

## Crops and Soils Research Paper

**Cite this article:** Piepho H-P *et al* (2022). One, two, three: portable sample size in agricultural research. *The Journal of Agricultural Science* 1–24. <https://doi.org/10.1017/S0021859622000466>

Received: 27 May 2022

Revised: 30 June 2022

Accepted: 13 July 2022

### Key words:



Experimental design; linear model; power; precision; replication

### Author for correspondence:

Hans-Peter Piepho,

E-mail: [piepho@uni-hohenheim.de](mailto:piepho@uni-hohenheim.de)

# One, two, three: portable sample size in agricultural research

Hans-Peter Piepho<sup>1</sup> , Doreen Gabriel<sup>2</sup>, Jens Hartung<sup>1</sup>, Andreas Büchse<sup>3</sup>,  
Meike Grosse<sup>4</sup>, Sabine Kurz<sup>5</sup>, Friedrich Laidig<sup>1</sup>, Volker Michel<sup>6</sup>, Iain Proctor<sup>3</sup>,  
Jan Erik Sedlmeier<sup>7</sup>, Kathrin Toppel<sup>8</sup> and Dörte Wittenburg<sup>9</sup> 

<sup>1</sup>Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Stuttgart, Germany; <sup>2</sup>Institute for Crop and Soil Science, Julius Kühn-Institut (JKI), Braunschweig, Germany; <sup>3</sup>BASF SE, Ludwigshafen am Rhein, Germany; <sup>4</sup>Research Institute of Organic Agriculture FiBL, Frick, Switzerland; <sup>5</sup>Fachgebiet Pflanzenbau, Hochschule für Wirtschaft und Umwelt Nürtingen-Geislingen (HfWU), Nürtingen, Germany; <sup>6</sup>Mecklenburg-Vorpommern Research Centre for Agriculture and Fisheries, Gülzow, Germany; <sup>7</sup>Applied Entomology, Institute of Phytomedicine, University of Hohenheim, Stuttgart, Germany; <sup>8</sup>Fachgebiet Tierhaltung und Produkte, Hochschule Osnabrück, Osnabrück, Germany and <sup>9</sup>Research Institute for Farm Animal Biology (FBN), Institute of Genetics and Biometry, Dummerstorf, Germany

### Abstract

Determination of sample size (the number of replications) is a key step in the design of an observational study or randomized experiment. Statistical procedures for this purpose are readily available. Their treatment in textbooks is often somewhat marginal, however, and frequently the focus is on just one particular method of inference (significance test, confidence interval). Here, we provide a unified review of approaches and explain their close interrelationships, emphasizing that all approaches rely on the standard error of the quantity of interest, most often a pairwise difference of two means. The focus is on methods that are easy to compute, even without a computer. Our main recommendation based on standard errors is summarized as what we call the 1-2-3 rule for a difference of two treatment means.

### Introduction

One of the most common questions in the design of experiments and observational studies is: how many replications or samples do I need? Answers to this key question are well established (e.g., Rasch *et al.*, 2011; Welham *et al.*, 2015, Chapter 10), and good software tools are available as well (Stroup, 2002; Rasch *et al.*, 2011; Green and MacLeod, 2015). At the same time this important topic is treated only tangentially in many textbooks, and often times the material is somewhat dispersed throughout the text. This makes it difficult to recommend a single source to practitioners wanting quick advice and having little time to delve into the underlying mathematical theory. Also, decisions on sample size require prior information on variance, which researchers may sometimes find hard to come by, but only if such prior information is furnished can the sample size question be settled. This may require rough estimates to be derived on the spot, and good illustrations with real examples for this in the agricultural sciences remain sparse. Moreover, much of the material on sample size calculation focuses on significance testing, whereas one may also determine sample size based on considerations of precision alone, without having a specific significance test in mind. The purpose of this tutorial paper, therefore, is to provide a compact overview of the most basic procedures and the underlying key concepts, showing how they are all intimately related and giving particular emphasis to procedures based solely on precision requirements. Several practical examples are used for illustration. While sample size calculations are usually implemented using statistical software, we here emphasize the utility of simple equations that allow a quick determination of appropriate sample size. Wheeler (1974, 1975) denoted such equations as ‘portable in the sense that one can use (them) in the midst of a consultation with no tools other than perhaps a pocket calculator.’ This was written before personal computers but we think the term ‘portable’ is still very apt for this type of equation, so we use it freely throughout the paper. If we factor in the availability of portable computers and phones, as well as of free software and programming environments, portability comes within reach even for more advanced methods, which we cover briefly in the later part of the paper.

The term sample size is mostly synonymous with the term replication. The latter term is mainly used in reference to randomized experiments, whereas the former is used more broadly, also in reference to observational studies and surveys. In this paper, we mostly use the term sample size, but occasionally use the terms replication or replicate when the context is a designed experiment. In surveys, units in the sample are randomly selected from a well-defined parent population. In designed experiment, treatments are randomly allocated to experimental units. Random sampling in surveys and randomization in designed experiments

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

are the prerequisites underlying all methods for statistical inference and for determining sample size considered in this paper.

The rest of the paper is structured as follows. In the next section, we consider inference for a single mean, followed by a section on the comparison of two means. These two sections cover the basic concepts, and provide a set of equations which in our experience fully cover the majority of applications occurring in practice. Thus, a first reader may focus attention on these two sections. In both sections, we consider several alternative ways to determine sample size, showing how these alternatives all depend on the standard error and are therefore intimately connected. The core idea put forward is that all methods can be formulated in terms of a specification of the standard error of a mean (SEM) or of a difference alone. Our focus is mainly on responses that are approximately normally distributed, but we also touch upon count data. Subsequently we consider several important advanced cases for which portable equations are available as well, including regression, sub-sampling (pseudo-replication), and series of experiments. In a further section, we briefly review two general approaches to determine sample size, both of which involve the use of a linear model (LM) package. The paper concludes with a brief general discussion.

## Estimating a single mean

### Determining sample size based on a precision requirement

We here consider three different types of specifications for the precision of a mean that lead to a determination of sample size. To illustrate these, we will consider the following example.

*Example 1:* Assume that we want to estimate the mean milk yield (in kg day<sup>-1</sup>) per animal in a dairy cow population. The population mean is denoted here as  $\mu$ . This mean may be estimated based on a random sample of  $n$  cows. The sample mean is defined as  $\bar{y}_\bullet = n^{-1} \sum_{j=1}^n y_j$ , where  $y_j$  ( $j = 1, \dots, n$ ) are the milk yields of the  $n$  cows in the sample, and it provides an estimate of the population mean  $\mu$ . If we assume that the individual milk yields  $y_j$  are independent with mean (expected value)  $\mu$  and variance  $\sigma^2$ , it follows that the sample mean  $\bar{y}_\bullet$  has expected value  $E(\bar{y}_\bullet) = \mu$  and variance  $\text{var}(\bar{y}_\bullet) = n^{-1}\sigma^2$ , which is inversely proportional to the sample size  $n$ . This crucial fact is well-known, and it forms the basis of all approaches to determine sample size.

### Precision requirement specified in terms of the standard error of a mean

A common measure of precision for a mean estimate is its standard error (SEM), defined as the square root of the variance of a mean (VM):

$$\text{SEM} = \sqrt{\frac{\sigma^2}{n}} \quad (1)$$

An important feature of the SEM, distinguishing it from the VM, is that it is on the same scale as the mean itself, making it attractive for a specification of the precision requirement. Thus, Eqn (1) may be solved for  $n$  as:

$$n = \frac{\sigma^2}{\text{SEM}^2} \quad (2)$$

*Example 1 (continued):* Assume that the mean milk yield per day is expected to be in the order of 30 kg day<sup>-1</sup> and that from prior analyses the variance is expected to be  $\sigma^2 = 88.4 \text{ kg}^2 \text{ day}^{-2}$  (see Table 1). We would like to estimate the mean  $\mu$  with a standard error of  $\text{SEM} = 2 \text{ kg day}^{-1}$ . To achieve this, the required sample size as per Eqn (2) is:

$$n = \frac{88.4}{2^2} = 22.1 \Rightarrow 23$$

Note that Eqn (2) does not usually return an integer value for  $n$ , so rounding to a near integer is necessary. If we want to be on the conservative side and ensure that the SEM is no larger than the targeted value, we need to round up as a general rule, which in our example yields  $n = 23$ . Equation (2) is exact, but some of the equations that follow are approximations, erring on the optimistic side, which is a further reason to generally round up.

### Precision requirement specified in terms of the allowable deviation of a mean estimate from its true parameter value

Using the SEM for specifying the desired precision requires having a sense of the interpretation of this quantity. This is facilitated if we can assume an approximate normal distribution for the sample mean. This assumption requires either normality of the individual responses  $y_j$ , or it requires the sample size to be sufficiently large for the central limit theorem to kick in. This theorem implies that the sum, and hence the mean of independently and identically distributed random variables has an approximate normal distribution when the sample size becomes large, independently of the shape of the distribution of the individual random variables from which it is computed (Hogg *et al.*, 2019, p. 341). It is not possible to give a general rule of thumb on how large a sample size is large enough. A common recommendation is that  $n \geq 30$  is required, but it really depends on the shape of the distribution what sample size is required for a sufficient approximation to normality (Montgomery and Runger, 2011, p. 227). If in doubt and the non-normal distribution from which the data stem can be specified, alternative methods may be employed, particularly the model-based simulation approach depicted later in the paper. It may be added that even if the sample mean is not perfectly normal, equations that assume normality still can give a useful rough indication of the necessary sample size, also in cases where the sample size is small.

**Table 1.** Mean, variance, smallest relevant difference, and required sample size per treatment for  $\alpha = 5\%$  and a power of 85% for four traits in a dairy cow population

Trait (units)	Mean	Variance ( $\sigma^2$ )	Smallest relevant difference ( $\delta$ )	Required sample size ( $n$ )	Standard error of a difference (SED)
Milk yield (kg day <sup>-1</sup> )	31.6	88.4	5.0	64	1.66
Fat (%)	3.79	0.464	0.5	34	0.165
Protein (%)	3.46	0.103	0.2	47	0.0663
Laktose (%)	4.83	0.204	0.2	92	0.0666

Under the assumption of approximate normality, we expect that over repeated sampling about 68% of the sample means  $\bar{y}_\bullet$  will fall within the interval  $\mu \pm SEM$ . Likewise, we may say that a single sample mean  $\bar{y}_\bullet$  is expected to fall within the range  $\mu \pm SEM$  with a probability of 68%. Thus, the SEM gives some indication of the expected closeness of  $\bar{y}_\bullet$  to  $\mu$ . The main limitation of the  $\mu \pm SEM$  interval is that the probability 68% is pretty low, leaving a probability of 32% that the sample mean  $\bar{y}_\bullet$  falls outside this interval. Thus, for specifying the sample size, we may consider increasing the probability by widening the interval. For example, further exploiting the properties of the normal distribution, we may assert that the sample mean falls within the interval  $\mu \pm 2SEM$  with a probability of approximately 95%.

To formalize and generalize this approach, we may consider the deviation between the sample and population mean:

$$d = \bar{y}_\bullet - \mu \tag{3}$$

This deviation has expected value zero and variance  $n^{-1}\sigma^2$ . The precision requirement may now be specified by imposing a threshold  $\tau$  on the size of the absolute deviation  $|d|$  that we are willing to accept. This threshold may be denoted as the *allowable absolute deviation* of the estimate  $\bar{y}_\bullet$  from the population mean  $\mu$ . Specifically, we may require that the probability that  $|d|$  exceeds  $\tau$  takes on a specific value  $\alpha$ , which we want to be small, e.g. 5%. Thus, we require:

$$P(|d| > \tau) = \alpha \tag{4}$$

where  $P(\cdot)$  denotes the probability of the event given in the brackets. This requirement may be rearranged slightly as:

$$P(|d| > \tau) = 2P(d > \tau) = 2P\left(\frac{d}{\sqrt{n^{-1}\sigma^2}} > \frac{\tau}{\sqrt{n^{-1}\sigma^2}}\right) = \alpha \tag{5}$$

Now observing that  $d/\sqrt{n^{-1}\sigma^2}$  has a standard normal distribution, it can be seen that  $\tau/\sqrt{n^{-1}\sigma^2} = \tau/\sqrt{\text{var}(\bar{y}_\bullet)}$  must be the  $(1 - \alpha/2) \times 100\%$  quantile of the standard normal distribution, denoted as  $z_{1-\alpha/2}$  (for  $\alpha = 5\%$  we have  $z_{1-\alpha/2} \approx 2$ ). Equating the two and solving for  $n$  yields:

$$n = \frac{\sigma^2 z_{1-\alpha/2}^2}{\tau^2} \tag{6}$$

Thus, if we accept a probability of  $\alpha$  for the sample mean  $\bar{y}_\bullet$  to deviate from the population mean  $\mu$  by more than  $\tau$  units, we need to choose  $n$  according to Eqn (6). An equivalent interpretation is that choosing  $n$  as per Eqn (6) ensures that the sample mean  $\bar{y}_\bullet$  will deviate from the population mean  $\mu$  by no more than  $\tau$  units with pre-specified probability  $1 - \alpha$ . A very common choice for  $\alpha$  is 5%, in which case  $z_{1-\alpha/2} \approx 2$  and hence:

$$n \approx \frac{4\sigma^2}{\tau^2} \tag{7}$$

*Example 1 (cont'd):* If we want to ensure that the sample mean for milk yield is within  $\tau = 2 \text{ kg day}^{-1}$  of the population mean with

a probability of 95%, we need to choose:

$$n \approx \frac{4 \times 88.4}{2^2} = 88.4 \Rightarrow 89$$

which is about four times the sample size we need when our requirement is  $SEM = 2 \text{ kg day}^{-1}$ . With this sample size, we achieve  $SEM \approx 1 \text{ kg day}^{-1}$ , which is half the desired  $\tau$ . This observation is no coincidence, as can be seen by comparing (7) with (2), which essentially just differ by a factor of 4 when choosing the same value for the desired SEM and  $\tau$ , translating as a factor of 2 when comparing the resulting SEM. Note that here we have specifically chosen the same required value for  $\tau$  as for SEM in the example immediately after Eqn (2) to illustrate this important difference in impact on the necessary sample size.

*Precision requirement specified in terms of an allowable half width of a confidence interval for a mean*

Recalling that  $(\bar{y}_\bullet - \mu)\sqrt{n/s^2}$  is  $t$ -distributed when  $y_j$  is normal (Hogg *et al.*, 2019, p. 215), a confidence interval for  $\mu$  with  $(1 - \alpha) \times 100\%$  coverage probability can be computed as

$$\bar{y}_\bullet \pm t_{n-1;1-\alpha/2} \sqrt{\frac{s^2}{n}} \tag{8}$$

where  $t_{n-1;1-\alpha/2}$  is the  $(1 - \alpha/2) \times 100\%$  quantile of the  $t$ -distribution with  $n - 1$  degrees of freedom and  $s^2 = (n - 1)^{-1} \sum_{j=1}^n (y_j - \bar{y}_\bullet)^2$  is the sample variance, estimating the population variance  $\sigma^2$ . The half width of this interval is  $HW = t_{n-1;1-\alpha/2} \sqrt{s^2/n}$ , which may be used to make a specification on precision. The challenge here compared to the approaches considered so far is that even for given values of the population variance  $\sigma^2$  and sample size  $n$ ,  $HW$  is not a fixed quantity but a random variable. Thus, for a specification of precision, we need to consider the expected value of  $HW$ , i.e.

$$EHW = E\left(t_{n-1;1-\alpha/2} \sqrt{\frac{s^2}{n}}\right) \tag{9}$$

This expected value, in turn, is not a simple function of  $n$ , because both  $t_{n-1;1-\alpha/2}$  and  $s^2$  involve  $n$ . Hence there is no explicit equation for  $n$  that can be derived from (9). Instead, numerical routines need to be used to solve (9) for  $n$  for given population variance  $\sigma^2$ ,  $\alpha$  and specification of  $EHW$ , for example in SAS (PROC POWER) or R (Rasch *et al.*, 2011). Alternatively, one may obtain an approximate solution by making two simplifying assumptions: (i) The sample variance  $s^2$  is replaced by the population variance  $\sigma^2$  and (ii) the quantile  $t_{n-1;1-\alpha/2}$  of the  $t$ -distribution is replaced by the corresponding quantile  $z_{1-\alpha/2}$  of the standard normal distribution, assuming that  $n$  will not be too small. These two simplifications lead to the approximation:

$$HW \approx z_{1-\alpha/2} \sqrt{\frac{\sigma^2}{n}} \tag{10}$$

Here, the approximation on the right-hand side is no longer a random variable, so we can use this to approximate the desired  $EHW$  and solve for  $n$  to obtain the approximation:

$$n \approx \frac{\sigma^2 z_{1-\alpha/2}^2}{EHW^2} \tag{11}$$

This equation is equivalent to (6) when replacing  $\tau$  with  $EHW$ . It will tend to yield smaller sample sizes than the exact numerical solution. When also taking into account the probability that the realized HW remains within pre-specified bounds (Beal, 1989), a larger sample size would be required, but this is not pursued here.

*Example 1 (cont'd):* If we want to ensure that a 95% confidence interval for the population mean of milk yield per day has an  $EHW$  of 2 kg day<sup>-1</sup>, we need to choose:

$$n \approx \frac{88.4 \times 2^2}{2^2} = 88.4 \Rightarrow 89.$$

This is the same result as per Eqn (6), and the  $SEM \approx 1$  kg day<sup>-1</sup>, which is half the desired  $EHW$ . Again, this equality is no coincidence, as can be seen from the equivalence of (6) and (11), if we equate  $\tau$  and  $EHW$ .

