A note on the sample size calculation for a Poisson distributed endpoint with overdispersion

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 Poisson distributed outcome, in the absence or presence of (additional) variability (overdispersion) as a result of having clustered observations. First of all, we introduce the basic concepts related to the Poisson 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) can be formulated as (for independent observations $(Y_i, \boldsymbol{x}_i)$ with $i = 1, \ldots, n$):

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

where f_Y is a density (e.g., a 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. Finally, we will describe in detail the different sample size calculation steps required to obtain the optimal sample size for a Poisson GLM and for a GL(M)M accounting for additional variation (i.e., overdispersion, for example, by clustering).

Important properties of the Poisson distribution

A discrete random variable Y follows a Poisson distribution with parameter $\lambda > 0$, if it has a probability mass function given by

$$f(y; \lambda) \equiv f(y) = P(Y = y) = \frac{\lambda^y \exp(-\lambda)}{y!}.$$

The most important property of a Poisson distributed random variable is:

$$\lambda = E(Y) = \mathsf{Var}(Y).$$

Consequently, the mean-variance relationship in a Poisson model is rather restrictive as a single parameter describes both the mean as well as the variance. Other more flexible discrete non-negative distributions are available to encompass overdispersion (i.e., when the observed variability in the data exceeds the one implied by the restrictive Poisson mean-variance relation), including the negative binomial or quasi-Poisson distributions.

Generalized linear mixed model with Poisson 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. Assume now that additional variation in the endpoint is introduced as a result of a random rater effect (i.e., different raters assessing the longevity of the roses introduce additional measurement variation). The aforementioned GLM can be extended as follows:

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

where $\boldsymbol{b} = (b_1, \ldots, b_R)$ for R raters and $b_r \sim N(0, \sigma_b^2)$, $r = 1, \ldots, R$, for rose i being assessed by rater r (i.e., implying that the random effects' design vector $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ir-1}, z_{ir}, z_{ir+1}, \ldots, z_{iR}) = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$). The variance σ_b^2 represents the between-rater variability. This model is referred to as a generalized linear mixed model (GLMM) with a random intercept for raters. Note that the model can be extended to include other sources of variability, including but not limited to, for example, bush and subplot effects, which are potentially nested.

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

$$\begin{aligned} H_0 &: \lambda_0 = \lambda_1 = \ldots = \lambda_C \\ H_1 &: \exists j \in \{1, \ldots, C\} : \lambda_j > \lambda_0, \end{aligned}$$

with λ_j representing the mean number of days the flower stays fresh when being preserved in compound j, $j = 1, \ldots, 14$, and λ_0 the mean for distilled water (reference category). In order to power our study, we will consider the reduced testing problem, comparing a single compound j^* with distilled water as follows:

$$\begin{aligned} H_0 &: \lambda_0 = \lambda_{j^*} \\ H_1 &: \lambda_{j^*} > \lambda_0, \end{aligned}$$

which can be reformulated in terms of the parameter $\delta = \lambda_{j^*} - \lambda_0$

$$\begin{aligned} H_0 &: \delta = 0 \\ H_1 &: \delta > 0. \end{aligned}$$

The final sample size n will be equal to the estimated sample size times 15 different solutions (14 compounds plus distilled water). Based on the calculated sample size, all differences between the compound-specific means and the reference mean λ_0 larger than δ will be able to be detected. In order to compare two means, we will rely on a Wald test (see below). 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. For the purpose of this technical note, we select the effect size to be equal to $\delta = 1$ day albeit that other values could be considered. Hence, a difference of one day will turn out to be statistically significant, given that the SSC is based on the underlying study design. 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). However, as described previously, in the Poisson case the mean-variance equivalence implies that estimating the mean number of days a rose stays fresh in distilled water, i.e., λ_0, immediately implies the variance of the Poisson distributed endpoint, at least in the absence of overdispersion (i.e., additional variation imposed by the raters - see below). Given the effect size δ selected in Step 3, the mean number of days for compound j^* is equal to $\lambda_{j^*} = \lambda_0 + \delta$. In order to estimate λ_0 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 using 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 Poisson setting we are considering in this note, we will need to rely on simulations (unless we rely on a normal approximation of the underlying Poisson distribution which is only valid if the mean is sufficiently large). As

mentioned previously, pilot data providing information with regard to the variance immediately provide knowledge about the mean. In order to clarify the different steps in the sample size calculation, we first introduce the SSC in case of no overdispersion. Next, we will discuss the SSC in the context of additional sources of variability.

SSC in the absence of overdispersion

Consider a pilot study including information about n_p roses being preserved in distilled water. Consequently, the following GLM can be fitted to the observed data (without correcting for potential systematic (fixed) species and garden effects):

$$Y_i \sim \text{Poisson}\,(\lambda_0)$$
$$\log(\lambda_0) = \beta_0$$

An estimate of the mean λ_0 is obtained using maximum likelihood (ML) inference, i.e., $\hat{\lambda}_0 = \exp(\hat{\beta}_0)$ is the ML estimate for λ_0. Based on the Poisson density specified previously, the loglikelihood function for this model is given by:

$$ll(\beta_0|\boldsymbol{y}) = \sum_{i=1}^{n_p} \{y_i \text{log}(\lambda_0) - \lambda_0 - \text{log}(y_i!)\} = \sum_{i=1}^{n_p} \{y_i\beta_0 - \exp(\beta_0) - \text{log}(y_i!)\},$$

in which the latter term does not depend on β_0. The ML estimate for β_0, denoted by $\hat{\beta}_0$, can be obtained as the solution of the equation:

$$\sum_{i=1}^{n_p} \left\{y_i - \exp(\hat{\beta}_0)\right\} = n_p \left\{\bar{y} - \exp(\hat{\beta}_0)\right\} = 0.$$

