Statistical Justification of Animal Numbers Used in Research and Teaching

By Dr. Shelly Naud, Research Biostatistician, COM Biostatistics Unit, UVM

<u>Overview:</u>

- Federal Mandate
- Writing a Rationale Justifying Animal Numbers: Studies not requiring a statistical justification
  - Examples of language justifying sample size for non-statistical experiments
- Writing a Rationale Justifying Animal Numbers: Studies Requiring a Statistical Justification Sample Size or Power Calculation)
  - Examples of sample size justifications
  - Example of the decision making process for determining the sample size
  - Concepts underlying sample size calculations
- Repeating experiments
- Advanced topics and special issues
- Resources

Federal Mandate:

A key principle governing the ethical use of animals in research, testing and teaching is that no animal life is wasted; the number of animals used in each project must be the minimum necessary to obtain valid and meaningful results. (1)

By Federal regulation, the IACUC must review the number of animals requested in each protocol and agree that the number is appropriately justified in terms of the stated goals of the project. (2), (3)

Researchers are required to consider the replacement of animal models (in vitro, computer models, etc.) to accomplish the objectives of his/her proposed research study. When it is determined that reasonable alternatives cannot replace the use of in vivo models, the IACUC is responsible for ensuring that investigators have adequately assessed their study requirements to ensure that the number of animals are appropriate to accomplish the proposed research objectives.

Specifically, the investigator must demonstrate that he/she is requesting the minimum number of animals necessary (avoid overpowering the study) while requesting an adequate number of animals (avoid underpowering the study) to provide data that will be relevant to fulfill his/her experimental objectives without unnecessary repetition of the study. (4)

References

1. Public Health Service. (1996) U.S. Government Principles for the Utilization and Care of Vertebrate Animals Used in Testing, Research and training. PHS Policy on Humane Care and Use of Laboratory Animals. Washington, D.C.
2. National Research Council. (1996) *Guide for the Care and Use of Laboratory Animals.* National Academy Press, Washington, D.C.
3. CFR (Code of Federal Regulations) (1985) Title 9 (Animals and Animal Products), Subchapter A (Animal Welfare). Washington, D.C.: Office of the Federal Register.
4. Laboratory Animal Sciences Program, Center for Cancer Research, National Cancer Institute

Writing a Rationale Justifying Animal Numbers: Studies not requiring a statistical justification

All studies where there are statistical inferences being made (i.e., are generalizing to a population rather than reporting descriptive statistics) are expected to provide a sample size analysis as described in the next section. However, there are many types of studies where this approach is not appropriate. The justification of animal numbers is done differently depending on the type of study. A protocol with multiple steps or experiments should include a justification for each sample of animals.

- *Teaching:* Use of animals in conjunction with a course for the purpose of teaching students. Numbers are determined by a specified student-to-animal ratio. The choice of the specified ratio must be explained in the justification statement. (see example 1)
- Field studies.
- Breeding protocols to maintain a certain strain of animals available for future research activities. The PI should provide information on how many animals are required to keep the strain of animal available. Protocols that include breeding as part of a larger research objective may need to provide statistical justification for the research objective in addition to the justification for the number of animals required for maintaining the breeding stock.
- *Antibody/tissue production:* Use of animals to produce antibodies or tissue. Numbers are determined by the amount of antibody/tissue required, the ability of an individual animal to provide the needed amount, etc. Provide details on the determining factors in the justification statement. (Example 2)
- *Studies with success/failure as the outcome:* Use of animals to demonstrate success or failure of a desired goal (e.g., production of transgenic animals, to determine whether a poisonous plant induces cleft palate in grazer species). Numbers are determined by the probability of success of the experimental procedure. Provide details in the justification statement.
- *Feasibility or pilot study:* Use of animals in a preliminary study to refine procedures and equipment and to discover problems before the main study begins, in which any results of interest will be verified by additional inferential studies. When any statistical analysis is done, it is primarily descriptive or exploratory in nature. Numbers are determined by the investigator using experience and judgment; generally the number is relatively small but should be large enough to provide needed estimates for future sample size analyses. (Example 3)
- *Exploratory studies:* Performed to generate new hypotheses (e.g., carrying out multiple genetic or biochemistry assays without a driving hypothesis). Often many variables are considered. These are often secondary aims in studies. The sample size is usually based on previous experience or, optimally, will make use of the same animals used for the primary aims by collecting additional data.

