Department of Primary Industries
and Regional Development

Animal Research Review Panel Guideline 32

Guideline for use by Animal Ethics Committees on the Determination of Animal Numbers and Experimental Design

Introduction

August 2024

Animals are used in scientific research across a wide range of fields from basic and preclinical science to wildlife ecology and veterinary studies. However, their usage has faced challenges concerning reproducibility, replicability, and the broader applicability of findings. These issues are complex, with numerous contributing factors (1). Key elements in addressing these challenges are the careful determination of study design and appropriate sample sizes. Accurate sample size calculation is crucial, and the use of statistical methods, such as power calculations, is increasingly being recognised as essential in the planning stages of research. The adoption of these statistical approaches not only improves the reliability and validity of research outcomes but also supports ethical and responsible use of animals across diverse scientific domains.

The determination of the total number of animals required for a study is crucial in adhering to the 3Rs principles and poses a challenge for Animal Ethics Committees (AECs). Approving too few animals could result in inconclusive results, potentially necessitating the use of more animals in follow-up studies. Conversely, approving an excessive number of animals could lead to unnecessary wastage. It is important to recognise that there may not be a single, universally accurate method for precisely calculating the number of animals needed for each experiment. However, it is imperative that investigators provide a comprehensive justification regarding how animal numbers were decided on to enable AECs to make informed decisions when reviewing applications.

Following the Animal Research Review Panel (ARRP) webinar on statistics (*August 9, 2022 https://www.dpi.nsw.gov.au/dpi/animals/animal-ethics-infolink/arrp-webinar-seminars/arrp-webinar-series/2022-webinar-series*), guidance in understanding methods for calculating the number of animals required for a research project was identified. This guideline then, aims to enhance the ability of AECs in Australia to effectively assess and validate the determination and justification of the animal numbers proposed for research projects.

A complete guide to appropriate methods for determining animal numbers is outside the scope of this guideline. For in-depth information, readers are encouraged to consult resources such as the freely available Experimental Design Assistant from the National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs) UK (2) or other relevant literature (3-5). Consultation with biostatisticians is also recommended for both researchers and committee members if such a resource is available to them. This guideline will however, provide an introduction to key elements that investigators should include in justification of animal numbers, enabling AECs to make informed decisions.

The Australian Code and ARRIVE Guidelines:

The determination of animal numbers and proper experimental design are intrinsically linked. Consequently, accurate determination of animal numbers is addressed in at least 4 sections of *The Australian Code for the Care and use of Animals for Scientific Purposes* (6) these include but are not limited to:

1.1 (v)(b) Respect for animals must underpin all decisions and actions involving the care and use of animals for scientific purposes. This respect is demonstrated by the reduction in the number of animals used.

1.5 (iii) Evidence to support a case to use animals must demonstrate that the project involves the minimum number of animals required to obtain valid data.

1.15 Regardless of the potential benefits of a project, the methods used must be scientifically valid, feasible, well designed and carefully conducted so that there is a reasonable expectation that the aims of the project will be achieved. Projects that are not scientifically valid must not be performed, no matter how mild the impact on the wellbeing of the animals.

1.21 The number of animals used in a project must be the minimum necessary to achieve the proposed aim(s) and to satisfy good statistical design. The use of too few animals may invalidate the experimental result and result in wastage of animals.

Additionally, the NHMRC recommend Institutions, AECs and investigators consider the Animal Research: Reporting of *In Vivo* Experiments (ARRIVE) guidelines (7) in their practical strategies to enhance best practice in the use of animals for scientific purposes (1; https://arriveguidelines.org/). Specific sections of the ARRIVE guidelines relevant to determination of animal numbers include but are not limited to:

Methods for determining animal numbers:

When deciding on the number of animals needed for a study, researchers generally use one of three methods. (3-5, 8). Each method has its advantages and disadvantages, as is the case for any type of statistical analysis and experimental design. The key point is that the best method depends upon the specific design and goals of the study as well as any prior knowledge from previous studies.

As such, it is the responsibility of the investigator to explain clearly which method they are using and why. This clarity helps the AEC to evaluate the scientific validity and need for the number of animals requested. The three commonly utilised methods are listed here and described in more detail below:

- Power Calculations: Mainly for hypothesis testing, these calculations require specifying elements such as effect size, variability, significance level, statistical power, and the direction of the effect. They're ideal for studies testing a specific biological effect of an intervention.

- Resource Method: Suitable for exploratory studies without clear primary outcomes, this method focuses on the number of animals, treatment questions, environmental factors, and a measurement of error, following a simple formula.
- Preliminary studies or Systematic Reviews of prior literature: Typically used when entering new research areas, or significant advances have been made in a pre-existing area. This approach relies on reviewing existing literature to justify the number of animals needed. It's crucial for researchers to provide a robust rationale based on systematic reviews or relevant studies when other methods aren't applicable.