Consequently, $\hat{\beta}_0 = \text{log}(\bar{y})$ with $\bar{y}$ the arithmetic mean of the n_p observations. Given an effect size δ, the mean in the *treatment* group is equal to $\hat{\lambda}_0 + \delta$. For this exercise, $\hat{\lambda}_0 = 7.6$ and $\delta = 1$. The following simulation code can be used (for both one- and two-sided testing problems):

```
Poisson_sims <- function(n_grid, lambda0, lambda1, alpha, test = "two_sided",
                         n_sims = 10000, seed_nr = 1234){

  power_vec <- matrix(nrow = 1, ncol = length(n_grid))

  for (j in 1:length(n_grid)){
    # 1. Choose sample size per group
    N <- n_grid[j]

    # 2. Select parameters
    lambda.control = lambda0
    lambda.treated = lambda1
    alpha = alpha

    # 3. Simulate huge number of experiments and test
    numberSimulation <- n_sims
    pval <- numeric(numberSimulation)
    zval <- numeric(numberSimulation)

    set.seed(seed_nr)
    for (i in 1:numberSimulation){
      # We simulate from Poisson distribution
```

```
      controlGroup <- rpois(N, lambda = lambda.control)
      treatedGroup <- rpois(N, lambda = lambda.treated)
      simData <- data.frame(response = c(controlGroup, treatedGroup),
      treatment = rep(c(0, 1), each = N))

      # We use a GLM model for Poisson regression to test effect of treatment
      # (Wald test)
      glm_fit <- summary(glm(response ~ treatment, data = simData, family=poisson()))
      pval[i] <- glm_fit$coeff["treatment", "Pr(>|z|)"]
      zval[i] <- glm_fit$coeff["treatment", "z value"]

      if (test == "greater" & zval[i] > 0){
        pval[i] <- pval[i]/2
      }
      if (test == "greater" & zval[i] < 0){
        pval[i] <- 1 - (pval[i]/2)
      }
      if (test == "less" & zval[i] < 0){
        pval[i] <- pval[i]/2
      }
      if (test == "less" & zval[i] > 0){
        pval[i] <- 1 - (pval[i]/2)
      }

    }

    # 4. Estimate power
    power_vec[j] = sum(pval < alpha)/numberSimulation
  }

  return(list(n_grid = n_grid, power_vec = power_vec))
}
```

Based on the simulations (with $\alpha = 0.05$, right-tailed one-sided testing and 1000 simulation runs with default seed number 1234), we obtain the following power-curve as a function of the sample size (per group) (see Figure 1):

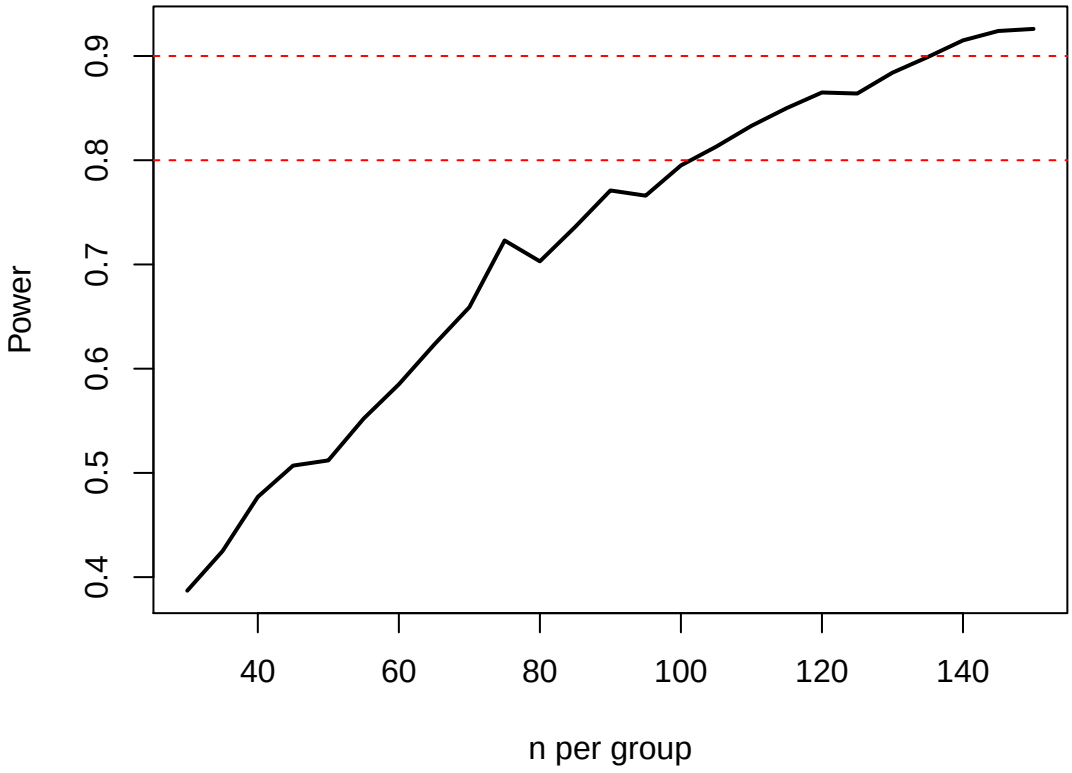

Figure 1: Estimated power-curve showing the relationship between the sample size per group (denoted by n) and the achieved power in the absence of overdispersion. Significance level $\alpha = 0.05$, right-tailed one-sided hypothesis testing and 1000 simulations.

In order to achieve a power of 80%, a sample size of at least $n \approx 100$ 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).

SSC in the presence of overdispersion

Finally, we study the sample size calculation in the presence of overdispersion. As described previously, such additional variability can be due to, for example, different raters that assess the freshness of flowers. Even in case of objective evaluation criteria, different raters could value freshness of the same flower in a different way. We will now highlight different ways of dealing with this overdispersion in a sample size calculation setting.

Approach 1: Ignoring overdispersion

One way to deal with the additional variation is to completely ignore the (potential) overdispersion introduced by the raters while estimating the variance (mean) from pilot data coming from different raters. In order to study the impact thereof on the sample size calculation, we start from the GLMM formulation introduced earlier to have

$$\log\left[\lambda(b_r)\right] = \beta_0 + b_r,$$

with $b_r \sim N(0, \sigma_b^2)$. This implies that the conditional expectation of Y_i, given the random rater effect b_r, is equal to

$$E(Y_i|b_r) = \lambda(b_r) = \exp(\beta_0 + b_r) = \exp(\beta_0)\exp(b_r).$$

Furthermore, the unconditional mean $E(Y_i)$, marginalized over the normal random effects distribution is given by

$$E(Y_i) = \int_{-\infty}^{\infty} \lambda(b) f_b(b) db = \int_{-\infty}^{\infty} \exp(\beta_0)\exp(b) f_b(b) db = \exp(\beta_0)\exp(\sigma_b^2/2).$$

From this derivation, one can conclude that when fitting a GLM to overdispersed Poisson data we will overestimate the control mean λ_0 (for distilled water) for an average rater with a factor $\exp(\sigma_b^2/2) > 1$. Consequently, the increase in variability will increase the sample size. More specifically, fitting the model

$$\begin{aligned} Y_i &\sim \text{Poisson}\,(\lambda_0) \\ \log(\lambda_0) &= \gamma_0, \end{aligned}$$

leads to $\hat{\lambda}_0 = \exp(\hat{\gamma}_0) = \exp(\hat{\beta}_0 + \hat{\sigma}_b^2/2)$ implying that

$$\delta = \hat{\gamma}_{j^*} - \hat{\gamma}_0 = \exp(\hat{\sigma}_b^2/2)\left\{\exp(\hat{\beta}_{j^*}) - \exp(\hat{\beta}_0)\right\}.$$

Given that $\exp(\hat{\sigma}_b^2/2) > 1$ and for a specific sample size, the true effect size $\exp(\hat{\beta}_{j^*}) - \exp(\hat{\beta}_0)$ when correcting for overdispersion, is smaller than δ. Hence, this approach is a **conservative one** (i.e., leading to a higher sample size) towards the sample size calculation, given the fact that the sample size is overestimated as compared to the optimal sample size (see also Approach 3). In order to visualize the impact of this approach on the sample size, we graphically depict the power-curve, given that the true underlying between-rater variability σ_r^2 is equal to 0.5 (see Figure 2). Clearly, the sample size required to power the study is larger than in the previous setting, given the inflation of the variability.

