e., the number of days a rose stays fresh) discussed elsewhere, one could opt to consider the freshness variable measured each day at the level of each individual rose as outcome variable in subsequent statistical analyses. In this note, we describe the relationship between the two approaches and focus on sample size computations in case of longitudinal Bernoulli distributed endpoints with an explicit link to the black tulip project.

First, we introduce the basic concepts related to the Bernoulli distribution. Second, we relate this to generalized linear (mixed) models in which the (transformed) mean of a random variable Y, conditional on covariates, is regressed linearly as a function of these covariates. In general, a generalized linear model (GLM) for independent observations $(Y_i, \boldsymbol{x}_i)$, with $i = 1, \ldots, n$, can be formulated as:

$$\begin{aligned} Y_i|\boldsymbol{x}_i &\sim f_Y(\cdot) \\ g\left[E(Y_i|\boldsymbol{x}_i)\right] &= \boldsymbol{\beta}\boldsymbol{x}_i^T \end{aligned}$$

where f_Y is a density (e.g., a Bernoulli probability mass distribution, Poisson probability mass distribution or Gaussian probability density function) from the exponential family, representing the (1) distributional assumption in a GLM. Moreover, $g(\cdot)$ represents the (2) link-function and $\boldsymbol{\beta}\boldsymbol{x}_i^T$ the (3) systematic component of a GLM. A generalized linear mixed model (GLMM) provides an extension of a GLM to accommodate clustering, for example, as a result of repeated measurements on the same subject (e.g., rose), through the specification of subject-specific random effects. The GLMM framework for binary data is outlined below. Finally, we will describe in detail the different sample size calculation steps required to obtain the optimal sample size for a Bernoulli GLMM for longitudinal binary data.

Important properties of the Bernoulli distribution

A Bernoulli experiment is a chance with two potential outcomes, i.e. failure or success. Hence, a discrete random variable Y follows a Bernoulli distribution with parameter $0 \leq \pi \leq 1$, i.e., the probability of success, if it has a probability mass function for $y \in \{0, 1\}$ given by

$$f(y;\pi) \equiv f(y) = P(Y = y) = \begin{cases} \pi & \text{if y = 1} \\ 1 - \pi & \text{if y = 0} \end{cases}$$

The probability mass function can also be expressed as follows:

$$f(y;\pi) = \pi^y \left(1 - \pi\right)^{1-y},$$

for $y \in \{0, 1\}$. The expectation of a Bernoulli distributed random variable is equal to $E(Y) = \pi$ and the variance equals Var$(Y) = \pi(1 - \pi)$.

Furthermore, the Bernoulli distribution is a special case of the binomial distribution characterizing a chance experiment in which a Bernoulli experiment is repeated n times with the same probability of success and independently of each other.

Generalized linear mixed model with (longitudinal) Bernoulli outcome

We will discuss the different steps of the sample size calculation in the context of a potential primary endpoint for the study identifying the best compound to preserve a freshly cut rose (cfr. *Project Discovering Associations* within the course P-MHD). More specifically, let Y_i represent the number of days flower i, $i = 1, \ldots, n$, stays fresh. Given covariate information $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, with p the number of covariates, Y_i follows a Poisson distribution with conditional mean $\lambda(\boldsymbol{x}_i) \equiv E(Y_i|\boldsymbol{x}_i)$. A GLM with log-link (natural logarithm) can be formulated as follows:

$$\begin{aligned} Y_i|\boldsymbol{x}_i &\sim \text{Poisson}\left(\lambda(\boldsymbol{x}_i)\right) \\ \log\left[\lambda(\boldsymbol{x}_i)\right] &= \boldsymbol{\beta}\boldsymbol{x}_i^T = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \end{aligned}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ the (row) vector of model parameters. Rose-specific covariate information can include the compound chosen to preserve the rose (recall that we consider $C = 14$ compounds next to distilled water in the experiment), the species of the rose (Floribunda or Hybrid Tea), the garden in which the rose has grown (northern or southern garden), etc.