### Summary and the 1-2 rule

We can summarize the procedures under the three types of specification for the precision in the previous three sub-sections as shown in Table 2. Importantly, all procedures involve the  $SEM$ , so the rules based on specifications for  $\tau$  and  $EHW$  can be cast as rules for the choice of  $SEM$ :

$$SEM = \frac{\tau}{z_{1-\alpha/2}} = \frac{EHW}{z_{1-\alpha/2}} \quad (12)$$

For  $\alpha = 5\%$  this amounts to the simple rule that  $SEM$  should be no larger than  $\tau/2$  or  $EHW/2$ . It also emerges that the precision measures  $\tau$  and  $EHW$  are exchangeable from a practical point of view, even though they have somewhat different underlying rationales. We can also turn this around and first just compute the  $SEM$  for a given design to evaluate its precision. Then  $2 \times SEM$  is the allowable deviation  $\tau$  or  $EHW$  the design permits to control. Because of the factors involved (1 for the  $SEM$  itself, and 2 for  $\tau$  or  $EHW$ ), we call this the *1-2 rule for a mean*.

### How to get a prior value for $\sigma^2$

**General:** The ideal is to find reports on similar studies as the one planned that report on the variance. Alternatively, a pilot study may be conducted to obtain a rough estimate of  $\sigma^2$ . Desirable though this may be, it is not always easy to get such information quickly.

A rule of thumb that may be useful here and does not make any distributional assumptions, is that the range in a sample of  $n$  observations, defined as the difference between the largest and smallest observed value in the sample, can be used to derive

upper and lower bounds on the sample standard deviation  $s = \sqrt{s^2}$  (van Belle, 2008, p. 36):

$$\frac{Range}{\sqrt{2(n-1)}} \leq s \leq \frac{n}{n-1} \frac{Range}{2}, \quad (13)$$

This rule is most useful in making quick assessments of problems in a given dataset, but it may also be useful in deriving a rough estimate of the standard deviation  $\sigma$ .

**Normality:** Welham *et al.* (2015, p. 245) propose to approximate the standard deviation by:

$$\sigma \approx \frac{max - min}{4} \quad (14)$$

where *min* and *max* are 'the likely minimum and maximum value for experimental units receiving the same treatment'. The actual rationale of this equation stems from the normal assumption and the fact that 95% of the data are expected within two standard deviations from the mean. This means that for this approximation to work well, the data must have an approximate normal distribution, and *min* and *max* must be estimates of the 2.5 and 97.5% quantiles of the distribution. In other words, *min* and *max* must be the bounds of an interval covering about 95% of the expected data from experimental units receiving the same treatment.

*Example 2:* It is not easy to accurately guess the 2.5 and 97.5% quantiles. To illustrate the difficulty, consider random samples of different sizes  $n$  from a normal distribution. Of course, if such samples were available when planning an experiment, the sample variance could be computed directly and used in place of  $\sigma^2$ , and this would be the preferred thing to do. However, for the sake of illustrating the challenge with (14), imagine that we determine the observed minimum and maximum value in a sample of size  $n$  and plug these into the equation. Table 3 shows results for  $n = 4, 8, 15, 30, 50, 100$ . It is seen that with a sample size of  $n = 30$  the expected range and median range come closest to the value of 4 that is postulated in (14). It emerges that a smaller sample size leads to under-estimation and a larger sample size to over-estimation of the standard deviation as per (14), if we simply plug in the observed minimum and maximum. But unless the sample size is very small, the approximation will be in the right ballpark, and that is usually sufficient for most practical purposes.

**Binary:** Up to here, for the most part, we have assumed the normality of the response  $y$ . Often, the observed data are counts, and these are not normal. The simplest case is binary data, where the count is either 0 or 1. The response is then said to have a binary distribution, which has one parameter, the probability that the response is 1. This probability, in turn, is equal to the mean  $\mu$  of

**Table 2.** Overview of procedures for determining sample size for a single mean

Parameter for specification of precision	Interpretation of precision parameter	Solution(s) for $n$ in main text	Further prior specification needed
$SEM$	Standard error of a mean	(2)	$\sigma^2$
$\tau$	Allowable absolute deviation from the population mean	(5), (6)	$\sigma^2, \alpha$
$EHW$	Expected half width of confidence interval for a mean	Exact only numerically, use (11) for an approximation	$\sigma^2, \alpha$

**Table 3.** Expected and median of range (*maximum – minimum*) for samples of different sample size *n* from standard normal distribution

Sample size <i>n</i>	Expected range		Median range	
	Numerical <sup>a</sup>	Simulated <sup>b</sup>	Numerical <sup>c</sup>	Simulated <sup>b</sup>
4	2.09	2.06	1.98	1.99
8	2.85	2.85	2.79	2.76
15	3.46	3.48	3.42	3.45
30	4.06	4.08	4.04	4.04
50	4.47	4.52	4.45	4.46
100	4.98	5.04	4.97	5.00

<sup>a</sup>Approximated as  $2\Phi^{-1}(0.5264^{1/n})$  (Chen and Tyler, 1999).

<sup>b</sup>Simulated based on 1000 runs, computing the mean in each run, and then taking the mean or median.

<sup>c</sup>Computed using the PROBMC function of SAS (Westfall *et al.*, 1999, p. 45) as the 50% quantile of the studentized range distribution with infinite degrees of freedom.

the binary random variable. For this distribution the variance equals:

$$\sigma^2 = \mu(1 - \mu) \tag{15}$$

with  $0 < \mu < 1$ . Thus, to approximate the variance in this case, we need a guess of the mean  $\mu$ . To be on the conservative side, we may consider the worst case with the largest variance  $\sigma^2$ , which occurs when  $\mu = 0.5$ .

*Example 3:* A research institute conducts an opinion poll and considers the worst-case scenario that the proportion of voters favouring a particular party is  $\mu = 0.5$ , in which case  $\sigma^2 = 0.25$ . The proportion of each party is to be estimated with precision  $SEM = 0.01$ . Thus, using (2), the sample size is chosen as:

$$n = \frac{\sigma^2}{SEM^2} = \frac{0.25}{0.01^2} = 2500.$$

*Example 4:* Monitoring foot pad health is an important task in rearing turkey. The prevalence of foot pad dermatitis in a given flock may be estimated by random sampling. Any animal with symptoms on at least one foot is considered as affected (Toppel *et al.*, 2019; for details on the scoring system see Hocking *et al.*, 2008). Typical prevalences range around 0.5, so it is suitable to determine the sample size under the worst-case scenario  $\mu = 0.5$ . If we set the allowable deviation from the true mean at  $\tau = 0.1$  with  $\alpha = 5\%$ , the sample size based on Eqn (6) is

$$n = \frac{z_{1-\alpha/2}^2 \sigma^2}{\tau^2} = \frac{1.96^2 \times 0.5^2}{0.1^2} = 96.04 \Rightarrow 97$$

Note that in using (6), we have assumed that the sample size *n* will be large enough for the central limit theorem to apply. We have further assumed that sampling is without replacement and that the population from which we are sampling is large relative to sample size (but see next sub-section entitled ‘Finite populations’).

**Binomial:** If on each experimental unit, we have *m* observational units, each with a binary response with the expected value  $\mu$  (a proportion or probability), then the binomial distribution may be assumed, which has variance:

$$\sigma^2 = \mu(1 - \mu)/m \tag{16}$$

for the observed proportion  $y = c/m$ , where *c* is the binomial count based on the *m* observational units. Thus, to approximate the variance in this case, we also need a guess of the mean  $\mu$ . Again, the worst-case scenario is  $\mu = 0.5$ . In practice, the data may show over-dispersion relative to the binomial model. A simple way to model this is to assume variance:

$$\sigma^2 = \phi\mu(1 - \mu)/m \tag{17}$$

where  $\phi$  is an over-dispersion parameter (McCullagh and Nelder, 1989, p. 124). In this scenario, estimating the mean alone does not help in approximating the variance; we also need an estimate of the over-dispersion, and this puts us back to the general case, where independent prior information on the variance needs to be obtained.

*Example 5:* In a large potato field, *n* = 347 control points were distributed to assess the abundance of the potato weevil (*Leptinotarsa decemlineata*) (Trommer, 1986). At each control point, *m* = 20 potato plants were assessed for the presence or absence of the weevil. The counts of affected plants (*c*) per control point are reproduced in Table 4.

For each control plot, we can compute the observed proportion  $y = c/m$ . The sample mean of *y* across the *n* = 347 control points is  $\bar{y}_\bullet = 0.2987$ , which is an estimate of the proportion  $\mu$  of infested plants on the field. Under a binomial distribution, the variance of *y* would be estimated as  $\bar{y}_\bullet(1 - \bar{y}_\bullet)/m = 0.2987 \times 0.7013/20 = 0.010474$ . This is considerably smaller than the sample variance  $s^2 = 0.10788$ . From this, the overdispersion is estimated as  $\hat{\phi} = s^2 / [\bar{y}_\bullet(1 - \bar{y}_\bullet)/m] = 10.2999$ . This value corresponds to the one obtained from Pearson’s chi-squared statistic for over-dispersed binomial data (McCullagh and Nelder, 1989, p. 127). Thus, the variance is about ten times the variance expected under a binomial model. The reason for this overdispersion is the clustering of patches of infested plants amidst areas of plants infested little or not at all, which is typical of crop diseases and pests. Incidentally, the hat symbol on  $\phi$  indicates that this is the corresponding sample estimator of the parameter. We will subsequently use the hat notation in several places, also for other parameters. Further note that we could use  $\hat{\mu}$  and  $\hat{\sigma}^2$  in place of  $\bar{y}_\bullet$  and  $s^2$  to denote the sample mean and variance, respectively.

Now assume that we go to a new field and want to determine the number of control points (each with *m* = 20 plants) needed to achieve a half width of  $HW = 0.05$  for a 95% confidence interval. A first rough assessment suggests that the infestation in the new

**Table 4.** Frequency distribution of the number of plants infested with the potato weevil (*c*) in samples of  $m = 20$  plants at  $n = 347$  control points (Trommer, 1986)

Count ( <i>c</i> )	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Frequency	104	33	20	16	20	21	9	8	7	9	11	9	10	6	11	8	5	4	8	13	15

field is in the order of  $\mu = 0.1$ . The variance is  $\sigma^2 = \phi\mu(1 - \mu)/m = 10.299 \times 0.1 \times 0.9/20 = 0.04635$ . Using this in Eqn (11), we find:

$$n \approx \frac{\sigma^2 z_{1-\alpha/2}^2}{EHW^2} = \frac{0.04635 \times 1.96^2}{0.05^2} = 71.22 \Rightarrow 72$$

Thus,  $n = 72$  control points would be required to achieve this precision.

**Poisson:** Under the Poisson model for counts, the count  $y$  can take on any non-negative integer value (0, 1, 2, ...). The variance of  $y$  is:

$$\sigma^2 = \mu \quad (18)$$

So again, a rough estimate of the mean is needed to approximate the variance. There is no worst-case scenario that helps as the variance increases monotonically with the mean  $\mu$ . Moreover, it needs to be considered that there is often over-dispersion relative to the mean, so the variance is:

$$\sigma^2 = \phi\mu \quad (19)$$

where  $\phi$  is an over-dispersion parameter (McCullagh and Nelder, 1989, p. 198). As with the over-dispersed binomial distribution, in this scenario, estimating the mean alone does not help in approximating the variance; we also need an estimate of the over-dispersion, and this, yet again, puts us back to the general case.

It is stressed that exact methods should be used for small binomial sample sizes  $m$  and also for small means in case of the Poisson. These exact methods, which are somewhat more involved (see, e.g., Agresti and Coull, 1998; Chen and Chen, 2014; Shan, 2016), will not be considered here.

**Example 6:** Inoculum density of *Cylindrocladium crotalariae* was assessed on 96 quadrats in a peanut field. On each quadrat, the number of microsclerotia was counted (Hau, Campbell and Beute, 1982). The frequency distribution is given in Table 5.

The mean count is  $\bar{y}_\bullet = 7.990$ , whereas the sample variance is  $s^2 = 30.47$ , showing substantial over-dispersion. The overdispersion is estimated as  $\hat{\phi} = s^2/\bar{y}_\bullet = 30.47/7.990 = 3.841$ . Again, this corresponds to the value obtained from Pearson's generalized chi-squared statistic (McCullagh and Nelder, 1989, p. 328). Now a new field is to be assessed on which first inspection by eyeballing suggests a mean infestation of  $\mu = 20$  microsclerotia per quadrat. We would like to estimate the population mean of the field with a precision of  $SEM = 2$ . The variance is expected to be  $\sigma^2$

$= \phi\mu = 3.841 \times 20 = 76.28$ . From Eqn (2) we find:

$$n = \frac{\sigma^2}{SEM^2} = \frac{76.28}{2^2} = 19.07 \Rightarrow 20$$

Thus,  $n = 20$  quadrats are needed to achieve the targeted precision.

### Finite populations

So far we have assumed that the population from which we are sampling (without replacement) is infinite, or very large. When the population is small, it is appropriate to consider a *finite population correction*, meaning that the VM equals (Kish, 1965, p. 63):

$$\text{var}(\bar{y}_\bullet) = \frac{N - n}{N - 1} n^{-1} \sigma^2 \quad (20)$$

where  $N$  is the population size. The methods for sample size determination in the previous sub-sections are applicable with this modification. Note that when  $n = N$ , i.e. under complete enumeration of the population, the variance in (20) reduces to zero as expected, because the finite population correction  $(N - n)(N - 1)^{-1}$  is zero in this case. For illustration, we consider the specification of an allowable absolute deviation  $\tau$  of the sample mean from the population mean with probability  $\alpha$ . Thus, we may equate  $\tau^2/\text{var}(\bar{y}_\bullet) = z_{1-\alpha/2}^2$  as we have done in the previous sub-sections. Solving this for  $n$  using (20) yields:

$$n = \frac{N\sigma^2}{(N - 1)(\tau^2/z_{1-\alpha/2}^2) + \sigma^2} \quad (21)$$

Note that this is equal to (6) when  $N$  approaches infinity. Applying this to the binary case with  $\sigma^2 = \mu(1 - \mu)$  yields (Thompson, 2002, p. 41)

$$n = \frac{N\mu(1 - \mu)}{(N - 1)(\tau^2/z_{1-\alpha/2}^2) + \mu(1 - \mu)} \quad (22)$$

**Example 3 (cont'd):** In opinion polls, the population from which a sample is taken usually has size  $N$  in the order of several millions, which is huge compared to the customary sample sizes in the order of  $n = 2500$ . In this case, the finite population correction may safely be ignored.

**Example 4 (cont'd):** If the flock size is  $N = 4000$  (Toppel et al., 2019), and we use the same specifications as before, the sample size as per (22) is  $n = 94$ , down from  $n = 97$  when the finite population correction is ignored (Eqn 6).

**Table 5.** Frequency distribution of the number  $y$  of microsclerotia (*Cylindrocladium crotalariae*) per quadrat on  $n = 96$  quadrats in a peanut field (Hau et al., 1982; Figure 3(b))

Count ( $y$ )	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16	17	18	20	21	22	26
Frequency	2	5	5	13	5	6	9	6	7	4	5	9	6	2	2	1	1	2	2	1	2	1

**Comparing two means**

In comparative studies and experiments, the objective is usually a pairwise comparison of means (Bailey, 2009). Thus, we are interested in estimating a difference  $\delta = \mu_1 - \mu_2$ , where  $\mu_1$  and  $\mu_2$  are the means to be compared. Here, we consider the case where the observations of both groups are independent. In this case, the variance of a difference (VD) between two sample means  $\bar{y}_{1\bullet}$  and  $\bar{y}_{2\bullet}$  equals  $\text{var}(\bar{y}_{1\bullet} - \bar{y}_{2\bullet}) = n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2$ , where  $n_1$  and  $n_2$  are the sample sizes and  $\sigma_1^2$  and  $\sigma_2^2$  are the variances in the two groups. If we assume homogeneity of variance, this simplifies to  $\text{var}(\bar{y}_{1\bullet} - \bar{y}_{2\bullet}) = (n_1^{-1} + n_2^{-1})\sigma^2$ , where  $\sigma^2$  is the common variance. Further, if the sample size is the same ( $n$ ) in each group, which is the optimal allocation under homogeneity of variance, this further simplifies to  $\text{var}(\bar{y}_{1\bullet} - \bar{y}_{2\bullet}) = 2n^{-1}\sigma^2$ . Here, we make this assumption for simplicity. Note that the variance is just twice the variance of a sample mean. Thus, apart from this slight modification, all methods in the previous section can be applied without much further ado, so the exposition of these methods can be brief here.

**Determining sample size based on a precision requirement**

In this section, we assume approximate normality of the response or sufficient sample size for the central limit theorem to ensure approximate normality of treatment means.