Note that this approach implies the specification of the effect size δ at the level of the population-averaged mean number of days a flower stays fresh in distilled water as compared to a specific compound.

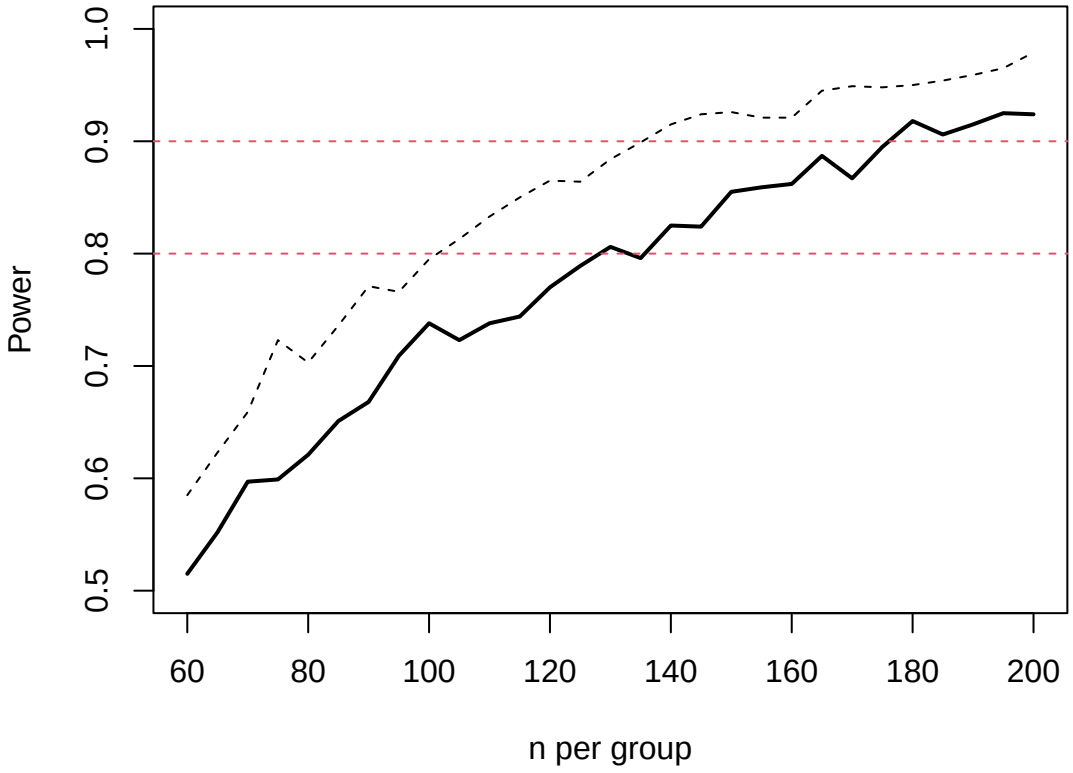

Figure 2: Estimated power-curve (black solid line) showing the relationship between the sample size per group (denoted by n) and the achieved power in the presence of overdispersion while ignoring it in the SSC. Significance level $\alpha = 0.05$, right-tailed one-sided hypothesis testing and 1000 simulations. The black dashed line shows the estimated power-curve under the baseline scenario without overdispersion.

Approach 2: Use pilot data from a single rater

Confining attention to a single rater in the pilot data leads to the estimation of a single rater-specific mean λ_{0r} (for rater r) with

$$\lambda_{0r} = \exp(\beta_0)\exp(b_r) = \exp(\gamma_0).$$

Depending on the sign of b_r we either overestimate (if $b_r > 0$ implying that the rater is less strict in defining freshness as compared to the average rater) or underestimate (if $b_r < 0$ implying that the rater is more strict as compared to the average rater) the average number of days that a rose stays fresh when assessed by an average rater. As a consequence, the obtained sample size is either too large, or too small (in view of the selected effect size and power, and relative to the optimal sample size). This approach is therefore not preferred.

Approach 3: Accommodate overdispersion in the SSC

The third option is to quantify the between-rater variability σ_b^2 based on the pilot data and either (1) use an ad hoc way of adjusting the sample size given the additional variability (cfr. use $\exp(\hat{\beta}_0)$ and $\exp(\hat{\beta}_0) + \delta$ as means in the sample size computations), or (2) introduce the observed variability explicitly in the simulation procedure. The latter approach is preferred given the fact that random variation in rater-specific random effects (intercepts) (as well as uncertainty in estimated means/variances) can easily be accounted for in the simulation approach. Moreover, it is also preferred since the model will closely match the model used in the final analysis.

Note that in this approach the specification of the effect size δ is done at the level of the mean number of days a flower stays fresh in distilled water as compared to a specific compound when assessed by an average rater (i.e., $\delta = \exp(\beta_{j^*}) - \exp(\beta_0)$), hence, correcting for overdispersion will decrease the sample size (as compared to suboptimal though conservative Approach 1).