As an alternative to the aggregated count outcome, one could opt to rely on the daily freshness evaluations for each of the individual roses until withering. More specifically, let W_{it} represent the freshness indicator (1 = flower is fresh, 0 = flower is withered) for flower i, $i = 1, \ldots, n$, at time point (day) t with $t = 1, \ldots, D$ and D represents the maximum preservation time in the study, i.e., the day the last rose withers. In this case, we can formulate the following (random intercept) generalized linear mixed model for the freshness indicator W_{it} conditional on flower- and time-specific covariate information $\boldsymbol{x}_{it} = (x_{it1}, \ldots, x_{itq})$, with q the number of covariates:

$$\begin{aligned} W_{it}|\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b} &\sim \text{Bernoulli}\left(\pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b})\right) \\ \text{logit}\left[\pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b})\right] &= \boldsymbol{\beta}\boldsymbol{x}_{it}^T + \boldsymbol{b}\boldsymbol{z}_i^T = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_q x_{itq} + b_i, \end{aligned}$$

where $\boldsymbol{b} = (b_1, \ldots, b_n)$ are random intercepts for the n different flowers, $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ii-1}, z_{ii}, z_{ii+1}, \ldots, z_{in}) = (0, \ldots, 0, 1, 0, \ldots, 0)$ and $b_i \sim N(0, \sigma_b^2)$. The flower-specific random effects accommodate for the association between measurements of freshness on the same flower. The variance σ_b^2 represents the between-flower variability. This model is referred to as a generalized linear mixed model (GLMM) with a random intercept for flowers. Note that the model can be extended to include other sources of variability, including but not limited to, for example, rater, bush and subplot effects, which are (potentially) nested. Finally, the covariates in $\boldsymbol{x}_{it}$ are rose- and time-specific covariates such as the ones in $\boldsymbol{x}_i$, including the day at which covariates are evaluated in a flower.

Some distributional reflections

Consider now the following random variable:

$$Y_i^* = \sum_{t=1}^{D} W_{it}$$

for $i = 1, \ldots, n$. Clearly, the conditional mean and variance (i.e., conditional on covariate information $\boldsymbol{x}_i^*$, including all flower-specific covariate information related to flower i, and random effects) of the random variable Y_i^* are equal to

$$\begin{aligned} E(Y_i^*|\boldsymbol{x}_i^*, \boldsymbol{z}_i, \boldsymbol{b}) &= \sum_{t=1}^{D} E(W_{it}|\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b}) = \sum_{t=1}^{D} \pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b}) = \lambda^*(\boldsymbol{x}_i^*, \boldsymbol{z}_i, \boldsymbol{b}) \\ \text{Var}(Y_i^*|\boldsymbol{x}_i^*, \boldsymbol{z}_i, \boldsymbol{b}) &= \sum_{t=1}^{D} \text{Var}(W_{it}|\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b}) = \sum_{t=1}^{D} \pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b})\left[1 - \pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b})\right] = \lambda^*(\boldsymbol{x}_i^*, \boldsymbol{z}_i, \boldsymbol{b}) - \sum_{t=1}^{D} \pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b})^2, \end{aligned}$$

for the latter expression relying on conditional independence of the random variables W_{it} given the random effects. The (conditional) distribution of the random variable Y_i^* is known as the Poisson-binomial distribution, a generalisation of the binomial distribution to the sum of (independent) Bernoulli distributed random variables that do not necessarily have the same success probabilities. Furthermore, the conditional variance is bounded above by the mean $\lambda^*(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{b})$. Despite the fact that the Poisson-binomial distribution is not exactly equal to a Poisson distribution in general, the Poisson distribution approximates the Poisson-binomial distribution for small probabilities $\pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b})$ and $D \to \infty$ similar to the binomial case. Consequently, the distributional assumptions in the longitudinal binary case and Poisson case are not exactly the same.

One can easily express the mean number of days a flower stays fresh, in terms of the aggregate endpoint Y_i^* defined above, conditional on covariate information $\boldsymbol{x}_i^*$, in terms of the probability of freshness of flower i, $i = 1, \ldots, n$, at time t, $t = 1, \ldots, D$ as follows:

$$\lambda^*(\boldsymbol{x}_i^*) = E(Y_i^* | \boldsymbol{x}_i^*) = E\left(\sum_{t=1}^{D} W_{it} \middle| \boldsymbol{x}_i^* \right) = \sum_{t=1}^{D} E\left(W_{it} | \boldsymbol{x}_i^* \right) = \sum_{t=1}^{D} \pi(\boldsymbol{x}_{it}),$$

where $\pi(\boldsymbol{x}_{it})$ represents the marginal probability of freshness defined as

$$\pi(\boldsymbol{x}_{it}) = \int_{-\infty}^{\infty} \pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b}) f_{b_i}(b_i) db_i.$$

This is particularly useful when relating the effect size in terms of mean number of days of freshness to the effect size at the level of the (evolution of the) probability of freshness over time.