*Precision requirement specified in terms of the standard error of a difference*

The standard error of a difference (SED) of two sample means equals:

$$SED = \sqrt{\frac{2\sigma^2}{n}} \tag{23}$$

Equation (23) may be solved for  $n$  to yield:

$$n = \frac{2\sigma^2}{SED^2} \tag{24}$$

*Example 7:* Ross and Knodt (1948) conducted a feeding experiment to assess the effect of supplemental vitamin A on the growth of Holstein heifers. There was a control group and a treatment group, both composed of 14 animals. The allocation of treatments to animals followed a completely randomized design. One of the response variables was weight gain (lb.). The pooled sample variance was  $s^2 = 2199 \text{ lb.}^2$ , and treatment means were in the order of 200 lb. Suppose a follow-up experiment is designed to compare the control to a new treatment with an improved formulation of the vitamin A supplementation with an SED of 20 lb. Setting  $\sigma^2 = 2199 \text{ lb.}^2$  based on the prior experiment, the required sample size is

$$n = \frac{2\sigma^2}{SED^2} = \frac{2 \times 2199}{20^2} = 11$$

*Precision requirement specified in terms of the allowable deviation of estimate of difference from its true parameter value*

The sample size required per treatment to ensure with probability  $1 - \alpha$  that the deviation of the estimated difference from the true

difference is no larger than  $\tau_\delta$  is:

$$n = \frac{2\sigma^2 z_{1-\alpha/2}^2}{\tau_\delta^2} \tag{25}$$

For  $\alpha$  is 5%, this is approximately:

$$n \approx \frac{8\sigma^2}{\tau_\delta^2} \tag{26}$$

*Example 7 (cont'd):* Suppose we are prepared to allow a deviation of  $\tau_\delta = 20 \text{ lb.}$  Thus, using  $\sigma^2 = 2199$  we require:

$$n \approx \frac{8\sigma^2}{\tau_\delta^2} = \frac{8 \times 2199}{20^2} = 43.98 \Rightarrow 44$$

This is four times the sample size required to achieve an SED of 20 lb. The precision achieved here is  $SED = 10 \text{ lb.}$ , which is half the desired  $\tau_\delta$ .

*Precision requirement specified in terms of the allowable half width of a confidence interval for a difference*

A confidence interval for  $\delta$  with  $(1 - \alpha) \times 100\%$  coverage probability can be computed as:

$$\bar{y}_{1\bullet} - \bar{y}_{2\bullet} \pm t_{w;1-\alpha/2} \sqrt{\frac{2s^2}{n}}, \tag{27}$$

where  $t_{w;1-\alpha/2}$  is the  $(1 - \alpha/2) \times 100\%$  quantile of the  $t$ -distribution with  $w = 2(n - 1)$  degrees of freedom and  $s^2$  is the pooled sample variance, estimating the population variance  $\sigma^2$ . Again, the exact method to determine the sample size for the confidence interval of a difference requires numerical methods as implemented in software packages. Here, we consider an approximate method. The half width of the interval is  $HW = t_{w;1-\alpha/2} \sqrt{2s^2/n}$ . It is worth pointing out that this  $HW$  is equal to the least significant difference (LSD) for the same  $\alpha$  as significance level, a point which we will come back to in the next section. The approximation replaces  $t_{w;1-\alpha/2}$  with  $z_{1-\alpha/2}$  and  $s^2$  with  $\sigma^2$ , yielding an expected half width (EHW) of

$$EHW \approx z_{1-\alpha/2} \sqrt{\frac{2\sigma^2}{n}}, \tag{28}$$

which we may also regard as the expected LSD (ELSD). Then solving for  $n$  yields:

$$n \approx \frac{2\sigma^2 z_{1-\alpha/2}^2}{EHW^2} \tag{29}$$

This equation is seen to be equivalent to (25) when replacing  $\tau_\delta$  with  $EHW$ . This approximate solution will tend to yield somewhat smaller sample sizes than the exact numerical solution.

*Example 7 (cont'd):* Suppose we want to achieve  $EHW = 20 \text{ lb.}$  with  $\alpha = 5\%$ . This requires:

$$n \approx \frac{2\sigma^2 z_{1-\alpha/2}^2}{EHW^2} \approx \frac{2 \times 2199 \times 2^2}{20^2} = 43.98 \Rightarrow 44$$

which is the same sample size as requires to achieve an allowable deviation of  $\tau_\delta = 20 \text{ lb.}$ , and also leads to  $SED = 10 \text{ lb.}$ , half the desired  $EHW$ .

*Precision requirement specified in terms of difference to be detected by a t-test*

A  $t$ -test may be used to test the null hypothesis  $H_0: \delta = 0$  against the alternative  $H_A: \delta \neq 0$ . The  $t$ -statistic for this test is:

$$t = \frac{\bar{y}_{1\bullet} - \bar{y}_{2\bullet}}{\sqrt{2s^2/n}} \quad (30)$$

This has a central  $t$ -distribution on  $w = 2(n - 1)$  degrees of freedom under  $H_0: \delta = 0$  and a non-central  $t$ -distribution under the alternative  $H_A: \delta \neq 0$ . There are two error rates to consider with a significance test, i.e.  $\alpha$ , the probability to falsely reject  $H_0$  when it is true, and  $\beta$ , the probability to erroneously accept  $H_0$  when it is false. The complement of the latter,  $1 - \beta$ , is the power of the test, i.e., the probability to correctly reject  $H_0$  when it is false. To plan sample size, we need to make a choice for the desired values of both  $\alpha$  and  $\beta$ . Moreover, prior information on the variance  $\sigma^2$  is needed, as well as a specification of the smallest relevant value of the difference  $\delta$  that we want to be able to detect with the test. These choices then determine the required sample size. Again, an exact solution for  $n$  requires numerical integration using the central  $t$ -distribution under  $H_0$  and the non-central  $t$ -distribution under  $H_A$  (Welham et al., 2015, p. 248). Some authors approximate the non-central  $t$ -distribution with the central one (Cochran and Cox, 1957, p. 20; Bailey, 2009, p. 36). A more portable approximate solution that replaces the central and non-central  $t$ -distributions with the standard normal, and the sample variance  $s^2$  with the population variance  $\sigma^2$ , is obtained as (van Belle, 2008, p. 30):

$$n \approx \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2} \quad (31)$$

where  $z_{1-\alpha/2}$  is as defined before and  $z_{1-\beta}$  is the  $(1 - \beta) \times 100\%$  quantile of the standard normal distribution. This equation is easily derived by observing that under  $H_0$ ,  $t$  is approximately standard normal, with the critical value for rejection at  $\pm z_{1-\alpha/2}$ , the  $(1 - \alpha/2) \times 100\%$  quantile of the standard normal. Under  $H_A$ ,  $t$  is approximately normal with unit variance and mean  $\delta/SED$ , with  $SED$  depending on  $n$  as shown in (23). This distribution has its  $\beta \times 100\%$ -quantile at  $\delta/SED - z_{1-\beta}$ . These two quantiles under the  $H_0$  and  $H_A$  distributions must match exactly for the desired  $\alpha$  and  $\beta$ , so we can equate them and solve for  $n$ , yielding Eqn (31).

A conventional value of  $\alpha$  is 5%, but 1% or 10% are also sometimes used. Typical choices for  $\beta$  are 5%, 10% and 20%. For routine application, it is convenient to define  $C_{\alpha,\beta} = (z_{1-\alpha/2} + z_{1-\beta})^2$  and compute this for typical choices of  $\alpha$  and  $\beta$  (Table 6). These values of  $C_{\alpha,\beta}$  can then be used in the equation:

$$n \approx \frac{2\sigma^2 C_{\alpha,\beta}}{\delta^2} \quad (32)$$

A portable version of (32) for the very common choice  $\alpha = 5\%$  and  $\beta = 10\%$  is:

$$n \approx \frac{21\sigma^2}{\delta^2} \quad (33)$$

Other portable equations can of course be derived for other desired values of  $\alpha$  and  $\beta$ . It is instructive to compare Eqn (33) for the  $t$ -test with the precision-based Eqn (26). Importantly, unless the power we desire is small, (33) yields a considerably

**Table 6.** Values of  $C_{\alpha,\beta} = (z_{1-\alpha/2} + z_{1-\beta})^2$  for typical choices of  $\alpha$  and  $\beta$

		$\beta$			
		5%	10%	15%	20%
$\alpha$	1%	17.8	14.9	13.0	11.7
	5%	13.0	10.5	9.0	7.8
	10%	10.8	8.6	7.2	6.2

larger sample size when we use the same values for the difference to be detected ( $\delta$ ) in (33) and the allowable deviation of the estimated from the true difference ( $\tau_\delta$ ) in (26). As we have explained, the latter can also be equated to the desired  $ELSD$ . Specifically, this means that choosing sample size so that a desired value for  $\tau_\delta$  or  $ELSD$  is achieved does not at all ensure sufficient power ( $1 - \beta$ ) to detect a critical difference  $\delta$  of the same size. In fact, if  $\delta = ELSD$ , the  $t$ -test has an expected power of 50% only, which will hardly be considered satisfactory. We will need an  $ELSD$  substantially smaller than  $\tau_\delta$  to achieve a reasonable power. For our portable example with a power of 90%, the ratio of required  $ELSD$  over  $\delta$  can be approximated by dividing (33) by (26) and solving for the ratio:

$$\frac{\tau_\delta}{\delta} = \frac{ELSD}{\delta} = \sqrt{\frac{8}{21}} \approx 0.62.$$

It may also be noted that if the desired power indeed equalled 50%, we would have  $z_{1-\beta} = z_{0.5} = 0$ , in which case (31) takes the same form as (25). So in this special case of a power of 50%, we may say that the specification of a value for  $\delta$  in (31) is equivalent to specifying the same value for  $\tau_\delta$  (equivalent to  $ELSD$ ) in (25). This coincidence is of little practical use, however, because a power of 50% is rarely considered sufficient. The more important point here is that in all other cases, specifying the same value for  $\delta$  in (31) and for  $\tau_\delta$  in (25) does not lead to the same sample size.

*Example 7 (cont'd):* A difference of  $\delta = 20$  lb. is to be detected at  $\alpha = 5\%$  with a power of 90%. This can be achieved with an approximate sample size of:

$$n \approx \frac{21\sigma^2}{\delta^2} = \frac{21 \times 2199}{20^2} = 115.4 \Rightarrow 116$$

As expected, this sample size is larger still than when we required  $EHW = 20$  lb. or  $\tau_\delta = 20$  lb.. The precision attained here is better as well, amounting to  $SED \approx 6$  lb.

We also use this example to assess the degree of the approximation involved by replacing the central and non-central  $t$ -distributions with the standard normal in (31). For the case at hand, the exact result (obtained with PROC POWER in SAS) yields  $n = 117$ , which is very close to the approximate result of  $n = 116$ . To explore this further, we also did the exact and approximate calculations for a range of larger values of the relevant difference  $\delta$ . The results in Table 7 show that the approximation is very good, even when the exact sample size is as small as  $n = 3$ . It emerges that if one wants to be on the safe side, adding one or two to the approximate sample size per group should suffice.

The equations considered so far require specifying the difference to be detected ( $\delta$ ) in absolute terms. It is sometimes easier for



**Table 7.** Required sample size for unpaired *t*-test at  $\alpha = 5\%$  with a power of 90% for  $\sigma^2 = 2199 \text{ lb.}^2$  and a range of values for the smallest relevant difference  $\delta$  in lb

Relevant difference ( $\delta$ )	Required sample size ( $n$ )	
	Approximate <sup>a</sup>	Exact <sup>b</sup>
20	116	117
30	52	53
40	29	30
50	19	20
60	13	14
70	10	11
80	8	9
90	6	7
100	5	6
120	4	5
150	3	4
200	2	3

<sup>a</sup>Using Eqn (31).

<sup>b</sup>Using PROC POWER of SAS.

researchers to specify this in relative terms instead, i.e., as a proportion or percentage difference. For this purpose, Eqn (31) can be slightly rewritten. To do so, we define the relative difference as:

$$\delta_r = \frac{\delta}{\mu} \tag{34}$$

where  $\mu = (\mu_1 + \mu_2)/2$  is the overall mean. The standard deviation can also be expressed in relative terms, and this is known as the coefficient of variation,  $CV = \sigma/\mu$ . With these definition, Eqn (31) can be rearranged to yield the approximation:

$$n \approx \frac{2CV^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta_r^2} \tag{35}$$

The portable version of (35) for the very common choice  $\alpha = 5\%$  and  $\beta = 10\%$  is (see Table 6):

$$n \approx \frac{21CV^2}{\delta_r^2} \tag{36}$$

These equations work equally with  $\delta_r$  and  $CV$  expressed as proportions or as percentages.

*Example 7 (cont'd):* The means for the control and treatment groups were 187.6 and 235.9 lb. From this, the coefficient of variation is computed as  $CV = 22.15\% = 0.2215$ . Suppose that in a new experiment we want to be able to detect a relative treatment difference of  $\delta_r = 10\% = 0.1$  compared to the overall mean at  $\alpha = 5\%$  with a power of 90%. Here we need a sample size of:

$$n \approx \frac{21CV^2}{\delta_r^2} = \frac{21 \times 22.15\%^2}{10\%^2} = 103$$

*Example 8:* Four traits are to be assessed to compare two different milking methods, i.e. a milking robot and a milking parlour. Long-term records on these four traits are available from 142 cows of the same population from which the animals

for the experiments are to be drawn and allocated to the two treatments at random. Sample means and variances are reported in Table 1. Discussions with the animal scientists conducting this experiment identified the smallest relevant differences  $\delta$  for the four traits as shown in Table 1. Based on these specifications, the sample size  $n$  per treatment for an unpaired *t*-test for  $\alpha = 5\%$  and a power of 85% were determined using Eqn (32) for each trait. It is seen that the sample size differs between traits, illustrating that when an experiment involves several traits, a compromise must be struck regarding a common sample size. We also note that the *SED* achieved with these sample sizes is about 1/3 of the smallest relevant difference  $\delta$  for each trait, a point we will take up again in the next section.

We note in passing that it is quite common to express effect size not relative to a mean but relative to the standard deviation ( $d = \delta/\sigma$ ; Cohen, 1977, 1992), a measure also known as Cohen's *d*, but agree with Lenth (2001) that it is difficult, if not misleading, to think about effect size in these terms.

### Summary and the 1-2-3 rule

We can summarize the procedures under the four types of specification in this section so far as shown in Table 8. It is instructive at this point to highlight that all procedures involve the *SED*. This important fact can be exploited to convert all procedures into simple rules in terms of the choice of *SED*. Thus, for achieving a desired value of  $\tau_\delta$ , *EHW* or *ELSD*, we need to choose:

$$SED = \frac{\tau_\delta}{z_{1-\alpha/2}} = \frac{EHW}{z_{1-\alpha/2}} = \frac{ELSD}{z_{1-\alpha/2}}. \tag{37}$$

It is also seen that in practice, the three precision measures  $\tau_\delta$ , *EHW* or *ELSD* are exchangeable, despite differences in their derivation. For detecting a minimal effect size  $\delta$ , we need to choose:

$$SED = \frac{\delta}{z_{1-\alpha/2} + z_{1-\beta}} \tag{38}$$

This latter fact led Mead (1988, p. 126; also see Mead *et al.*, 2012, p. 137) to suggest the rule of thumb that *SED* should be no larger than  $|\delta|/3$ , corresponding to an approximate power of  $1 - \beta = 0.85$  at  $\alpha = 5\%$  because  $z_{1-\alpha/2} + z_{1-\beta} \approx 3$ . By comparison, using  $z_{1-\alpha/2} \approx 2$  for  $\alpha = 5\%$  in Eqn (37) yields the rule that *SED* should be no larger than  $\tau_\delta/2 = EHW/2 = ELSD/2$ . As in the case of a single mean (previous section), we can turn this around and first compute the *SED* for a given design to evaluate its precision. Then  $2 \times SED$  is the allowable deviation  $\tau_\delta$  (*EHW*, *ELSD*) the design permits to control. Similarly,  $3 \times SED$  is the smallest absolute difference  $|\delta|$  the design can detect. Because of the divisors and multipliers involved (1 for *SED* itself, 2 for  $\tau_\delta$ , *EHW* or *ELSD*, and 3 for  $\delta$ ), we refer to this set of portable equations and rules as the *1-2-3 rule for a difference*.

### Procedures for counts

As was already pointed out in the previous sub-section, the common distributional models for counts (e.g., binary, binomial, Poisson) imply that the variance depends on the mean. When it comes to the comparison of means between two groups, the consequence is that there is the heterogeneity of variance between the groups unless the means are identical. Therefore, all

**Table 8.** Overview of procedures for determining sample size for a mean difference

Parameter for specification of precision	Interpretation of precision parameter	Solution(s) for $n$ in main text (equation numbers in brackets)	Further prior specification needed
$SED$	Standard error of a difference	(24)	$\sigma^2$
$\tau_\delta$	Allowable absolute deviation from the true difference	(24), (26)	$\sigma^2, \alpha$
$EHW$	Expected half width of confidence interval for the difference	Exact only numerically; (28) for an approximation	$\sigma^2, \alpha$
$\delta$	Smallest absolute difference to be detected	Exact only numerically; (31), (32) and (33) for an approximation	$\sigma^2, \alpha, \beta$
$\delta_r$	Smallest relative difference to be detected	Exact only numerically; (35) and (36) for an approximation	$CV, \alpha, \beta$

specifications for sample size need to be made explicitly in terms of the two means, and not just their difference, which is a slight complication compared to the normal case assuming homogeneity. As a result of this slight complication, there are several approximate approaches for determining the sample size. Most of them rely on the approximate normality of estimators of the parameters, which is a consequence of the central limit theorem. This is not the place to give a full account of all the different options. Many of these are nicely summarized in van Belle (2008).