Power Calculations

Power calculations are a statistical technique increasingly being recommended for determining the necessary sample size in research studies to ensure the ethical and effective use of animals (1, 7). Their most appropriate use is for hypothesis testing study designs which aim to assess the specific effect of an intervention on a predetermined primary outcome. “Power” in this context refers to the likelihood that a statistical test will correctly identify a true biological effect.

To conduct a power calculation researchers need to define at least six critical parameters of their proposed study. These include:

- The primary outcome. This is the outcome of greatest importance and that which the entire study will be powered upon.
- The mean and standard deviation (or variability) of the primary outcome for both the intervention and comparator group.
- The effect size, which is the expected difference in outcomes between the control and intervention group (often measured as a mean score difference). This is either calculated from pilot data or estimated from prior studies.
- The significance level (also referred to as *Alpha*), which represents the chance of a false positive result. There is no standard convention for what value is used for alpha. This is typically set at 5% or 0.05 and represents a 1 in 20 probability that a positive result is false. Alpha can range between 1% and 10% (or 0.01 and 0.10).
- Power (also referred to as *Beta*) is defined as the probability of obtaining a true positive (an outcome that is genuinely statistically significant). There is no standard convention for what level of power should be chosen, however studies will typically be considered adequately powered at a level of 0.8. A high-powered study would have a value of 0.95 and as such power is usually set between 80-95% or 0.8 – 0.95.
- Direction of effect (also referred to as tail) and indicates whether or not the intervention is expected to increase or decrease the outcome.

An example of how these parameters may be used in a simple two group design is outlined in Appendix 1.

Resource Method

The resource method (9) is a relatively straightforward method for determining the number of animals needed in exploratory and hypothesis generating studies. This method is particularly useful when there's limited or no prior information about the effect size or variability of the outcome in the study. Such studies typically do not have a single primary outcome but instead aim to measure multiple outcomes. It is important to note that the resource method is not a commonly used or published method and owing to its simplicity is not considered as robust as a conventional power calculation.

To apply the Resource Method, researchers need to consider 4 key parameters:

- The number of animals (N): This is the total number of animals planned to be used in the experiment.
- Treatment component (T): This refers to the number of different treatments or questions the study intends to explore.
- Blocking component (B): This accounts for the number of environmental or experimental variables that might influence the study and need to be controlled for.
- The Error of measurement component (E): This is a measure of the variability or uncertainty in each experiment. It is proposed that for any given experiment, the E should reside between 10 and 20.

In its simplest form, the formula for the Resource Method is expressed as $E = N – B – T$, which helps in balancing the studies scientific need with practical resources, enabling researchers to systematically estimate an appropriate number of animals for their study, considering various factors impacting the experiment.

A significant limitation of the resource method is that it is largely based on a provisional estimate by the researcher on the number of animals required and an explicit description of how the study will be implemented to account for environmental factors that may impact the final error estimate. For example sex, strain, and housing (single or group). An explicit description of the experimental unit determining the N is also required. As such, while the formula appears simple, the level of diligence on the investigator to ensure all appropriate blocking factors and treatment factors have been accounted for is high. In this instance committee members must be provided with these details clearly and any justification for the numbers requested should include a description of how they have been incorporated into the ultimate sample size requested.

An example of how these parameters may be used in a simple design is outlined in Appendix 2. References 4, 5 and 8 provide further examples.

Preliminary Studies and Systematic Reviews of Literature

These methods are used for determining sample size, especially in new areas of research where research teams might not have direct experience with the specific intervention being studied. In such cases, it is not considered sufficient to estimate the number of animals needed based on experiences with different experiments or interventions. Instead, researchers are encouraged to rigorously review prior published work on similar or closely related interventions and outcomes and present this information to the AEC. Systematic Reviews are particularly valuable in this context. They involve a clearly articulated search strategy, with clear criteria for including and excluding studies, along with assessment of potential biases in the studies reviewed. This approach is considered the most thorough for identifying relevant existing published data.

If investigators choose not to use Power Calculations or the Resource Method for sample size determination, they need to provide AECs with a solid rationale based on their literature review. This justification should detail how previous studies influenced the design and size of the proposed study. It is crucial to understand that merely citing past observations of statistically significant results does not suffice as evidence of a well-planned and valid sample size determination.

Determining animal numbers and the interaction with study design:

Good ethical review of research applications must rigorously evaluate the scientific merit which includes a well-defined scientific method encompassing valid hypotheses, aims, unbiased experimental design and analysis. However, as identified by the NHMRC, studies using animals for scientific purposes often fail to meet minimum standards (1).