For the calculation of the variance of Y_i^*, conditional on the random effect b_i, we rely on the assumption of conditional independence of the random variables W_{it}. However, this assumption implies that

$$\begin{aligned} P(W_{it} = w_t, W_{i(t+1)} = w_{t+1} \mid b_i) &= P(W_{it} = w_t \mid b_i) P(W_{i(t+1)} = w_{t+1} \mid b_i) \\ &= \pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b})^{w_t} \left[1 - \pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b})\right]^{1-w_t} \pi(\boldsymbol{x}_{i(t+1)}, \boldsymbol{z}_i, \boldsymbol{b})^{w_{t+1}} \left[1 - \pi(\boldsymbol{x}_{i(t+1)}, \boldsymbol{z}_i, \boldsymbol{b})\right]^{1-w_{t+1}}, \end{aligned}$$

for $w_t, w_{t+1} \in \{0, 1\}$. Consequently, we have under (conditional) independence

$$P(W_{it} = 0, W_{i(t+1)} = 1 \mid b_i) = \left[1 - \pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b})\right] \pi(\boldsymbol{x}_{i(t+1)}, \boldsymbol{z}_i, \boldsymbol{b})$$

which is not equal to zero unless $\pi(\boldsymbol{x}_{i(t+1)}, \boldsymbol{z}_i, \boldsymbol{b}) = 0$. Needless to say that after a flower withers at time t, $W_{i(t+1)}$ can only take value zero, hence, the aforementioned probability should be equal to zero. This induces a temporal association in repeated measurements on the same flower which can be modelled either directly or by including more complicated random effects (e.g., random slopes) into the model. We consider the latter approach beyond the scope of this technical note. A direct modelling approach could consider the formulation of a model in which the likelihood contribution for flower i, $i = 1, \ldots, n$ becomes

$$L_i(\boldsymbol{\beta} \mid \boldsymbol{w}_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{b}) = f(w_{i1}, \ldots, w_{iD} \mid \boldsymbol{b}) = f_{D|D-1}(w_{iD} \mid w_{i(D-1)}) \times \ldots \times f_{2|1}(w_{i2} \mid w_{i1}),$$

where $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iD})$ and relying on the Markov property, i.e., W_{it} only depends on $W_{i(t-1)}$ (and not on other W_{it^*}). Furthermore, one can easily show that

$$f_{t|t-1}(w_{it} \mid w_{i(t-1)}) = (1 - w_{it})^{1-w_{i(t-1)}} \left[\left(\frac{\pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b})}{\pi(\boldsymbol{x}_{i(t-1)}, \boldsymbol{z}_i, \boldsymbol{b})} \right)^{w_{it}} \left(1 - \frac{\pi(\boldsymbol{x}_{it}, \boldsymbol{z}_i, \boldsymbol{b})}{\pi(\boldsymbol{x}_{i(t-1)}, \boldsymbol{z}_i, \boldsymbol{b})} \right)^{1-w_{it}} \right]^{w_{i(t-1)}}.$$

A marginal likelihood can be obtained by numerically integrating out the random effects $\boldsymbol{b}$ in the likelihood function $L(\boldsymbol{\beta} \mid \boldsymbol{w}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{b}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta} \mid \boldsymbol{w}_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{b})$. Again we will not discuss this further here.

Interpretation of the fixed treatment effect

In the context of finding the best compound (as compared to distilled water) among the $C = 14$ candidate compounds, $C = 14$ dummy variables could be included as covariates in the GLMM. However, these fixed compound effects enable the probability of freshness to be different as compared to distilled water at baseline (time $t = 0$) rather than implying a differential evolution in performance over time. Due to randomization of flowers to different compounds, we do not expect differences between compounds and distilled water at baseline and one could opt to exclude fixed compound effects from the model.

