



## 2. Sample Size for a Microbiome Experiment

George Savva



Quadram Institute Best Practice in Microbiome Research: Sample Size for a Microbiome Experiment v1.0 by [George M Savva](#) is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

### 2.1. Introduction and Importance

Choosing the number of experimental units (i.e. the 'sample size') is an important part of any research study design. The larger the sample size, the more precise the estimates from your study will be, and the study will have more power to detect smaller effects.

Having too many samples has ethical and resource implications, so you may need to trade off this precision against ethics and cost.

However, under-powered studies are one of the main causes of irreproducibility in bioscience because they lead to high false positive rates and a low chance of detecting true effects.

Funding bodies are increasingly concerned with the detail of sample size calculations, particularly for animal and human studies, before they will fund research. Ethics committees also scrutinise sample size calculations, and reporting guidelines for both animal and human studies require that the rationale for determining study sample size is included in reports.

Having said this, sample size calculation is not an exact science, since it relies on many assumptions regarding the distribution of data and expected effects; and an assumed analysis plan that may not ultimately be the one that is followed.

In practice, a sample size calculation is required to (i) ensure that you have thought through the design and analysis of your study in relation to its aims and (ii) that your study is large enough to have a good chance of answering the question that it sets out to answer, to the level of certainty required, under a set of realistic assumptions about how the data might turn out.

The literature on power and sample size determination in microbiome studies is extremely limited (Appendix A). Although some microbiome-specific tools exist, general power and sample size tools can often be applied successfully, as well as ideas from studies based on gene expression datasets which often have a similar structure and aims.

This document sets out possible approaches to sample size calculation for microbiome studies and issues to be aware of. **If you are not confident in calculating a sample size for your study, then always seek help from a statistician.**

## 2.2. Basic principles

A sample size calculation finds the smallest sample size needed to give you an acceptable chance of meeting your study aims. Hence your study aims, and what constitutes an 'acceptable' risk of failure (typically failure to detect a true effect) must both be clearly defined.

### 2.2.1. Setting the question and power required

- Be clear about what the goal of your study is, that is, what question(s) must your study be able to answer. Be clear whether your study is confirmatory (aiming to test a specific hypothesis or estimate a specific quantity such as a prevalence) or exploratory, with no preconception about what might be found. Some studies have both confirmatory and exploratory aims, and more than one of each; in these cases, the aims that define the study's 'success' should be used for the power calculation. By success here we mean that the study answers the questions it sets out to, not that there is a 'positive' result.
- Be clear about what power is required, that is, what is the probability that you successfully detect an effect, should it exist.
- For studies testing many possible effects, power is analogous to the proportion of true effects that are successfully detected.
- For studies aiming to estimate something, rather than test a theory, consider the level of precision you require [1].
- Make sure your calculation matches your design and your objectives. For example, if you are mainly interested in a mediation effect or interaction, or you are interested in identifying 'responders' to a treatment, make sure you power to test this, not just a simple two-group comparison. If your aim is to identify differentially-abundant microbes, don't power on a test for alpha-diversity.

### 2.2.2. Approaches to sample size calculation (more detail below):

Any justifiable method for calculating sample sizes is acceptable, but important considerations are:

- For a confirmatory study, that is a study to test specific hypotheses or estimate specific quantities, identify the elements required for a sample size calculation (described below).
- For studies targeting specific univariate features of the microbiome, standard approaches to sample size calculations (power calculations or calculations based on precision of effect estimates) can be adopted.
- For studies testing the more general hypothesis that the microbiomes from two groups will be 'different' (for example, as defined by beta-diversity), power and sample size can be calculated by simulation via R packages, e.g. *micropower*, so long as a target effect size can be identified.
- For exploratory studies you should explore the likely effect sizes detectable at different sample sizes, or the proportion of effects of a certain size that are likely to be statistically significant at different sample sizes.

## 2.3. Methods for sample size calculation

### 2.3.1. Analytical methods

For simple research questions based on one or a few microbial features (such as a specific diversity measure or presence or abundance of a particular species), required sample sizes can be calculated using standard formulae, so long as each of the elements required above for the calculation can be identified. (See Appendix B for a general overview of sample size calculations).

For example, if your aim is to compare alpha diversity between two independent groups, a power calculation for a simple unpaired t-test would be appropriate. Standard free-to-use tools such as G\*power or functions from various R packages (*pwr* or *pwr2*) can be used depending on the study design.