The method for determining animal numbers is intrinsically connected to the study design. When reviewing applications, AECs should consider whether the project is hypothesis generating or hypothesis testing. If the latter, researchers should clearly specify a primary outcome on which all statistical analyses, species selection, and sample size calculations are based. Additionally, any secondary or confounding outcomes to be assessed should be identified. A crucial, yet often overlooked parameter in animal studies is the definition of the experimental unit. This unit could be the individual animal, or in some cases, the cage. For example, in a dietary study with animals housed in groups, the consumption measured per cage should be considered as a single experimental unit, not multiple units based on the number of animals.

Some studies might aim to develop a new surgical procedure or be for teaching and training purposes rather than testing a specific biological outcome. Here, sample size requested may be better determined on resources and technical requirements or in the case of educational activities, number of students in a course. Regardless, the rationale for the requested number of animals should always be clearly communicated to the AEC. If there's any uncertainty, AECs should seek additional information for clarity before approving any applications.

The NC3Rs offers a free and valuable tool, the Experimental Design Assistant (EDA; https://eda.nc3rs.org.uk/) that aids researchers in experiment design, including statistical analysis guidance. The EDA also provides informative fact sheets on key aspects of good scientific methodology, like sample size determination and analysis.

Recommendations:

To ensure clarity, detail, and evidence-based justification in the process of requesting animal numbers as the foundation for ethical and scientific integrity in study design, we recommend the following key considerations for AECs:

1. Study design and Primary Outcome: Investigators must specify whether their study is hypothesis-testing, and if so, clearly define the primary outcome that will guide all sample size determination and statistical analyses.
2. Details in Power Calculations: When using power calculations, investigators are required to provide detailed information on each specific parameter used in determining the sample size. These details should be sufficient to allow the committee to verify and understand the basis of the calculations.
3. Justification for not using a Power Calculation. If a power calculation is not used, investigators must explain why it was not appropriate for their study. This explanation

should reference the study design and the alternative method used to determine the sample size.

4. Transparent justification of animal numbers. Regardless of the method used for determining the number of animals, investigators should provide clear, evidence-based justifications to the AEC. This is essential to ensure the AEC can make informed decisions on the appropriateness of the requested number.

Additionally, an optional checklist for review of applications is provided in Appendix 3.

References:

1. National Health and Medical Research Council (NHMRC) 2017, Best practice methodology in the use of animals for scientific purposes. (www.nhmrc.gov.au/guidelines-publications/ea20)
2. NC3Rs UK, Experimental Design Assistant (https://eda.nc3rs.org.uk/)
3. Percie du Sert N, Ahluwalia A, Alam S, Avey MT et. al. (2020). Reporting animal research: Explanation and elaboration to the ARRIVE guidelines 2.0. PLoS Biol 18(7): e3000411. https://doi.org/10.1371/journal.pbio.3000411
4. Gaskill BN and Garner JP. (2020). Power to the people: power, negative results and sample size. Journal of the American Association for Laboratory Animal Science. 59(1): 9-16. DOI: 10.30802/AALAS-JAALAS-19-000042
5. Charan J and Kantharia ND. (2013). How to calculate sample size in animal studies? J Pharmacol Pharmacother. 4(4): 303-306. DOI: 10.4103/0976-500X.119726.
6. National Health and Medical Research Council (2013) Australian code for the care and use of animals for scientific purposes 8th edition. Canberra: National Health and Medical Research Council.
7. Percie du Sert N, Ahluwalia A, Alam S, Avey MT et. al. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. PLoS Biol 18(7): e3000410. https://doi.org/10.1371/journal.pbio.3000410
8. Festing MF and Altman DG. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J. 43:244-258.
9. Mead, R. (1988). The design of experiments, Statistical principles for practical applications. Cambridge: Cambridge University Press. 620p.

Appendix 1:

What does a power calculation used to determine sample size look like?

Power calculations can change depending on the design of an experiment and the number of groups being compared. The example below identifies a simple 2 group design in which there is a control group and an intervention group for a study that aims to determine if a single intervention (for example injection of a drug every day for 2 weeks) changes body weight.

In order to perform power calculations, investigators can access a number of free online software packages. One of the more commonly accessed packages is called G*Power; . (https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower) With further resources and manuals available here: https://stats.oarc.ucla.edu/other/gpower/

Another freely accessible online resource is called Statulator; (https://statulator.com/) with a blog and additional support available: https://www.statulator.com/blog/

To our scenario, in order to determine how many animals are needed, investigators need to identify:

- The primary outcome: In this scenario is body weight.
- Mean and standard deviation: The researchers would need to provide a value for the average body weight of the control group with standard deviation as well as a value for the average body weight of the intervention group and standard deviation after being injected with the drug every day for 2 weeks. Using these values, an effect size can be calculated either manually or using a statistical software package.
- Effect size:
    - There are typically three main sources from which effect size can be calculated:

        1. Previously completed studies in which the same drug was injected into a similar species of animal.

Or if pre-existing data does not exist

2. A systematic review of published literature of similar design and purpose as the proposed study.
3. An expected or predicted effect based on the knowledge of the investigator.