Interpretation of the interaction between treatment and time

The compound effect of interest is quantified as model parameters associated with the interaction between compound and time. More specifically, the interaction between the dummy variables associated with each of the $C = 14$ compounds and time allow for a differential evolution of the probability of freshness over time, i.e., ignoring other covariate information without loss of generality, the model becomes:

$$W_{it}|\boldsymbol{x}_{it}, b_i \sim \text{Bernoulli}\left(\pi(\boldsymbol{x}_{it}, b_i)\right)$$
$$\text{logit}\left[\pi(\boldsymbol{x}_{it}, b_i)\right] = \beta_0 + \sum_{j=1}^{C} \beta_j x_{itj} x_{it15} + \beta_{15} x_{it15} + b_i,$$

where

$$x_{itj} = \begin{cases} 1 & \text{if compound of flower } i = j \\ 0 & \text{otherwise} \end{cases}$$

and $x_{it15} = t$ represents the time at which flower i is evaluated for its freshness. Note that β_{15} is the model parameter quantifying the time effect for distilled water. Furthermore, time-specific parameters $\beta_j x_{it15}$ represent the difference in log odds of flower freshness between compound j and distilled water (as reference category) for a fixed time point t given that

$$\log\left[\text{odds}_j(t)\right] - \log\left[\text{odds}_0(t)\right] = \left[\beta_0 + (\beta_j + \beta_{15})x_{it15} + b_i\right] - \left[\beta_0 + \beta_{15} x_{it15} + b_i\right] = \beta_j x_{it15},$$

hence β_j refers to the difference in *slopes* between compound j and distilled water at the linear predictor scale.

Given that the probability of freshness decreases with increasing time, β_{15} is expected to have a value smaller than zero. Consequently, assessing whether a specific compound j, $j = 1, \ldots, C$, performs better than distilled water (as a reference) implies testing whether β_j is significantly larger than zero, implying a slower decrease in probability of freshness over time as compared to distilled water (see below).

Steps in sample size calculation (SSC)

First of all, we summarize the different steps required to perform any sample size calculation:

1. Specify parameter of interest, statistical hypothesis and test
2. Specify the significance level
3. Specify the effect size (or equivalence limit)
4. Obtain values or estimates of other parameters needed (e.g., variance parameters)
5. Specify a target value for the power

Step 1: Specify the parameter and statistical hypothesis

The primary objective of this study is to identify the *best* compound among the $C = 14$ candidate compounds to preserve the cut rose. In order to do so, we consider the following (superiority) testing problem

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_C = 0$$
$$H_1 : \exists j \in \{1, \ldots, C\} : \beta_j > 0,$$

with β_j representing the difference in log odds of freshness between compound j, $j = 1, \ldots, 14$, and distilled water at a fixed time point t. In order to power our study, we will consider the reduced testing problem, comparing a single compound j^* with distilled water as follows:

$$H_0 : \beta_{j^*} = 0$$
$$H_1 : \beta_{j^*} > 0.$$

The final sample size n will be equal to the estimated sample size times 15 different solutions (14 compounds plus distilled water). Note that our testing problem is a one-sided problem which is important to consider in sample size calculations.

Step 2: Specify the significance level

The **significance level**, denoted here as α, is equal to selected threshold for the (family-wise) Type I error probability. More specifically, $\alpha = P(H_1|H_0)$ and is usually considered to be equal to 5% (in statistical literature; see, e.g., Fisher, 1925, 1935). Changes in significance level(s) for individual testing problems in the context of multiple testing are discussed below.

Step 3: Specify the effect size

The **effect size** δ is defined as the minimal biologically relevant (i.e., of scientific interest and clinical importance) (treatment) effect that one wants to be able to detect given the study design. One way to define δ is directly in terms of the relative increase in *slope* for the evolution of the probability of freshness for a compound relative to distilled water. For example, one could argue that a 10% increase in *slope* should be detectable based on the experiment. Here, *slope* refers to the coefficient corresponding to the linear time effect in the systematic component (hence, at the transformed scale). More specifically, we have then

$$\frac{\beta_{15} + \beta_{j^*}}{\beta_{15}} = 0.90,$$

implying that $\delta = \beta_{j^*}/\beta_{15} = -0.10$, hence, $\beta_{j^*} = -0.10\beta_{15} > 0$, if $\beta_{15} < 0$. Needless to say that the interpretation of the relative difference in slopes between the best compound and distilled water in terms of the difference in mean duration of preservation of the flower is difficult. Alternatively, one could translate the effect size δ in terms of the relative difference in slopes to a direct impact in terms of the difference in mean number of days a flower stays fresh in the best compound, say $\lambda^*_{j^*}$, as compared to distilled water, denoted by λ^*_0. For example, a one day difference in means could be considered and a translation to a relative difference in slopes could be performed by solving the following equation for β_{j^*}:

$$\lambda^*_{j^*} - \lambda^*_0 = \sum_{t=1}^{D} (\pi_{tj^*} - \pi_{t0}) = 1,$$

where

$$\pi_{tj^*} = \int_{-\infty}^{\infty} \pi_{tj^*}(b) f(b) db$$

and $\pi_{tj^*}(b)$ is the conditional probability of freshness for a subject with random effect b. Note that the aforementioned integral has no closed form expression, hence, the solution of this equation can only be determined numerically.

The effect size is one of the crucial ingredients driving the sample size in the sense that a too small effect size could lead to an extreme inflation of the sample size. Hence, be reasonable when selecting the effect size, and determine it together with the cross-functional team.

Step 4: Obtain values or estimates of other parameters needed

Often you need to specify values for other important parameters such as **variance parameters** (cfr. when conducting a two-sample t-test to compare two means one needs to determine an estimate of the (un)common variance of measurements in both groups). In this case, we need information about the between-flower variability, i.e., the random effects variance σ_b^2, next to estimates for β_0 and the *slope* β_{15} quantifying the evolution of the probability of freshness over time for distilled water. In order to estimate β_0, β_{15} and σ_b^2 we can rely on pilot data (or historical data, if available).

Step 5: Specify a target value for the power

The **power** (of a test) is defined as the probability to reject the null hypothesis given that the alternative hypothesis is true. More specifically, it is equal to the probability with which the desired effect size will be detected, i.e., power $= 1 - \beta = P(H_1|H_1)$ (with β the type II error probability). Typical values for the power range between 80% and 90%. An increase in power will result in an increase in the required sample size.

Performing the sample size calculation

In many (complicated) settings, no analytical sample size formulas are directly available. In such settings, the only way to compute the sample size is by means of simulations. A simulation approach can be applied in all circumstances, even for study designs for which sample size formulas based on parametric distributional assumptions are available in statistical software. Needless to say, a simulation approach is computationally more expensive than computations based on analytical formulas and it requires programming skills. For the longitudinal binary (Bernoulli) setting we are considering in this technical note, we will need to rely on simulations.

Consider a pilot study including information about n_p roses being preserved in distilled water and followed up longitudinally over time (with daily recording of the freshness of each of the flowers). Consequently, the following GLMM can be fitted to the observed data (without correcting for potential systematic (fixed) species and garden effects), while preserving the aforementioned notation:

$$\begin{aligned} W_{it}|x_{it15}, b_i &\sim \text{Bernoulli}\,(\pi_{t0}) \\ \text{logit}\,(\pi_{t0}) &= \beta_0 + \beta_{15} x_{it15} + b_i, \end{aligned}$$

where $b_i \sim N(0, \sigma_b^2)$. Estimates of β_0, β_{15} and σ_b^2 are obtained using, for example, maximum likelihood (ML) inference.

Poisson outcome data generation

For the first exercise (simulation setting 1), we start from aggregated Poisson data with the mean in the *control* group equal to $\hat{\lambda}_0 = 7.6$ and an effect size in terms of an increase in mean number of days of freshness of one day.

GLM approach

Here, we fit a GLM for binary data to the generated datasets (with a specified sample size n) and test for the significance of the interaction effect between treatment and time. The estimated power curve based on these simulations (with $\alpha = 0.05$, right-tailed one-sided testing and 1000 simulation runs with default seed number 1234) as a function of the sample size (per group) is graphically depicted in Figure 1.

In order to achieve a power of 80%, a sample size of about $n \approx 50$ is required for each (compound) group based on these simulations. Clearly, a larger group-specific sample size is required if an achieved power of 90% is preferred. Recall that the total sample size is obtained by multiplying n with 15 (i.e., the number of groups including $C = 14$ compounds + distilled water). The rather limited sample size is obtained as a result of ignoring dependence between the observations, thereby overestimating the effective sample size.