As an illustration, we assume that we have a total of ten raters, each of them rating an equal number of roses (implying that the sample size increases with steps of size 10). Again, the between-rater variability is assumed to be equal to 0.5. As an additional step in the simulations, we will generate rater-specific random effects from a normal distribution with mean 0 and variance σ_b^2.

```
Poisson_sims_ext <- function(n_grid, lambda0, lambda1, alpha, n_clusters, sigma2_b,
                             test = "two_sided", n_sims = 10000, seed_nr = 1234){

  power_vec <- matrix(nrow = 1, ncol = length(n_grid))

  for (j in 1:length(n_grid)){
    # 1. Choose sample size per group
    N <- n_grid[j]

    # 2. Select parameters
    lambda.control = lambda0
    lambda.treated = lambda1
    alpha = alpha

    # 3. Simulate huge number of experiments and test
    numberSimulation <- n_sims
    pval <- numeric(numberSimulation)
    zval <- numeric(numberSimulation)

    set.seed(seed_nr)
    for (i in 1:numberSimulation){
      # Rater-specific random effects (n_clusters)
      b <- rep(rnorm(n_clusters, mean = 0, sd = sqrt(sigma2_b)), each = N/n_clusters)
      cluster_id <- rep(1:n_clusters, each = N/n_clusters)
      # We simulate from Poisson distribution taking into account random effects b
      # (N per group)
      controlGroup <- rpois(N, lambda = lambda.control*exp(b))
      treatedGroup <- rpois(N, lambda = lambda.treated*exp(b))
      simData <- data.frame(response = c(controlGroup, treatedGroup),
                            treatment = rep(c(0, 1), each = N),
                            cluster_id = rep(cluster_id,2))
      # We use a GLMM model for Poisson regression to test effect of treatment
      # (Wald test)
      glmer_fit <- summary(glmer(response ~ treatment + (1 | cluster_id),
                           data = simData, family=poisson()))
      pval[i] <- glmer_fit$coeff["treatment", "Pr(>|z|)"]
      zval[i] <- glmer_fit$coeff["treatment", "z value"]

      if (test == "greater" & zval[i] > 0){
        pval[i] <- pval[i]/2
      }
      if (test == "greater" & zval[i] < 0){
```

```
      pval[i] <- 1 - (pval[i]/2)
    }
    if (test == "less" & zval[i] < 0){
      pval[i] <- pval[i]/2
    }
    if (test == "less" & zval[i] > 0){
      pval[i] <- 1 - (pval[i]/2)
    }

  }

  # 4. Estimate power
  power_vec[j] = sum(pval < alpha)/numberSimulation
 }

 return(list(n_grid = n_grid, power_vec = power_vec))
}
```

In Figure 3, we graphically depict the power-curve showing the relation between the group-specific sample size n and the achieved power (taking into account the experimental study design with roses being evaluated by one of ten potential raters). We clearly observe a decrease in required sample size as compared to Approach 1 (see black dashed line in Figure 3).

Next to treating the rater effect as a random effect, one could also opt to correct for the variation imposed by different raters evaluating the roses through the inclusion of a fixed effect for raters in the model. In that way, the generalized linear mixed model reduces to a generalized linear model with 8 additional parameters (i.e., 10 raters implying 9 model parameters associated with 9 dummy variables indicating which rater assessed a specific flower, while the random effects variance does not require estimation in the GLM as compared to the GLMM). In Figure 4, we quantify the impact of switching from a fixed effects to random effects perspective while correcting for overdispersion imposed by raters. Clearly the achieved power in relation to the sample size is not affected by the choice between a random or fixed effect for rater, at least for the range of group-specific sample sizes required to achieve a sufficiently high power.

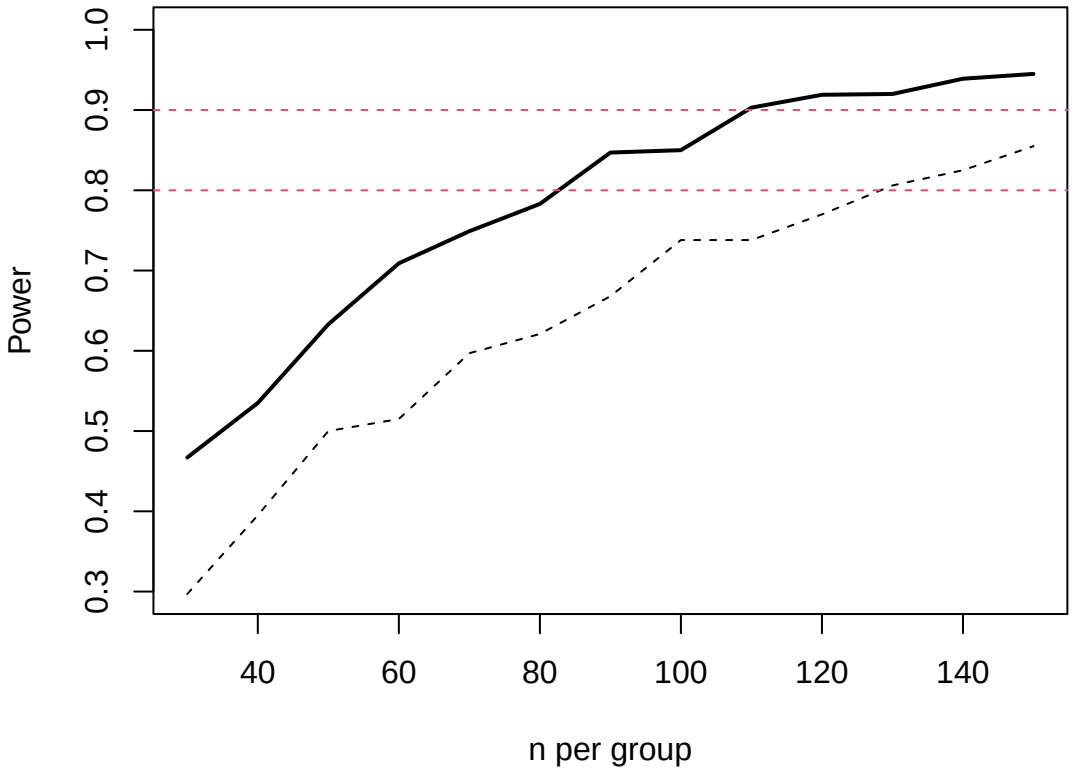

Figure 3: Estimated power-curve (black solid line) showing the relationship between the sample size per group (denoted by n) and the achieved power in the presence of overdispersion while accounting for it in the SSC. Significance level $\alpha = 0.05$, 10 clusters (raters) with an equal number of roses per cluster, right-tailed one-sided hypothesis testing and 1000 simulations. The black dashed line shows the estimated power-curve when ignoring the overdispersion (Approach 1).