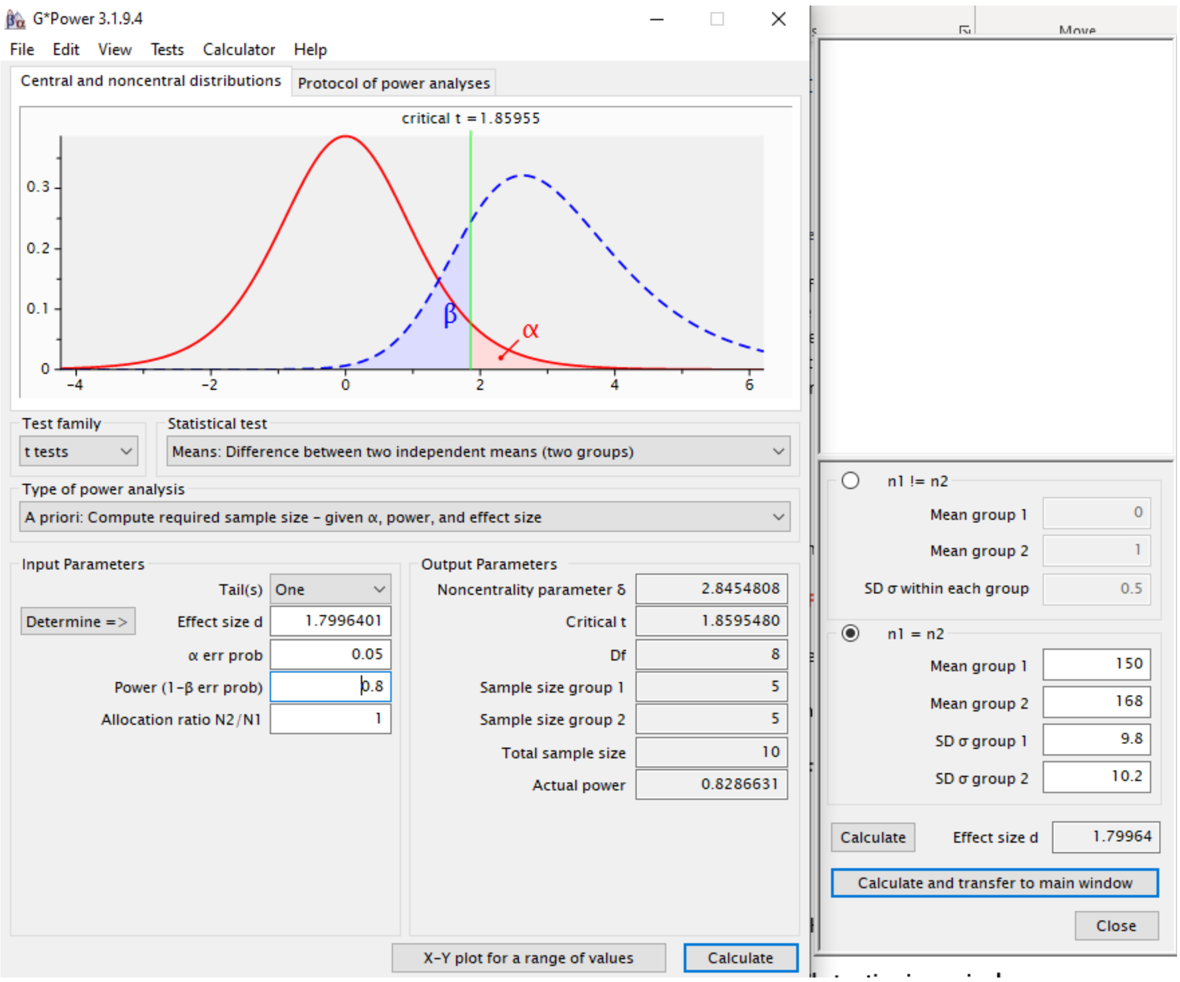
It should be noted that power calculations that do not use an effect size calculated from previous studies or published literature are the weakest and most prone to bias and should be scrutinised more thoroughly by committee members.

Effect size values are typically identified as small, medium or large and the values that represent these categories will vary depending on the study design. In our simple example here, effect size values that represent small, medium or large effect are 0.2, 0.5 and 0.8 respectively. However, the effect size is a consequence of the main effect of the intervention and the consistency of that effect in all animals tested. As such, if the intervention has a large and consistent impact on the outcome then the effect size will be large and may reach values greater than 1.

- The significance level (Alpha): Alpha is fixed to a probability of a 1 in 20 chance that the result is a false positive and as such is given the value of 0.05.
- Power (Beta): The researchers will need to indicate what they have defined as the probability of obtaining a true positive and is usually set between 80-95% or 0.8 – 0.95. Power in this scenario will be 0.80
- Direction of effect (tail): The investigator needs to determine if the change in body weight was expected or if it was unknown. If the assumption is that body weight is expected to change (either increase or decrease) then the power analysis is said to be 1 tailed. If the direction of the effect of change on body weight was unknown, then it would be a 2 tailed test.

