International Journal of Technology and Applied Science (IJTAS)
E-ISSN: 2230-9004 ● Website: www.ijtas.com ● Email: editor@ijtas.com

Experimental Design and Sample Size Estimation in Microbiological Research: A Practical Review

Dr. B. Chandrasekhara Reddy[1], Dr. H. Pallavi[2]

[1]Department of Statistics, Government College, Autonomous, Anantapuramu -515001, India.
[2]Department of Microbiology, Government College, Autonomous, Anantapuramu -515001, India.

Abstract:
Microbiology produces data of many kinds: colony counts, zones of inhibition, minimum inhibitory concentrations (MICs), survival of infected animals, qPCR cycle thresholds and sequencing read tables. Whatever the data type, the reliability of a conclusion depends on how the experiment was designed and whether enough independent samples were used. Weak design, very small sample sizes and pseudoreplication (treating repeated measurements as independent samples) are common in the microbiological literature and reduce reproducibility. This review explains, in a practical manner, the core principles of experimental design (replication, randomization, blocking and controls), the correct identification of the experimental unit, and the common designs used in the laboratory (completely randomized, randomized block, factorial, split-plot and response-surface designs). It then describes sample size estimation using power analysis, with worked examples for log-transformed counts, proportions, count data and microbiome studies. Special issues such as plate-count statistics, MIC data, animal infection models, contamination controls in low-biomass work and reporting standards are discussed. A step-by-step workflow and a checklist are provided so that researchers can plan, justify and report their experiments with confidence.

Keywords: Experimental design; sample size; statistical power; pseudoreplication; randomization; microbiology; microbiome; reproducibility.

1. Introduction

Most microbiological experiments are small. A typical study may compare a few treatments with three plates each, or two or three biological replicates of a culture. Such designs are often chosen from habit or convenience rather than from a calculation. Yet the same experiment may be used to support claims about antimicrobial activity, pathogen survival or community change.

Poor design and low statistical power are well documented problems across the life sciences. Studies with low power have a lower chance of detecting true effects, and the effects they do report are more likely to be exaggerated or false [1,2]. Pseudoreplication, which is the use of repeated measures from the same unit as if they were independent, inflates false-positive results and has been shown to be widespread in laboratory research [3,4,5]. In microbiome research, the problems are increased by

compositional data, contamination and very large numbers of variables, and several authors have called for better design and reporting in the field [6,7].

The foundations of experimental design were laid by R. A. Fisher, who emphasized replication, randomization and local control [8]. These ideas are as relevant to a Petri dish as to a field plot. This review translates them into practical guidance for microbiologists. It is aimed at students, laboratory researchers and reviewers who need a clear, working understanding rather than a mathematical treatment.

The objectives of this review are: (i) to explain the key principles of experimental design; (ii) to show how to identify the true experimental unit; (iii) to describe common designs; (iv) to explain how to estimate sample size, with worked examples; (v) to highlight domain-specific issues in microbiology; and (vi) to give a practical workflow and checklist.

2. Basic Principles of Experimental Design

2.1 Replication

Replication means applying each treatment to more than one independent experimental unit. It allows the natural variation between units to be estimated, and it increases the precision of the estimated treatment effect. Replication should not be confused with repeated measurement of the same unit (see Section 3).

It is useful to separate three levels:

- **Technical replicates:** repeated measurements of the same sample (for example, three plates from one dilution, or triplicate qPCR wells). They measure the precision of the method, not biological variation.
- **Biological replicates:** independent samples or independently grown cultures. They measure biological variation and support generalization.
- **Independent experiments:** the whole experiment repeated on different days, with fresh media and reagents. They show whether the result is reproducible.

2.2 Randomization

Randomization means assigning treatments to units by chance, for example using a random number table or software. It protects against systematic bias, such as plates at the edge of an incubator being warmer, or treatments always being handled in the same order. It also provides the basis for valid statistical tests [8,9]. In practice, microbiologists should randomize the position of plates, flasks and tubes in incubators and shakers, the order of processing and, where possible, the order of measurement.

2.3 Blocking and local control

Blocking groups similar units together so that known sources of variation are removed from the comparison. Common blocks in microbiology are the day of experiment, batch of medium, incubator shelf, operator and sequencing run. Each block contains all treatments, and the block effect is removed in the analysis. This increases power without increasing the number of units.