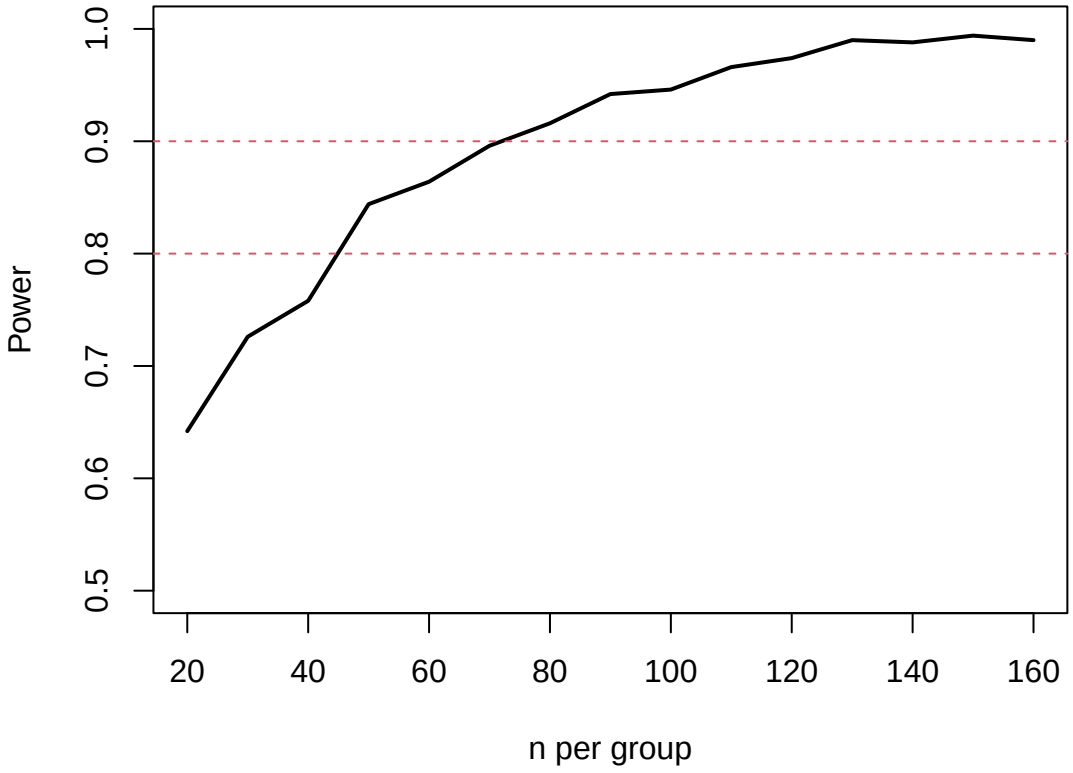

Figure 1: Estimated power-curve (GLM approach; simulation setting 1) showing the relationship between the sample size per group (denoted by n) and the achieved power. Significance level $\alpha = 0.05$, right-tailed one-sided hypothesis testing and 1000 simulations.

Generalized Estimating Equations (GEE) approach

In this subsection, we show the results of fitting a model relying on the GEE approach to estimate robust standard error estimates. More specifically, we consider an exchangeable working correlation structure and estimate the model parameters based on the simulated data. Clearly, the required sample size to reach the targeted power level(s) is much larger than the sample size obtained by presuming the observations to be independent (compare Figure 2 with Figure 1). Again this is related to the fact that the precision in the GLM approach is overestimated as compared to the inflation of the standard error estimates in the GEE approach.

Bernoulli outcome data generation

In the second simulation setting, we directly generate longitudinal binary outcome data (simulation setting 2). More specifically, we consider $\hat{\beta}_0 = 7.500$, $\hat{\beta}_{15} = -1.300$ and $\hat{\sigma}_b^2 = 0.500$. The effect size, being the relative difference in slope (at the linear predictor scale), is considered to be $\delta = -0.100$ (scenario 1; solid line in Figure 3) or $\delta = -0.075$ (scenario 2; dashed line in Figure 3). The estimated power curves showing the relationship between the sample size per group and achieved power are based on fitting a binary GLMM to the generated data and performing a right-tailed one-sided Wald-based hypothesis test at 5% significance level. Note that the data generation requires a regularization step to ensure that withered flowers are not considered fresh at a later stage in the generated longitudinal profile.