In our scenario, the researchers have completed a pilot study that has previously injected the drug in question into a similar species of animal. That pilot study reported average body weight of the control group at the end of the two weeks was 150g with a standard deviation of 9.8 while the intervention group had an average body weight of 168g with a standard deviation of 10.2.

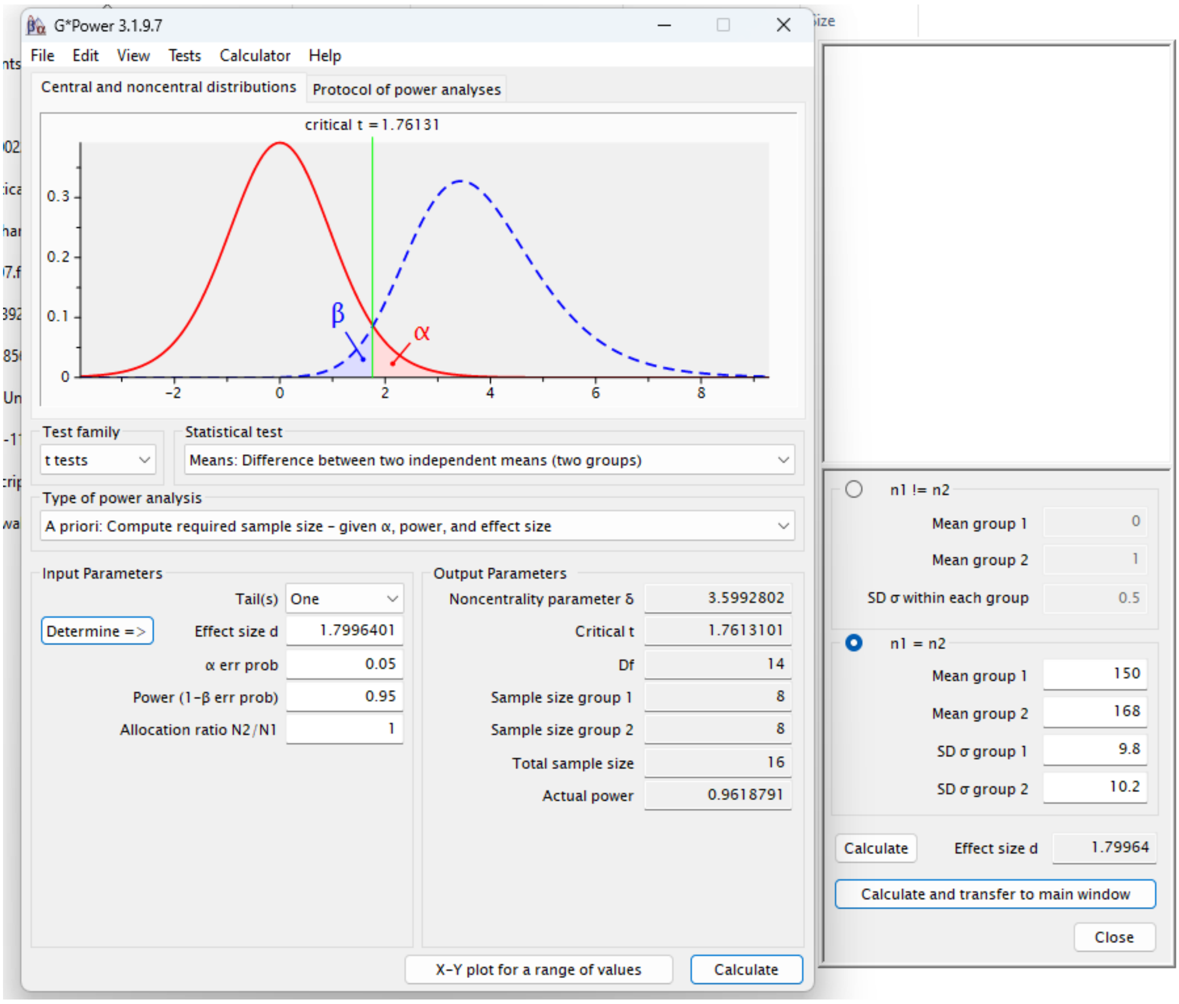
Using the online G*power tool set, the investigators use these numbers to assess the effect size of difference between two groups. It is calculated as 1.799. This is a very high ES (Figure 1). This value is then used to determine the sample size using all of this information. That is, they decide to set an alpha of 0.05, a power value of 0.8 and given we know weight is expected to increase (from the pilot data) a one-tailed t-test. The result produced by the software is that 5 animals are required in each group (control and intervention).