Here, we will just mention one particularly handy approximate approach that employs a variance-stabilizing transformation of the response variable  $y$ . For the Poisson distribution with large mean, the square root transformation  $z = \sqrt{y}$  stabilizes the variance at  $\text{var}(z) \approx 1/4$  (McCullagh and Nelder, 1989, p. 196). For the binomial distribution with large  $m$  (number of observational units per sample), the angular transformation  $z = \arcsin\{(c/m)^{1/2}\}$ , where  $c$  is the binomial count, approximately stabilizes the variance at  $\text{var}(z) \approx 1/(4m)$  (McCullagh and Nelder, 1989, p. 137). Allowing for over-dispersion, which is the rule rather than the exception in comparative experiments in agriculture (Young *et al.*, 1999), the variance needs to be adjusted to  $\text{var}(z) \approx \phi/4$  and  $\text{var}(z) \approx \phi/(4m)$  for the Poisson and binomial distributions, respectively. The advantage of using these variances on the transformed scale is that they are independent of the mean, simplifying the sample size calculation a bit. Thus, we can use:

$$\sigma^2 = \phi/4 \quad (39)$$

for the over-dispersed Poisson and:

$$\sigma^2 = \phi/(4m) \quad (40)$$

for the over-dispersed binomial distribution in equations in the preceding sub-sections. At the same time, however, the specifications for  $\tau_\delta$  and  $\delta$  need to be made on the transformed scale, and this, in turn, requires that the two means need to be specified explicitly, rather than just their difference. For example, under the (over-dispersed) Poisson model we use:

$$\delta = \sqrt{\mu_1} - \sqrt{\mu_2} \quad (41)$$

and under the (over-dispersed) binomial model we use:

$$\delta = \arcsin(\mu_1^{1/2}) - \arcsin(\mu_2^{1/2}) \quad (42)$$

where the means correspond to the binomial probabilities being

compared (Cohen, 1977, p. 181, 1992). These expressions can be inserted in Eqns (31)–(33), leading to explicit equations for the Poisson and binomial models if desired (Cochran and Cox, 1957, p. 27; van Belle, 2008, p. 40 and p. 44). With over-dispersion, which should be the default assumption for replicated experiments, a prior estimate of the over-dispersion parameter  $\phi$  will be required to evaluate the variances in (39) and (40) for use in the expressions in the preceding sub-sections. Such an estimate can be obtained via Pearson's chi-squared statistic or the residual deviance based on a generalized linear model (GLM; McCullagh and Nelder, 1989). Later in this paper, we will consider a simulation-based approach that can be applied for count data when the simplifying assumptions made here (e.g., large binomial  $m$  or large Poisson mean) are not met. Also, we note in passing that the angular transformation, originally proposed for binomial proportions, may sometimes work for estimated (continuous) proportions, but see Warton and Hui (2011) for important cautionary notes, Piepho (2003) and Malik and Piepho (2016) for alternative transformations, and Duoma and Weedon (2018) on beta regression as an alternative.

*Example 9:* A field experiment is to be conducted in randomized complete blocks to compare a new herbicide against the weed grass *Bromus sterilis* to a control treatment. The number of weed plants per  $m^2$  will be assessed by sampling five squares of  $0.25 m^2$  per plot and dividing the total count of weed plants by  $1.25 m^2$ . A previous trial with the same kind of design yielded the total counts per plot shown in Table 9. Analysis of this trial using a GLM for overdispersed Poisson data using a log-link (McCullagh and Nelder, 1989) yielded the overdispersion estimate  $\hat{\phi} = 2.59$ . For the future trial, the smallest relevant effect is specified in terms of the mean  $\mu_1 = 15$  plants per  $1.25 m^2$  for the control and the mean  $\mu_2 = 3$  plants per  $1.25 m^2$  for a new herbicide, corresponding to a reduction of weed infestation by 80%. Using a square-root transformation with variance in (39) and effect size in (41), we find from (31) for  $\alpha = 5\%$  and a power of 90% that  $\delta = \sqrt{15} - \sqrt{3} = 2.14$ ,  $\sigma^2 = 2.59/4 = 0.648$  and:

$$n \approx \frac{2 \times 0.648 \times (1.96 + 1.28)^2}{2.14^2} = 2.97 \Rightarrow 3$$

Incidentally, the variance specification based on a GLM ( $\sigma^2 = 0.648$ ) is quite close to the estimate by an analysis of the square-root transformed counts ( $s^2 = 0.607$ ), confirming the utility of the simple data transformation approach. It may be added that this sample size is smaller than the one actually used ( $n = 4$ ). That sample size, however, would not have been sufficient to detect a weed reduction by 50% at the same level of significance and power.

**Table 9.** Total counts of *Bromus sterilis* on 1.25 m<sup>2</sup> per plot in an experiment laid out as a randomized complete block design with four treatments and four blocks (Büchse and Piepho, 2006)

Treatment	Block				Sample mean m <sup>-2</sup>
	I	II	III	IV	
1 (control)	20	20	6	14	12.0
2	6	5	0	3	2.8
3	13	3	0	0	3.2
4	10	11	9	5	7.0

*Example 10:* A field experiment is to be conducted to assess the effect of a neonicotinoid on the abundance of an insect species. The expected abundance for the control is in the order of ten individuals per trap. The smallest relevant difference for the treatment corresponds to a 25% drop in abundance, amounting to 7.5 individuals per trap for the treatment. We set  $\mu_1 = 10, \mu_2 = 7.5, \alpha = 5\%$  and  $\beta = 20\%$ . Assuming a Poisson distribution, we initially set  $\sigma^2 = 0.25$  as per (39) on the optimistic assumption of no overdispersion. We find  $\delta = \sqrt{\mu_1} - \sqrt{\mu_2} = 0.4237$  and use all of these specifications in Eqn (31), finding  $n = 21.86 \Rightarrow 22$ . From a previous study, we expect an overdispersion of  $\phi = 1.3$ . Hence, we adjust our sample size upwards to  $n = \phi \times 21.86 = 1.3 \times 21.86 = 28.42 \Rightarrow n = 29$ .

*Example 11:* Diseases or traits due to hereditary defects can often be detected by gene tests which require a population-wide evaluation. The principal idea is to test for an association between the status of a gene (which may have three outcomes/genotypes in a diploid organism) and the occurrence of a disease. For instance, being horned or hornless in cattle is caused by a mutation at the ‘polled locus’, and a gene test has already been established to increase the frequency of polled cattle in future through selective mating. To test whether the horned or hornless phenotype in cattle is caused by a specific variant in the genome, we distinguish factor level A (genotype pp) and B (genotype Pp or PP) and consider the 2 × 2 classification in Table 10.

The task is to approximate sample size allowing the detection of differences in probabilities  $\mu_1$  and  $\mu_2$  between groups. Table 10 reflects the expected counts and an obvious solution is to continue with the binomial model using, e.g.,  $\mu_1 = 0.9$  and  $\mu_2 = 0.5$  in Eqn (42) and  $\sigma^2 = 1/4$ . Using  $\sigma^2 = 1/4$  instead of  $1/(4m)$ , as you would expect from the binomial model, is justified by treating the binary variable as a limiting case here (see Paulson and Wallis, 1947; cited in Cochran and Cox, 1957, p. 27; also see the Appendix).

**Table 10.** Expected counts in a 2 × 2 classification of groups and treatments

	Factor level (e.g. treatment or status)		Sum total
	A	B	
Group 1 (e.g. affected individuals)	$n_1 \times \mu_1$	$n_1 \times (1 - \mu_1)$	$n_1$
Group 2 (e.g. unaffected individuals)	$n_2 \times \mu_2$	$n_2 \times (1 - \mu_2)$	$n_2$
Sum total	$n_A$	$n_B$	$n_1 + n_2 = n_A + n_B$

The parameters  $\mu_1$  and  $\mu_2$  are the probabilities of occurrence of level A in the two groups, and  $n_1$  and  $n_2$  are the sample sizes for the two groups.

Assuming a balanced design and  $\delta = 0.4637$  from (42), yields  $n_1 = n_2 = n = 25$  using Eqn (31) with  $\alpha = 5\%$  and a power of 90% (also see Cochran and Cox, 1957, p. 27). Using a more specifically tailored formula due to Fleiss (1981; also see eq. 2.50 in Rasch *et al.*, 2011) yields  $n_1 = n_2 = 26$ , and further using a correction due to Casagrande *et al.* (1978; also see eq. 2.51 in Rasch *et al.*, 2011, p. 49) yields  $n_1 = n_2 = 30$ , so our approximate approach lands us in the right ballpark. We conclude by noting that the task above could also be tackled with a chi-square test of independence on 1 degree of freedom or by a test of the log-odds ratio, as will be discussed later.

**Paired samples**

This section so far focused on unpaired, i.e. independent samples. When paired samples are considered, we may resort to procedures for a single mean, replacing the observed values  $y_j$  with observed paired differences  $d_j$  between two treatments or conditions. Accordingly, mean and variance of  $y_j$  need to be replaced by mean and variance of paired differences  $d_j$ . Because of this one-to-one relation with the case of a single mean, procedures for paired samples are slightly simpler than for unpaired samples. We note that the section for a single mean does not explicitly consider significance tests, but the confidence interval for a difference may be used to conduct a significance test of  $H_0$  that the expected difference equals zero, exploiting the close relation between both procedures. The  $H_0$  is rejected at the significance level  $\alpha$  when the  $(1 - \alpha) \times 100\%$  confidence interval for the difference does not include zero. Thus, all options can be implemented with the procedures for a single mean. As regards significance testing, Eqn (31) needs to be modified as:

$$n \approx \frac{\sigma_d^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2} \tag{43}$$

where  $\sigma_d^2$  is the variance of pairwise differences  $d_j$  and approximate normality is assumed.

*Example 12:* An experiment was conducted with Fleckvieh dairy cows to compare the lying time per day indoors and outdoors on an experimental farm (Benz *et al.*, 2020). Sufficient lying time is an important trait for claw health. A total of 13 cows, sampled randomly from the current population at the farm (~50 animals), were included in the experiment and their average lying time per day assessed in both phases using pedometers (Table 11). The indoor phase in the barn took place in early September 2017 and the outdoor phase was conducted in late September 2017.

From these data, the variance  $\sigma_d^2$  of the 13 pairwise differences ( $d_1 = 745 - 614 = 113$ , etc.) is estimated at  $s_d^2 = 7355$ . If for a new study we want to be able to detect a lying time difference of  $\delta = 40$  min day<sup>-1</sup> at  $\alpha = 5\%$  with a power of 80%, the required sample size is:

$$n \approx \frac{\sigma_d^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2} = \frac{7355 \times (1.96 + 0.84)^2}{40^2} \approx 28$$

Thus, we would require 28 cows. This sample size is expected to yield  $SED = 14.29$  (Eqn 1) and  $EHW = \tau_\delta = 28.01$  (Eqns 6 and 10).

It is noted that the size of the population at the farm from which the cows are to be sampled, is relatively small. One might therefore consider a finite-sample correction to account for this as in described in the previous section for a single mean, which would

**Table 11.** Lying times (min day<sup>-1</sup>) of 13 cows indoors and outdoors

Cow	Lying time (min day <sup>-1</sup> )		
	Indoors	Outdoors	Difference ( $d_i$ )
6_Eva_158	745	614	131
14_Fiury_951	678	561	117
18_Olympia_048	682	568	114
19_Gitti_184	738	551	187
26_Zirbel_507	819	861	-42
27_Mila_172	548	615	-67
33_Mirzl_031	688	768	-80
34_Distel_077	631	551	80
36_Olina_214	621	630	-9
37_Olga_179	612	586	26
39_Frieda_239	742	612	130
52_Alma_540	716	691	25
53_Gundi_025	605	494	111
Mean	678.8	623.2	55.6

lead to a slightly smaller sample size, but would restrict the validity of the results to the farm population studied. An alternative view is that the ~50 animals at the farm are themselves a sample from the much larger Fleckvieh population and that the objective of the study is not to characterize the limited population at the farm, but to characterize conditions at the farm itself. In this view, which provides a somewhat broader inference, it makes sense to regard the  $n=28$  animals as a sample from the broader Fleckvieh population, raised under the conditions of the farm at hand. In this case, a finite-population correction is not needed.

### More advanced settings

#### More than two means

When more than two means are to be compared, the same methods as in the previous section can be used, as in the end we usually want to compare all pairs of means. The only additional consideration is that there may be a need to control the family-wise Type I error rate in case of multiple pairwise tests. In case of normal data, this means using Tukey's test rather than the  $t$ -test (Bretz *et al.*, 2011), and sample size calculations may be adjusted accordingly (Horn and Vollandt, 1995; Hsu, 1996). A simple approximation is afforded by the Bonferroni method which prescribes dividing the targeted family-wise  $\alpha$  by the number of tests. That number equals  $\nu(\nu-1)/2$  for all pairwise comparisons among  $\nu$  treatments, so the  $t$ -tests would be conducted at significance level  $\alpha' = \alpha/[\nu(\nu-1)/2]$ . In a similar vein, the Tukey or Bonferroni methods can also be used when considering confidence intervals desired to have joint coverage probability of  $(1-\alpha) \times 100\%$ ; with the Bonferroni method this is achieved by computing  $(1-\alpha') \times 100\%$  confidence intervals using the method in the previous section for the individual comparisons.

We note here that pairwise comparison of means are usually preceded by a one-way analysis of variance (ANOVA)  $F$ -test of the global null hypothesis of no treatment differences (though this is not strictly necessary when pairwise comparisons are done

controlling the family-wise Type I error rate). It is also possible to determine sample size based on the power of the one-way ANOVA  $F$ -test (Dufner *et al.*, 1992, p. 196; Dean and Voss, 1999, p. 49f.; Rasch *et al.*, 2011, p. 59), and we will come back to this option in the next section. It is emphasized here that we think the consideration of pairwise comparisons is usually preferable for determining sample size also when  $\nu > 2$ , because it is more intuitive and easier in terms of the specification of the precision required. When the focus is on individual pairwise mean differences, all equations for unpaired samples remain valid with the variance  $\sigma^2$  for the design in question. Specifically, these equations can be applied with the three most common and basic experimental designs used in agricultural research, i.e. the completely randomized design, the randomized complete block design, and the Latin square design (Cochran and Cox, 1957; Dean and Voss, 1999).

*Example 7 (cont'd):* Suppose we want to add three further new formulations with vitamin A, increasing the total number of treatments to  $\nu = 5$ . If the specifications for the required precision or power remain unchanged, so does the required sample size per treatment group. Only the total sample size increases from  $2n$  to  $\nu n = 5n$ . If we want to cater for a control of the family-wise Type I error rate at the  $\alpha = 5\%$  level, we may consider a Bonferroni approximation of the Tukey test and use a pairwise  $t$ -tests at  $\alpha' = \alpha/[\nu(\nu-1)/2] = 5\%/[5 \times 4/2] = 0.5\%$ . Thus, we would replace  $z_{1-\alpha/2}$  with  $z_{1-\alpha'/2}$ , which for  $\alpha' = 0.5\%$  equals  $z_{1-\alpha'/2} = 2.81$ , compared to  $z_{1-\alpha/2} = 1.96$  for  $\alpha = 5\%$ . Thus, sample size requirement according to (31) for  $\delta = 20$  lb. at  $\alpha = 5\%$  with a power of 90% would increase from  $n = 115$  to  $n = 184$  per group.

### Regression models

The simplest case in regression is a linear regression of a response  $y$  on a single regressor variable  $x$ . Apart from sample size, the placement of the treatments on the  $x$ -axis needs to be decided. For a linear regression, the optimal allocation is to place half the observation at the lower end,  $x_L$ , and the other half at the upper end,  $x_U$ , of the relevant range for  $x$  (Rasch *et al.*, 1998, p. 273; Rasch *et al.*, 2011, p. 127f.). Optimal allocation for higher-order polynomials or intrinsically nonlinear models is more complex and will not be elaborated here (see, e.g., Dette, 1995). But it is stressed that the optimal allocation for such models will almost invariably involve more than two  $x$ -levels.

The simplest linear case, however, can be used to make a rough assessment of the required sample size. The optimal design with observations split between  $x_L$  and  $x_U$  essentially means that at both points the mean needs to be estimated. If we denote these means by  $\mu_L$  and  $\mu_U$ , the linear slope is given by  $\gamma = (\mu_U - \mu_L)/(x_U - x_L)$ , showing that estimating the slope is indeed equivalent to comparing the two means. This, in turn, suggests that we can use the methods in Section 'Comparing two means' to determine the sample size. We just need to quantify the relevant change in the response from  $x_L$  to  $x_U$ , given by  $\delta = \mu_U - \mu_L$ .

With many nonlinear models, more than two  $x$ -levels will be needed, and the definition of a relative effect may be more difficult. Often, parts of the expected linear response can be well approximated by linear regression, and relevant changes defined by parts. Therefore, consideration of the simplest linear case may be sufficient to determine a suitable number of observations per  $x$ -level. It may also be considered that higher-order polynomials (quadratic or cubic) can be estimated based on orthogonal polynomial contrasts (Dean and Voss, 1999, p. 261), which also

constitute a type of mean comparison, giving further support to our portable approach.

Our considerations here do not imply that we would normally recommend doing a linear regression with just two  $x$ -levels, unless one is absolutely sure that the functional relationship will indeed be linear. In order to be able to test the lack-of-fit of any regression model (Dean and Voss, 1999, p. 249; Piepho and Edmondson, 2018), one or two extra  $x$ -levels will be needed. Also, for each parameter in a nonlinear regression model to be estimated, one additional  $x$ -level will be required. This leads to the following rule-of-thumb for the number of  $x$ -levels: (i) Determine the number of parameters of the most complex nonlinear model you are intending to fit (there may be several). (ii) That number plus one or two should be the number of  $x$ -levels in the experiment. The number of replications per  $x$ -level can then be determined as in the previous section. More sophisticated approaches for allocating samples to  $x$ -levels are, of course, possible, especially when several  $x$ -variables are considered, and these may involve unequal sample sizes between  $x$ -levels (Box and Draper, 2007). Also, the  $x$ -levels are usually equally spaced, even though depending on the assumed model unequal spacing may sometimes be preferable (Huang *et al.*, 2020). These more sophisticated approaches are not considered here, however, because they are less portable.