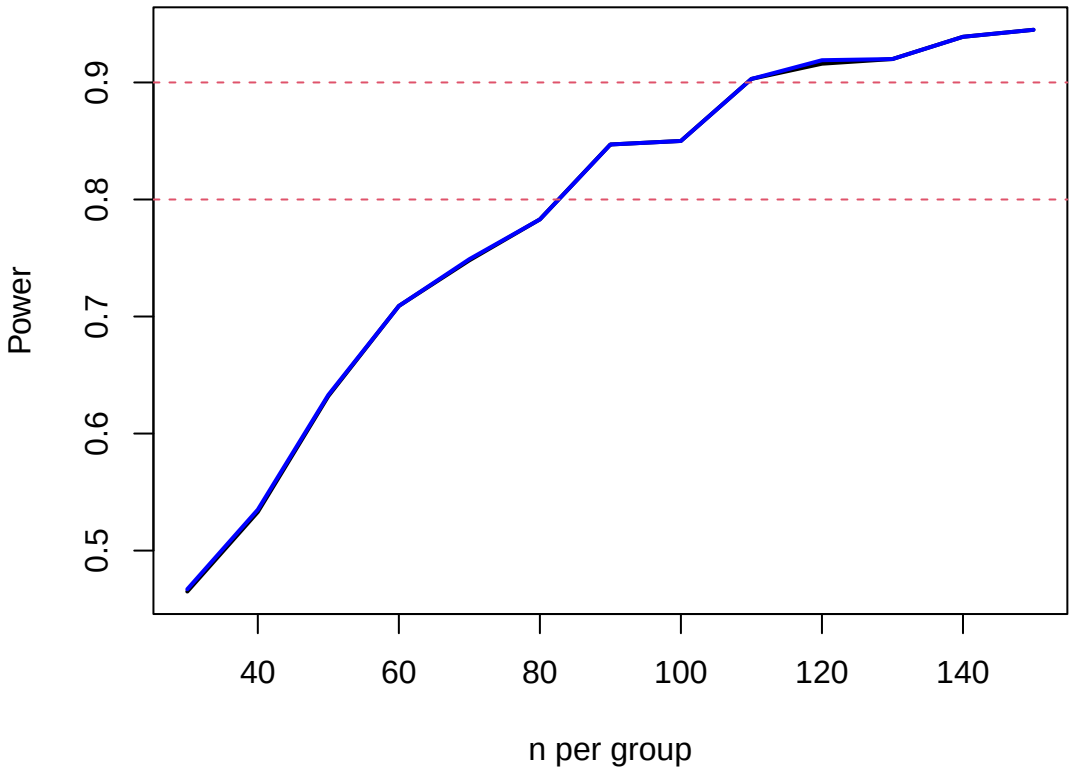

Figure 4: Estimated power-curve showing the relationship between the sample size per group (denoted by n) and the achieved power in the presence of overdispersion while accounting for it in the SSC using either a random effect (intercept) for raters (GLMM approach; blue line) or a fixed effect for raters (GLM approach; black line). Significance level $\alpha = 0.05$, 10 clusters (raters) with an equal number of roses per cluster, right-tailed one-sided hypothesis testing and 1000 simulations.

Multiplicity adjustments

Multiple testing is not accounted for in the previous (sub)sections. A straightforward way of adjusting the sample size to correct for the number of comparisons done in the final analysis (e.g., as a result of repeatedly comparing each of the compounds with distilled water) is by directly altering the significance level. More specifically, a Bonferroni-corrected significance level for each of the individual testing problems is equal to the selected family-wise α-level divided by the total number of tests (i.e., $\alpha^* = \alpha/K$ with K the number of comparisons). In order to illustrate the impact of including a Bonferroni correction, we performed the sample size calculation in absence of overdispersion (cfr. results depicted in Figure 1) with a significance level $\alpha^* = \alpha/C = 0.05/14 \approx 0.00357$ and other parameters fixed to aforementioned values.

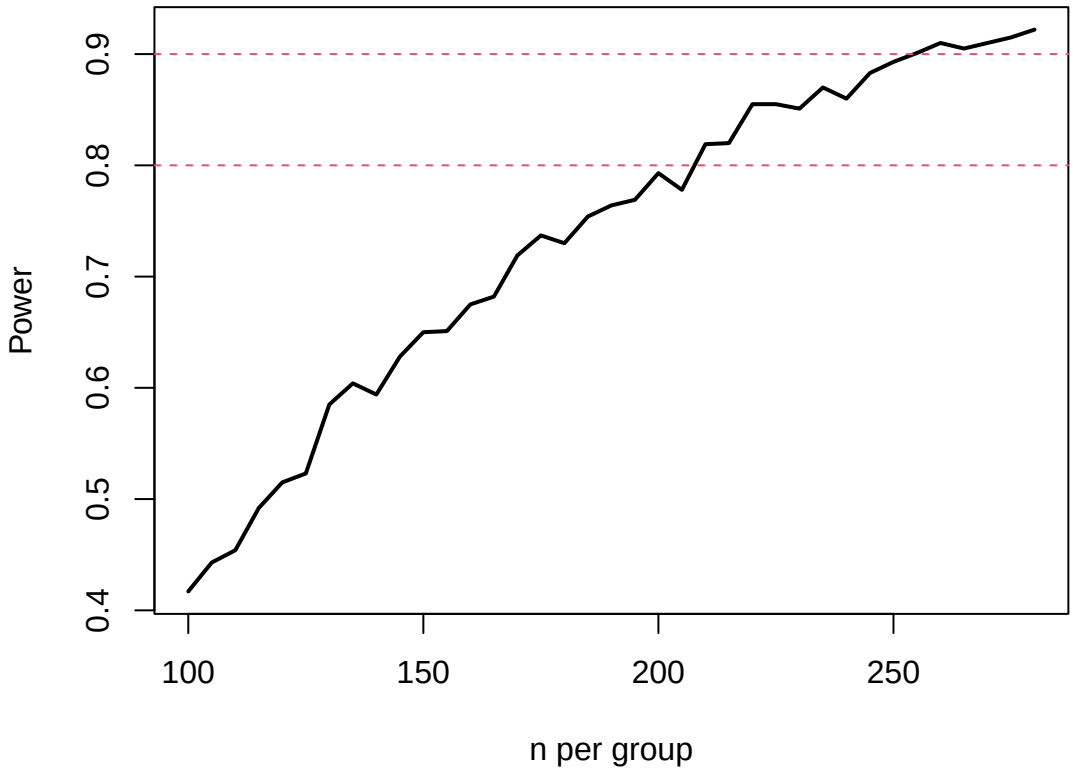

Figure 5: Estimated power-curve showing the relationship between the sample size per group (denoted by n) and the achieved power in the absence of overdispersion. Significance level $\alpha = 0.05/14 \approx 0.00357$, right-tailed one-sided hypothesis testing and 1000 simulations.

Bonferroni method

The Bonferroni method provides a direct and easy way of correcting for multiple testing. However, the approach is conservative in the sense that the nominal family-wise significance level is often strictly smaller than the target α. More specifically, the method provides a correct upper bound on the overall (family-wise) Type I error probability of falsely rejecting at least one hypothesis in the whole set of hypotheses considered. Although this upper bound holds regardless of the relationships between the hypotheses, in many cases the bound may be too high.

Formally, consider C statistical hypotheses problems $H^{(1)}, \ldots, H^{(C)}$ with corresponding p-values $p_1, \ldots, p_C$, respectively. Let C_0 represent the number of true null hypotheses (presumably unknown to the researcher).