### 2.3.2 Power and sample size by simulation

For more complex designs (for example longitudinal designs) or outcomes (such as comparisons of beta-diversity), where formulae are not available, sample sizes can be calculated by simulation [2]. This proceeds as follows:

- Simulate datasets that are as similar as possible to the real data that you are likely to observe, with typical levels of variation and including the effects that you wish to be able to detect.
- Analyse these using your planned methods, and find the proportion of datasets in which you were able to successfully detect the effect(s) of interest.
- Vary the sample size and any other assumptions to determine how the power varies.

This method relies on being able to simulate realistically-structured datasets with realistic, but artificially induced, effects, and knowing what your method of analysis is going to be.

Several *R* packages are available to help with this.

For complex study designs with simple outcomes the *simr* package [3] can perform power calculations by simulation based on linear mixed models.

For simple designs with comparison of beta-diversity as the primary outcome, the *micropower* package [4] can simulate microbiomes based on data from the *human microbiome project*, and test the power to detect differences for a variety of sample sizes and effect sizes.

For univariate differential abundance and a range of multivariate analyses the recently released *powmic* package [5] promises power and sample size calculations based on real datasets; interested users should study the primary paper and tutorials.

It is not recommended that you use the *hmp* package [6] for power calculations, since this does not correspond to a downstream statistical analysis that you are likely to perform and may be based on unrealistic distributions of parameters.

### 2.4. Repeated measures, longitudinal studies or clustering in design

Where observations are not independent of each other (for example, when repeated measures are taken on the same unit, or when units are housed or treated in groups), then adjustments must be made to sample size calculations as follows:

### 2.4.1 Cross-overs, paired designs or repeated measure designs with the comparison *within* groups.

If sources of variation can be removed from estimates of the comparisons of interest, then studies will be more efficient and the power required will be reduced. The simplest example of this is in paired experiments. For example, in a paired study, the variance in outcomes between pairs is not important, since we make our comparison within pairs.

Suppose we have  $N$  observations per group, and we are interested in a between groups comparison of means. Then, if the experimental units vary with a standard deviation of  $sd_{total}$ , then the standard error of the mean difference is:

$$se = \sqrt{2} \times \frac{sd}{\sqrt{N}}$$

If we can pair our observations and divide the between-group variance into the variance between pairs and the variance within pairs, and we assign one of each pair to each group, then our estimate has variance:

$$se = \sqrt{2} \times \frac{sd_{within}}{\sqrt{N}}$$

That is, the precision has been improved by a factor of

$$\frac{sd^2}{sd_{within}^2} = \frac{1}{1 - ICC}$$

where  $ICC$  is intra-class correlation within pairs of observations (see 2.4.2 for a discussion of  $ICC$ ).

As the effective sample size has been increased by a factor of  $\frac{1}{1-ICC}$ , so we can reduce the sample size by this same factor.

Note the benefit of using a paired design depends on the  $ICC$ . If there is little intra-class correlation, (e.g. if pre- and post- measures are largely uncorrelated in a random controlled trial (RCT) then nothing is gained by taking the extra measurement. If the pairs are very highly correlated, and the  $ICC$  is close to 1, then the sample size can be dramatically reduced.

### 2.4.2 Cage-effects, intra-class correlation and repeated measures designs with the comparison across clusters.

Statistical power (and precision) is determined by the number of *independent* experimental units. Where units are not treated independently of each other, or could influence each others outcomes, so called ‘intra-class correlations’ (ICC) can be introduced, and this affects the calculation of standard error and hence the power of the study and the required sample sizes.

For a comprehensive discussion of this issue with practical recommendations, particularly as it relates to animal experiments, see Basson et al., [7].

ICC between experimental units can be caused by factors that influence all individuals within a group (such as a shared environment, shared genetic background, or contamination within a particular group). This can cause outcomes within groups to be more similar than we would expect by chance, and so each unit contributes less information to the study than it would otherwise do.

For example, we have observed strong ICCs in microbiomes of mice sharing cages, and of farmed fish sharing water systems. This is unsurprising given the sensitivity of microbiomes to environmental factors. We cannot analyse animals independently without taking account of this shared variance, and this is reflected in the required sample size.

The ‘design effect’ of a study is the factor by which we need to inflate the sample size to take into account ICC. The design effect is calculated by:

$$\text{Design effect} = 1 + ICC \times (n - 1)$$

Where ICC is the intra-class correlation, and  $n$  is the number of animals per group. ICC is the proportion of total variation that is attributable to group membership as opposed to individual variation. It is calculated by:

$$ICC = \frac{sd_{between}^2}{sd_{between}^2 + sd_{within}^2}$$

For example, an ICC of 0.3 is not uncommon for features of microbiomes of mice housed in the same cages (i.e. 1/3 of variance being cage-specific rather than mouse-specific). So, if mice are to be housed five to a cage, then required sample sizes should be inflated by a factor of around  $1 + .3 * (5 - 1) = 2.2$  to account for this.