Continuing with the idea that a rough assessment of sample size is possible by considering the linear case and the comparison of the response at  $x_L$  and  $x_U$ , and keeping in mind that usually we want to test more than just the two extreme levels  $x_L$  and  $x_U$ , we may consider the case of  $v$  treatments equally spaced on the  $x$ -axis. Generally, the variance of the estimate of the linear slope  $\gamma$  equals  $\sigma^2$ , divided by the sum of squares of the  $x$ -levels. If we assume that the  $v$   $x$ -levels  $x_1, x_2, \dots, x_v$  are equally spaced and each level is replicated  $n$  times, the standard error of a slope (SES) equals:

$$SES = \sqrt{\sigma^2 \frac{D_v}{n(x_U - x_L)^2}} \tag{44}$$

where  $D_v = [12(v - 1)^2] / [v(v^2 - 1)]$ . Values of  $D_v$  for typical values of  $v$  are given in Table 12. An interesting limiting case occurs when  $v = 2$ , for which  $D_v = 2$ . If, without loss of generality, we assume  $x_L = 0$  and  $x_U = 1$ , then SES equals SED in (23), confirming our suggestion that considering a comparison of the means at  $x_L$  and  $x_U$  provides a useful rough guide to sample size per treatment level for linear regression. Further considering (44) and the values of  $D_v$  for  $v > 2$  in Table 12 confirms that this provides a conservative estimate of sample size. Solving (44) for  $n$  yields the sample size per treatment required to achieve a preset value of SES:

$$n = \sigma^2 \frac{D_v}{SES^2(x_U - x_L)^2} \tag{45}$$

Similar derivations give the equations based on the allowable deviation of the estimate of  $\gamma$  from its true value,  $\tau_\gamma$ , as:

$$n = \sigma^2 \frac{z_{1-\alpha/2}^2 D_v}{\tau_\gamma^2 (x_U - x_L)^2} \tag{46}$$

and based on the EHW as:

$$n = \sigma^2 \frac{z_{1-\alpha/2}^2 D_v}{EHW^2 (x_U - x_L)^2} \tag{47}$$

Finally, the sample size based on a  $t$ -test of  $H_0: \gamma = 0$  v.  $H_A: \gamma \neq 0$  is:

$$n \approx \frac{D_v \sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{\gamma^2 (x_U - x_L)^2} \tag{48}$$

where  $\gamma$  is the smallest absolute value of the slope that we consider relevant. By way of analogy to procedures in the two previous sections (1-2 and 1-2-3 rules), all of these rules could be converted into specifications in terms of the required SES, but this is not detailed here for brevity.

*Example 7 (cont'd):* In the experiment conducted by Ross and Knodt (1948), the basal ration contained 114 000 USP units of vitamin A per daily allowance per heifer (USP is a unit used in the United States to measure the mass of a vitamin or drug based on its expected biological effects). This was supplemented with 129 400 USP units of vitamin A for the 14 heifers in the vitamin A group. Now assume a follow-up experiment is planned in which  $v = 5$  equally spaced levels of supplementation between  $x_L = 0$  and  $x_U = 129\,400$  USP units are to be tested. We use  $\sigma^2 = 2199$  as before. When illustrating Eqn (31) for a  $t$ -test to compare two means, we had considered the difference of  $\delta = 20$  lb to be the smallest relevant effect size. In our regression, this increase corresponds to a linear slope of  $\gamma = (\mu_U - \mu_L) / (x_U - x_L) = \delta / 129\,400 = 20 / 129\,400$ . Also note that  $\gamma(x_U - x_L) = \delta = 20$  lb. Hence, using (48) the sample size needed per treatment for linear regression with  $v = 5$  levels at  $\alpha = 5\%$  with a power of 90% is:

$$\begin{aligned} n &\approx \frac{D_v \sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{\gamma^2 (x_U - x_L)^2} = \frac{1.60 \times 2199 \times (1.96 + 1.28)^2}{20^2} \\ &= 92.42 \Rightarrow 93 \end{aligned}$$

This is somewhat smaller than the sample size needed per treatment for comparing two means, where we found  $n = 116$ . The ratio of these two sample sizes,  $93/116 \approx 0.8$ , equals the ratio of the values for  $D_v$  for  $v = 5$  and  $v = 2$  (Table 12),  $1.60/2.00 = 0.8$ , which is no coincidence. The example confirms our assertion that sample size based on comparing two means provides a rough guide to sample size for linear regression based on more than two  $x$ -levels.

### Pseudoreplication

The number of replications in a randomized trial is given by the number of experimental units to which treatment is independently and randomly allocated. In a field trial, the experimental

**Table 12.** Values of  $D_v$  (see near Eqn 44) for  $v = 2, 3, \dots, 8$  treatment levels

$v$	2	3	4	5	6	7	8
$D_v$	2.00	2.00	1.80	1.60	1.43	1.29	1.17

unit is the plot. In some cases, there may be multiple observations per experimental unit so the number of observational units exceeds that of experimental units. It is generally important to bear in mind that observational units and experimental units do not necessarily coincide (Bailey, 2009). If there are multiple observations per experimental unit, these are denoted as sub-samples (Piepho, 1997; Welham *et al.*, 2015, p. 47) or pseudo-replications (Hurlbert, 1984; Davies and Gray, 2015).

There are two ways to properly analyse such data: (i) Compute means per experimental unit and then subject these to ANOVA in accordance with the randomization layout. (ii) Fit a mixed model in accordance with the randomization layout that has two random effects, one for experimental units and one for observational units. In case the number of observations per experimental unit is constant, both analyses will yield identical results, unless the variance for experimental units is estimated to be zero under option (ii), in which case option (i) is preferable for better Type I error control. With an unequal number of observations per experimental unit, option (ii) is to be preferred because it gives proper weights to experimental units depending on the respective number of observational units (Piepho, 1997). In this section, we will restrict attention to the equi-replicated case, which is also preferable in terms of overall precision.

The sample size here has two components, i.e., the number of experimental units ( $n_e$ ) and the number of observational units per experimental unit ( $n_o$ ). The latter number may be a given, depending on the circumstances. Where  $n_o$  can be freely chosen, its optimal value depends on the variance components for two sources of variation, i.e. the variance between experimental units ( $\sigma_e^2$ ) and the variance between observational units within the same experimental unit ( $\sigma_o^2$ ). A linear model (LM) for the  $j$ th observation on the  $i$ th experimental unit ( $y_{ij}$ ) can be written as:

$$y_{ij} = \mu + e_i + o_{ij}, \tag{49}$$

where  $e_i$  is the error of the  $i$ th experimental unit, having variance  $\sigma_e^2$ , and  $o_{ij}$  is the observational error of  $y_{ij}$ , having variance  $\sigma_o^2$ . The variance of a mean (VM) is:

$$\text{var}(\bar{y}_{\bullet\bullet}) = VM = \frac{\sigma_e^2}{n_e} + \frac{\sigma_o^2}{n_e n_o} \tag{50}$$

where  $\bar{y}_{\bullet\bullet} = (n_e n_o)^{-1} \sum_{i=1}^{n_e} \sum_{j=1}^{n_o} y_{ij}$  is the mean. This equation shows that increasing  $n_e$  reduces the impact of both  $\sigma_e^2$  and  $\sigma_o^2$ , whereas increasing  $n_o$  only reduces the impact of  $\sigma_o^2$ . Hence, an additional experimental unit is worth more than an additional observational unit. If  $c_e$  is the operational cost for one experimental unit and  $c_o$  is the cost for one observational unit, both measured in the same units (e.g., in monetary terms or as working time), the optimal number of observational units per experimental unit is given by (Snedecor and Cochran, 1989, p. 448):

$$n_o = \sqrt{\frac{c_e \sigma_o^2}{c_o \sigma_e^2}} \tag{51}$$

For given sample size  $n_e$ , this minimizes the total cost  $C = n_e c_e + n_e n_o c_o$ . Once the optimal  $n_o$  is determined, the required  $n_e$  can be determined using the methods in the previous two sections

(single mean, two means), setting  $n_e = n$  and

$$\sigma^2 = \sigma_e^2 + \frac{\sigma_o^2}{n_o}. \tag{52}$$

We also note that for completely randomized designs or randomized complete block designs or Latin square designs, the variance of a difference (VD) will be  $VD = 2 \times VM$  and the standard error of a difference will be  $SED = \sqrt{VD} = \sqrt{2} \times SEM$ , where  $SEM = \sqrt{VM}$ .

*Example 13:* A trial was conducted to assess the merit of four different intermediate crops following the main crop oats. The crops were tested with three different sowing methods, allocated to main plots in four complete blocks. The five different intermediate crops were allocated to the subplots. Cover by these intermediate crops was assessed based on five randomly placed counting frames (0.1 m<sup>2</sup> each) per plot. One of the traits was the visually estimated percentage of ground cover by the intermediate crop. Inspection of the residuals indicated that an angular transformation would stabilize the variance. For the transformed data, the variance for the main plots was estimated to be zero. The variance for subplots was 0.000318 and that for samples within subplots was 0.00840.

For a future trial of the same kind, to be laid out as a randomized complete block design, the optimal number of samples per plot ( $n_o$ ) is to be determined. For this purpose, we use values  $\sigma_e^2 = 0.000318$  and  $\sigma_o^2 = 0.00840$ , showing that the within-plot variance dominates. Cost for plots and samples are not quantified here. Instead, we try different feasible values for  $n_o$  as shown in Table 13 and compute the associated value of  $\sigma^2$  as per (52). With these variances, we determine the optimal sample size  $n_e$  using Eqn (31) with  $\alpha = 5\%$  and  $\beta = 20\%$ . As we used a data transformation, specification of  $\delta$  is based on Eqn (42), where  $\mu_1$  and  $\mu_2$  are coverages, expressed as proportions, reflecting the smallest relevant difference at expected orders of magnitude for both proportions. Here, we set  $\mu_1 = 0.1$  and  $\mu_2 = 0.2$ .

If we use two samples per plot, we need four replicates. If we use only a single sample per plot, seven replicates are needed. Conversely, we may conclude that for a given trial with four replicates, a sample size of  $n_o = 5$  as used in the current trial is more than required at the pre-specified level of power.

*Example 14:* In a field experiment with spring barley, the number of ears per two metres within rows is to be determined at the start of ripening (BBCH stage 83; Bleiholder *et al.*, 1989). The experiment has eight treatments and  $n_e = 4$  replications, arranged in complete blocks. The number of ears is determined by sampling sections of two metre length within rows. A decision needs to be taken on the spot as to the number  $n_o$  of sections to be sampled per plot. To this end, initially two sections are

**Table 13.** Number of plots ( $n_e$ ) required as depending on number of samples per plot ( $n_o$ ) for  $\alpha = 5\%$ ,  $\beta = 20\%$ ,  $\mu_1 = 0.1$  and  $\mu_2 = 0.2$  based on angular transformation for intermediate crop experiment (Example 13)

$n_o$	$\sigma^2$	$n_e$ as per (31)	$n_e$ rounded up
1	0.00872	6.80	7
2	0.00452	3.52	4
3	0.00312	2.43	3
4	0.00242	1.89	2
5	0.00200	1.56	2

sampled per plot and the data analysed to estimate the variance components for plots and sections within plots. The data is given in Table 14. The response is a count but inspection of residual plots (Kozak and Piepho, 2018) indicated no serious departure from the homogeneity of variance assumption. Thus, the untransformed data were analysed using a linear mixed model (LMM) with fixed effect for treatments and replicates and random effects for plots and sections (residual effect). The resulting variance component estimates, obtained using a mixed model package on a laptop computer during the coffee break, are  $\hat{\sigma}_e^2 = 12.00$  and  $\hat{\sigma}_o^2 = 19.98$ . These estimates are now used to plan the final sample size  $n_o$  for a  $t$ -test using Eqn (31). Observing that the plot mean has the variance given in (52), we plug this equation into (31). Solving for  $n_o$  yields:

$$n_o \approx \sigma_o^2 \left( \frac{n_e \delta^2}{2(z_{1-\alpha/2} + z_{1-\beta})^2} - \sigma_e^2 \right)^{-1} \tag{53}$$

It is important to note that this equation can return a negative value for  $n_o$ . If this happens, the number of replications ( $n_e$ ), which is a fixed quantity for an ongoing field experiment, is too small to achieve the desired precision. By way of illustration, assume that we want to plan for a  $t$ -test at  $\alpha = 5\%$ , a power of 90% and a relevant difference of  $\delta = 5$  ears per section. The resulting approximate sample size is:

$$n_o \approx 19.98 \times \left( \frac{4 \times 5^2}{2(1.96 + 1.28)^2} - 12.00 \right)^{-1} = -2.76$$

A negative sample size is not feasible, of course. It can be shown algebraically that the reason for this result is that  $n_e = 4$  are too few replications to achieve the required precision; even a very large number of sections ( $n_o$ ) would not achieve this, because as per Eqn (52) the variance  $\sigma^2 \approx \sigma_e^2 = 12.00$ , which is too large. In the next experiment of this kind, it will therefore be prudent to increase the number of replications ( $n_e$ ). To illustrate further, consider the less ambitious choice of  $\delta = 10$  ears per section. Using Eqn (53), this yields  $n_o = 2.84 \Rightarrow 3$ .

Example 15: A long-term three-factorial experiment is conducted with the factors tillage, fertilization and biodynamic preparations. The experimental design is a strip-split plot design with four replicates and eight treatments. The total number of

plots is 32 with a size of  $12 \times 12 \text{ m}^2$  each. In 2021 spelt was grown, which started to lodge due to several heavy rain events. After a visual assessment of the lodging, it appeared that there was a significant effect of biodynamic preparations on the lodging resistance. Therefore, the stability of the plants was examined more closely. The specific stem weight (in mg per 10 cm stem) and circumference (in mm) were selected as parameters for stability according to Zuber *et al.* (1999). Forty stems per plot were chosen randomly and harvested by hand. Both traits were measured on each stem. In addition,  $n_o = 2$  quadrats of  $50 \times 50 \text{ cm}^2$  each were assessed on each plot for average plant height (cm) and number of culms. The data were analysed using a LMM with random effects for plots ( $e_i$ ) and stems within plots ( $o_{ij}$ ) and fixed effects for treatments and blocks. This model is used in order to obtain variance components for a planned design to be laid out in randomized complete blocks. The variance components for all four traits are reported in Table 15.

We first consider the optimal allocation using Eqn (51). The variable costs per sample ( $c_o$ ) range between 0.05 and 0.25 € for the four traits (Table 15), while the fixed costs per plot and year (without examinations of samples) are estimated at about 300 €. To illustrate the calculation for stem circumference, we use  $c_o = 0.10$  € and  $c_e = 300$  €, yielding the optimal allocation:

$$n_o = \sqrt{\frac{c_e \sigma_o^2}{c_o \sigma_e^2}} = \sqrt{\frac{300 \times 2.4979}{0.10 \times 0.1671}} = 211.77 \Rightarrow 212.$$

This optimal allocation, and also that for the other three traits (Table 15), are considerably larger than the ones used in the trial, suggesting that an increase of sample size  $n_o$  would be worthwhile. However, the optimal allocations seem unrealistically high. This is mainly a result of the minute cost per sample ( $c_o$ ), compared to the fixed cost per plot ( $c_e$ ). Moreover, the optimal allocation assumes that we are free in choosing the number of plots ( $n_e$ ), but in the current trial this is a given. Nevertheless, the optimal allocation is still instructive, tentatively pointing in the direction of higher sample size per plot for all traits.

The effect of increasing  $n_o$  and  $n_e$  on the *SED* is shown in Figs 1–4. The set of values for  $n_o$  always starts with the one used in the current trial and ends with the optimal allocation. For traits having a very large variance for samples ( $\sigma_o^2$ ) compared to the variance for plots ( $\sigma_e^2$ ), i.e. stem weight and circumference,

Table 14. Number of ears per section within rows (2 metres) (Example 14)

Block	Treatment							
	1	2	3	4	5	6	7	8
1	45	48	32	31	35	72	51	45
	47	39	36	45	37	66	49	38
2	51	31	45	49	38	63	53	43
	45	42	56	47	42	58	62	37
3	40	37	46	38	40	70	55	54
	47	45	51	43	42	65	58	51
4	47	38	54	44	45	62	52	51
	42	41	48	50	52	64	56	43

**Table 15.** Variance component estimates (obtained by residual maximum likelihood) and treatment mean estimates (Example 15)

Trait	$\sigma_e^2$	$\sigma_o^2$	Range of treatment means in trial	Cost per sample $c_o$ (€)	Optimal allocation for $n_o$ (Eqn 51)
Stem circumference (mm)	0.1671	2.4979	11.8 – 13.1	0.10	212
Stem weight (mg)	7.4258	1116.62	96.6 – 116.6	0.10	672
Plant height (cm)	68.19	39.13	132 – 152	0.05	59
Culm number per 50 × 50 cm <sup>2</sup>	123.53	106.49	54 – 83	0.25	33

Fixed cost per plot:  $c_e = 300$  €.

increasing sample size  $n_o$  for fixed plot numbers ( $n_e$ ) has the most notable effect.

