Biostats VHM 801/802 Courses
Fall 2007/2021, Atlantic Veterinary College, PEI
Henrik Stryhn

Notes on statistical sample size calculations

These notes are intended to supplement, *not replace*, material in statistics textbooks ([1],[4],[11],[13]) about sample size determination used in the biostatistics courses at the AVC. Textbooks specialized on sample size determination exist as well (e.g. [14]), and when the sample size question is discussed across many designs the coverage necessarily becomes rather specific and technical. All these details may overshadow the basic ideas, which are the same across all designs, and our purpose here is to provide a reasonably non-technical overview of those ideas; specifically, to

1) to review ways of and arguments for choosing sample size for new studies,

2) to show how non-standard sample size questions may often be rephrased in terms of simple questions in designs with accessible answers (formulas or software).

These purposes are separated into Sections 1 and 2. The statistical prerequisites for the notes are basic notions of confidence intervals and test statistics[1]. However, for Section 2 to be of interest, acquaintance with the partly complex designs and models reviewed there by example, probably corresponding to the level of a second biostatistics course (e.g., VHM 802), would be required.

1. Introduction to sample size calculations

The primary purpose of sample size calculations is for planning of experiments or studies where statistical data analysis is contemplated: how many "subjects" or "experimental/observational units" are needed to achieve the study objectives? Many different considerations play into this: cost and study logistics to name a few. Also the statistical considerations can differ substantially between study types and objectives, but the basic premise behind them is an attempt to ensure a reasonable (desired) precision for the study results. Obviously, too few subjects may not suffice to override the random variation in the data, and nobody wants to waste resources having too many subjects, so this is about finding a suitable balance. Although statistical considerations can help, one should realize from the onset that the final answer will typically be quite subjective rather than a definite answer from some calculation. This is largely due to the fact that all statistical calculations will be based on information about the results expected from the study (before it is carried out).

The specific meaning of "subjects" depends on the actual study, in particular its design, and in complex designs there may be more than one type of "experimental/observational unit" (e.g., the two-level designs discussed in Section 2). In simple designs, the meaning is often intuitive and straightforward; for example, in a study comparing treatments given to dogs to alleviate a disease, the experimental units are the dogs. Turning next to the question about the desired study "precision", there are two major ways of quantifying or specifying "precision", namely:

– the *length of a confidence interval* (of chosen coverage/confidence level, say 0.95) for a parameter of interest, or almost equivalently the standard error (the most common usage of "precision") of the corresponding parameter estimate — for situations where the primary interest is in the parameter and its estimate, and not necessarily in testing a hypothesis for the parameter;

[1] Although not stated explicitly, our discussion will be in the context of classical or "frequentist" statistical methodology; the ideas can certainly be applied to Bayesian analysis as well, but the key concepts confidence interval and statistical test have different meanings.

– the *power of a statistical test* to demonstrate a (significant) difference of a certain magnitude. Specifically, the power is the probability of the outcome of the statistical test to be significant at a prescribed significance level (e.g., 0.05) given that the null hypothesis being tested is false. This probability naturally depends on what is actually true — intuitively, how far the null hypothesis is off the true situation — which must therefore be specified to perform any power calculations. For example, if the null hypothesis is that a treatment and a control group have equal means, power calculations require specification of the true difference between the means in the two groups.

The second approach is probably the best known, but certainly not always the best way to go ([2]). Confidence interval methods are most natural for some study objectives, e.g. pertaining to reference intervals ([8]).

We will review the two resulting approaches for sample size calculations in the context of normal distribution models for the basic experimental designs: one, two and multiple samples. For non-normal data, the concepts are similar but in most cases all calculations require specialized software or computation. The fact that some explicit formulas exist for normal distribution models is helpful for understanding the principles of calculation, but generally speaking the use of suitable statistical software is much recommended over application of "simple" formula found scattered across the literature. Many of those formulas were developed when more precise exact calculations were not feasible (e.g. [3]), and nowadays it may be difficult to justify the use of an approximation when a more exact calculation is readily available (in software). Furthermore, such approximations tend to have their assumptions and limitations less well described, and well-developed software menus will make the user more aware of the actual choices made. Some pointers to statistical software are given in Note 1.9 below.

Setting 1.1: One-sample model/design

Consider a statistical model which assumes observations $y_1, \ldots, y_n$ to be independent and distributed according to $\mathrm{N}(\mu, \sigma^2)$. The primary interest is usually in the mean parameter μ, although both the mean and the standard deviation σ are assumed unknown. The $(1-\alpha)$ confidence interval for μ takes the well-known form:

$$\mu: \quad \bar{y} \pm t(1-\tfrac{\alpha}{2}, n-1)\, s/\sqrt{n}, \tag{1}$$

where $t(1-\frac{\alpha}{2}, n-1)$ is the $(1-\frac{\alpha}{2})$-percentile in a t-distribution with $n-1$ degrees of freedom. In order to achieve a confidence interval of a certain, predescribed length (say, not exceeding L), the obvious procedure is to solve the equation $L \geq 2 \times t(1-\frac{\alpha}{2}, n-1)s/\sqrt{n}$ with respect to n. Note that the "2" on the right hand side stems from the length of the interval being twice its margin of error $(t(1-\frac{\alpha}{2}, n-1)s/\sqrt{n})$. It gives of course the same result to solve for the margin of error (say $M = L/2$), where the equation becomes: $M \geq t(1-\frac{\alpha}{2}, n-1)s/\sqrt{n}$. To make the mechanics of solving these equations a bit easier, one usually approximates the t-distribution percentile by the corresponding value from a N(0,1) distribution. This approximation is valid only when n is large, and the calculation should be redone with a suitable t-distribution percentile if the resulting n is not so. Also, the value of s is unknown and must be substituted by an assumed/estimated/guessed value of the true standard deviation (for simplicity denoted also by σ), usually an estimate from previous data, a pilot study or a literature search. The (approximate) equation for n becomes

$$n \geq [z(1-\tfrac{\alpha}{2})\, \sigma/M]^2, \tag{2}$$

where $z(1-\frac{\alpha}{2})$ is the $(1-\frac{\alpha}{2})$-percentile in a the standard normal distribution, $\mathrm{N}(0,1)$. Choice of sample size based on power calculations is usually less natural for one-sample models because of lack

of interesting null hypotheses. We therefore defer both that discussion and the illustration of the formulas above to the next setting below with paired samples, which effectively are treated as a single sample.

Setting 1.2: Two paired (matched, dependent) samples

One standard way of analyzing paired samples is to create differences, say $d_i = y_{1i} - y_{2i}$, where y_{1i} and y_{2i} constitute the pair of values in the i^{th} pair, and to assume the differences form a single sample of independent observations from $\mathrm{N}(\mu, \sigma^2)$. The pairs may be two values from the same subject[2], and interest lies in comparing the two values in the pair because they have been treated differently, usually to give a treatment and a control measurement in each pair. The parameter of interest is μ, the difference in means between the first and second value in a pair. Sample size may be selected based on a desired margin of error (M) for a confidence interval for the mean difference, using formula (2). In the present context, n is the number of pairs, and σ is the standard deviation of *the differences* within pairs.

For power calculations, we need a null hypothesis of interest, most naturally $\mathrm{H}_0 : \mu = 0$, and we also need an alternative hypotheiss H_a, which may be taken as either one- or two-sided. Our test statistic for H_0 is $t = \bar{d}/(s_d/\sqrt{n})$, and the power for a test at significance level α and a two-sided H_a ($\mathrm{H}_a : \mu \neq 0$) given a true difference in population means of δ is the probability $\Pr_\delta(|t| \geq t(1-\frac{\alpha}{2}, n-1))$, where subscript δ refers to the true distribution of $d_1, \ldots, d_n$ being the $\mathrm{N}(\delta, \sigma^2)$. Unfortunately, under this assumption the distribution of t is somewhat complex — a so-called non-central t-distribution, requiring special tables or software to compute probabilities. Therefore, one typically resorts to statistical software for power calculations and for determining the required sample size to obtain a desired power.

> **Example 1** A small numerical example illustrates the concepts. Assume differences in blood pressure before and after an intervention to be normally distributed with an unknown mean and an unknown standard deviation, which is guessed to be 10 units (mm Hg). To achieve a 95% confidence interval for the mean difference with a margin of error of at most 3 units, requires at least $(1.96 \cdot 10/3)^2 = 42.7 \approx 43$ subjects. Redoing the calculation with $t(0.975, 42) = 2.018$ (from statistical software), we get: $(2.018 \cdot 10/3)^2 = 45.2 \approx 46$ subjects.
>
> With 46 subjects, the probability/power of detecting a true difference before and after the intervention of 3 units, using a 5% significance level and a two-sided alternative hypothesis, is 0.512, and to achieve a power of 0.80 would require a minimum sample size of 90 subjects (values obtained from statistical software). For this example, a one-sided alternative could be argued, because presumably the interest is only in the intervention leading to a lower blood pressure. With the one-sided alternative, the power for 46 subjects increases to 0.640, and the required sample size to achieve a power of 0.80 drops to 71. The reason for these changes is that with a one-sided alternative, it gets "easier" to reject the null hypothesis (specifically, the P-value gets halved when the observed effect is in the direction of H_a).

In the example, we used the same value (3) for the desired margin of error (M) of the confidence interval and the true difference of interest (δ) for the power calculation, even if these two quantities have different meanings. It is not easy to offer general advice on their choice. For the CI margin of error, one would need to consider when the CI was sufficiently small to make the estimate useful in practice; estimates with too large CIs are not particularly informative. Because we can use a CI for a hypothesis test, the value of M can also be interpreted as the value for the estimate that

[2] Examples are abundant: measurements of the left and right legs (arms, eyes, lungs) from human or animal patients, or measurements before/after an intervention.

falls exactly on the significance threshold (i.e., $P = 0.05$) for a test with a two-sided H_a. For the true difference (sometimes called "effect size") of interest, one would need to consider what a relevant (e.g., of biological or clinical interest) effect (in the example, change in blood pressure due to the intervention) would be. By using the same value for both calculations and noting that the required sample size from the power calculation is much larger, we get an illustration of the fact that an observed effect size of M by no means guarantees (with probability 0.8) that the true (population) effect size (δ) is of at least the same magnitude. To achieve the latter, we need a larger sample size.