- When your design will include samples that are not independent, find the relevant ICC and adjust the sample size accordingly (or calculate the sample size for several reasonable guesses at what it might be). Alternatively, if you know the likely contribution from different sources of error you can estimate power by simulation using e.g. the *simr* R package.
- ICCs are rarely reported, but can be calculated using prior datasets from studies with similar designs.

## 2.5. Important considerations

### 2.5.1 Use of previous literature/data for inputs

- To conduct sample size calculations you will need to know the likely variability of your outcome measures, and the magnitude of the effect size that you anticipate.
- For variance estimates, seek out previous similar data (ideally the same population measured using the same technology) or published literature.
- For effect sizes, consider primarily what is the ‘minimum important difference’ that you want your study to be able to detect.
- Be wary of *significant* estimates arising from previously published work. We know that published effect sizes are prone to upward bias. Therefore, taking, as an example, 50% of the published effect size as the likely effect size for replication, is reasonable.
- Note also that, if a previous study returned a p-value of exactly or very close to 0.05, then the study had only 50% power to detect the effect that it reported. This can be easily seen if you consider that an exact replication of this study would be equally likely to return a p-value above or below the p-value of the original, that is above or below 0.05. You need to roughly double the size of the previous study to have well-powered replication in this case.
- Consider how ‘generalisable’ effect sizes are, and whether standard errors are likely to be as a result of one technology over another, or one setting over another. For example, you might be planning a human study, but have information on likely effect sizes and variance from *in vitro* colon model studies. This is likely to be useful, but not directly transferable, so consider what proportion of the effect observed *in vitro* would be expected, or would still be worth pursuing if observed in the follow-up study, and calculate power/sample size according to this.

### **2.5.2 Be honest about sample size requirements and the feasibility of your study.**

- Many sample size calculations are fudged to fit budget constraints, by using favorable estimates of standard deviation and effect size. If your study must be underpowered by resource constraints, it is better to be honest about this and the risk of failure.

### **2.5.3 Do not plan to replicate experiments – conduct single larger experiments instead.**

- If you have the resources to conduct two identical independent experiments, then it is much better from a power and precision point of view to consider these as one large experiment, divided into two blocks, and to analyse them together. Replication in this way tells you very little about the reliability of a finding.

### **2.5.4 Rules of thumb and common practice**

- Some commonly used 'rules-of-thumb' regarding sample sizes have a sound basis, whereas others do not. For example, clinical studies of well-studied univariate clinical outcomes in certain populations will all have roughly the same sample size, since the variation in the outcome measure and target effect sizes will likely be the same.
- However, microbiome studies are often very underpowered, and it is not appropriate to directly mirror sample sizes seen in other studies, even with 'significant' findings. Instead use data from these studies as inputs to conduct your own sample size calculations.

### **2.5.5 Account for loss of data**

- Always make an allowance for loss of data, either through drop-out of participants, failure to collect a sample, removal of outliers etc. Typically sample sizes should usually be inflated by 10-30% for human studies depending on the design. Population-representative cohorts will typically have higher drop-out rates than volunteer cohorts, similarly studies with older, less healthy people or with long periods between baseline and follow-up will have higher drop-out rates than shorter studies with younger, healthier people.

### **2.5.6 Be realistic about likely recruitment**

- Whatever the required sample size for a study, be realistic about the chance of recruiting this number of participants. Prior evidence of recruitment from similar settings and similar studies should be sought, or careful calculations about the likely number of

eligible participants meeting inclusion criteria in the planned research setting and anticipated recruitment rate should be made.

### 2.5.7 Consider carefully whether your study is worth doing

- A power calculation might reveal that your study cannot be feasibly conducted within your resource constraints.
- However, even if you cannot recruit the number of participants a power calculation suggests, you might still decide to conduct a research study, if there is no alternative and the research question is worth pursuing. Sample size calculations are only estimates, and underpowered studies (if analysed correctly, that is, i.e. not relying on p-values to draw conclusions) can still provide valuable information. These can also contribute to future meta-analyses to provide more definitive results.

### 2.5.8 Exploratory or preliminary work

For exploratory or pilot work, it can be more difficult to determine a 'required' sample size, since the goals of the study in terms of its success or failure are ambiguous.

However, an exploratory study does need to be big enough to meet its own objectives. These objectives might be screening to see if significant effects are likely to exist across a range of possible outcomes, or to get information on standard deviations or preliminary effect sizes.

If screening for possible effects is the aim, then power is very important, as you do not want to miss potentially important effects. However, you might tolerate a higher risk of false positives (so a higher critical p-value) which will reduce the required sample size.

The greater the sample size, the more precise the estimates from the research will be, and the more subtle the effects that can be detected.