2.4 Controls

A valid experiment needs suitable controls:

- **Negative control:** no treatment, or the vehicle or solvent alone (for example, DMSO in an antimicrobial assay).
- **Positive control:** a treatment known to work (for example, a standard antibiotic).
- **Sterility control:** uninoculated medium to detect contamination.
- **Growth control:** inoculated medium without the test agent.
- **Extraction and no-template controls** in molecular work, to detect reagent contamination [10,11].

2.5 Blinding

Where measurement involves judgement, such as reading zone edges, counting colonies by eye or scoring histology, the person measuring should not know the treatment. Blinding reduces observer bias and is simple to arrange by coding the plates.

3. The Experimental Unit and Pseudoreplication

The experimental unit is the smallest unit to which a treatment is independently applied. It is the unit that determines the true sample size, n. Hurlbert defined pseudoreplication as the use of inferential statistics to test for treatment effects when treatments are not replicated or the replicates are not independent [3].

Examples in microbiology

Situation	Wrong n	Correct experimental unit
One flask of culture treated with a compound, then 5 plates counted from it	5 plates	1 flask (n = 1)
One mouse, with 4 tissue samples plated	4 samples	1 mouse
One cell-culture plate, with wells treated with the same drug stock	All wells	Often the plate or the independent experiment, depending on how the treatment was applied
One soil sample divided into 6 subsamples	6 subsamples	1 soil sample
10 stool samples from 3 patients	10 samples	3 patients (if samples are from the same patients, use a repeated-measures or mixed model)

Lazic and colleagues have discussed how often the question of what N is goes unanswered in cell-culture and animal work, and recommend deciding it before the experiment starts [4,5]. When data have a hierarchical structure (for example plates within flasks within experiments), the analysis should reflect it using nested analysis of variance or mixed-effects models, with the highest independent level as the unit of inference.

4. Common Experimental Designs

4.1 Completely randomized design (CRD)

Treatments are assigned entirely at random to units. It is the simplest design and is suitable when units are homogeneous, for example comparing three disinfectants using identical, randomly arranged tubes. It is analyzed by one-way ANOVA (or the t-test for two groups). Its weakness is that all unexplained variation goes into the error term.

4.2 Randomized complete block design (RCBD)

Units are grouped into blocks, and every treatment appears once in each block. For example, a comparison of four media can be run on five different days, with each day as a block. The block effect is removed, which usually gives better precision than a CRD when days differ. It is analyzed by two-way ANOVA without interaction, or by a mixed model with block as a random effect [9,12].

4.3 Factorial designs

Factorial designs study two or more factors together, for example temperature (3 levels) and pH (3 levels) on growth. They allow detection of interactions, such as an antibiotic working better only at a particular pH. They are more efficient than varying one factor at a time, which uses more runs and cannot reveal interactions [13].

4.4 Split-plot and nested designs

When one factor is hard to change (for example incubator temperature) and another is easy (for example medium type), the hard-to-change factor is applied to whole plots and the other to subplots. The analysis must use two error terms. Nested designs are used when units are naturally grouped, for example plates within flasks.

4.5 Latin square

The Latin square controls two sources of variation at once, for example incubator shelf and operator, using a small number of runs. It requires the number of treatments to equal the number of rows and columns and assumes no interactions.

4.6 Response surface and screening designs

In media optimization, fermentation and process development, many factors may influence yield. Screening designs (such as Plackett-Burman or fractional factorial) identify the important factors with few runs. Response surface designs (central composite or Box-Behnken) then find the optimum [9,12,13].

4.7 Paired and repeated-measures designs

When the same unit is measured before and after treatment, or at several times, the measurements are correlated. Paired tests or repeated-measures and mixed models should be used. This is common in longitudinal clinical and microbiome studies.

Table 1. Choosing a design

Situation	Suggested design	Typical analysis
Homogeneous units, one factor	CRD	One-way ANOVA / t-test
Known source of variation (day, batch)	RCBD	Two-way ANOVA / mixed model
Two or more factors, interaction of interest	Factorial	Factorial ANOVA
Hard-to-change factor	Split-plot	Mixed model
Two nuisance factors, few units	Latin square	ANOVA
Optimization of many variables	Screening + response surface	Regression / RSM
Same unit measured repeatedly	Repeated measures	Mixed model

5. Sample Size Estimation

5.1 Why sample size matters

If the sample is too small, a real effect may be missed (Type II error), and any significant result is more likely to be a false or exaggerated one [1,2]. If it is too large, resources and animals are wasted, which is an ethical concern. A sample size should therefore be justified before the experiment, not after.