Three examples of how ethical decision making may influence sample size:

Example 1: adjusting power level:

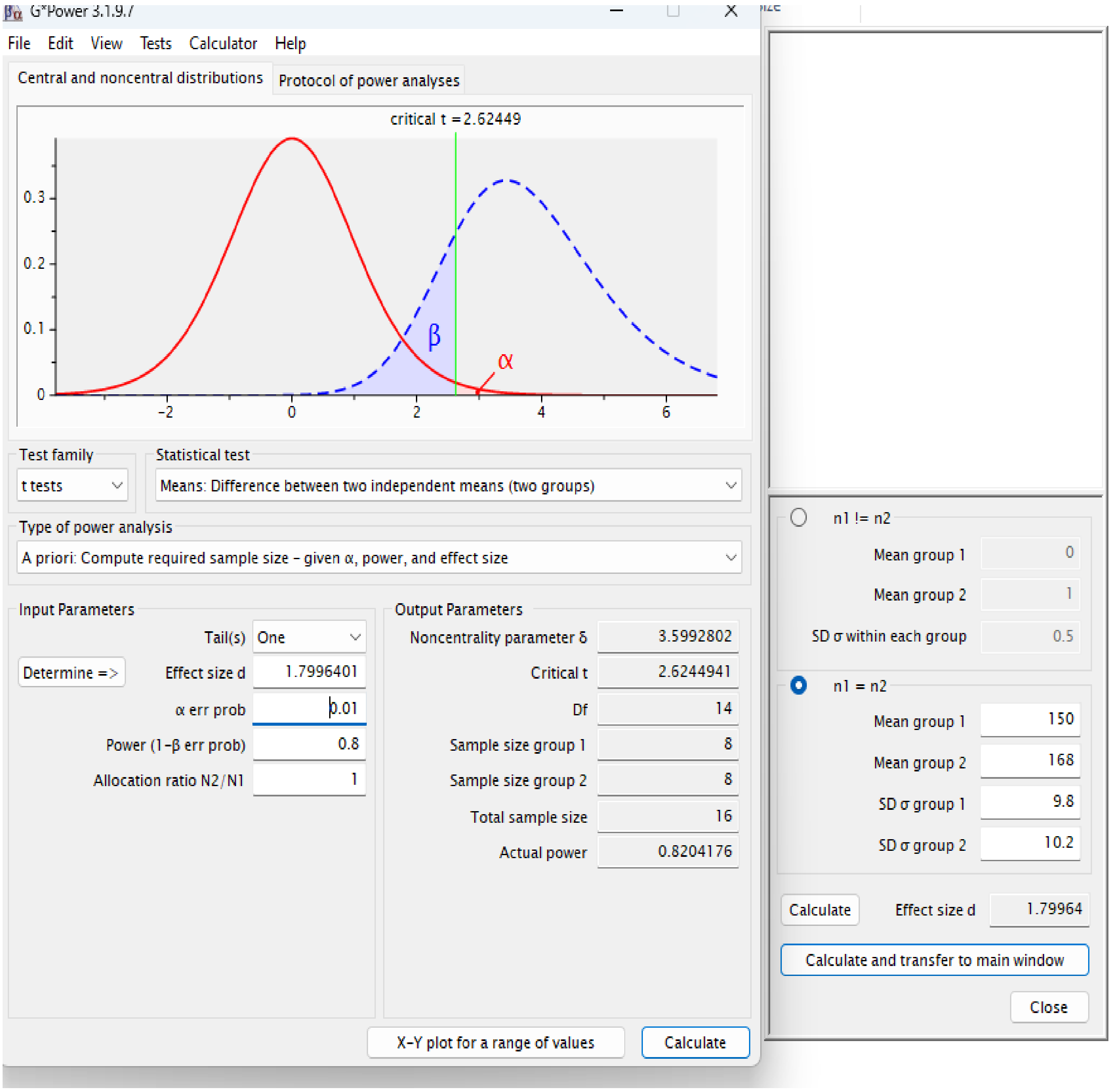
Let's assume the committee does not accept 80 % power to be adequate for the purposes of ensuring scientific validity in keeping with the code. They recommend a power level of 95%. In this instance, using the exact same data for Primary Outcome; Mean and Standard Deviation; Effect Size; and Alpha; however adjusting power to 0.95, the new sample size requested would be 8 animals per group and as such the researcher would require an additional 6 animals for a total sample of 16



Example 1 Figure

Example 2: adjusting confidence (alpha) level:

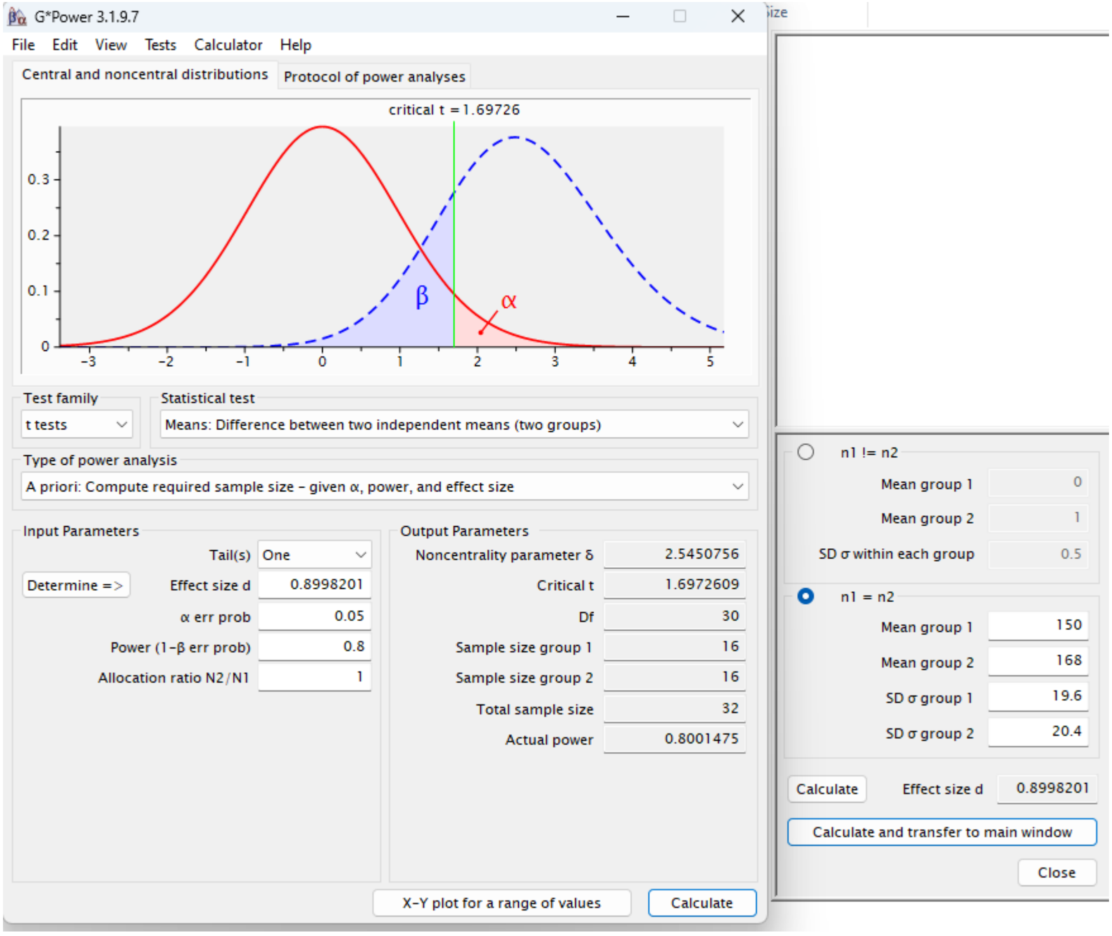
Let's assume the committee was accepting of a power level at 0.8, however does not accept that a significance level of 0.05 (1 in 20 chance of rejecting a false positive) to be adequate for the purposes of ensuring scientific validity in keeping with the code. They instead would like the confidence of a 1 in 100 chance of there being a false positive result. And in this case recommend an alpha level of 0.01. In this instance, using the exact same data for Primary Outcome; Mean and Standard Deviation; Effect Size; and power as in the original sample size calculation; however now adjusting alpha to 0.01 instead of 0.05, the new sample size requested would be 8 animals per group and as such the researcher would require an additional 6 animals for a total sample of 16.



Example 2 figure

<u>Example 3: adjusting Effect Size:</u>

Recall that Effect size is determined based on the mean and standard deviation of the primary outcome. As such, if your primary outcome is highly variable either within a group of animals or between groups of animals (intervention and control) the ES can be dramatically differently. As an example, let's assume the body weight of animals in our scenario is more variable than initially reported. And in fact, the body weight of our two groups in the pilot study are 150 ± 19.6g and 168 ± 20.4g (not 150 ± 9.8g and 168 ± 10.2g as in the original scenario). The ES calculated is now only 0.8998 (where it was originally 1.7996). If we now re-run the sample size calculation as we did originally (Alpha is 0.05; and Power is 0.8) the sample size required is now 16 animals per group. This is more than triple the original number of animals requested.



Example 3 Figure

These three alternate scenarios in which the number of animals requested is variable, despite using the exact same pilot data should highlight to committees that the reliance on power calculations to confidently justify animal numbers is dependent on the confidence of the decisions made by investigators on their level of significance and power, and the accuracy of their a priori data to determine ES. As such, committee members should interrogate these justifications to ensure confidence in the sample size requested.