The analyses so far have provided some insights, but have not settled the question of optimal sample size. We may consider two specific questions: (i) Was the chosen sample size  $n_o$  sufficient for the current trial, where  $n_e = 4$ ? (ii) What are the best numbers of plots ( $n_e$ ) and samples per plot ( $n_o$ ) for a future trial? These are two different questions, and the answer in each case depends on the precision requirement. We will consider both questions in turn, using one trait at a time. For stem circumference, the smallest difference considered relevant is taken to be  $\delta = 1$  mm. For  $n_o = 40$  as used in the current trial, the variance of a plot mean equals:

$$\sigma^2 = \sigma_e^2 + \frac{\sigma_o^2}{n_o} = 0.1671 + \frac{2.4979}{40} = 0.2295$$

To detect a difference of  $\delta = 1$  mm at  $\alpha = 5\%$  with a power of 80%, we need:

$$n_e \approx \frac{2 \times 0.2295 \times (1.96 + 0.84)^2}{1^2} = 3.60 \Rightarrow 4$$

plots per treatment (Eqn 31), which is the number of plots in the

current trial. Conversely, using Eqn (53), we find that for a given number of  $n_e = 4$  plots:

$$n_o \approx \sigma_o^2 \left( \frac{n_e \delta^2}{2(z_{1-\alpha/2} + z_{1-\beta})^2} - \sigma_e^2 \right)^{-1} = 2.4979 \times \left( \frac{4 \times 1^2}{2(1.96 + 0.84)^2} - 0.1971 \right)^{-1} = 28.49 \Rightarrow 29$$

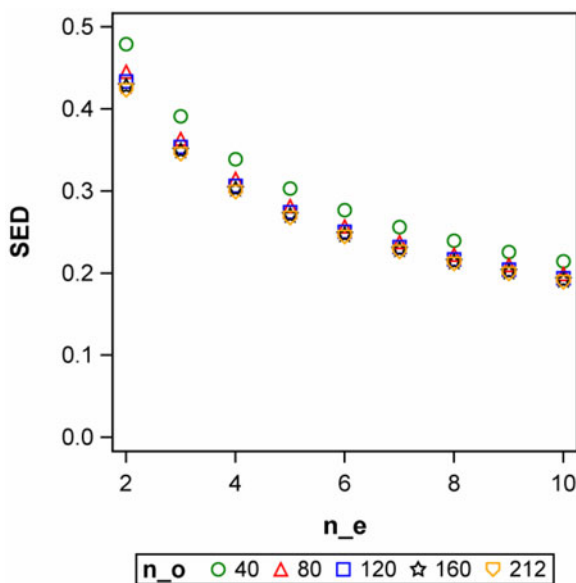
samples would be needed, which is less than the number in the current trial, so there is scope for reducing sample size a bit for this trait.

Turning to the second question, we use the optimal allocation  $n_o = 212$ , for which:

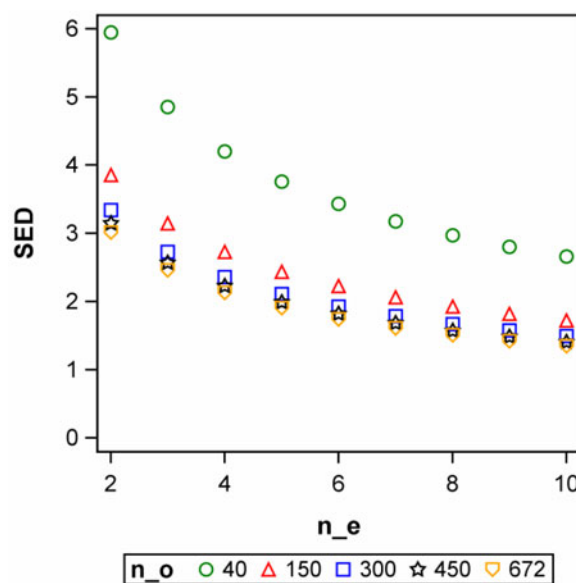
$$\sigma^2 = \sigma_e^2 + \frac{\sigma_o^2}{n_o} = 0.1671 + \frac{2.4979}{212} = 0.1789$$

Here, we would only need:

$$n_e \approx \frac{2 \times 0.1789 \times (1.96 + 0.84)^2}{1^2} = 2.81 \Rightarrow 3$$



**Fig. 1.** Colour online. Plot of *SED* versus  $n_e$  for  $n_o = 40, 80, 120, 160, 212$ . Trait: Stem circumference.



**Fig. 2.** Colour online. Plot of *SED* versus  $n_e$  for  $n_o = 40, 150, 300, 450, 672$ . Trait: Stem weight.



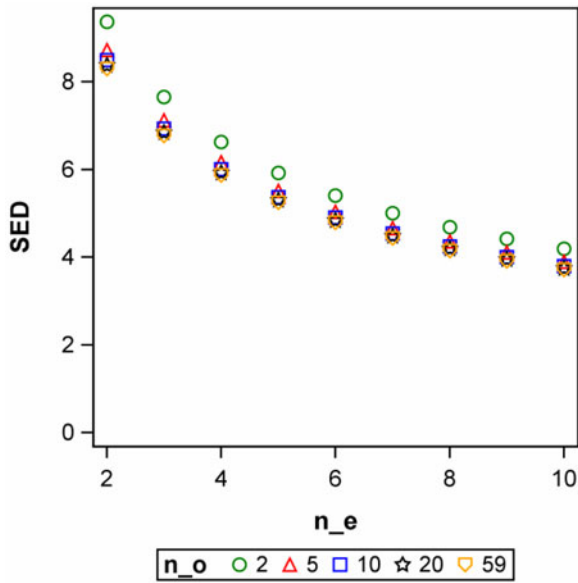


Fig. 3. Colour online. Plot of *SED* versus  $n_e$  for  $n_o = 2, 5, 10, 20, 59$ . Trait: Plant height.

plots per treatment. If in consideration of other traits, however, we decide to stick with  $n_e = 4$  plots, we would need to use the sample size  $n_o = 29$  found above.

Next, consider the trait stem weight. The individual specific stem weights were measured here so the two variance components for plots and samples could be estimated. Otherwise, for statistical analysis comparing treatments, we only need the plot means. This suggests that the 10 cm stem sections harvested for a plot can be pooled and the bulk weight determined and divided by the number of stems, thus reducing  $c_o$  and facilitating an increase in  $n_o$ . The fact that the sample variance is very much larger than the plot variance suggests that a large number of samples per plot is indeed very advantageous, and the plot in Fig. 2 also bears this out, considering the marked drop in *SED* when increasing  $n_o$  from 40 to 150. If  $\delta = 10$  mg is considered as the smallest

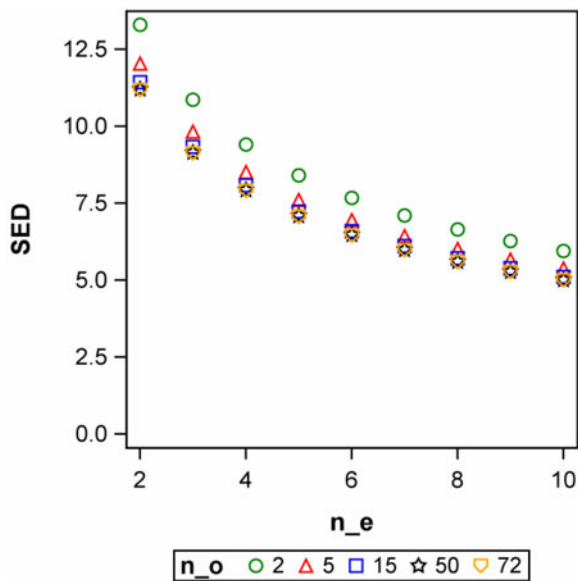


Fig. 4. Colour online. Plot of *SED* versus  $n_e$  for  $n_o = 2, 5, 15, 50, 72$ . Trait: Culm number.

relevant difference, then  $n_e = 6$  plots per treatment are required to achieve a power of 80% at  $\alpha = 5\%$  when  $n_o = 40$ . By comparison, when  $n_o = 70$  the required plot number is  $n_e = 4$ , and when  $n_o = 100$ , only  $n_e = 3$  plots are needed.

To conclude the example, we briefly consider the remaining two traits. For plant height,  $\delta = 10$  cm may be considered the smallest relevant difference. Applying Eqn (53) yields a negative value for  $n_o$ , indicating that the number of  $n_e = 4$  plots is not sufficient to detect such a difference. Hence, we determine the variance of a plot mean with the optimal allocation  $n_o = 59$  as  $\sigma^2 = 68.85$  and find the necessary number of plots per treatment from (31) as  $n_e = 11$  ( $\alpha = 5\%$ , power = 80%). This is much larger than the number of plots usually used, and we may not be prepared to increase the trial to this size. The calculations indicate, however, that with a smaller plot number we cannot expect sufficient precision for this trait. To improve precision (reduce  $\sigma_o^2$ ), one could also consider estimating individual stem heights rather than that of a collection of plants on a  $50 \times 50$  cm<sup>2</sup> sub-plot. Measuring the height of 30 or 40 individual stems per plot seems feasible. For culm number, we may use  $\delta = 10$  as the smallest relevant difference ( $\alpha = 5\%$ , power = 80%). As before, Eqn (53) shows that  $n_e = 4$  plots is not a sufficient number of replications. Applying the optimal allocation yields  $\sigma^2 = 126.76$ , and using this in (31) yields  $n_e = 20$  plots, which is clearly beyond feasible limits, indicating that it will not be possible to detect relevant treatment differences for this trait. Again, it may be worth devising an improved method for assessing this trait at the plot level.

### Series of trials

#### Several sites

Series of trials have a long history, and in agriculture, perhaps the most prominent use case is variety of trials (Yates and Cochran, 1938). Consider a set of variety trials conducted at several sites and each laid out in randomized complete blocks. The purpose is to assess the mean performance of the tested varieties in a target population of environments (Atlin *et al.*, 2000), and the sites used to conduct the trials are thought to be representative of this population. Hence, sites are modelled as a random factor and the following LMM is used for analysis:

$$y_{ijk} = \mu + g_i + s_j + (gs)_{ij} + b_{jk} + e_{ijk} \tag{54}$$

where  $\mu$  is an intercept,  $g_i$  is the main effect of the  $i$ th genotype,  $s_j$  is the main effect of the  $j$ th site,  $(gs)_{ij}$  is the interaction of the  $i$ th genotype and the  $j$ th site,  $b_{jk}$  is the effect of the  $k$ th replicate at the  $j$ th site, and  $e_{ijk}$  is the plot error. Assuming that the data is entirely balanced and all effects indexed by sites are random with constant variance, the variance of a difference (VD) of two genotype means ( $\mu + g_i$ ) is given by (Talbot, 1984):

$$VD = 2 \left( \frac{\sigma_{gs}^2}{n_s} + \frac{\sigma_e^2}{n_s n_r} \right) \tag{55}$$

where  $\sigma_{gs}^2$  is the genotype-by-site (genotype-by-environment) interaction variance,  $\sigma_e^2$  is the plot error variance,  $n_s$  is the number of sites and  $n_r$  is the number of replicates per site. The *SED* is then given by  $SED = \sqrt{VD}$ . Notice the similarity of the expression in brackets in Eqn (55) with the variance given in Eqn (50), which is no coincidence. In fact, we can exploit this analogy to find the optimal number of replications per site, provided that variable

costs can be obtained for each additional plot ( $c_r$ ) and for each additional trial site ( $c_s$ ). Then the optimal number of replications per site is given by (Snedecor and Cochran, 1989, p. 448):

$$n_r = \sqrt{\frac{c_s \sigma_e^2}{c_r \sigma_{gs}^2}} \tag{56}$$

Either with this number of replications per site, or the number of replications determined by other considerations, the required number of sites  $n_s$  can be determined using the methods for two means in the third section, setting  $n_s = n$  and:

$$\sigma^2 = \sigma_{gs}^2 + \frac{\sigma_e^2}{n_r} \tag{57}$$

If cost plays no crucial role, returning to Eqn (55), it may be observed that for a given total number of plots,  $n_r n_s$ ,  $VD$  would be maximized by setting  $n_r = 1$  and thus maximizing the number of sites,  $n_s$ . It must be borne in mind, however, that with one replication only it is impossible to assess the precision of an individual trial or identify outliers based on an analysis of residuals.

We also note the close links of  $VD$  in Eqn (55) with broad-sense heritability, which is often used by breeders to assess the efficiency of their trialling system (Atlin *et al.*, 2000). Assuming that genotypic main effects  $g_i$  have variance  $\sigma_g^2$ , this is given by (Piepho and Möhring, 2007):

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + (1/2)VD} \tag{58}$$

This equation may appear a bit unusual but is, in fact, easily seen to be equivalent to the common equation for broad-sense heritability (Nyquist, 1991):

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{gs}^2/n_s + \sigma_e^2/(n_s n_r)} \tag{59}$$

The main advantage of (58) is that it is amenable to straightforward generalization to all kinds of departures from the simple assumptions made here, including heterogeneity of variance, incomplete block designs, unbalanced data, or spatial analysis (Piepho and Möhring, 2007). We are making this link with heritability here to point out that maximizing  $H^2$  for given  $\sigma_g^2$  is much the same thing as minimizing  $VD$  or  $SED$ . The simple Eqn (59) can be used with pre-determined value of  $n_r$  to yield a portable expression for  $n_s$  as a function of the desired value of  $H^2$ :

$$n_s = \frac{\sigma_g^2}{\sigma^2} \frac{H^2}{1 - H^2} \tag{60}$$

where  $\sigma^2$  is as given in (57). This kind of equation was considered by Yan (2021), who further suggested to generally require  $H^2 = 0.75$ , in which case (60) simplifies to  $n_s = 3\sigma_g^2/\sigma^2$ . As convenient as this may seem, we think this latter requirement for  $H^2$  is taking the idea of portability one step too far, as the desirable and achievable level of  $H^2$  is usually quite context-specific and, among other things, depends on the trait and the stage of the breeding programme.

*Several sites crossed with years*

A second important case occurs when the trials are also replicated across years at the same sites. In this case, again assuming balanced data, the  $VD$  in (52) extends to (Talbot, 1984; Casler, 2015):

$$VD = 2 \left( \frac{\sigma_{gs}^2}{n_s} + \frac{\sigma_{gy}^2}{n_y} + \frac{\sigma_{gsy}^2}{n_s n_y} + \frac{\sigma_e^2}{n_s n_y n_r} \right) \tag{61}$$

where  $\sigma_{gy}^2$  is the genotype-by-year interaction variance,  $\sigma_{gsy}^2$  is the genotype-by-site-by-year interaction variance, and  $n_y$  is the number of years. The optimal allocation problem now involves three variables,  $n_r$ ,  $n_s$  and  $n_y$ , which is a bit more complex at first sight. In practice, however, the number of years,  $n_y$ , is usually fixed, or only has very few options, so we can regard this as a given in most applications. In this case, the term in (61) involving  $\sigma_{gsy}^2$  is a constant and does not affect the optimization problem. In fact, it is sufficient to consider the same optimization as in Eqn (56) using  $\tilde{\sigma}_{gs}^2 = \sigma_{gs}^2 + n_y^{-1} \sigma_{gsy}^2$  and  $\tilde{\sigma}_e^2 = n_y^{-1} \sigma_e^2$  in place of  $\sigma_{gs}^2$  and  $\sigma_e^2$ , respectively.

It is also worth stressing that  $n_y$  is typically smaller than  $n_s$ , resulting from the fact that it is more difficult to add more years than to add more sites. For the same reason, there is sometimes a suggestion to replace years by sites. Such suggestions are unhelpful, however, when the genotype-year variance is non-negligible. This is because the only way to reduce its effect on  $VD$  is to increase  $n_y$ , as is apparent from Eqn (61).

*Example 16:* Post-registration variety trials conducted in the German federal state of Mecklenburg-Vorpommern typically comprise 25–30 varieties and fall into two broad categories, depending on the crop: (i) single-factor experiments with variety (genotype) as the only treatment factor, and (ii) two-factor experiments with management intensity as a second factor. Trials in category (i) usually have four replications, sometimes three, whereas trials in category (ii) typically have two replications, sometimes three. We here consider the design for a single-factor scenario. Variance components for yield ( $\times 10^{-1}$  t/ha) were obtained from regional wheat variety trials (Table 16). The trials were two-factorial but these were analysed by intensity level, amounting to a single-factor analysis as appropriate for our purpose here. These analyses are based on a two-stage approach, in which variety means and associated standard errors ( $SEM$ ) are computed in the first stage and then are submitted to a weighted mixed model analysis in which the inverse of the squared standard errors are used as weights (Piepho and Michel, 2000). This approach accounts for the heterogeneity of error variances between trials. Here, for the purpose of planning trials, we are assuming a constant variance as an approximation and that designs will be laid out in complete blocks.