Setting 1.3: General calculations for confidence intervals

The formula (2) is the simplest example of a general method for computing confidence intervals in normal distribution models[3]. For a parameter of interest (Par, in the above one-sample setting: μ), an estimate of the parameter (Est, above: $\bar{y}$), a standard error of the estimate (SE(Est), above: $s/\sqrt{n}$) and the degrees of freedom for the variance estimate in the model (df, above: $n-1$), a $(1-\alpha)$ confidence interval takes the form (using the notation from [4]),

$$Par: \quad Est \pm t(1-\tfrac{\alpha}{2}, \mathrm{df})\,\mathrm{SE}(Est). \tag{3}$$

Furthermore, the standard error SE(Est) takes the form $s \times \sqrt{constant}$, where s is the estimated (error) standard deviation in the model, and the constant depends only on the design and the estimate used. In the above one-sample setting, the constant equals $1/n$. Generally, the constant will involve the dimensions of the design. Therefore, *the method of inverting the confidence interval with respect to n applies generally to models of this type*, when the constant has been figured out.

It is important to remember that such sample size calculations (as well as those based on power) rely on the specific model assumptions, i.e. the normally distributed errors with the same variance. If those assumptions can be improved by transformation of the outcome, it may be preferable to carry out the sample size calculation on transformed scale.

Using the t-distribution in (1) and (3) ensures that the coverage of the confidence interval is correct when estimating the standard deviation from the data, and accounts for the variability arising from this estimation by using a larger reference distribution than $\mathrm{N}(0,1)$, which corresponds (unrealistically) to σ being known. These estimated standard deviations will differ from the true value and as well from our guessed value used in the calculation. The practical implication is that some observed confidence intervals will be larger than prescribed, and some will be smaller. If we want to achieve a certain confidence that the width of the confidence interval will not exceed the prescribed size, that needs to be accounted for *additionally* in the calculation ([9]). We would then introduce an additional desired confidence (probability) that the CI does not exceed the desired size, similar to the desired power (probability). Without this added feature, one might expect that probability to be around 0.5. Unfortunately, the resulting calculations get more complex and typically need to be carried out in statistical software (and not all software implementations have this option).

Setting 1.4: Two independent samples

Consider a statistical model with independent samples $(y_{11}, \ldots, y_{1n_1})$ and $(y_{21}, \ldots, y_{2n_2})$ from normal distributions $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$, respectively. A common simplification of the setup is to take the standard deviations in the two populations as equal (i.e., $\sigma_1 = \sigma_2 = \sigma$), and plan for equal sample sizes in the two samples (i.e., $n_1 = n_2 = n$). Then similar procedures as above apply to selecting the

[3] More precisely, it applies to confidence intervals for mean parameters in models with a single error term assumed to follow $\mathrm{N}(0, \sigma^2)$.

sample size based either on a desired confidence interval size for the mean difference, or on the power of a t-test for the null hypothesis $H_0: \mu_1 = \mu_2$. The relevant quantities for the confidence interval are therefore: $Par = \mu_1 - \mu_2$, $Est = \bar{y}_{1\cdot} - \bar{y}_{2\cdot}$ and $\mathrm{SE}(Est) = s \times \sqrt{2/n}$ (because $\mathrm{Var}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = \sigma^2 \times 2/n$). The resulting (approximate) formula for n, analogous to (2), is

$$n \geq 2\,[z(1-\tfrac{\alpha}{2})\,\sigma/M]^2. \tag{4}$$

Beware again the different interpretations of σ: here σ is the guessed, common value of the standard deviations in the two populations. Calculations of power or sample size to achieve a desired power require additional values of a true (non-zero) difference between population means and of the significance level (α).

Example 2 Also here we include a small numerical example, from ([2]). Blood-clotting times were measured for adult male rabbits that were randomized onto two groups that received different drugs. In a small pilot trial, the pooled within-group standard deviation was estimated at 0.72 (min). From [2], "We wish to test at the 0.05 level of significance with a 90% chance of detecting a true difference between population means as small as 0.5 min." The description is for a sample size calculation based on a desired power of 0.90 for an effect size of 0.5, and because there is no indication of a particular direction of interest, we would assume a two-sided alternative H_a. This leads to a sample size of 45 animals per group, if groups are equally sized (statistical software).

If we instead insert $M = 0.5$ into (4), we get: $n \approx 16$ and in a second calculation with $t(0.975, 30) = 2.042$, the revised bound becomes: $n \geq 17.3$, or $n = 18$. Note that we used 30 degrees of freedom in the t-distribution, corresponding to the pooled standard deviation in a two-sample design. As in the previous 1-sample example, this calculation does not include the variability in the estimated standard deviation.