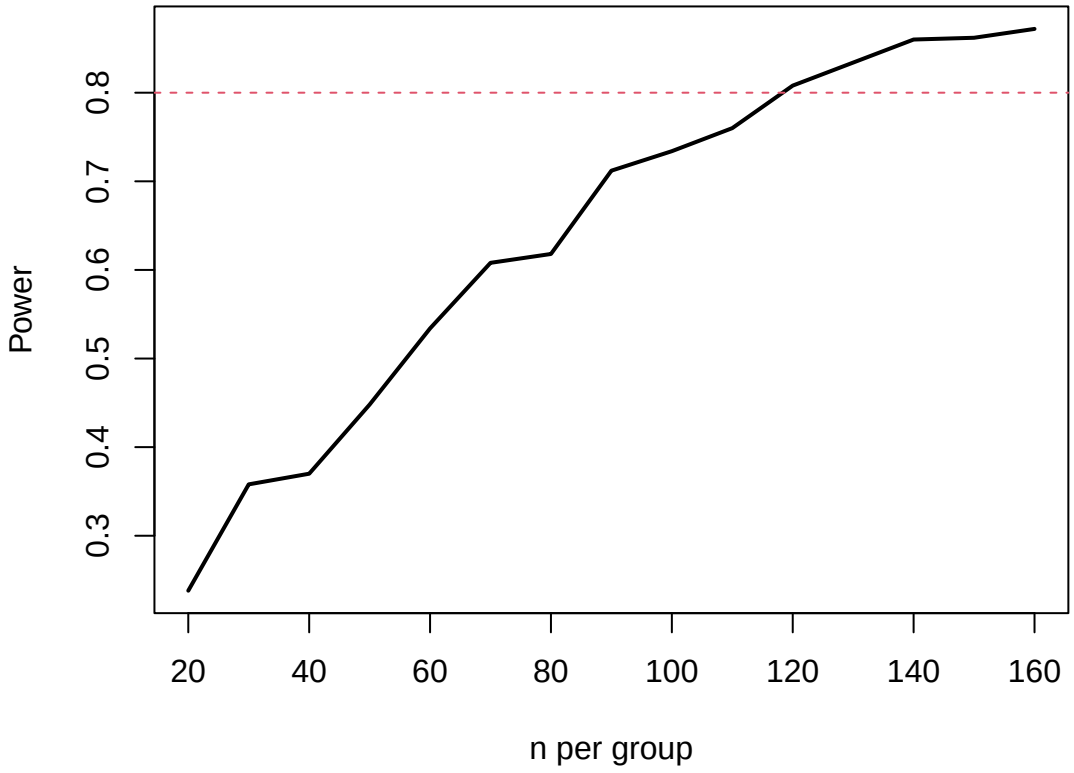

Figure 2: Estimated power-curve (GEE approach; simulation setting 1) showing the relationship between the sample size per group (denoted by n) and the achieved power. Significance level $\alpha = 0.05$, right-tailed one-sided hypothesis testing and 1000 simulations.

GLMM approach

From the figure it is clear that the required sample size to achieve a power of 80% is about 40 flowers per group for the scenario with $\delta = -0.100$ and around 60 flowers with $\delta = -0.075$. Needless to say, the latter scenario implies a smaller difference in mean number of days of freshness. In order to compare these scenarios with the count data case, we calculated the simulation-based mean duration of freshness in both the control and treatment group for the two scenarios. For scenario 1, the mean duration of freshness based on the simulated data is 6.268 and 6.913 days for the control and treatment group, respectively. In scenario 2, the difference is only about half a day (with the mean in the treatment group equal to 6.741 days). Consequently, a larger sample size is needed to be able to detect a difference of half a day in terms of the mean duration of freshness based on the longitudinal binary data at hand.