5.2 Key concepts

- **Significance level (alpha):** probability of a false positive, usually 0.05.
- **Power (1 - beta):** probability of detecting a real effect, usually 80% or 90%.
- **Effect size:** the smallest difference that is biologically meaningful. It should be decided from biology, not from what is convenient.
- **Variability (standard deviation, SD):** taken from pilot data, earlier publications or laboratory records.
- **Other factors:** one-sided or two-sided test, number of groups, expected loss of samples.

Power analysis was formalized by Cohen [14] and is described in clinical and laboratory contexts by several authors [15,16].

5.3 Comparing two means

For two groups of equal size, with a two-sided test, the number per group is approximately:

$$\mathbf{n = 2\ (Z_{\alpha/2} + Z_{\beta})^2\ x\ (SD)^2 / d^2}$$

where d is the difference to be detected. For alpha = 0.05 and 80% power, $\boldsymbol{Z_{\alpha/2}}$ = 1.96 and $\boldsymbol{Z_{\beta}}$ = 0.84, so $(Z_{\alpha/2} + Z_{\beta})^2 = (1.96 + 0.84)^2 = 7.84$. For 90% power, Z_{β} = 1.28 and the sum of squares is 10.51.

Worked example (log counts). A researcher wants to show that a disinfectant reduces bacterial counts by at least 1 log10 CFU compared with the control. From pilot work, the SD of log10 CFU between independent replicates is 0.5. Then n = 2 x 7.84 x 0.25 / 1 = 3.9, so about 4 per group. Because this formula uses the normal distribution, one or two extra units are usually added when n is small (so 5 per group). If the difference of interest is only 0.5 log, n = 2 x 7.84 x 0.25 / 0.25 = 15.7, so about 16 to 17 per group. The example shows how the required sample size rises steeply when the effect to be detected is halved: halving d multiplies n by four.

5.4 Comparing two proportions

For binary outcomes such as survival, infection or presence of a gene, the number per group is approximately:

$$n = (Z_{\alpha/2} + Z_{\beta})^2 \times [p_1(1-p_1) + p_2(1-p_2)] / (p_1 - p_2)^2$$

Worked example (survival). If the expected survival is 80% in a treated group and 50% in the control, with alpha = 0.05 and 80% power: n = 7.84 x (0.16 + 0.25) / 0.09 = 35.7, so about 36 animals per group. This shows that binary outcomes need far larger samples than continuous measurements, and where possible a continuous outcome (for example bacterial load) is more efficient than survival or no survival.

5.5 More than two groups

For one-way ANOVA, sample size depends on the effect size f (Cohen's f), with 0.1, 0.25 and 0.4 conventionally called small, medium and large [14]. Free software such as G*Power can calculate it [17]. When many pairwise comparisons are planned, the significance level should be adjusted (for example with Tukey or Bonferroni methods) and sample size increased accordingly.

5.6 Count data

Counts such as colonies, plaques or reads often follow a Poisson distribution, where the variance equals the mean. In practice, biological counts are usually over dispersed (variance greater than mean), and the negative binomial distribution or quasi-Poisson models are more suitable. Simulation-based power analysis for generalized linear mixed models is available, for example the SIMR package in R [18]. For simple comparisons, log transformation of counts (log10 of count plus a small constant when zeros occur) and the normal-based formula above often give a reasonable first estimate.

5.7 Adjusting for dropouts and practical limits

The calculated n should be inflated for expected losses. If 10% of units may be lost to contamination, spoiled plates or animal death, the planned number is n / (1 - 0.10). Where only a very small n is possible, the study should be presented honestly as exploratory, and effect sizes with confidence intervals should be reported rather than relying on p-values alone.

5.8 When there is no pilot data

If the SD is unknown, options include: a small pilot (at least 5 to 6 units per group, remembering that its SD is itself uncertain), use of published values from similar systems, or use of conventional effect sizes [14,15]. A sensitivity analysis, which shows the effect size detectable with the sample size that is feasible, is a good alternative.

Table 2. Approximate sample size per group
(two-sided alpha = 0.05, 80% power, normal approximation)

Effect size d / SD	n per group
0.5 (small-medium)	63
0.8 (large)	25
1.0	16
1.5	7

Effect size d / SD	n per group
2.0	4

These values follow from the formula in Section 5.3 (n = 15.7 / (d/SD) ^2) and should be rounded up.