Sources for text:
iacuc.usu.edu/files/uploads/animal_numbers.doc
http://www.unco.edu/osp/ETHICS/iacuc/CornellJustNumbers.pdf
ILAR

Examples of language justifying sample size for non-statistical experiments

1. Example of teaching protocol (numbers based on students):
“We plan to have up to six 3-person teams of students per lab section. We plan to have two lab sections for one semester (Spring Semester) each year. We will design these labs to enable each team to observe up to 6 rats per lab for 15 minute observational period. (When the PI executed this type of 3-hr student lab at a previous institution, he found than when the lab is carefully planned, teams of students can observed up to 6 rats, each for 18-minute observation periods.) Thus we will require 72 rats per Spring Semester when the class is scheduled.”

2. Example based on cell yield:

“The number of animals to be used in each experiment is based on the expected number of different cells which can be isolated from individual animals and has been published by this laboratory [Reference cited.]. The cell populations and anticipated cell yield/mouse from heart are as follows: γδ+ cells, 3 x 105 cells; CD4+ T cells, 6 x 106 cells; and bone marrow cells, 9 x 107 cells.
“Numbers of mice required/group is determined by the formula: #cells needed for an experiment/# cells which can be obtained per mouse. The experiments should need to be done only once as the statistics will be obtained from replicate samples in tissue culture. It may be necessary to divide experiments in half (use half of the required number of mice at one time for one set of in vitro experiments and the other half at a second time for the remaining experiments) because of the extensive amount of in vitro studies which are planned.”

3. Example of a pilot study:

“We are requesting 30 animals for the initial (developmental) phase of this study. The first 15 wil be used to work out the details of the procedure for inducing (variable #1); once this procedure has been established, the next 15 animals will be used to (examine a range of the variable – a “dose-response curve.”) These numbers are estimates since we are trying to develop a new procedure, and can not begin this process until this protocol has been approved, hence, we have no preliminary data with which to work. As indicated in the protocol, we will use cadaver specimens to determine an anticipated range for the measurement.”

Writing a Rationale Justifying Animal Numbers:

Studies Requiring a Statistical Justification (Sample Size or Power Calculation)

The objective of a power analysis is to estimate the minimum number of animals needed to reliably detect the expected effect size. Too often studies are underpowered. In one review it was found that 70% of the studies with negative results had sample sizes too small to detect a 50% improvement. Using too few animals is as wasteful as using too many.

Saying that a previous study found statistically significant differences with a certain number of animals is not an acceptable rationale. The information from the previous study, however, should be used in carrying out a sample size analysis. Please include references for these studies in your IACUC protocol. When a protocol describes multiple experiments with different outcomes, a sample size analysis should be carried out for each one. When there are several related experiments planned, use the data from the earlier experiments for making better sample size estimates for later experiments. For the IACUC justification, provide a rationale for the first set of experiments.

For more complex designs, consult a statistician.

Elements of a sample size analysis. These should be included in the rationale. If the terminology is not familiar or if the relationship with sample size is not understood, then please refer to the separate section: Concepts underlying sample size calculations.

- The level of significance/ alpha/ Type I error. The alpha is usually set at 5% or 1%.
- Power or Type II error/ beta. Typically the power is set to 80%. In some fields, the power needs to be set higher.
- Outcome, study design, and statistic. It should be clear what is the outcome and whether it is categorical, ordinal (e.g. rating scales), or continuous. Is the outcome measured more than once? How many groups are there? Based on this information, a statistical test is chosen.
- Effect size or alternative hypothesis. How large of a difference does the researcher hope to be able to detect? There are specific formulas for calculating the effect size depending on the statistic of interest. Once the effect size is estimated, the sample size can be determined from available tables or using specialized software. This is the most problematic aspect of calculating a sample size since it is not generally known at the start of a study and must be guesstimated based on available data. There are two acceptable strategies.