Appendix 2:

What does the resource method look like?

Resource equations are best utilised in exploratory studies or those where power calculation may not be practical. It is important to note that the resource method is not a commonly used or published method and owing to its simplicity is not considered as robust as a conventional power calculation. The equation seeks to determine an error component depending on the number of treatment and environmental factors that need to be considered. The example below identifies a relatively simple 4 group design in which the amount of sugar consumption on body weight is investigated.

In this scenario, investigators have proposed a study in which there are 4 groups of animals (1 control and 3 treatment groups each receiving the same diet that differs only in increasing amounts of sugar) for the purposes of determining the minimum amount of sugar consumption that increases body weight. The investigators have proposed that they require 8 animals per group because that is what they have always used and have provided a review of the literature to confirm that approximately 8 animals per group is standard for publication. As such they seek approval to purchase 32 animals. They also propose to purchase all of the animals at once and complete the experiment in one go, in one strain of mouse and only in females.

For the resource equation we need to identify the:

- Number of animals (N) – proposed by the researchers.
- Treatment component (T) – The number of groups in this instance.
- Blocking components to consider (B) – environmental influences.
- And the error of measurement component (E) – our indicator of whether or not the proposed number of animals is appropriate.

In this scenario for the resource equation looks like this:

The number of animals (N) = 32

The treatment component (T) = 4

The blocking component (B) = 1 (being that the study is being completed in one go and there are no strain or sex differences to correct for)

The resource equation requires these to be represented as “degrees of freedom” which is an adjustment of the subtraction of 1 unit. As such the units for the equation are now N = 31; B = 0; and T = 3.

Using the formula E = N-B-T

E = 31 – 0 – 3**(28).**

Mead recommends that E should be between 10-20 and as such this proposal would be overpowered for their aims. It may be that a sample size of only 6 animals per group will suffice as this would adjust E to 23 – 0 – 3**(20)** and within the upper limit.

An alternative approach may have been to complete this experiment in 8 blocks using 4 animals at a time (1 from each group) over 8 subsequent cohorts. In this instance there is an environmental factor that could influence the design (8 repeats). N remains as 31; T remains as 3; however, B is now 8 and correcting for degrees of freedom 8-1 = 7. The new formula is now

E = 31 – 7 – 3 **(21).** And as such the original request for 32 animals is appropriate.

APPENDIX 3:

CHECKLIST FOR AECs

There is no one uniformly accepted method for determining sample size in animal studies. However, despite the variable methods, there are some core common components that should appear in any attempt by investigators to justify the number of animals they propose for their study. The following is a guide of some of the elements committee members may look for in applications. If the answer to any question is "NO" then investigators should be asked to provide the information.

1. Has the study identified a design type (e.g. exploratory; hypothesis testing; or observational; with or without a control group)
2. For either study type, has a primary outcome upon which the significance, design and implementation of the study been justified? Note. This is essential for hypothesis testing studies.
3. Has an estimated sample size been requested using any of the 3 methods described in this guide?
4. For studies utilising power calculations:
    a. Have each of the 6 critical parameters listed in this guide been articulated?
    b. Has the mean and standard deviation, effect size and assumption of direction been based on animal data from a similar age, sex, and strain?
        i. If not, has a reasonable attempt been made to justify similarity between the data used and the animal requested?
    c. Has the experimental unit for analysis of the primary outcome been articulated?
5. For studies utilising the resource method:
    a. Has an estimated sample size per group been provided?
        i. If so, is the estimation appropriately justified?
    b. Have all treatment groups been articulated?
    c. Has a comprehensive list of potential environmental / blocking factors been considered (e.g. sex, strain, housing condition etc.)?
6. For studies using preliminary studies and systematic reviews:
    a. If presenting a review, was it conducted in a systematic approach?
        i. If not, have steps regarding how the investigator ensured their justification based on the totality of evidence available and not just a selection of supporting studies been provided?
    b. Does the literature provided investigate the same primary outcome as that proposed?
        i. If not, has an adequate attempt to justify why the literature used is relevant been made?
    c. Does the literature provided investigate the same primary outcome as that proposed, in a similar animal model to that requested?

    i. If not, has an adequate attempt to justify why the literature used is relevant been made?

7. Does the committee member feel that the justification for the number of animals requested is in keeping with the Code and/or ARRIVE Guidelines?

Published by the Department of Primary Industries and Regional Development

Title: Guideline for use by Animal Ethics Committees on the Determination of Animal Numbers and Experimental Design

First published: August 2024

Department reference number: INT24/80144

© State of New South Wales through DPIRD 2024. The information contained in this publication is based on knowledge and understanding at the time of writing August 2024. However, because of advances in knowledge, users are reminded of the need to ensure that the information upon which they rely is up to date and to check the currency of the information with the appropriate officer of the DPIRD or the user's independent adviser.