Using the variance components in Table 16, Fig. 5 shows plots of  $SED$  for different values of  $n_y$  (1, 2, 3, 4, 5),  $n_s$  (1, 3, 5, 7, 9), and  $n_r$  (1, 2, 3, 4). The choice  $n_r = 2$  is a frequently used number of replications used in a series of variety trials, whereas  $n_r = 4$  is less common. The figures show two horizontal reference lines, one at  $SED = 4$  ( $\times 10^{-1}$  t/ha) and one at  $SED = 2$  ( $\times 10^{-1}$  t/ha). These two lines are based on the researchers' assessment and delineate the range of precision they require, first to make tentative recommendations and then making final recommendations at a later stage. The results show that the most crucial factor is the number of years  $n_y$ . With only one year of testing, at least nine sites are required to achieve  $SED = 4$ , the number of replications

**Table 16.** Variance component estimates (obtained by residual maximum likelihood) for yield ( $\times 10^{-2} \text{ t}^2/\text{ha}^2$ ) in post-registration wheat variety trials in Mecklenburg-Vorpommern (Landesforschungsanstalt für Landwirtschaft und Fischerei)

Variance component	Estimate ( $\times 10^{-2} \text{ t}^2/\text{ha}^2$ )
$\sigma_{gs}^2$	2.36
$\sigma_{gy}^2$	6.27
$\sigma_{gsy}^2$	9.21
$\sigma_e^2$ <sup>a</sup>	13.78

<sup>a</sup>To obtain an estimate of the error variance  $\sigma_e^2$ , we computed the average SEM, squared this and multiplied by 2.5, the average number of replications across all trials.

having relatively little impact. By comparison, with two years of testing, about the same precision of  $SED = 4$ , considered appropriate for preliminary recommendations, can already be achieved with three or four sites. Sufficient precision for final recommendations at  $SED = 2$  is only achieved after four to five years with more than five sites.

We have assumed here that the error variance is constant and does not change with the number of replications. In this regard, it is worth adding that the current practice in the post-registration trials conducted by the federal state is to use row-column designs, where complete blocks can be formed by groups of rows or columns, or by single rows or columns (Piepho *et al.*, 2021). A standard analysis according to a randomized complete block design can always be used if the additional row or column blocking proves ineffective. A model selection routine has been implemented that checks this via information criteria. In addition, spatial covariance structures are fitted, and this add-on, as well as the incomplete column blocks often improve precision. This is particularly true of larger trials showing marked irregular spatial heterogeneity throughout the field, and the chances of an improved fit with such more complex models increase with the number of replications. Even though this has not yet been comprehensively evaluated, in practice, there is a tendency for the  $SED$  of individual trials to drop with increased  $n_r$  at a somewhat higher rate than expected from the simple Eqn (23). Thus, our usage of the simple Eqn (61) constitutes an approximation. We also stress the importance of having  $n_r > 1$  so that individual trials can be analysed in their own right, including critical scrutiny of outliers, and trials can be weighted by their precision in an analysis across environments (Piepho and Michel, 2000).

### Several sites nested within years

A third case occurs when the sites change completely each year, meaning that sites are nested within years. In this case, the two-way interaction variance  $\sigma_{gs}^2$  is confounded with  $\sigma_{gsy}^2$  and so cannot be separately estimated. The VD becomes:

$$VD = 2 \left( \frac{\sigma_{gy}^2}{n_y} + \frac{\sigma_{gs}^2 + \sigma_{gsy}^2}{n_s n_y} + \frac{\sigma_e^2}{n_s n_y n_r} \right) \quad (62)$$

where  $n_s$  is the number of sites used in each year, so the total number of sites used in  $n_y$  years is  $n_s n_y$ . The most important point here is that this VD will always be smaller than or equal to that in (61) for the same values of  $n_r$ ,  $n_s$  and  $n_y$ , and the same values of the variance components, showing that a change of sites each year is generally desirable in terms of efficiency.

The advantage is most pronounced when  $\sigma_{gs}^2$  is large relative to the other variance components. Again, the term in (62) involving  $\sigma_{gy}^2$  is a constant and it is sufficient to consider the same optimization as before this time using  $\tilde{\sigma}_{gs}^2 = n_y^{-1}(\sigma_{gs}^2 + \sigma_{gsy}^2)$  and  $\tilde{\sigma}_e^2 = n_y^{-1}\sigma_e^2$  in place of  $\sigma_{gs}^2$  and  $\sigma_e^2$ , respectively.

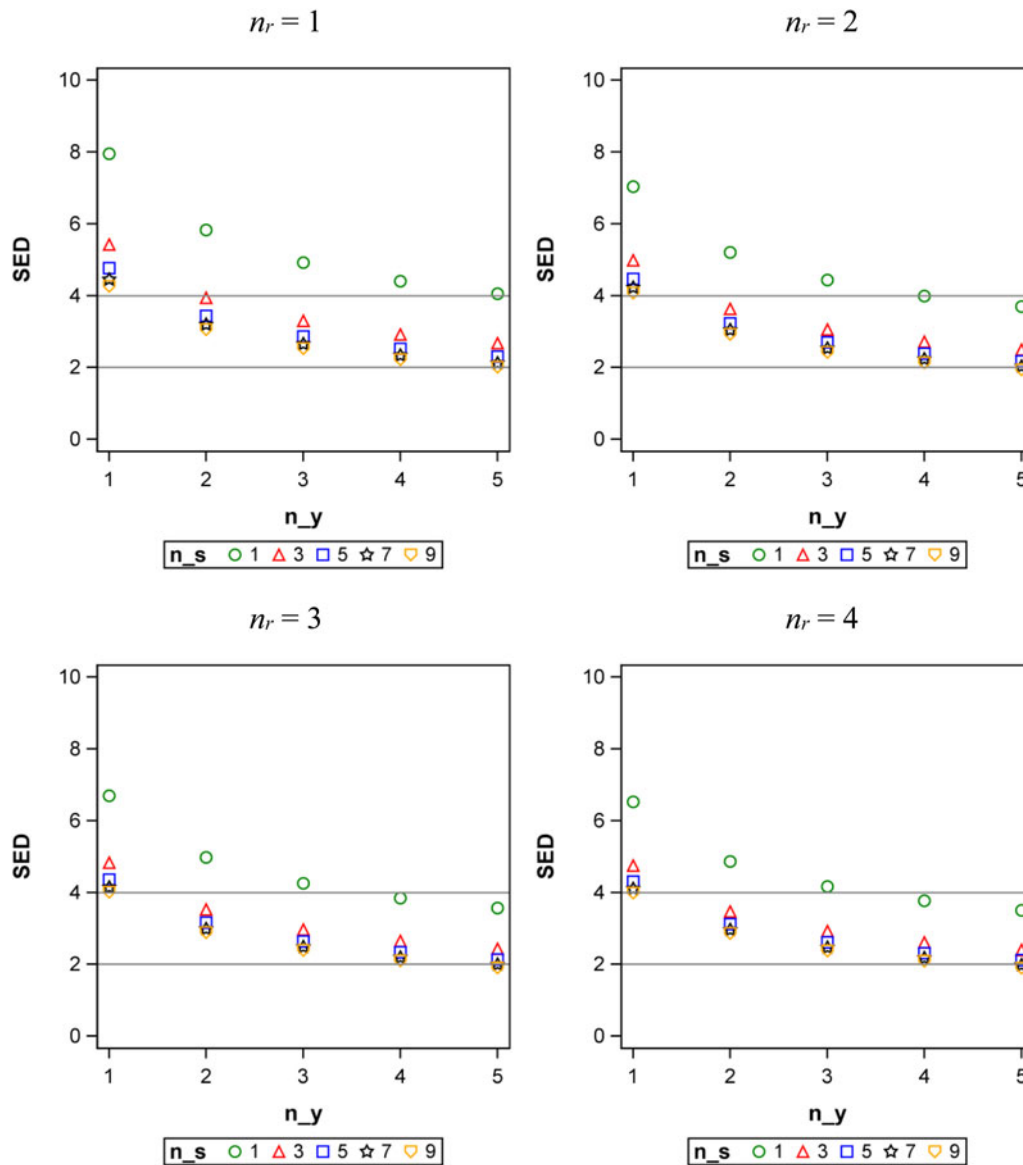
*Example 16 (cont'd):* Using the variance components in Table 16 with  $n_r = 2$ ,  $n_s = 7$ , and  $n_y = 5$ , we find  $SED = 2.03 (\times 10^{-1} \text{ t/ha})$  for the crossed design (Eqn 61) and  $SED = 1.89 (\times 10^{-1} \text{ t/ha})$  for the nested design, demonstrating the advantage of the latter. Whereas Eqns (61) and (62) represent the two competing design options of sites crossed with, or nested within years in pure form, in practice there is often a mix between both, with some sites used in more than one year and others used only in single years within a series of trials. This is especially true when transitioning from a design where all sites are used in all years, to a trialling system with new sites being added each year and the number of sites increasing towards later years, as is the current practice with post-registration trials in Mecklenburg-Vorpommern. Explicit equations evaluating the design efficiency are more complex (Laidig and Utz, 1992), and a computer-based approach as discussed in the next section is more convenient.

### Using a linear model package to compute precision and power

All examples considered so far were tackled by easy-to-compute (portable) equations. In this section, we briefly review generalizations of the methods considered so far that allow determining sample size, precision and power for any design based on a LM even when simple explicit equations are not available, e.g. when using designs with incomplete blocks rather than complete blocks. As this more general approach is not usually amenable to manual computation, it is certainly not as portable as the approaches reviewed so far and so will not be considered in detail here. However, the underlying principles are the same, and with a good computer and statistical package computation is straightforward. Detailed illustrations will be found in the cited references and in a companion paper, which is under preparation.

### Numerical approximations

In all cases considered so far, simple explicit equations were available for the  $SED$  as a function of  $n$ , which was the basis for obtaining explicit equations for the necessary sample size. Quite often, however, the model or the structure of the design may mean that such convenient expressions for sample size are not available. In these cases, one can always resort to the matrix formulation of the LM at hand and corresponding matrix expressions for  $SED$ . We will not consider these matrix expressions here and refer the interested readers to pertinent textbooks and papers (McLean *et al.*, 1991; Searle *et al.*, 1992). Suffice it to say that for LM, having the residual error term as the only random effect, exact calculations are always possible, whereas with LMM some kind of approximation is usually needed (Kenward and Roger, 1997). We skip these details because the computation can be left to a good mixed model package that has a facility to fix variance components at pre-specified values. The main task for the researcher then is to set up a dataset for the contemplated design and fit the model corresponding to the design at hand to a dummy response variable. The structure of the dataset must be the exact same as would be used for analysis of data obtained in the planned experiment. Then for pre-specified variance



**Fig. 5.** Colour online. Standard error of a difference ( $SED$ ; in  $10^{-1}$  t/ha) as per (61) with  $SED = \sqrt{VD}$  for variance components  $\sigma_{gs}^2$ ,  $\sigma_{gy}^2$ ,  $\sigma_{gsy}^2$ , and  $\sigma_e^2$  as shown in Table 16 for different values of  $n_y$  (1, 2, 3, 4, 5) = number of years,  $n_s$  (1, 3, 5, 7, 9) = number of sites, and  $n_r$  (1, 2, 3, 4) = number of replicates. Reference lines on ordinate at  $SED = 4$  ( $\times 10^{-1}$  t/ha) (desirable for early assessment) and  $SED = 2$  ( $\times 10^{-1}$  t/ha) (desirable for recommendations).

component values the package will compute the  $SED$  for the comparisons or contrasts of interest (Stroup, 2002; Casler, 2015). The main point to appreciate here is that the  $SED$  really only depends on the design and the variance components, but not on the actual data. Thus, we can set the dummy response in our dataset to any numerical value. This is what makes the approach feasible at the design stage. In the special cases considered in the preceding sections, this package-based approach will return exactly the same  $SED$  as the explicit equations given in this paper. So we are still doing the same thing as before but just with some more flexibility as regards the model, the structure of the data and the design.

The only slight inconvenience of this whole approach is that it does not give us the optimal sample size in one go. Instead, it returns the  $SED$  for a given design, which among other choices entails a specific choice of sample size. So what we need to do is try designs of increasing size until we find one with approximately the desired  $SED$ .

We may also consider planning sample size with a significance test in mind. Here, we need to compute the power,  $1 - \beta$ , for designs with increasing sample size until we achieve the desired power (Stroup, 2002). As before, we will focus on the comparison of two means ( $\mu_1$  and  $\mu_2$ ), which will always be done by a  $t$ -test. Hence, the exact power calculation involves the central and non-central  $t$ -distributions of the test statistic under the null and alternative hypotheses, whereas the approximate calculation uses the standard normal distribution to approximate both distributions. Here, we will first consider the normal approximation and then look at the exact solution. The test statistic is:

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{SED} \quad (63)$$

where  $\hat{\mu}_1$  and  $\hat{\mu}_2$  are the two mean estimates under the fitted LMM. Equation (63) assumes that the  $SED$  is known, in which

case  $t$  has a normal distribution. When  $SED$  is replaced by an estimate, this turns into a  $t$ -distribution. Note that this reduces to (30) for the simple  $t$ -test, which involves arithmetic means of the data in each group, whereas the model-based mean estimates here can take different forms and do not usually have such simple algebraic expressions. Again, we can rely entirely on the package to do the computations for us.

Under the null hypothesis  $H_0$ ,  $t$  in (63) has an approximate standard normal distribution with unit variance and mean zero. Hence, we will reject the null hypothesis when  $|t| > z_{1-\alpha/2}$  as usual. Under the alternative hypothesis  $H_A$ , when the difference between the means equals  $\delta = \mu_1 - \mu_2$ ,  $t$  has an approximate normal distribution with unit variance and mean  $\delta/SED$ . All we need to do then is work out the probability under  $H_A$  that  $|t| > z_{1-\alpha/2}$ , i.e. either  $t < -z_{1-\alpha/2}$  or  $t > z_{1-\alpha/2}$ . This probability is the power  $1 - \beta$  for the given design. For this calculation, we just need a function for the cumulative distribution function of the standard normal, denoted as  $\Phi(\cdot)$ . Then the two required rejection probabilities are:

$$P(t < -z_{1-\alpha/2} | H_A) = \Phi(-z_{1-\alpha/2} - \delta/SED) \text{ and} \\ P(t > z_{1-\alpha/2} | H_A) = 1 - \Phi(z_{1-\alpha/2} - \delta/SED)$$

Adding up, the power is:

$$\text{Power} = \Phi(-z_{1-\alpha/2} - \delta/SED) + 1 - \Phi(z_{1-\alpha/2} - \delta/SED) \quad (64)$$

This approximate calculation is entirely general and works for any LMM. All we need to do is get the  $SED$  from our package as described above and plug it into (64). The more exact calculation replaces  $z_{1-\alpha/2}$  with  $t_{1-\alpha/2}$ , the critical value of a central  $t$ -distribution with appropriate residual degrees of freedom (Welham *et al.*, 2015, p. 248). Likewise, the normal distribution under  $H_A$  is replaced by the noncentral  $t$ -distribution with appropriate degrees of freedom and noncentrality parameter  $\delta/SED$ . The appropriate degrees of freedom will be the residual degrees of freedom for LM, in which case the power calculation is indeed exact, whereas for LMM the degrees of freedom often have to be approximated (Kenward and Roger, 1997), meaning that the whole power calculation will still be approximate. A more accurate power calculation, properly dealing also with the degrees-of-freedom issue, can be conveniently obtained by simulation, and this will be discussed in the next sub-section. Here, we restrict attention to the normal approximation, which is in line with all the portable equations given so far, and which will be sufficient for most practical purposes, unless the residual degrees of freedom are very small.

The general approach just described may also be turned into a generalization of the simple rule given in Eqn (38), where we pre-specify the desired power. Even the simple 1-2-3 rule of thumb can still be applied, i.e. make sure  $SED$  is no greater than  $|\delta|/3$  when desiring a power of 85% at  $\alpha = 5\%$ . Similarly, if the focus is on the estimation of effect size, make sure the  $SED$  is no greater than  $\tau_\delta/2$  or  $EHW/2$  (Eqn 37). These rules are portable indeed, as they work for any LMM. Also, they obviate the need to explicitly calculate the power but only require evaluating the  $SED$  for candidate designs.

### Simulating power and precision

The analytical approach described in the previous sub-section often works fine for LMM when only the  $SED$  is required. By

contrast,  $EHW$  and power calculations require appropriate denominator degrees of freedom, and these may need to be approximated depending on the design and model (Kenward and Roger, 1997). Generally, in small samples, the plug-in approach of the previous sub-section may not be sufficiently accurate. In such settings, a simulation approach may be preferred (Gbur *et al.*, 2012; Green and MacLeod, 2015). Also, generalized LMMs involve an extra layer of approximation because the likelihood needs to be approximated, and all inference (significance tests, confidence intervals) is only approximate as well (Wolfinger and O'Connell, 1993; Bolker *et al.*, 2009; Stroup, 2013). Thus, simulation is often the method of choice. An important added benefit of simulation is that it allows assessing the validity of nominal Type I error rates for significance tests and confidence intervals, which may be worth checking in their own right, especially when one or several random effects are associated with only limited degrees of freedom.

The simulation approach works similar to the analytical approach in that a data frame with the same structure as the contemplated design is generated. Next, a large number of datasets with this structure are simulated according to the same model that will be used for analysing the data. The intended analysis is then applied to all simulated datasets and measures of precision ( $SED$ ,  $EHW$ ) and power (significance) are computed. These measures (means, quantiles, etc.) may then be summarized across all simulated datasets. For assessing the  $SED$ , we may compute the root mean squared deviation of the estimated and true difference. Thus, in each simulation run (i.e. for each simulated dataset), we compute  $(\delta - \hat{\delta})^2$  and then these squared deviations are averaged across simulation runs. In general, the mean squared deviation would also comprise bias, but from the properties of generalized least squares estimation, we may assume that the estimators are unbiased for LMM (Kackar and Harville, 1981). Thus, the root mean squared deviation assesses the  $SED$ . This estimate of  $SED$  can be compared to the model-based estimates of  $SED$ . Moreover, for each simulation run, we may obtain the  $HW$  of a confidence interval and average this over the simulation runs. Power may be assessed by simply computing the proportion of simulation runs in which the simulated test rejected the null hypothesis at the nominal significance level  $\alpha$ . Since simulations are based on independent datasets, a confidence interval for the power of the test can be computed based on a binomial distribution.