Setting 1.5: Multiple independent samples ~ one-way ANOVA

For two samples, the estimate of interest and the associated test were obvious, but with more than two samples several options exist. A single parameter could be chosen as the basis for the calculation, e.g. a group mean, a difference between two group means, or more generally a contrast across several (possibly all) groups. The general approach from Setting 1.3 would then apply, once the necessary calculations for the standard error have been managed. Note that the relevant standard deviation is the within-group standard deviation from the ANOVA. Calculations based on power are typically also feasible, once it has been established how the group sizes enter into the corresponding one-parameter test. We give some examples of this approach in Section 2 for two-way ANOVAs.

If the interest is in the full comparison between groups, it is most natural to base the calculation on the power of the overall F-test. This leads to calculations in so-called non-central F-distributions, and is typically done with statistical software. The effect size can either be completely specified by all individual group means, or by the largest difference between two group means.

With multiple groups it is common to seek equal sample sizes for all groups, and thus a balanced design. This may however not be the most efficient choice. A contrast between two groups (i.e., a pairwise difference) will, if the two within-group standard deviations are the same, have the lowest standard error when the group sizes are equal. On the other hand, if the within-group standard deviations differ, the lowest standard error is obtained when the ratio between the group sizes equals the ratio of the standard deviations. For example, if $\sigma_1/\sigma_2 = 2$, then the optimal choice is to have $n_1/n_2 = 2$, that is, twice as many observations in group 1 than group 2. However, for a one-way ANOVA the within-group standard deviations are generally assumed to be equal.

Another situation where unequal sample sizes may be advantageous, arises in a one-way design with multiple treatment groups and a single control group. If the primary interest is in comparing each of the treatment with the control (and there is less interest in comparisons among the treatments), then it is more efficient to take the control group larger than each of the treatment groups. We can think of this intuitively as reflecting that the control group is used for every comparison, whereas each treatment group is used only once. The best choice is to take $n_c/n_t = \sqrt{g-1}$, where $n_c \sim$ control group, $n_t \sim$ any other treatment group, and $g =$ number of treatments (including control). For example, with $g = 5$ (i.e., four treatments and one control) we should take the control group of double size compared to each of the treatment groups.

Setting 1.6: Non-normal (in particular, binomial) data

The general principles for sample size calculation still apply, but exact calculations become more difficult and even in the simplest designs typically require statistical software. Approximate methods relying on normal approximations of estimates and corresponding z- and χ^2-statistics can be used, and considering the many other uncertainties involved in sample size calculations these may be fully satisfactory. Knowledge of the approximations available for the specific statistics is however needed. For example, many textbooks ([1],[11]) give a sample size formula for a single proportion based on the normal approximation confidence interval, from a similar reasoning as we used in Settings 1.1 and 1.3. As the normal approximation CI is only recommended/valid under certain conditions, knowledge of those conditions is helpful to assess the usefulness of the results of such a sample size calculation. For sample size calculations based on power, use of statistical software is generally recommended over any explicit formulas relying on normal approximations.

Many non-normal distributions, such as the binomial distribution, do not have a separate variance parameter, and the variability instead depends on the mean parameter (e.g., the probability parameter in the binomial distribution). Not having to specify a variability parameter in order to carry out a sample size calculation is certainly a relief, but it also makes the dependence of both confidence intervals and tests on the parameter of interest more complex, and it may also increase the risk of using the wrong distribution. For example, in the Poisson distribution for counts the variance is equal to the mean, but many practical examples of count data show larger dispersion than that, having motivated extensions of the Poisson distribution, such as to the negative binomial distribution. Clearly, if the variability in the resulting data is underestimated, one will tend to underestimate the required sample size as well.

Setting 1.7: Sampling to detect disease

A specialized objective for collection of binary data is to quantify the probability that a certain disease (or another binary condition) does not exist in the target population for the sampling. The sample size question for such studies should tell us how many individuals from the population to sample (and test negative) in order to become confident that the disease is absent in the population. Clearly, if we want to be totally sure that the probability of disease (or prevalence) equals zero, all individuals in the population should be tested (and test negative). Therefore, we will settle for (high) confidence that the prevalence is below a certain (small) threshold. In a classical (frequentist) statistical framework, this can be achieved by solving for the upper bound of a one-sided confidence interval based on observing a sample of only negative results. Because of the very simple outcome (only negatives), the "exact binomial" (Clopper-Pearson) confidence interval has as its upper bound the value p, so that $\alpha = \mathrm{P}(X = 0) = (1-p)^n$, where $(1-\alpha)$ is the desired confidence and $X \sim \mathrm{Bin}(n, p)$. Repeated sampling from a population only follows a binomial distribution exactly for an infinite population or

sampling with replacement, neither of which are likely to be of practical interest. For sampling without replacement, one can either approximate by the binomial distribution when the sample size is small relative to the population size, or instead use a hypergeometric distribution for a finite population of size N and D infected individuals (so that $p = D/N$). The hypergeometric distribution also allows exact calculation of the sample size corresponding to a desired $p_{\min}$-value; the resulting formulas are:

$$\text{infinite population} \quad : \quad n = \ln(\alpha)/\ln(1 - p_{\min}), \tag{5}$$

$$\text{finite population} \quad : \quad n = (1 - \alpha^{1/D}) \times (N - (D-1)/2). \tag{6}$$