6. Domain-Specific Considerations in Microbiology

6.1 Plate counts

Plate counts are subject to sampling error that follows approximately a Poisson distribution. The relative error of a count of N colonies is about 1 over the square root of N, so a plate with 30 colonies has a coefficient of variation of about 18%, and a plate with 300 colonies about 6%. This is part of the reason that plates with roughly 30 to 300 colonies are preferred for counting, although the exact range depends on the method and standard used [19,20]. Counts should be log-transformed before applying t-tests or ANOVA, since bacterial growth is multiplicative. Serial dilution errors also accumulate, so independent dilution series for each biological replicate are preferable to repeated plating from a single series.

6.2 Antimicrobial susceptibility and MIC data

MIC values are obtained in doubling dilutions, so they are ordinal and discrete rather than continuous, and the usual variability is plus or minus one dilution. MIC data are usually log2-transformed for analysis. Large numbers of isolates are needed to characterize wild-type MIC distributions and to estimate epidemiological cutoff values, and statistical methods for this purpose have been described [21]. For comparisons of resistance rates between groups, the proportion formula in Section 5.4 applies, and isolates from the same patient or outbreak should not be treated as independent.

6.3 Disk diffusion and zone-of-inhibition assays

Zone diameters are continuous and generally suitable for ANOVA. The plate (not the disk) is often the experimental unit when several disks are placed on one plate, so plate effects should be handled using blocks or a mixed model. Agar depth, inoculum density and incubation conditions must be standardized because they affect zone size.

6.4 Animal infection models

Animal studies need ethical justification of numbers (replacement, reduction and refinement). The animal or the cage is the experimental unit, depending on how treatment is applied; if treatment is given in the drinking water of a cage, the cage is the unit. Randomization, blinding and pre-specified endpoints are essential, and the ARRIVE 2.0 guidelines give a complete reporting list [22,23]. Using bacterial load as a continuous outcome reduces the number of animals needed compared with using survival alone.

6.5 Microbiome and metagenomic studies

Microbiome studies pose special problems: data are compositional (relative abundances sum to a constant), sparse (many zeros), high-dimensional and strongly affected by batch and contamination [6,7,24]. Key points are:

- **Sample size:** The true unit is the individual, site or environmental sample, not the number of sequencing reads. Deep sequencing of few samples does not make up for few biological replicates.

Power calculation methods exist for community-level tests such as PERMANOVA [25,26] and for taxon-based comparisons [27]. Because power depends strongly on effect size, the within-group dispersion and the diversity metric used, calculations should use pilot or public data from a similar sample type.

- **Confounding:** Diet, age, medication, geography and sampling method strongly influence microbial communities. Cases and controls should be matched or the variables recorded and adjusted for.
- **Batch effects:** DNA extraction kit, sequencing run and operator can be larger than the biological effect. Samples from different groups should be randomized across extraction batches and sequencing runs, and not processed group by group.
- **Contamination:** In low-biomass samples, contaminant DNA from reagents can dominate the signal. Negative controls (extraction blanks and no-template controls) should be included in every batch and sequenced, and contaminants can be identified statistically [10,11,28].
- **Analysis:** Methods that respect compositionality, such as log-ratio approaches and ANCOM, are recommended, and the choice of method affects results [24,29,30].

6.6 Quantitative PCR and molecular assays

Technical replicates of qPCR wells (usually two or three) only improve measurement precision. Biological replicates determine sample size. Standard curves, no-template controls and inhibition controls should be included, and reference genes should be validated for the condition tested.

7. A Practical Workflow for Planning an Experiment

1. **State a clear, testable hypothesis** and define the primary outcome (for example log10 CFU/mL at 24 hours).
2. **Identify the experimental unit** and the level at which treatments will be applied.
3. **Choose the design** (Table 1) and decide on blocks, controls and blinding.
4. **Decide the smallest biologically meaningful effect** and obtain an estimate of variability from pilot data or literature.
5. **Calculate the sample size** with a stated alpha, power and test; increase it for expected losses.
6. **Randomize** allocation, positions and order of processing, and record the method used.
7. **Write an analysis plan before collecting data**, including transformations, tests, handling of outliers and correction for multiple comparisons.
8. **Run the experiment** with standardized protocols and complete records.
9. **Analyse and report** effect sizes with confidence intervals, not p-values alone, and state the exact n and what it represents.
10. **Repeat independently** to confirm that the result is reproducible.