    - Pilot study or results from a comparable study
    - Minimal clinical relevance

Note that the sample size should adjust for the anticipated animal losses.

Strategies to increase power and/or minimize the number of animals.

- Increase alpha. In the preliminary or exploratory stages of a set of experiments, it can be justifiable to use .10 instead of .05.
- Appropriate variables. Categorical data need larger sample sizes than continuous outcomes.
- Covariates. Including covariates that explain some of the variation in the outcome can increase power. See references for possible covariates. Important ones to consider are: baseline measures of the outcome, sex, age, litter, researcher carrying out the measurements.

- Improving measurement precision. Reducing the outcome variable's variability increases the effect size, and consequently power.
- Pre-post designs control for inter-individual variability. Crossover designs do the same by having all animals get both treatments, although at different times.
- Longitudinal studies. Increasing the number of times measurements are made can help improve power, however, the rate of improvement in power drops off after 4 measures (Zhang & Ahn).
- Certain statistical approaches are more flexible and powerful than repeated measures, such as ANOVA, growth curves, multilevel regression, time to reach a peak or area under the curve.

Examples of sample size justifications

"Fourteen (14) rats per group are required based on the following power calculation and our estimation of attrition due to surgical failure. We have conducted similar studies in the past investigating the effects of these lesions on stress-induced weight change. Using the means from this study (31.89 for control/sham, 31.14 for control/lesion, 3.25 for stress/sham and 15.07 for stress lesion with an average standard deviation of 7.397), we calculated an effect size of 1.565. Hence, conducting a power analysis (using GPOWER software) for a one-way ANOVA with this effect size revealed a sample size of 12 per experimental group. This value is consistent with numbers that I have used in the past for similar studies. Because rats will undergo surgery to implant bilateral guide cannulae aimed at a specific anatomic area of the brain, we anticipate that as many as 15% of rats will have one or both cannulae misplaced (outside the desired structure). Thus, we are requested 2 additional rats per group for these experiments. Thus, 14 rats per group are requested for all experiments. A description of the numbers of animals used for each experiment is shown in section D3."

"A total of 16 male minipigs will be needed. Data will be analyzed using a one way ANOVA with a type i error level of 5%. We assumed that (measurement) results 9 weeks after (treatment) in untreated controls will be comparable to the results in untreated animals previously studied by (Principal Investigator), i.e., .27:1.06. An improvement in (variable) of 0.10 is considered a moderate improvement and would be considered clinically significant. Based on those data to detect a .10 improvement in (variable) in the treatment group compared with the shams with a power of90% will require 7 animals per group. We have requested 2 additional animals due to the potential loss due to intra-operative complications."

Example of the decision making process for determining the sample size

Results of urine analyses in 16-week old rats (means and standard errors with 8 rats per group)

	Normal	Impaired kidneys	S.D. estimate (based on SE from published study)	Effect size= 75% of difference/ s.d.
Urine Volume	71 ± 8	58 ± 6	7 * SQRT(8)= 20	0.65
Na+ excretion	3.7 ± 0.6	4.8 ± 0.4	0.5 * SQRT(8)= 1.4	0.78
Creatine clearance	60 ± 1	51 ± 2	1.5 * SQRT(8)= 4.2	

The researcher would like to evaluate a drug's impact on impaired kidneys' function, specifically he would like to see an improvement in diuretic and natriuretic effects without a further reduction in the glomerular filtration rate (evaluated indirectly by measuring creatinine). For the power calculation, the following was decided.