Consequently, the family-wise error rate (FWER), defined as the probability of rejecting at least one true null hypothesis, that is, making at least one type I error, can be controlled using the Bonferroni correction. More specifically, for each individual testing problem $H^{(j)}$, $j = 1, \ldots, C$, we reject the null hypothesis if and only if the corresponding p-value p_j is smaller or equal than $\alpha^* = \alpha/C$. Following Boole's inequality, we have for p-values $p_{(k)}$, $k = 1, \ldots, C_0$, corresponding to the null hypotheses being true:

$$\text{FWER} = P\left[\bigcup_{k=1}^{C_0}\left(p_{(k)} \leq \frac{\alpha}{C}\right)\right] \leq \sum_{k=1}^{C_0}\left[P\left(p_{(k)} \leq \frac{\alpha}{C}\right)\right] = C_0\frac{\alpha}{C} \leq \alpha.$$

From the previous inequality, it is clear that the Bonferroni correction does not require additional assumptions regarding the dependence among hypotheses $H^{(1)}, \ldots, H^{(C)}$ or about C_0, the number of null hypotheses being true. Consequently, if C is large and/or the test statistics for $H^{(1)}, \ldots, H^{(C)}$ are positively correlated, FWER $< \alpha$ and the Bonferroni correction is considered conservative. This implies that the correction potentially increases the probability of producing false negative test results, thereby reducing statistical power.

Bonferroni-Holm method

The Bonferroni-Holm method, also referred to as the Holm method, sequential Bonferroni or Holm-Bonferroni method, is an alternative approach to counteract the problem of multiple comparisons. Similar to the Bonferroni method, the Bonferroni-Holm method intends to control the FWER and is uniformly more powerful than the Bonferroni correction.

The Bonferroni-Holm method (for a FWER being not larger than α) is conducted according to the following steps:

1. Suppose we have C hypotheses $H^{(1)}, \ldots, H^{(C)}$ that correspond to the **ordered** (from lowest to highest) p-values $p_{(1)}, \ldots, p_{(C)}$
2. If $p_{(1)} \leq \alpha/C$, reject the null hypothesis for problem $H^{(1)}$ and continue to the next step; otherwise stop.
3. If $p_{(2)} \leq \alpha/(C-1)$, reject the null hypothesis for problem $H^{(2)}$ and continue to the next step; otherwise stop.
4. ...
5. If $p_{(k)} \leq \alpha/(C-k+1)$, reject the null hypothesis for problem $H^{(k)}$ and continue to the next step; otherwise stop.
6. ...
7. If $p_{(C)} \leq \alpha$, reject the null hypothesis for problem $H^{(C)}$ and stop.

One can mathematically show that the approach ensures that the FWER is smaller or equal to α (based on the Bonferroni inequalities; not shown here). Bonferroni-Holm corrected p-values can be calculated as follows:

$$\tilde{p}_{(k)} = \max_{j \leq k}\left[(C-j+1)p_{(j)}\right]_1,$$

where $[x]_1 = \min(x, 1)$.

How can we integrate this data-driven approach in the aforementioned sample size calculation example (in the absence of overdispersion)? In order to integrate the Bonferroni-Holm method in the simulation approach, we rely on the following reasoning. When considering the simple Bonferroni method, we powered our study for compound-specific means being at least one day larger than the mean freshness duration in distilled water. Indeed, the resulting sample size based on a significance level $\alpha^* = \alpha/C$ was multiplied with $C + 1$, thereby presuming that $\lambda_j \geq \lambda_0 + \delta$ for all $j = 1, \ldots, C$. In the extreme case that all compounds are performing better than distilled water, we can extend the simulation set-up assuming the inclusion of $C = 14$ *treatments* and computing the Bonferroni-Holm corrected p-values in each simulation run. The achieved power is then derived

as the probability of correctly rejecting a single null hypothesis obtained by averaging the power for each of the compound-specific testing problems $H^{(j)}$, $j = 1, \ldots, C$ (see R code below). Results of the Bonferroni-Holm method are presented in Figure 6.

```
Poisson_sims_ext2 <- function(n_grid, lambda0, lambda1, ngroups, alpha,
                              test = "two_sided", n_sims = 10000, seed_nr = 1234,
                              method = "Holm"){

  power_vec <- matrix(nrow = 1, ncol = length(n_grid))

  for (j in 1:length(n_grid)){
    # 1. Choose sample size per group
    N <- n_grid[j]

    # 2. Select parameters
    lambda.control = lambda0
    for (group_id in 1:ngroups){
     lambda.group = paste0("lambda.treated",group_id)
     assign(lambda.group, lambda1)}
    alpha = alpha

    # 3. Simulate huge number of experiments and perform tests
    numberSimulation <- n_sims
    pval <- matrix(0, nrow = numberSimulation, ncol = ngroups)
    zval <- matrix(0, nrow = numberSimulation, ncol = ngroups)

    set.seed(seed_nr)
    for (i in 1:numberSimulation){
      # We simulate from Poisson distribution
      controlGroup <- rpois(N, lambda = lambda.control)
      treatedGroup <- vector()
      for (group_id in 1:ngroups){
       lambda.group = paste0("lambda.treated",group_id)
       treatedGroup = c(treatedGroup, rpois(N, lambda = get(lambda.group)))}

      simData <- data.frame(response = c(controlGroup, treatedGroup),
                            treatment = rep(c(0, 1:ngroups), each = N))
      # We use a GLMM model for Poisson regression to test effect of treatment
      # (Wald test)
      glm_fit <- summary(glm(response ~ as.factor(treatment), data = simData,
                                    family=poisson()))
      pval[i,] <- glm_fit$coeff[-1, "Pr(>|z|)"]
      zval[i,] <- glm_fit$coeff[-1, "z value"]

      for (k in 1:ngroups){
       if (test == "greater" & zval[i,k] > 0){
         pval[i,k] <- pval[i,k]/2
       }
       if (test == "greater" & zval[i,k] < 0){
         pval[i,k] <- 1 - (pval[i,k]/2)
       }
```

```
       if (test == "less" & zval[i,k] < 0){
         pval[i,k] <- pval[i,k]/2
       }
       if (test == "less" & zval[i,k] > 0){
         pval[i,k] <- 1 - (pval[i,k]/2)
       }
      }

      # Multiplicity adjustment(s)
      if (method == "Bonferroni"){
           pval[i,] = p.adjust(pval[i,], method = "bonferroni")
      }

      if (method == "Holm"){
          pval[i,] = p.adjust(pval[i,], method = "holm")
      }

      if (method == "Benjamini-Hochberg"){
          pval[i,] = p.adjust(pval[i,], method = "BH")
      }

    }

    # 4. Estimate power
    power_vec[j] = mean(apply(pval, 2,
         FUN = function(x){sum(x < alpha)}))/numberSimulation
  }

  return(list(n_grid = n_grid, power_vec = power_vec))
}
```