- For exploratory work, calculate the proportion of the effects at given sizes that are likely to be detectable at the proposed sample size (using any of the methods above) and consider whether this is likely to be informative enough to proceed.
- For preliminary work, be clear about which outcomes of the study will inform the next investigation, and ensure your estimates of these are precise enough.

## Appendix A:

### A short scoping review of literature on power and sample size in microbiome studies.

#### Search terms and results:

A search was conducted (January 2021) using World of Knowledge: WoK: 18 results for (TI=microbiome AND (TI=sample size OR TI=power)) AND LANGUAGE: (English) Evaluation of abstracts reduced this to 11 papers. Full text search revealed four directly relevant papers [4–6, 8]:

## Appendix B

### How does sample size affect precision

Sample size governs the *precision* of the estimates from your study (reflected by, for example, their standard errors). If you are hypothesis testing, this precision in turn governs the *power* of your study to detect deviations from null hypotheses.

The power of a study to detect an effect is defined as the probability that the study will successfully identify it, if it in fact exists. We typically choose to design our studies with 80% or 90% power. This means that we accept a 10% to 20% risk of missing a true effect.

The simple equation:

$$s. e. \propto \frac{s. d.}{\sqrt{n}}$$

Describes the relationship between the standard error of an estimate, the standard deviation of the outcome measure used to calculate it, and the sample size used (with the exact relationship depending on study design).

So, for example, as sample size doubles, standard error of estimates (hence the width of confidence intervals) will decrease by a factor of  $\sqrt{2}$ . With respect to power, more precision means that smaller differences between groups can be detected. Doubling of sample size will enable detection of effects that are  $\sqrt{2}$  times smaller.

## Information needed for a sample size calculation

To conduct a traditional sample size calculation, you need to know:

- **The hypothesis (or hypotheses) that you are planning to test or the parameters you want to estimate.** These will follow from your study aims.  
The study design and analytical methods you will use to address these aims
- **The likely variation in the outcome measures across samples.** This will likely come from prior data on a similar outcome in a similar population.
- **The precision required, or the size of the effect that you want to be able to reliably detect.** The precision required, or minimum effect sizes, can be guided by a combination of prior data (to establish what might be a feasible difference) and your own study aims (to establish what is the smallest difference that would be considered important enough not to want to miss).
- **The threshold for statistical significance.** Our convention is to apply a threshold of  $p < 0.05$  for statistical significance. The most common reason for deviating from a critical threshold of 0.05 is an adjustment for multiple testing. If you are testing many hypotheses (such as many different candidate microbiome features) then you might make a correction to a p-value, controlling instead for a false discovery rate (FDR), in which case the required FDR should be known.
- **The required power.** How much risk of missing a true effect are you willing to accept, or what proportion of missed true effects are you willing to accept.



Quadram Institute Best Practice in Microbiome Research: Sample Size for a Microbiome Experiment v1.0 by [George M Savva](#) is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

## References

1. Rothman KJ, Greenland S (2018) Panning Study Size Based on Precision Rather Than Power. *Epidemiol Camb Mass* 29:599–603.  
<https://doi.org/10.1097/EDE.0000000000000876>

2. Arnold BF, Hogan DR, Colford JM, Hubbard AE (2011) Simulation methods to estimate design power: an overview for applied research. *BMC Med Res Methodol* 11:94. <https://doi.org/10.1186/1471-2288-11-94>
3. Green P, MacLeod CJ (2016) SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol Evol* 7:493–498. <https://doi.org/10.1111/2041-210X.12504>
4. Kelly BJ, Gross R, Bittinger K, et al (2015) Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. *Bioinforma Oxf Engl* 31:2461–2468. <https://doi.org/10.1093/bioinformatics/btv183>
5. Chen L (2020) powmic: an R package for power assessment in microbiome case–control studies. *Bioinformatics* 36:3563–3565. <https://doi.org/10.1093/bioinformatics/btaa197>
6. Rosa PSL, Brooks JP, Deych E, et al (2012) Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. *PLOS ONE* 7:e52078. <https://doi.org/10.1371/journal.pone.0052078>
7. Basson AR, LaSalla A, Lam G, et al (2020) Artificial microbiome heterogeneity spurs six practical action themes and examples to increase study power-driven reproducibility. *Sci Rep* 10:5039. <https://doi.org/10.1038/s41598-020-60900-y>
8. Casals-Pascual C, González A, Vázquez-Baeza Y, et al (2020) Microbial Diversity in Clinical Microbiome Studies: Sample Size and Statistical Power Considerations. *Gastroenterology* 158:1524–1528. <https://doi.org/10.1053/j.gastro.2019.11.305>