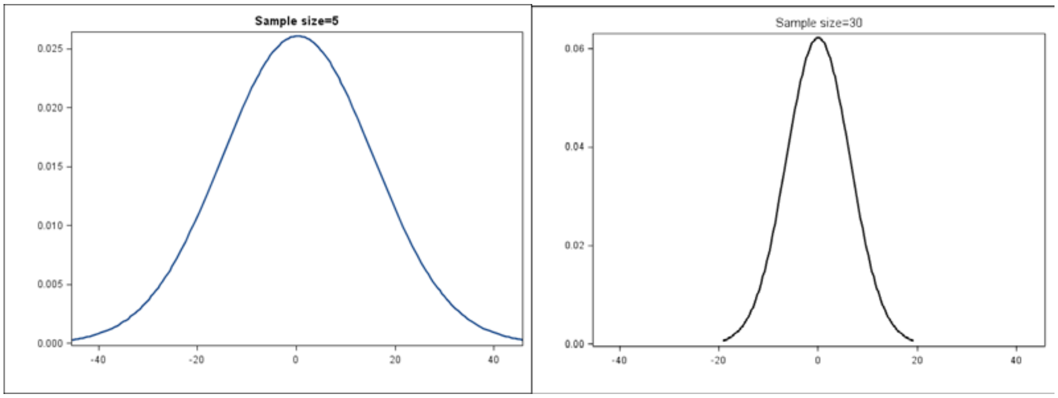
- The usual alpha (.05) and beta (.80) were accepted (elements 1 and 2 above).
- Using published data and a small pilot study, the expected effect sizes were estimated to be 75% of the difference between impaired and normal.
- The sample size was based on the smaller of the two effect sizes.
- No animal loss is expected.

(More details and website http://www.epibiostat.ucsf.edu/biostat/sampsize.html)

Concepts underlying sample size calculations

Inferential statistics involves generalizing from the experimental sample (or a specific result) to the population from which that sample was drawn (true value). Statistical theory describes the probability of making correct inferences.

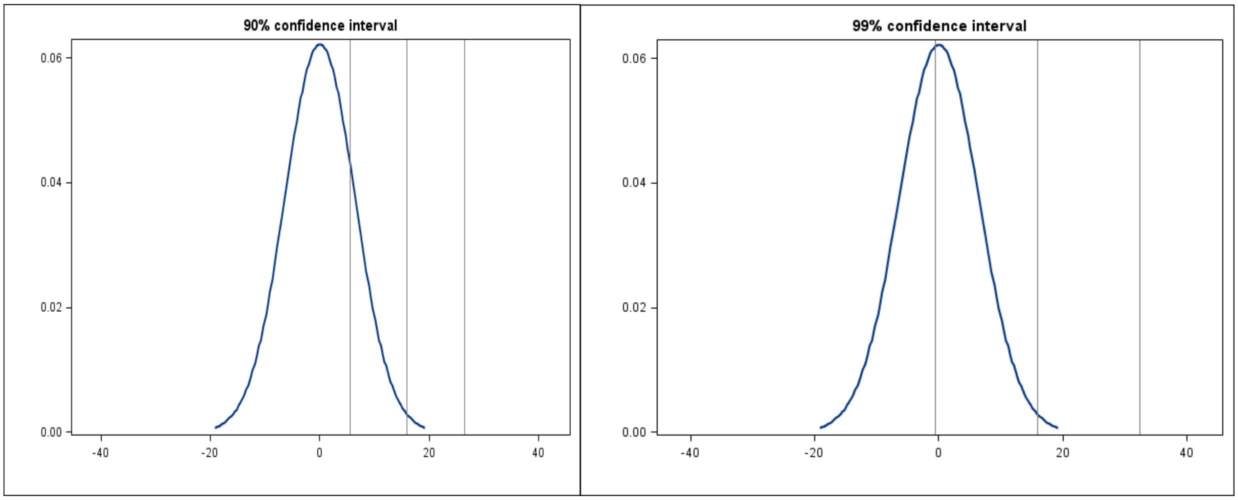
For example, the mean of a sample (X) will estimate the true or population mean but it is not expected to be exactly equal to the population mean (μ). The central limit theorem states that the means (X) of all possible samples with sample size n, as long as n is sufficiently large, will be normally distributed with a mean of μ and a standard error of s/n -1/2. Below are two distributions of sample means (X) from the same population: the sample sizes, however, differ.



What is important to notice is that as the sample size increases, the mean does not change. The precision of estimating the population mean, however, improves.

Alpha (level of significance), confidence interval

Given that a sample is not likely to estimate the population mean exactly, it is good practice to provide a range of plausible values. The width of the range or confidence interval is based on the significance level chosen. If alpha = .10, then one reports the range corresponding to a 90% confidence interval (the sample mean ±1.645 * the estimate of the standard error, i.e., the standard deviation of the distribution of sample means). There is a 90% probability that the true mean falls within the confidence interval. Inversely, there is a 10% probability that the population mean is not included in the confidence interval. Choosing an alpha of .01 will yield a confidence interval that is wider and there will be only 1% chance of excluding the population mean.