Benjamini-Hochberg method

In contrast to the previous methods, i.e., Bonferroni and Bonferroni-Holm corrections, the Benjamini-Hochberg (BH) procedure focuses on controlling the false discovery rate (FDR) rather than the FWER. More specifically, the FDR is defined as the (expected) ratio of the number of false discoveries (i.e., rejecting the null hypothesis while it is true) to the total number of discoveries (i.e., all rejections of the null hypothesis). FDR-controlling procedures provide less stringent control of the Type I errors of each of the individual tests as compared to FWER-controlling methods, including the ones described above. Consequently, FDR-controlling procedures have greater power at the cost of an increased FWER. Although not preferred in this context, the previous simulation code does allow for the specification of the BH-method with following steps (again based on a step-up procedure with p-values ordered from low to high):

1. For a given α, find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$;
2. Reject the null hypothesis (i.e., declare discoveries) for all $H_{(j)}$, $j = 1, \ldots, k$.

Although the BH procedure is valid if all tests are independent, it is not universally valid [3].

We demonstrate the results of different multiplicity adjustments under the scenario without overdispersion in Figure 6. More specifically, the achieved power as a function of sample size per compound (denoted by n in the figure) is shown for the Bonferroni method (gray line), the Bonferroni-Holm method (purple line) and the Benjamini-Hochberg procedure (orange line), the latter generally having a higher power at the cost of a higher FWER.

Note that the black solid line corresponds to the earlier results presented in Figure 5 based on a comparison of the mean freshness duration between a control group and a single treatment group with Bonferroni correction. Needless to say, these results are similar to those for a Bonferroni correction in a multiple group simulation setting (gray line).

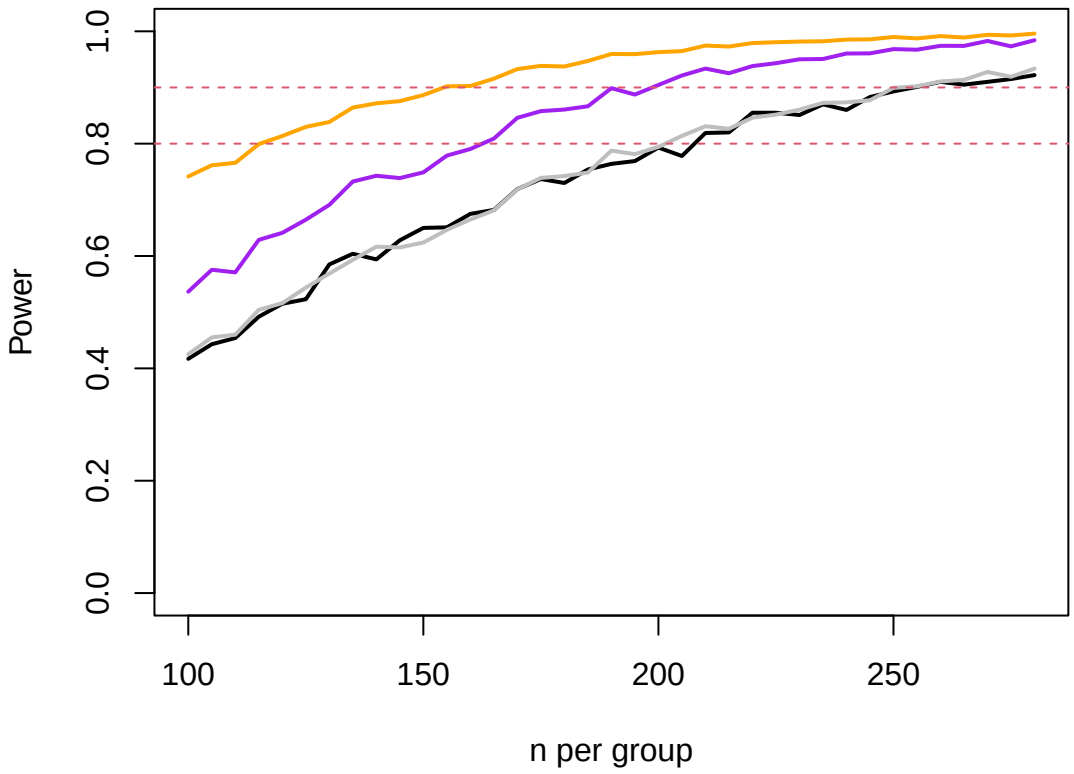

Figure 6: Estimated power-curves showing the relationship between the sample size per group (denoted by n) and the average achieved power (averaged over all comparisons) in the absence of overdispersion and relying on different multiplicity adjustments (Bonferroni - gray, Bonferroni-Holm - purple, Benjamini-Hochberg - orange). The black solid line represents the estimated power-curve in Figure 5 based on a single-test (single treatment group) simulation procedure. FWER $\alpha = 0.05$ or FDR $\alpha = 0.05$, right-tailed one-sided hypothesis testing and 1000 simulations.

Important remarks

Note that testing for the treatment (here: compound) effect is not always exactly the same as comparing the (population) means for the control and treatment group in a generalized linear mixed model. More specifically, the treatment effect quantified through, for example, the model parameter β_1 in the generalized linear mixed model does not necessarily has a marginal or population-averaged interpretation. One can easily show that in a linear mixed model, the subject-specific interpretation of the 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. Indeed, following the same reasoning as before, we have:

$$E(Y_i|\boldsymbol{x}_i) = \exp\left(\boldsymbol{\beta}\boldsymbol{x}_i^T\right)\exp(\sigma_b^2/2).$$

Consequently, in the overdispersed Poisson setting discussed in this technical note, the (subject-specific) treatment effect β_1 quantifies the impact of the treatment (or compound) on the population-averaged mean $E(Y_i)$ as well, i.e., $\exp(\beta_1)$ is the multiplicative effect of the treatment on the population-averaged mean of the control group or $\exp(\beta_1)$ is equal to the relative increase or decrease in mean following treatment as compared to the control.