Example 3 We illustrate by a textbook example ([5]). Consider a confined population of size 100 individuals (maybe a nursing home), and a contagious disease such as norovirus, for which we might assume that, if present, it would be detectable in at least 10% of the individuals. We desire a sample size large enough to be 95% confident that the disease is not present. Our settings should be: $\alpha = 0.05$, $p_{\min} = 0.1$, $N = 100$ and hence $D = 10$. The formulas (5) and (6) give: $n = 28.4$ and $n = 24.7$, respectively. In this instance, when being led to sampling a large proportion of the population, the infinite population approximation is not sensible, and our suggested sample size therefore comes from the calculation accounting for the finite population: we should take $n = 25$. Note however that the binomial approximation will always be conservative, in the sense of suggesting a larger sample size.

The simple setting discussed here can be extended in many relevant ways. The most important in practice is probably to be able to account for imperfect detection methods, i.e. diagnostic tests that may yield a false negative result (false positive results are often less of an issue, because if a false positive test outcome was suspected it could presumably be followed up with another more accurate diagnostic test). It may also be relevant to incorporate prior beliefs about the probability of disease, hence in a Bayesian framework. A specialized set of methods with strong links to Bayesian reasoning goes under the name "freedom of disease modelling".

Note 1.8: Further remarks

The preceding sections have covered only a few of the basic settings for sample size calculations. One may wonder how to proceed in different and potentially more complex situations (that are not readily available in statistical software). Some possibilities will be briefly discussed here. Before going into the details it is important to remind ourselves that we should not necessarily strive, or expect, for very specific and exact answers. With the many unknowns and subjective choices involved in sample size calculations, it is more realistic to seek rough practical guidance on whether a planned study might be totally off its requirements (in either direction), or on a sensible target range for study planning. Competing requirements for a study can arise from the analytical part, for example when multiple objectives are considered or when multiple outcomes are to be analysed. In such instances one may need to simplify by focusing on the most important aspect(s) of the study. Even if a complex statistical analysis is envisioned, it may be advantageous to base a sample size calculation on a simpler setting that reflects the primary study objectives, and where input parameter values for the calculation can more easily be set. For example, in a repeated measures study one may want to base the choice of sample size on a comparison between treatment groups at a specific time point rather than the full repeated measures analysis. Section 2 gives further examples of how complex situations may be reduced to simpler settings without substantial loss of information.

In some situations, one can use ad-hoc adjustments to account for complexities not accounting for within the standard settings. Two examples will be mentioned here. As we saw in Setting 1.7,

sampling from a finite population (size N) will typically translate into savings in sample size. One general way to adjust for a finite population is to replace the sample size n obtained for an infinite population by $n^* = [(1/n) + (1/N)]^{-1}$. The second ad-hoc adjustment is an inflation factor for sample size to account for lack of independence when experimental units are being "clustered"; a common example is that production animals (say cows) are "clustered" in herds, leading to some lack of independence in their outcomes. We will elaborate on this in Setting 2.3.

A general approach to study any particular question related to a statistical model is to explore the model's properties by simulation: use simulated random data to mimic outcomes from the model, and evaluate the performance of test statistics (or other statistics) by their outcomes among the simulated data (note that this involves generating more than one simulated dataset). This approach also applies to sample size calculations because one can try different sample sizes and investigate whether the desired precision (of confidence intervals) or power is achieved. In complex settings not covered by statistical software or ad-hoc adjustments, this may be the only way to realistically assess whether a planned study has adequately sample size(s). The practical limitations of the approach is that it must be possible to simulate random data from the model. Generally speaking, the complexity of the approach will make it better suited for specialists, even if some statistical software (e.g., Stata) include sophisticated features for simulation. The paper [6] describes a simulation approach using an early version of Stata.

We end this note by briefly mentioning two fundamental and unfortunately fairly common misconceptions about sample size calculations, largely following the excellent practical introduction to sample size calculations ([10]). To facilitate the specification of effect sizes for power-based sample size calculations, it has been suggested to generally quantify these as "small", "medium" and "large" relative to the variability (in normal models, standard deviation), with suggested corresponding numerical values. Because the dependence on effect size and variability in normal distribution models is only through their ratio, this effectively eliminates the need to give values for both effect size and variability, and hence makes calculations possible without any real knowledge of the measurements to be taken. The drawback is evidently that the calculations become detached from any practical objective, leaving it "fuzzy" what the resulting study is expected to achieve.

Another misconception is to perform and report so-called "retrospective" (or "post-hoc") power calculations for a study already done, with the effect size determined from the observed data. The intent is to offer insight into whether the study conducted had adequate power, typically in situations where no significant effect was obtained. However, once the data have already been collected, the best information about the unknown parameters is in the estimates and confidence intervals, so these should be reported instead of trying to extract supplementary information from calculations relevant to planning of studies. In fact, if no significant effect was obtained, it follows immediately that the study did not have sufficient power for the observed effect size, so the extra calculation is pointless (see also ([7],[15]) for further criticism of this practice). If the study's purpose was *not* to establish a difference, it is much better to perform an analysis for a so-called *equivalence* (or *non-inferiority*) hypothesis than to perform a retrospective power calculation. This topic is beyond the scope of these notes.