Increasing the sample size doesn't change the probability of making an incorrect inference given the same alpha, but the precision will change.

Alpha versus p value

The alpha is the criteria set a priori to conducting the experiment. It defines the cut-off outside of which the statistical results will be considered significant. The p value is the probability of obtaining ≥ the actual result under the assumption that the null hypothesis is true. The smaller the p value, the more confident a researcher can be that the decision to reject the null hypothesis is correct.

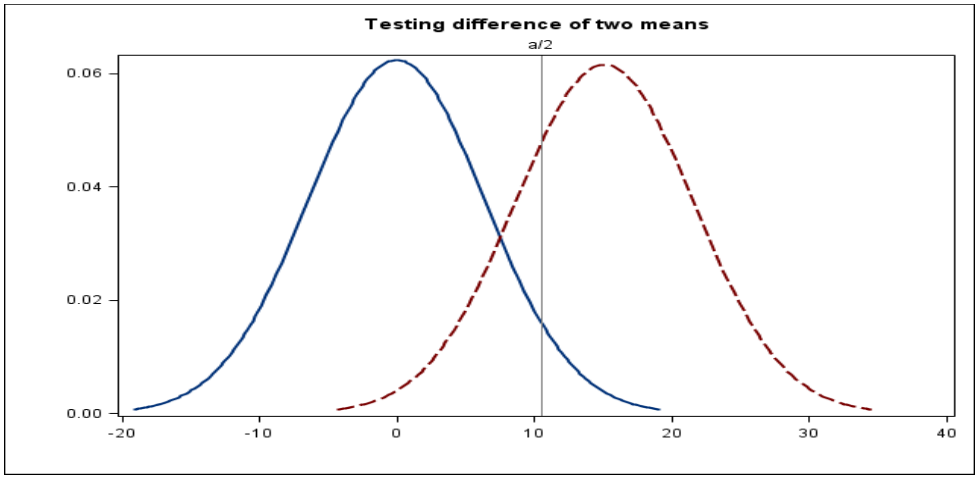
Type I error for a single sample test

In many single sample experiments the null hypothesis is that the population mean is 0. Another frequent single sample test evaluates the difference from a hypothesized mean (the sample mean – hypothesized mean = 0). The probability of rejecting the null hypothesis when it is correct, known as a Type I error, is equal to the level of significance chosen.

The alternative hypothesis and Type II error

In order to be able to estimate power, an alternative hypothesis has to be defined. The researcher uses the alternative hypothesis to specify the minimum difference that s/he wants to be able to detect. In the illustration below the alternative hypothesis is the curve with the expected mean of 15. Thus, the minimum difference that is scientifically relevant in this example is 15.

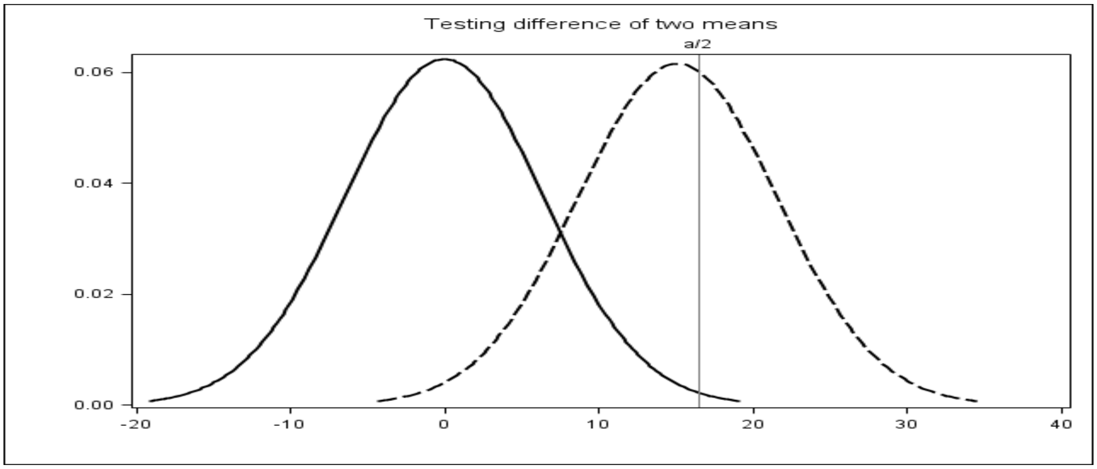
Given a two-sided alpha of .10, the confidence interval for accepting the null is ±10.53; i.e., the null hypothesis will be rejected if the experiment finds a difference > |10.53|. The probability of correctly rejecting the null hypothesis when the alternative hypothesis is true is the area of the curve above alpha/2. This is referred to the statistical test's power. In the illustration below, there is 75% power to correctly reject the null hypothesis if the alternative hypothesis is true. The Type II error corresponds to the area of the dotted curve that falls below alpha. This area of the curve is referred to as beta. There is a 25% probability of failing to reject the null hypothesis when the alternative hypothesis is correct.