In this tutorial we provide an overview of several multiplicity adjustments that can be made of which the Bonferroni and Bonferroni-Holm methods are controlling the family-wise error rate. Alternatively, the Benjamini-Hochberg method controls the false discovery rate [4]. Other methods have been proposed in the literature including the method by Hommel [5], the method by Hochberg [6] or the Benjamini-Yekutieli procedure [3]. These methods could be used as well though are valid under different conditions regarding dependence between tests. Note that the controlling the FDR guarantees control over the FWER in case all null hypotheses are true. However, if some true discoveries are to be made, implying that $C_0 < C$, the FWER $\geq$ FDR and improvements in power can be made. Hence, any procedure controlling the FWER will also control the FDR, but not the other way around. As an illustration, we depict the FWER and FDR in case of the Benjamini-Hochberg procedure in Figure 7 (for $C_0 = 14$, implying that $\lambda_0 = \lambda_j = 7.6$, for all $j = 1, \ldots, C$ and for $C_0 = 7$ with $\lambda_0 = \lambda_j = 7.6$, for $j = 1, \ldots, 7$ and $\lambda_j = 7.6 + 1$ for all other compounds). For $C_0 = 14$, we have that FWER (black line) and FDR (blue line) are equal, while for $C_0 = 7$ the FDR is controlled (i.e., smaller than $\alpha = 0.05$) and the FWER is uncontrolled (i.e., exceeds 5% - red dashed line).

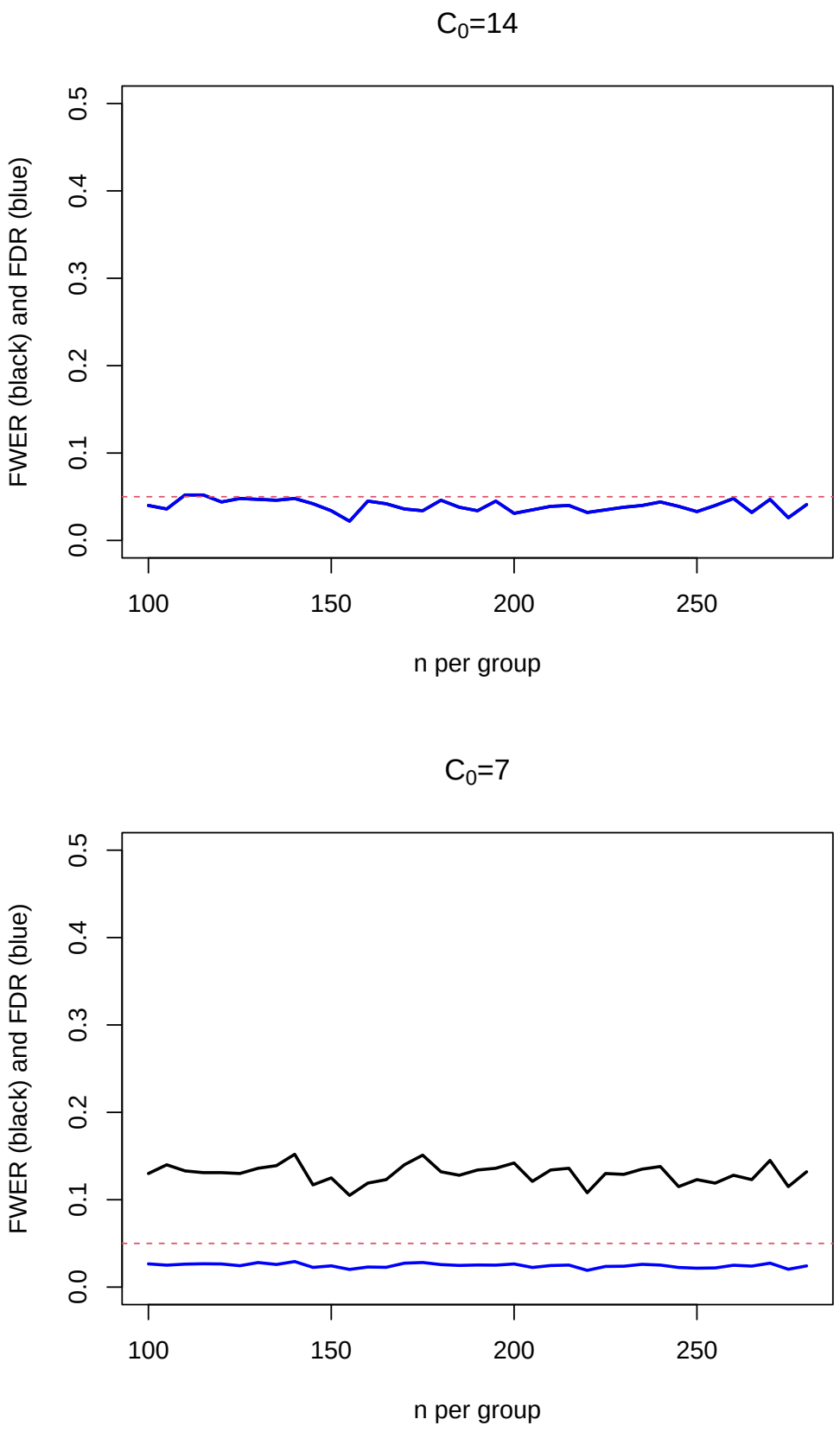

Figure 7: Benjamini-Hochberg procedure: FDR $\leq \alpha = 0.05$ for right-tailed one-sided hypothesis testing derived based on 1000 simulations under (1) a scenario with $C_0 = 14$ and $\lambda_0 = \lambda_j = 7.6$ for all $j = 1, \ldots, C$ (upper panel), and (2) a scenario with $C_0 = 7$ and $\lambda_0 = \lambda_j = 7.6$ for $j = 1, \ldots, 7$ while $\lambda_j = 7.6 + 1$ for $j = 8, \ldots, 14$ (lower panel).

The R code provided here is adapted from the code supplied in the course material of dr. Yannick Vandendijck. More specifically, the functions allow for the simulation-based computation (based on *n_sims* simulations) of the achieved power given a vector of possible sample sizes (specified in *n_grid*) and depending on the choice of the means *lambda0* and *lambda1* for the control and treatment groups. The significance level is specified in the function using the *alpha*-argument. The functions are constructed to be able to perform sample size calculations for both one- and two-sided testing problems (see *test* option). Finally, the *seed_nr* is included for reproducibility of the results. Additional arguments for the specification of clustering (*n_clusters* and $\sigma 2_b$) are available in the function *Poisson_sims_ext*. The functions *glm* and *glmer*, used in the R code above, are available in the stats and lme4 packages, respectively. The function *Poisson_sims_ext2* extends the function *Poisson_sims* in order to allow for multiplicity adjustments in the calculation of the required sample size. The function *p.adjust* used therein is available in the stats package and allows for (the calculation of adjusted p-values based on) alternative multiplicity adjustments as compared to the ones considered in this technical note. Please carefully check the documentation for the functions listed in this description when altering the arguments thereof.

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] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of Statistics, 29(4): 1165–1188.

[4] Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57, 289–300.

[5] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika, 75, 383–386.

[6] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika, 75, 800–803.

[7] Julious, S. A. (2009). Sample sizes for clinical trials. Chapman and Hall/CRC.