Important remarks

In this technical note, we focus on the use of subject-specific models to model the longitudinal binary outcome data. However, marginal alternatives are available as well in which the interpretation of the estimated (treatment) effects is population-averaged, rather than subject-specific, which is especially useful in the context of binary outcome data. More specifically, in case of a Bernoulli distributed endpoint the interpretation of the treatment effects in a GLMM is not marginal or population-averaged. If one is interested in such marginal effects, one requires a marginalization step in which one integrates out the random effects relying on the (estimated) random

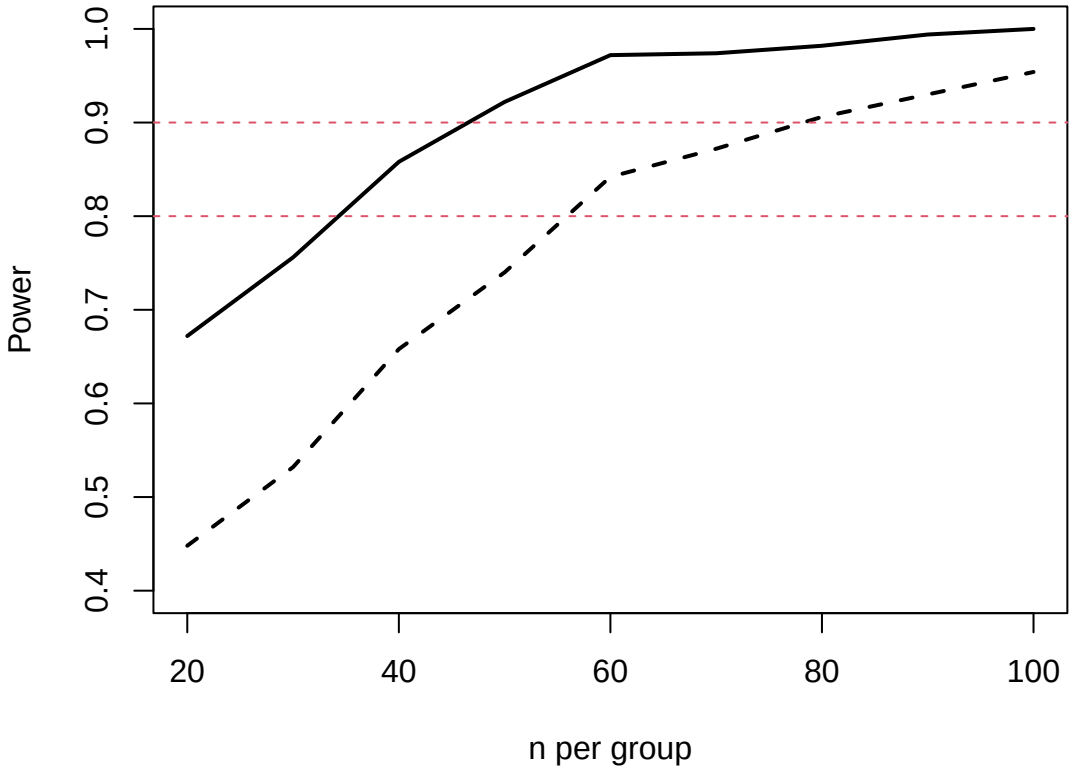

Figure 3: Estimated power-curve (GLMM approach; simulation setting 2) showing the relationship between the sample size per group (denoted by n) and the achieved power. Significance level $\alpha = 0.05$, right-tailed one-sided hypothesis testing and 1000 simulations (solid line: $\delta = -0.1$; dashed line: $\delta = -0.075$).

effects distribution. One can easily show that in a linear mixed model, the subject-specific interpretation of model parameters coincides with a population-averaged interpretation. The same applies for a Poisson outcome, albeit that the intercept does not have a population-averaged interpretation. Alternatively, a marginal GEE approach could be considered to account for the temporal association (for example, autoregressive working correlation structure for the association between consecutive measurements on the same flower). Again, the interpretation is different as compared to the interpretation of the effects in a GLMM for binary outcome data. In the sample size calculations shown here, we demonstrate the use of the GEE approach when simulating longitudinal binary outcome data starting from an underlying Poisson process. Needless to say, this exercise could be repeated when generating binary data directly.

Here we considered a longitudinal binary endpoint, as opposed to a count endpoint discussed in the other technical note, however, alternative formulations could be considered as well. For example, one could opt to model the time until withering of a flower as an (interval-censored) time-to-event process for which survival models could be used. Extensions of a Cox proportional hazards model or accelerated failure time model are available to encompass clustering, i.e., so-called frailty models. For more details, the reader is referred to [3].

References

[1] Fisher, R. A. (1925). Statistical methods for research workers. Oliver and Boyd Ltd, Edinburgh.

[2] Fisher, R. A. (1935). The Design of Experiments. Macmillan.

[3] Wienke, A. (2010). Frailty models in survival analysis. Chapman and Hall/CRC.