These concepts are often summarized in a table as below.

True mean of the population		Result of statistical test
The null hypothesis is correct	The alternative hypothesis is correct	
Correct inference Probability = 1 - α	Type II error Probability = β	**Accept null hypothesis**
Type I error Probability = α Probability = α	Correct inference Probability = 1 - β Power of a test	**Reject null hypothesis**

Alpha and beta have an inverse relationship. When alpha is decreased, beta increases. To compensate for loss of power, the sample size will need to be increased. Using the same example as above, the graph below shows what happens when you decrease alpha to .01. Note that this shifts the area under the curve that corresponds to alpha/2, with the resulting increase of beta and decrease in power. Type II error increases to over 50%.



Alpha and beta are typically set to .05 and .20 respectively. Mathematically, this is saying that a Type I error is four times more serious than a Type II error. This appears to be horrible, but the estimated power should be for the *minimum* difference that one hopes to be able to detect. The actual difference might be larger. In the very early stages of exploratory analyses, it might be better to have greater power and set a higher significance level (e.g., .10). That would limit the probability of rejecting a line of research that could turn out to be promising without increasing the sample size, but at the cost of increasing the chance of finding something significant that really isn't. Setting the level of significance at .10 is not appropriate for an experiment that is meant for publication.

We can only calculate the hypothetical power of a test since the actual mean of the population is unknown – otherwise there wouldn't really be a need to carry out the experiment. In the example above, the true mean could be 20, therefore the power of the statistical test will be greater than the estimated power.

Repeating experiments

Terminology.

The debate about the need for repeating studies has been confused to a certain extent by the use of terminology that can have different meanings. Following are the definitions of the terminology as used in this section.

Replication has a very specific meaning in the context of experimental design and statistics. It is the repetition of a basic experiment when ***one*** experimental unit is assigned to each of the treatment conditions (experimental groups). Its purpose is to estimate variability and increase precision. No statistical analyses can be done without replicates.

Each experimental unit should be independent and randomized to the treatment conditions. It can be a single animal, a patch of skin, a litter. Repeated measures of the same experimental unit at the same time and place are not replicates.

Repeating an experiment refers to what many are calling experimental replication. It refers to carrying out an experiment without any modifications to the experimental design or methods. The results are considered independent and analyzed separately.

The reproducibility of an experiment is an important part of the confirmation of the validity of its outcomes, but ideally it would occur in a totally different setting. It does not require an exact adherence to the original procedures. It is testing or extending the finding.

The Three Rs of Animal Experimentation

Government agencies, the Office of Laboratory Animal Welfare, IACUC committees, and most scientific journals adhere to the principles set by Russell and Burch in 1959.

- Replace: methods such as mathematical models, computer simulation, and in vitro biological systems should be considered.
- Refine: avoidance or minimization of discomfort, distress, and pain when consistent with sound scientific practices, is imperative.
- Reduce: select the minimum number of animals required to obtain valid results.