Note 1.9: Software for sample size calculation

A plethora of different statistical software exists for power and sample size calculations, but overviews of the software available at any point in time tend to quickly become obsolete. We confine ourselves to a few comments about implementations in major statistical packages and a few selected Internet links. Most major statistical packages, such as Stata and Minitab, have built-in routines for the simplest

designs (one- and two-sample models for discrete and continuous data). The Minitab implementation (from version 15 onward) is user-friendly and easy to use, although the simplicity is achieved by restricting the options available. Stata (from version 13 onward) offers a comprehensive suite of sample size calculations, both in terms of the designs covered and the features offered for each design. The menu interface is relatively easy to use, despite its many controls. The R programming platform includes many libraries for sample size calculation (simple designs: `pwr`), and this is a good place to look for a contributed library for a specialized design. The general interface is however less intuitive (not menu-based) and unlikely to appeal to novel R users.

In addition, a wide selection of web calculators and free programs exist for sample size calculation. Some caution is advised when using such tools because the responsibility for the correctness of their use and results generally lies with the user. We include a link to an actively updated overview page, as well as a few respected sources:

- `https://statpages.info/#Power` (list of interactive stats webpages),
- `http://www.openepi.com` (open source epidemiological tools),
- `https://epitools.ausvet.com.au/` (free epidemiological tools maintained by Ausvet, including in particular freedom of disease calculators),
- `https://homepage.divms.uiowa.edu/~rlenth/Power/index.html` (website of Russell Lenth, a major contributor to the field).

2. An approach to nonstandard sample size calculations

As has hopefully become apparent from the discussion in the previous section, the essential tool in sample size calculations is the ability to calculate the standard error of certain estimates — involved in either confidence intervals or test statistics. Such calculations are possible in a much wider range of models and designs than those reviewed so far, allowing us to handle such models and designs in a very similar way as the basic designs. Throughout only designs for a quantitative outcome with normally distributed errors are considered.

Setting 2.1: Balanced ANOVA model without interactions

Consider as an example the additive 2-way ANOVA model,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \quad i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, c, \tag{7}$$

with indices i and j corresponding to two factors A and B, and with index k corresponding to replications. If interest is in comparing two levels of factor A (e.g. if A has only 2 levels), the appropriate statistic is $Est = \bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}$. The variance of Est is $\sigma^2 \times 2/(bc)$, where σ is the standard deviation of the error terms ε_{ijk} in the model. Therefore, sample size calculations based on the precision of the two levels of factor A are almost entirely similar[4] to two-sample sample size calculations, only substituting the sample size n for the number of observations in each group (bc) and the sample standard deviation for the standard deviation of the error terms. However, the calculations will be based on an additional assumption — the lack of interaction in (7) — and the error standard deviation may

[4] More precisely, they are similar except for the impact of the degrees of freedom for the estimated variance in the model. This can usually be ignored for the calculations as long as the resulting value is checked afterwards and is not critically small.

be much less than the standard deviations among observations in each of the two groups (if factor B has a large effect). If it is desired to base sample size calculations on more than two levels of factor A, the reduction from model (7) is to a 1-way ANOVA. Models with more factors and unbalanced designs are dealt with along similar lines.

Setting 2.2: Balanced ANOVA model with interaction

We extend the model (7) with an interaction between factors A and B,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, c. \tag{8}$$

Assume that interest is in the interaction, and — for simplicity — that both factors A and B have only 2 levels ($a = b = 2$). Then the interaction has only a single degree of freedom and can be represented by the estimate (contrast) $Est = \bar{y}_{11\cdot} + \bar{y}_{22\cdot} - \bar{y}_{12\cdot} - \bar{y}_{21\cdot}$[5]. The interpretation of Est is that it estimates the difference between factor A differences at the two levels of factor B (recall, that the presence of interaction means exactly that factor A differences are not the same at the different levels of factor B). For sample size calculations based on Est, we calculate $\mathrm{Var}(Est) = \sigma^2 \times 4/c$. The formula tells us how to base sample size on desired confidence interval length for the interaction. For power calculations, we compare the formula to the two-sample situation: it has a "4" instead of a "2", because the interaction is estimated from 4 groups of size c instead of the usual 2 groups. However, we can "fix" this by a little rewriting: $\mathrm{Var}(Est) = (\sqrt{2}\sigma)^2 \times 2/c$. This shows, that sample size calculations may be based on two-sample formulae/software, if we take as the standard deviation the error standard deviation multiplied by $\sqrt{2}$ and as the group size the number of replications (c) within each group of the combined factor A×B.