### Discussion

Sample size, or the number of replications, concerns only one of the three basic principles of experimental design, otherwise known as Fisher's 3 R's (Preece, 1990), i.e., replication, randomization and ('r') blocking (local control). Thus, having determined a suitable sample size does not settle all design issues. In fact, good blocking may reduce the error variance  $\sigma^2$ , thus allowing a smaller sample size to be used than under complete randomization. When the design involves incomplete blocks, specialized methods can be used to generate good designs for an increasing series of replication numbers (e.g. Edmondson, 2020; Piepho *et al.*, 2021) and then the methods in the previous section can be used to evaluate the precision and power.

Depending on the textbook one picks up on sample size, one may end up with different formulae. This can be confusing both to the consulting statistician and the research scientist. One potential source of confusion is that equations are often presented

in isolation from alternative approaches. This tutorial has provided an overview of different approaches and elucidated the close interconnections between them. The key quantity that connects all procedures is the standard error of an effect size estimate. Most of the time, the effect size of interest will be a difference of means, possible on a transformed scale, in which case the focus is on the *SED*. The portable procedures we have reviewed provide approximate estimates of sample size, and we believe that this is usually sufficient. Our review does not claim any degree of completeness. While we have covered basic procedures for one or two means quite comprehensively, our treatment of other scenarios was necessarily selective and reflects our own backgrounds. The general procedures reviewed in the previous section can help practitioners to think effectively and decide about sample size in their own research or in the statistical consulting they provide, even if the problem at hand initially does not seem straightforward or covered by any canned solutions. A detailed exemplification of these computer-based approaches will be provided in a companion paper.

In the previous section, we also focused on the *SED* and pairwise comparisons, as in the sections before. The proposal of Stroup (2002) is focused on significance testing using the *F*-test. This is equivalent to a *t*-test when the *F*-statistic has a single degree of freedom for the linear hypothesis being tested, as is the case with a pairwise comparison. In keeping with our focus on pairwise comparisons throughout this paper, we have fully relied on that equivalence. The examples also showed that the same approach can easily be used to assess *SED*, and that applying the 1-2-3 rule to the *SED* of contending designs provides a convenient and simple approach to designing a trial.

Our treatment of significance tests has assumed that the test will be two-sided, i.e., there is a point null hypothesis such as  $H_0: \delta = 0$  and the alternative covers both sides of  $H_0$ , i.e.  $\delta > 0$  and  $\delta < 0$ . It is emphasized here, however, that in some applications it may be more appropriate to consider one-sided null hypotheses of the form  $H_0: \delta \leq 0$ , with corresponding alternative hypothesis  $H_A: \delta > 0$ . For example, in experiments with animals, it may only be of interest to demonstrate that a new treatment is superior to a control. In the same vein, one may also construct one-sided instead of two-sided confidence intervals. The practical consequence of a one-sided significance test or confidence interval is that for the same level  $\alpha$  the required sample size is lower than for the corresponding two-sided procedure. This has prompted some ethics committees and other official bodies deciding on the approval of experiments with animals to make the one-sided procedure the default assumption, requiring the experimenter to provide convincing arguments in case a two-sided procedure is proposed. Our view is that in most applications both sides of departure from a point null hypothesis are relevant, even though the consequences may depend on the side. This is why we have only considered two-sided tests or intervals. If one decides that the procedure needs to be one-sided, then the  $1 - \alpha/2$  quantile of the standard normal distribution can be replaced by the  $1 - \alpha$  quantile in the relevant equations given in this paper, with an associated drop in the necessary sample size. For details see, e.g., Rasch *et al.* (2011).

## Conclusion

Here is our portable take-home message in case estimated treatment means are approximately normal, the 1-2-3 rule: Design your sample size in terms of a suitable requirement for the

value of the standard error of the effect you are targeting. This will usually be a *SED* because most experiments are comparative. According to the 1-2-3 rule, there are three options for setting the *SED*:

- (1) You can define the required precision directly in terms of a targeted value of *SED* itself.
- (2) If you target a specific allowable deviation  $\tau_\delta$  of an estimator or *EHW* of a confidence interval, set  $SED = \tau_\delta/2$  or  $SED = EHW/2$ . Conversely, based on a given value for the *SED*, it may be stated that at  $\alpha = 5\%$  the *EHW* of a 95% confidence interval is  $2 \times SED$ .
- (3) If you are considering a significance test with the smallest relevant effect size  $\delta$ , set  $SED = \delta/3$ . Conversely, based on a given value for the *SED*, it may be stated that the smallest relevant difference that can be detected with a power of 85% is  $3 \times SED$ .

**Acknowledgements.** We thank Mario F. D'Antuono (Perth, Australia) for very helpful comments on an earlier draft of this paper. Thanks are also due to Georg Petschenka (University of Hohenheim, Stuttgart, Germany) for his help in putting together Example 10, and to Hans-Georg Schön (Hochschule Osnabrück, Germany) for his help with Example 4.

**Author contributions.** HPP conceived and designed the study. DG and JH implemented and checked all calculations using statistical packages and helped editing the paper. AB, MG, SK, FL, VM, IP, JES, KT and DW helped developing the examples and editing the text for these.

**Financial support.** This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

**Conflicts of interest.** The authors declare there are no conflicts of interest.

**Ethical standards.** Not applicable. Where data on vertebrates was used, this was published data.

## References

- Agresti A and Coull BA (1998) Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician* **52**, 119–126.
- Atlin GN, Baker RJ, McRae KB and Lu X (2000) Selection response in subdivided target regions. *Crop Science* **40**, 7–13.
- Bailey RA (2009) *Design of Comparative Experiments*. Cambridge: Cambridge University Press.
- Beal SL (1989) Sample size determination for confidence intervals on the population means and on the difference between two population means. *Biometrics* **45**, 969–977.
- Benz B, Kaess M, Wattendorf-Moser F and Hubert S (2020) Auswirkungen einer Stundenweide von Milchkühen auf Verhalten und Leistung in einem Praxisbetrieb. *Züchtungskunde* **92**, 159–171.
- Bleiholder H, van den Boom T, Langelüdecke P and Stauss R (1989) Einheitliche Codierung der phänologischen Stadien bei Kultur- und Schadpflanzen. *Gesunde Pflanzen* **41**, 381–384.
- Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Henry M, Stevens H and White JSS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution* **24**, 127–135.
- Box GEP and Draper NR (2007) *Response Surfaces, Mixtures, and Ridge Analysis*. New York: Wiley.
- Bretz F, Hothorn T and Westfall P (2011) *Multiple Comparisons Using R*. Boca Raton: CRC Press.
- Büchse A and Piepho HP (2006) Messen, Schätzen, Bonitieren: Konsequenzen des Skalenniveaus für die Auswertung und Interpretation von Versuchsergebnissen, Tagungsband der DLG-Technikertagung. 37. Arbeitstagung AG Feldversuche in Soest am 30 und 31. Januar 2006, S. 82–102.

- Casagrande JT, Pike MC and Smith PG** (1978) An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics* **34**, 483–486.
- Casler MD** (2015) Fundamentals of experimental design: guidelines for designing successful experiments. *Agronomy Journal* **107**, 692–705.
- Chen Z and Chen X** (2014) Exact calculation of minimum sample size for estimating a Poisson parameter. *Communications in Statistics - Theory and Methods* **45**, 4692–4715.
- Chen CC and Tyler CW** (1999) Accurate approximation to the extreme order statistics of Gaussian samples. *Communications in Statistics - Simulation and Computation* **28**, 177–188.
- Cochran WG and Cox GM** (1957) *Experimental Designs*, 2nd Edn. New York: Wiley.
- Cohen J** (1977) *Statistical Power Analysis for the Behavioral Sciences*, revised Edn. New York: Academic Press.
- Cohen J** (1992) A power primer. *Psychological Bulletin* **112**, 155–159.
- Davies GM and Gray A** (2015) Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). *Ecology and Evolution* **5**, 5295–5304.
- Dean A and Voss D** (1999) *Design and Analysis of Experiments*. Berlin: Springer.
- Dette H** (1995) Optimal designs for identifying the degree of a polynomial regression. *Annals of Statistics* **23**, 1248–1267.
- Dufner J, Jensen U and Schumacher E** (1992) *Statistik mit SAS*, 2nd Edn. Auflage, Stuttgart: Teubner.
- Duoma JC and Weedon JT** (2018) Analysing continuous proportions in ecology and evolution: a practical introduction to beta and Dirichlet regression. *Methods in Ecology and Evolution* **10**, 1412–1430.
- Edmondson R** (2020) Multi-level blocked designs for comparative experiments. *Journal of Agricultural Biological and Environmental Statistics* **25**, 500–522.
- Flleiss JL** (1981) *Statistical Methods for Rates and Proportions*, 2nd Edn. New York: Wiley.
- Gbur EE, Stroup WW, McCarter KS, Durham S, Young LJ, Christman M, West M and Kramer M** (2012) *Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences*. Madison: ASA-CSSA-SSSA.
- Green P and MacLeod CJ** (2015) SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution* **7**, 493–498.
- Hau FC, Campbell CL and Beute MK** (1982) Inoculum distribution and sampling methods for *Cylindrocladium crotalariae* in a peanut field. *Plant Disease* **66**, 568–571.
- Hocking PM, Mayne RK, Else RW, French NA and Gatcliffe J** (2008) Standard European footpad dermatitis scoring system for use in turkey processing. *World Poultry Science Journal* **64**, 323–328.
- Hogg RV, McKean JW and Craig AT** (2019) *Introduction to Mathematical Statistics*, Eighth Edn. Boston: Pearson.
- Horn M and Vollandt R** (1995) *Multiple Tests und Auswahlverfahren*. Stuttgart und Jena: Gustav Fischer Verlag.
- Hsu J** (1996) *Multiple Comparisons. Theory and Methods*. London: Chapman & Hall.
- Huang Y, Gilmour SG, Mylona K and Goos P** (2020) Optimal design of experiments for hybrid nonlinear models, with applications to extended Michaelis–Menten kinetics. *Journal of Agricultural Biological and Environmental Statistics* **26**, 601–625.
- Hurlbert SH** (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* **54**, 187–211.
- Kackar AN and Harville DA** (1981) Unbiasedness of two-stage estimation and precision procedures for mixed linear models. *Communications in Statistics A* **10**, 1249–1261.
- Kenward MG and Roger JH** (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983–997.
- Kish L** (1965) *Survey Sampling*. New York: Wiley.
- Kozak M and Piepho HP** (2018) What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. *Journal of Agronomy and Crop Science* **204**, 86–98.
- Laidig F and Utz HF** (1992) Combining results from nonorthogonal trial series over years. *Biuletyn Oceny Odmian* **24–25**, 55–68.
- Lenth RV** (2001) Some practical guidelines for effective sample size determination. *The American Statistician* **55**, 187–193.
- Malik WA and Piepho HP** (2016) On generalized exponential transformations for proportions. *Communications in Statistics - Theory and Methods* **45**, 5857–5870.
- McCullagh P and Nelder JA** (1989) *Generalized Linear Models*, 2nd Edn. London: Chapman & Hall.
- McLean RA, Sanders WL and Stroup WW** (1991) A unified approach to mixed linear models. *The American Statistician* **45**, 54–64.
- Mead R** (1988) *The Design of Experiments. Statistical Principles for Practical Application*. Cambridge: Cambridge University Press.
- Mead R, Gilmour SG and Mead A** (2012) *Statistical Principles for the Design of Experiments: Applications to Real Experiments*. Cambridge: Cambridge University Press.
- Montgomery DC and Runger GC** (2011) *Applied Statistics and Probability for Engineers*, 5th Edn. New York: John Wiley & Sons, Inc.
- Nyquist WE** (1991) Estimation of heritability and prediction of selection response in plant populations. *Critical Reviews in Plant Sciences* **10**, 235–322.
- Paulson E and Wallis WA** (1947) Planning and analysing experiments for comparing two percentages. In Eisenhart C, Hastay MW and Wallis WA (eds), *Selected Techniques of Statistical Analysis*. New York and London: McGraw-Hill (Chapter 7), pp. 247–265.
- Piepho HP** (1997) Analysis of a randomized complete block design with unequal subclass numbers. *Agronomy Journal* **89**, 718–723.
- Piepho HP** (2003) The folded exponential transformation for proportions. *The Statistician* **52**, 575–589.
- Piepho HP and Edmondson RN** (2018) A tutorial on the statistical analysis of factorial experiments with qualitative and quantitative treatment factor levels. *Journal of Agronomy and Crop Science* **204**, 429–455.
- Piepho HP and Michel V** (2000) Überlegungen zur regionalen Auswertung von Landessortenversuchen. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* **31**, 123–136.
- Piepho HP and Möhring J** (2007) Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* **177**, 1881–1888.
- Piepho HP, Williams ER and Michel V** (2021) Generating row-column field experimental designs with good neighbour balance and even distribution of treatment replications. *Journal of Agronomy and Crop Science* **207**, 745–753.
- Preece DA** (1990) R.A. Fisher and experimental design: a review. *Biometrics* **46**, 925–935.
- Rasch D, Herrendörfer G, Bock J, Victor N and Guiard V** (1998) *Verfahrensbibliothek: Versuchsplanung und -Auswertung*, Band 2. München: Oldenbourg.
- Rasch D, Pilz J, Gebhardt A and Verdooren RL** (2011) *Optimal Experimental Design with R*. Boca Raton: Chapman and Hall.
- Ross RH and Knodt CB** (1948) The effect of supplemental vitamin A upon growth, blood plasma carotene, vitamin A, inorganic calcium, and phosphorus of Holstein heifers. *Journal of Dairy Science* **31**, 1062–1067.
- Searle SR, Casella G and McCulloch CE** (1992) *Variance Components*. New York: Wiley.
- Shan G** (2016) Exact sample size determination for the ratio of two incidence rates under the Poisson distribution. *Computational Statistics* **31**, 1633–1644.
- Snedecor GW and Cochran WG** (1989) *Statistical Methods*, 8th Edn. Ames: Blackwell Publishing.
- Stroup WW** (2002) Power analysis based on spatial effects mixed models: a tool for comparing design and analysis strategies in the presence of spatial variability. *Journal of Agricultural Biological and Environmental Statistics* **7**, 491–511.
- Stroup WW** (2013) *Generalized Linear Mixed Models*. Boca Raton: CRC Press.
- Talbot M** (1984) Yield variability of crop varieties in the U.K. *Journal of Agricultural Science, Camb.* **102**, 315–321.
- Thompson SK** (2002) *Sampling*, 2nd Edn. New York: Wiley.
- Toppel K, Spindler B, Kaufmann F, Gaulty M, Kemper N and Andersson R** (2019) Foot pad health as part of on-farm-monitoring in Turkey flocks. *Frontiers in Veterinary Science* **6**, 25.

- Trommer R** (1986) *Anwendung mathematisch-statistischer Verfahren bei der Überwachung von Schaderregern der landwirtschaftlichen Produktion* (Dissertation B). Akademie der Landwirtschaftswissenschaften der DDR.
- van Belle G** (2008) *Statistical Rules of Thumb*, 2nd Edn. New York: Wiley.
- Warton D and Hui FKC** (2011) The arcsin is asinine: the analysis of proportions in ecology. *Ecology* **92**, 3–10.
- Welham SJ, Gezan SA, Clark SJ and Mead A** (2015) *Statistical Methods in Biology*. Boca Raton: CRC Press.
- Westfall PH, Tobias RD, Rom D, Wolfinger RD and Hochberg Y** (1999) *Multiple Comparisons and Multiple Tests*. Cary: SAS Institute.
- Wheeler RE** (1974) Portable power. *Technometrics* **16**, 193–201.
- Wheeler RE** (1975) The validity of portable power. *Technometrics* **17**, 177–179.
- Wolfinger R and O'Connell M** (1993) Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233–243.
- Yan W** (2021) Estimation of the optimal number of replicates in crop variety trials. *Frontiers in Plant Science* **11**, 590762.
- Yates F and Cochran WG** (1938) The analysis of groups of experiments. *Journal of Agricultural Science, Camb.* **28**, 556–580.
- Young LJ, Campbell NL and Capuano GA** (1999) Analysis of overdispersed count data from single-factor experiments: a comparative study. *Journal of Agricultural, Biological and Environmental Statistics* **4**, 258–275.
- Zuber U, Winzeler H, Messmer MM, Keller M, Keller B, Schmid JE and Stamp P** (1999) Morphological traits associated with lodging resistance of spring wheat (*Triticum aestivum* L.). *Journal of Agronomy and Crop Science* **182**, 17–24.

## Appendix

*Example 11:* We need to point out that on the surface we have determined sample size here based on the binary distribution for the transformed response of the individual animals, using  $\sigma^2 = 1/4$ . This is the limiting case of a binomial distribution with  $m = 1$ . However, the approximate variance for the angular-transformed binomial in (40), which is based on a Taylor-series expansion, requires  $m \gg 1$  (McCullagh and Nelder, 1989, p. 137). To justify our approach, consider the following argument (also see Paulson and Wallis, 1947; cited in Cochran and Cox, 1957, p. 27). The count of level A in a group may be regarded as a single binomial count  $c$  for sample size  $m$ . In this view,  $n = 1$  for both groups, and our task is to find the optimal binomial sample size  $m$ . So we may set  $n = 1$  and  $\sigma^2 = 1/(4m)$  in (31) with  $\delta$  as given in (42). Solving this for  $m$  yields

$$m \approx \frac{2(1/4)(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2}$$

justifying our use of Eqn (31) even for a binary variable.