Any policy that indiscriminately requires researchers to repeat their experiments is in violation of the principle of reduction. Are there journals that are in violation of these principals? Fitts (2011) examined the Instructions for Authors of the biomedical journals with high impact factor and summarizes:

> …no journal required a certain number of replications. All journals require a valid statistical argument, and some state that it is desirable to replicate the results. Generally the journal's reviewers, not the journal themselves, insist on and perpetuate the 'three times' rule. Is it a good rule? …. When the data are quantitative and are associated with probability values for type I and type II errors for each test, the 'three times' rule can be excessive.

Several researchers have mentioned the editorial written by Heather Van Epps. The first heading is telling: "One mouse ≠ one experiment." Further on: "… a single, independent experiment [is defined] as one in which experimental and control groups (comprising individual mice, culture wells, etc.) are tested contemporaneously to answer a specific question. Each independent experiment must be repeated a sufficient number of times to demonstrate the reproducibility of the data." The author seems to be most concerned about experiments without any replication (single mouse experiments), but then jumps to making a rather confusing argument for either increasing the number of replications within a study or to repeating studies. A study can have several "experiments" done over time and combined for a single analysis. This can be considered as sound advice but the indiscriminate repeating of studies that are analyzed separately cannot be justified

Currently, a more serious problem in studies utilizing laboratory animals is lack of sufficient power. The emphasis on repeating studies will acerbate rather than help solve this problem.

Finding a balance

The crux of the situation can be summarized as follows: routinely repeating experiments is a waste of animals; judiciously extending prior findings to prove its validity and/or generalizability is good science.

The significance of the results of the original experiment should be taken into consideration (Fitts 2011):

- If a study's p-value is < .005, it is likely to be significant if repeated. A repetition for the sole purpose of testing its reproducibility is not necessary.
- On the other hand, if the study's p-value is approximately .05, it's a good idea to consider reproducing it. The sample size will need to be carefully considered since there is only a 50-50 chance of getting a significant result.

A researcher may consider "repeating" an experiment if there is reason to believe that there are factors that were not controlled during the experiment that impacted the outcome of the study (e.g., questionable vaccine) or limits its generalizability (e.g., dose or mouse strain). These experiments, however, can probably define the factor(s) that will be different and can be considered as extensions.

Certain experiments that do not have statistical results (e.g., success/failure as the outcome) should probably be repeated.

In general, if multiple experiments are done, the results should be combined in a single statistical analysis.

The IACUC's objective is to do good science in an ethical manner.

References:

Festing MFW, Overend P, Gaines Das R, Cortina Borja M, Berdoy M. 2002. The Design of Animal Experiments: Reducing the Use of Animals in Research Through Better Experimental Design. London: Royal Society of Medicine Press Limited.

Fitts DA. 2011. Ethics and animal numbers: Informal analyses, uncertain sample sizes, inefficient replications, and type I errors. Journal of the American Association for Laboratory Animal Science, 50(4), 445-453.

Van Epps HL. 2009. JEM's 2009 Tune-up. JEMS, 2006(5), 968-969.

Advanced Topics & Special Issues

(a) One-sided alpha. If the alternative hypothesis states that the experimental group will differ in a specific direction (e.g., the treatment group's mean will be at least ½ standard deviation higher than the control group), then the study can be powered for a one-sided test.

(b) Study design.

(c) Longitudinal. Increasing the number of times measurements are made can help improve power. The rate of improvement in power drops off after 4 measures (Zhang & Ahn). Certain statistical approaches are more flexible and powerful than repeated measures ANOVA: growth curves, multilevel regression, time to reach a peak or area under the curve.

(d) Pre-post designs control for inter-individual variability. Crossover designs do the same by having all animals get both treatments, although at different times.

(e) Factorial design. Simpler factorial designs can evaluate whether two treatments have a synergistic effect or a whether a treatment interacts with an independent variable (e.g., sex and drug). More complex designs (i.e., fractional designs) can assess whether any of a number of factors influences the outcome (sex, strain, diet, method of measuring the endpoint). Factorial designs are much more efficient than using a separate experiment testing each factor separately.

(f) Multiple-stage design. Statistical tests are done at fixed intervals to determine whether the experiment should be continued or not. The usual version is to test for both futility and success at each interim step. If the effect size is much higher or lower than expected, the experiment can be terminated early. Another version is to only test only for futility at the interim steps. This approach can save considerable time and minimize the number of animals. It's appropriate when there is little or questionable preliminary data for calculating the effect size.