Setting 2.3: Two-level random effects models

This section assumes some familiarity with random effects (or multi-level) models. We restrict our discussion to two-level data structures, which could correspond to modelling data for animals within herds. It can be said generally that with more levels and hence an increasing number of random effects, the computation of variances for the estimates of interest becomes more difficult, and more importantly the assessment of estimates of the variation in the data becomes more speculative (and really needs to be based on a previous study of a similar type). The basic statistical two-level model is (using multiple indices here to denote the two levels of the model instead of the different factor levels),

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \ldots + \beta_r x_{rij} + A_i + \varepsilon_{ij}, \quad i = 1, \ldots, a;\ j = 1, \ldots, m, \tag{9}$$

where the *random effects* A_i are assumed to follow $\mathrm{N}(0, \sigma_A^2)$ and the errors ε_{ij} are assumed to follow $\mathrm{N}(0, \sigma_\varepsilon^2)$. In the context of animals within herds, the index i corresponds to herds and the index j corresponds to animals. For simplicity, we have taken the number of animals in each herd to be the same for all herds (m). The fixed part of the model, involving possibly both factors and regression variables, is contained in the formalism $\beta_0 + \beta_1 x_{1ij} + \ldots + \beta_r x_{rij}$, where β_0 is usually termed the "intercept", and $\beta_1, \ldots, \beta_r$ are the regression coefficients for the explanatory variables (possibly dummy variables) $x_1, \ldots, x_r$.

One important feature of two-level models is that explanatory variables may vary at different levels (in a designed experiment, treatments may be applied at different levels). For example, treatments may be applied to animals but herd management factors are applied to herds, not animals. Another

[5] In a parameterization with the parameter restrictions $\sum \alpha_i = 0$, $\sum \beta_j = 0$ and $\sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$, the contrast Est estimates $\gamma_{11} = \gamma_{22} = -\gamma_{12} = -\gamma_{21}$.

way of saying this is that there are different experimental units in the design/model — the lowest level units (animals) and the highest level units (herds). When carrying out sample size calculations for multi-level models, it is crucial to correctly identify the experimental unit for the factor of interest. If the experimental unit is at the lowest level, the upper level(s) may be ignored for sample size calculations, and the same precision is obtained if replications are obtained at the lowest or at higher levels. Again in our example, it thus makes no difference for a comparison of animal treatments, based on model (9), if replications involve more animals per herd or more herds. For explanatory variables at a higher level, the specific assumptions of the model (9) become important (see the examples below). Explanatory variables may also both vary at the lowest level and show considerable clustering at higher levels (typically in observational studies); in such situations, sample size calculations become quite difficult, and procedures will depend on the specific circumstances of the study design (and simulation may become an attractive option).

Setting 2.3.1 Two-level model with same correlations within a group

For model (9), if we assume that the random variables (A_i) and (ε_{ij}) are all independent, the model effectively assumes all observations within a level/group to be (positively) correlated and to the same degree.[6] This would seem a reasonable model for our example with animals within a herd if there are no apriori reasons to expect some animals to be more strongly correlated than others, after the fixed effects in the model are taken into account. Herd factor comparisons will be based on herd averages, so in order to assess precision we need to compute their variances. Considering the first herd, for ease of notation, we have from the model formula (9)

$$\bar{y}_{1\cdot} = \beta_0 + \beta_1 \bar{x}_{11\cdot} + \ldots + \beta_r \bar{x}_{r1\cdot} + A_1 + \bar{\varepsilon}_{1\cdot} = \mu_1 + A_1 + \bar{\varepsilon}_{1\cdot},$$

writing just μ_1 for the fixed part of the model. By the assumption of all random variables being independent, we compute

$$\mathrm{Var}(\bar{y}_{1\cdot}) = \mathrm{Var}(A_1) + \mathrm{Var}(\bar{\varepsilon}_{1\cdot}) = \sigma_A^2 + \sigma_\varepsilon^2/m, \tag{10}$$

where m is the number of animals in the herd. The formula shows that taking more animals within a herd will only partly reduce the precision of herd averages, and if the variation between herds (σ_A) is large relative to the variation within herds, it has only little impact. Therefore, the relevant parameter to adjust to achieve desired precision of herd comparisons, is the number of herds, and the appropriate standard deviation to use is given by (the square root of) formula (10). To exemplify, if two types of herds are to be compared, sample size calculations for a two-sample model apply to determine the number of herds of each type. Values of σ_A and σ_ε need to be assessed, possibly from previous studies. It is also possible to rewrite the right hand side of (10) as $\sigma^2(1+\rho(m-1))/m$, where $\sigma^2 = \sigma_A^2 + \sigma_\varepsilon^2$ is the total variance on each observation, and ρ is the (intraclass) correlation between two animals in the same herd. It may be easier or more intuitive to assess these values than the two variance components.

The quantity $[1+\rho(m-1)]$ is sometimes called a variance inflation factor for clustered data, because the calculation showed that it expresses how much the variance of a group (upper level) mean increases (multiplicatively) in the two-level model relatively to a situation where all observations are independent. A sample size calculation based on inference (estimation or test) at the upper level in a two-level model can be carried out in two steps. First, calculate a sample size while assuming all observations to be independent; second, multiply the resulting sample size by the variance inflation factor. This approach is effectively the ad-hoc adjustment for clustered data mentioned in Note 1.8.