8. Common Mistakes and How to Avoid Them

Mistake	Consequence	Remedy
Treating technical replicates as n	False positives	Define the unit; use nested or mixed models
No power calculation	Missed or exaggerated	Justify n before the experiment

Mistake	Consequence	Remedy
	effects	
No randomization	Hidden bias	Randomize positions and order
One experiment, no repeat	Unknown reproducibility	Repeat independently
Changing several factors at once	Cannot assign cause	Use factorial design
Ignoring batch effects	Confounded results	Block or randomize across batches
Multiple tests without correction	Inflated Type I error	Adjust for multiplicity
Choosing the test after seeing data	Biased results	Pre-specify the analysis
Reporting only p-values	Loss of information	Report effect size and confidence interval

9. Reporting and Reproducibility

Adequate reporting allows others to judge and repeat the work. Authors should state: the experimental unit and exact n per group; the way treatments were allocated and whether randomization and blinding were used; how sample size was determined; the statistical methods, software and version; criteria for excluding data; and the numbers of technical and biological replicates and independent experiments. For animal work, ARRIVE 2.0 should be followed [22]. For microbiome work, data and code should be shared, and controls reported [6,24]. Better design and open reporting are among the most effective ways of improving reproducibility in microbiology [24,31].

10. Conclusions

Good experimental design is the cheapest way to improve the quality of microbiological research. The essential steps are simple: identify the true experimental unit, replicate it adequately, randomize, block known sources of variation, include proper controls, and justify the sample size before collecting data. Different microbiological data types need different methods, but the same logic applies to all of them. Researchers who plan the number and structure of their experiments in advance will produce results that are more reliable, more reproducible and easier to publish. Journals, reviewers and teachers can help by asking authors to state what n represents and how it was chosen.

REFERENCES:

1. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013; 14:365-376.
2. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124.
3. Hurlbert SH. Pseudoreplication and the design of ecological field experiments. Ecol Monogr. 1984;54(2):187-211.
4. Lazic SE. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? BMC Neurosci. 2010; 11:5.

5. Lazic SE, Clarke-Williams CJ, Munafo MR. What exactly is 'N' in cell culture and animal experiments? PLoS Biol. 2018;16(4): e2005282.
6. Knight R, Vrbanac A, Taylor BC, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018; 16:410-422.
7. Schloss PD. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. mBio. 2018;9(3): e00525-18.
8. Fisher RA. The Design of Experiments. Edinburgh: Oliver and Boyd; 1935.
9. Montgomery DC. Design and Analysis of Experiments. 9th ed. Hoboken: Wiley; 2017.
10. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014; 12:87.
11. Eisenhofer R, Minich JJ, Marotz C, et al. Contamination in low microbial biomass microbiome studies: issues and recommendations. Trends Microbiol. 2019;27(2):105-117.
12. Box GEP, Hunter WG, Hunter JS. Statistics for Experimenters. 2nd ed. Hoboken: Wiley; 2005.
13. Czitrom V. One-factor-at-a-time versus designed experiments. Am Stat. 1999;53(2):126-131.
14. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale: Lawrence Erlbaum; 1988.
15. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials. 1981;2(2):93-113.
16. Festing MFW, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J. 2002;43(4):244-258.
17. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39(2):175-191.
18. Green P, MacLeod CJ. SIMR: an R package for power analysis of generalized linear mixed models by simulation. Methods Ecol Evol. 2016;7(4):493-498.
19. Tomasiewicz DM, Hotchkiss DK, Reinbold GW, Read RB, Hartman PA. The most suitable number of colonies on plates for counting. J Food Prot. 1980;43(4):282-286.
20. International Organization for Standardization. ISO 7218: Microbiology of food and animal feeding stuffs: general requirements and guidance for microbiological examinations. Geneva: ISO; 2007.
21. Turnidge J, Kahlmeter G, Kronvall G. Statistical characterisation of bacterial wild-type MIC value distributions and the determination of epidemiological cut-off values. Clin Microbiol Infect. 2006;12(5):418-425.
22. Percie du Sert N, Hurst V, Ahluwalia A, et al. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. PLoS Biol. 2020;18(7):e3000410.
23. Russell WMS, Burch RL. The Principles of Humane Experimental Technique. London: Methuen; 1959.
24. Weiss S, Xu ZZ, Peddada S, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5:27.
25. Kelly BJ, Gross R, Bittinger K, et al. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. Bioinformatics. 2015;31(15):2461-2468.

26. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 2001;26(1):32-46.
27. La Rosa PS, Brooks JP, Deych E, et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PLoS ONE. 2012;7(12):e52078.
28. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. 2018;6:226.
29. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8:2224.
30. Mandal S, Van Treuren W, White RA, et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis. 2015;26:27663.
31. Julious SA. Sample Sizes for Clinical Trials. Boca Raton: CRC Press; 2010.