(g) Sequential sampling. This is a more flexible alternative to multiple-stage design. Reference: http://www.nal.usda.gov/awic/newsletters/v7n1/7n1chamo.htm

(h) There are more powerful alternative methods for the median lethal dose (LD_{50}) test. Reference: http://dels-old.nas.edu/ilar_n/ilarjournal/43_4/v4304Rispin.pdf

(i) Controls for multiple experiments.

(j) Specifying the correct experimental unit. When the treatment is given to a group, for example, a new diet to all animals in a cage or to a pregnant animal with weight of newborns being the outcome, then the experimental unit is a group and not the individual animals. The statistical analysis should reflect this by nesting the group. This does not necessarily increase the number of animals needed. In some cases, the same animal can be used for more than one test, e.g., compounds tested on different patches of skin. In this case, the experimental unit would be the patches of skin but the analysis would still block on the animal. Check references ILAR p 204 Quality control techniques can monitor

(k) Adjusting alpha for multiple tests. A common statistical error that has been found in the published literature is in the analyses of multiple groups. It is inappropriate to do multiple pair-wise comparisons (e.g., two t-tests when there are three groups) without adjusting the alpha. Another strategy that results in more power is to do a one-way ANOVA and follow-up with LSD post-hoc tests if the overall test is significant. Most basic statistical textbooks cover this issue.

(l) Non-normality and outliers. Many of the best-known statistical methods are only appropriate when the data is normally distributed (bell-shaped curve) and no extreme outliers are present. If these assumptions are violated, the statistical result may be biased. It is possible for a single outcome to determine whether a statistical test is significant or not! Alternatives are non-parametric statistics which do not have the same assumptions, transformation of the data (e.g., log transformation), and/or sensitivity testing (redoing statistical tests without outliers).

(m) Known and unknown sources of variability. This is a vital part of any experiment. As much as possible all animals must be uniform (same strain, age, diet, housing). Bias due to uncontrolled sources of variability are offset by randomization or, where possible, measured and included in the statistical analysis as covariates. Blinding staff to treatment is recommended when possible. Measurement error should be minimized as much as possible. Transcription errors is another source of error. Raw data should be screened for errors. Double entry of data is recommended when creating a database for statistical analysis. When a single experiment is done over time (e.g., four animals per day over 8 days), a factorial design is recommended. Reference: Chapters 2, 3 & 4 of Festing et al. The Design of Experiments: Reducing the use of animals in research through better experimental design.

(n) Validity.

(o) Post-hoc power analysis. A researcher may be interested in evaluating the possibility that a prior negative result was due to lack of power, especially when there are conflicting results.

(p) Fixed sample size. When the number of available animals is pre-determined and beyond the control of the researcher, a power analysis can be substituted for a sample size analysis. Here also, there are two strategies. One is to calculate the power based on an estimate of the effect size. Another approach is to set the power level and to calculate the effect size. The researcher must then evaluate whether it is of clinical relevance.

(q) Quality control techniques. When similar experiments are performed repeatedly in the same laboratory, monitoring the outcomes continuously can raise flags when there is an unexpected result. REFERENCES:

   a. Hayashi M, Hasimoto S, Sakamoto Y, Hamada C, Sofuni T, Yoshimura I. 1994. Statistical analysis of data in mutagenicity assays: Rodent micronucleus assay. Environ Health Perspect 102 (Suppl 1): 49-52.
   b. Morrison V, Ashby J. 1995. High resolution rodent bone marrow micronucleus assays of 1,2-dimethylhydrazine: Implications of systemic toxicity and individual responders. Mutagenesis 10:129-135.

(r) Missing data. There are some alternatives to dropping an animal's results if there are missing data. If the study is longitudinal, newer statistical methodologies like multilevel regression can be used. If there are multiple correlated outcomes, imputation could be considered.

(s) Powering studies with aim of not rejecting the null hypothesis (equivalency studies).

(t) When a power analysis can't be done. Festing's book p 79: Resource equation method.