[6] The technical name for this assumption is a compound symmetry or exchangeable correlation structure.

Thus, the adjustment has a specific meaning and justification, but it does require the groups (say herds) to be of equal size (m). This is usually an unreasonable assumption in practice, so the formula is used with an average group size. Also the assumed equal within-cluster correlation ρ may be difficult to justify and specify a value for. More crucially, the adjustment only applies to group-level inference, and this restriction of its use may not be sufficiently transparent when the adjustment is used in practice.

Setting 2.3.2: Two-level model with autoregressive correlations over time

Model (9) is also useful for repeated measures data with a series of observations (typically over time) on a number of subjects (say, animals). Then the upper level (index i) corresponds to animals, and the lower level (index j) corresponds to time. It is often of interest to extend the model from Example 2.3 with additional correlations between the error terms ε_{ij} within each subject, and we consider here only the simplest of these — an autoregressive correlation structure (shown here for $n_i = 4$),

$$\mathrm{Corr}(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4}) = \begin{pmatrix} 1 & & & \\ \rho & 1 & & \\ \rho^2 & \rho & 1 & \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}.$$

The model contains as an additional parameter ρ — the correlation between errors of two observations adjacent in time. The variance of subject averages is still of the same form as in (10),

$$\mathrm{Var}(\bar{y}_{1\cdot}) = \sigma_A^2 + \mathrm{Var}(\bar{\varepsilon}_{1\cdot}),$$

and the last term can be computed to yield for the autoregressive correlation structure:

$$\mathrm{Var}(\bar{\varepsilon}_{1\cdot}) = \frac{\sigma_\varepsilon^2}{m} + \frac{2\sigma_\varepsilon^2 \rho\,(m - 1 - m\rho + \rho^m)}{m^2(1-\rho)^2}. \tag{11}$$

The remaining part of sample size calculations follows the same route as in Example 2.3. To illustrate the formulae, we show in the following table some estimated values of variations and the resulting subject mean variances; data are on milk yield with 6 measures per cow, from [12].

Model	Same correlations	Autoregressive
estimates	$\hat{\sigma}_A^2 = 21.4$ $\hat{\sigma}_\varepsilon^2 = 17.9$	$\hat{\sigma}_A^2 = 18.7$ $\hat{\sigma}_\varepsilon^2 = 20.1$ $\hat{\rho} = 0.239$
subject mean variance	24.4	23.7

References

[1] Baldi, B. & Moore, D. S. (2018), *The Practice of Statistics in the Life Sciences*, 4th ed., W. H. Freeman and Company, New York.

[2] Bland, J. M., Altman, D., Brown, M., Cullum, N., Raftery, J. & Torgerson, D. (2009), The tyranny of power: is there a better way to calculate sample size?, *BMJ: British Medical Journal* **339**, 1133–1135.

[3] Campbell, M. J., Julious, S. A. & , Altman, D. G. (1995), Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons, *BMJ: British Medical Journal* **311**, 1145–1148.

[4] Christensen, R. (1996), *Analysis of Variance, Design and Regression*, Chapman & Hall / CRC Press, Boca Raton.

[5] Dohoo, I. R., Martin, S. W. & Stryhn, H. (2012), *Methods in Epidemiological Research*, VER Inc., Charlottetown.

[6] Feiveson, A. H. (2002), Power by simulation, *The Stata Journal* **2**, 107–124.

[7] Hoenig, J. M. & Heisey, D. M. (2001), The abuse of power: the pervasive fallacy of power calculations for data analysis, *The American Statistician* **55**, 19–24.

[8] Jennen-Steinmetz, C. & Wellek, S. (2005), A new approach to sample size calculation for reference interval studies, *Statistics in Medicine* **24**, 3199–3212.

[9] Kupper, L. L. & Hafner, K. B. (1989), How appropriate are popular sample size formulas?, *The American Statistician* **43**, 101–105.

[10] Lenth, R. V. (2001), Some practical guidelines for effective sample size determination, *The American Statistician* **55**, 187–193.

[11] Moore, D. S., McCabe, G. P. & Craig, B. A. (2012), *Introduction to the Practice of Statistics*, 7th ed., W. H. Freeman and Company, New York.

[12] Nødtvedt A., Dohoo I. R., Sanchez J., Conboy G., DesCôteaux L. & Keefe G. (2002), Increase in milk yield following eprinomectin treatment at calving in pastured dairy cattle, *Vet. Parasitol.* **105**, 191–206.

[13] Oehlert, G. W. (2000), *A First Course in Design and Analysis of Experiments*, W. H. Freeman and Company, New York.

[14] Ryan, T. P. (2013), *Sample Size Determination and Power*, Wiley, New York.

[15] Smith, A. H. & Bates, M. N. (1992), Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies, *Epidemiology* **3**, 449–452.

[16] Zar, J. H. (2010), *Biostatistical Analysis*, 3rd ed., Pearson Printice Hall, New Jersey.