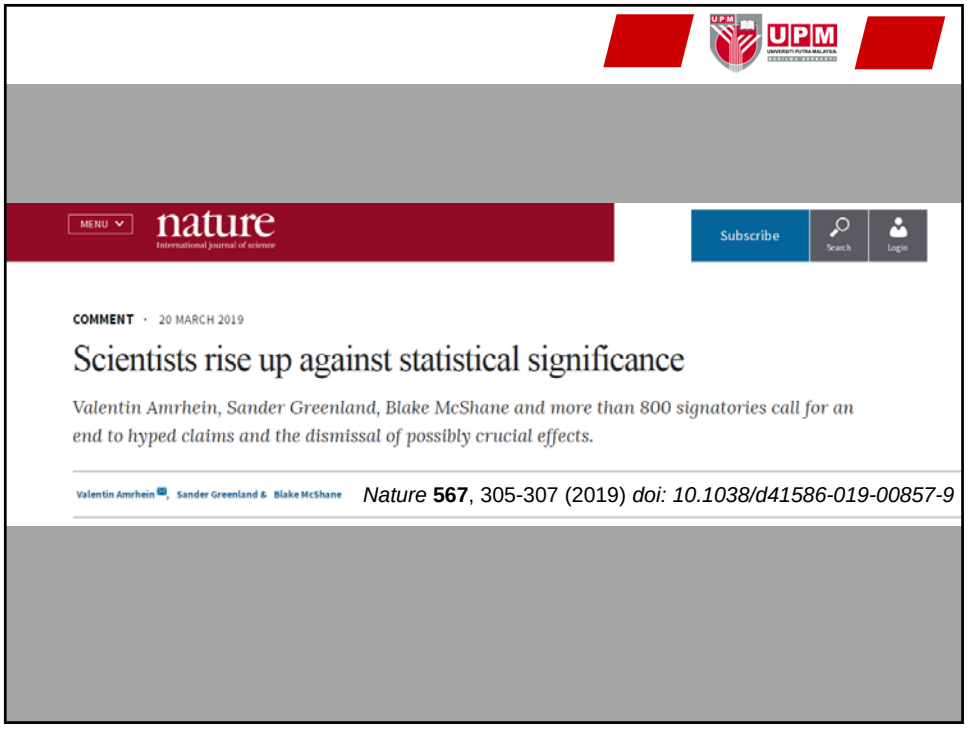
# Statistics Made Easy in Animal Experiments

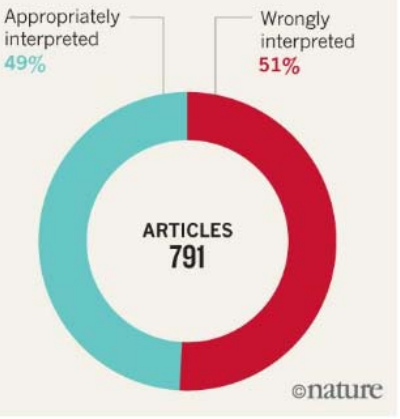
*Goh Yong Meng*
*Professor / Head of Department,*
*Department of Veterinary Preclinical Sciences,*
*Faculty of Veterinary Medicine,*
*Universiti Putra Malaysia.*

*ymgoh@upm.edu.my / gohyongmeng@gmail.com*



## STATISTICAL ABUSES & WRONG INTERPRETATIONS

An analysis of 791 articles across 5 journals* found that around half mistakenly assume non-significance means no effect.




*Data taken from: P. Schatz *et al. Arch. Clin. Neuropsychol.* **20**, 1053–1059 (2005); F. Fidler *et al. Conserv. Biol.* **20**, 1539–1544 (2006); R. Hoekstra *et al. Psychon. Bull. Rev.* **13**, 1033–1037 (2006); F. Bernardi *et al. Eur. Sociol. Rev.* **33**, 1–15 (2017).

Source: V. Amrhein *et al.* *Nature* **567**, 305-307 (2019) *doi: 10.1038/d41586-019-00857-9*

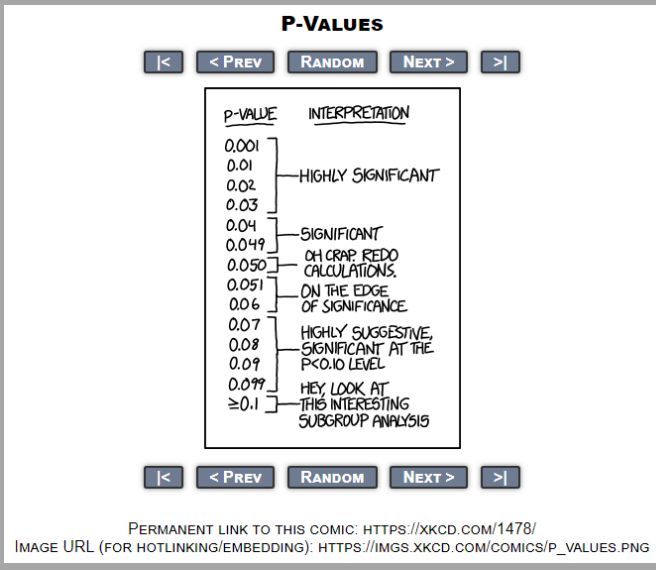
Remember "significance hunting"

***

**

*

in the 80's or even 90's ?



Flurry of responses in *Nature* to the proposal

consumption and disposal are not borne by chemicals producers, or shared down the value-chain" (see go.nature.com/2wonrvy).

Such costs should not be borne by taxpayers, the state or national treasury or by any other third party (see go.nature.com/2xzahnh). Rather, they should be met by producer industries to avoid market distortion.

**Pam Miller** *Alaska Community Action on Toxics, Anchorage, Alaska, USA.*
**Joe DiGangi** *International POPS Elimination Network, South Korea.*
*pamela@akaction.org*

**Retiring significance: a free pass to bias**

Statistical significance sets a convenient obstacle to unfounded claims. In my view, removing the obstacle (V. Amrhein *et al. Nature* **567**, 305–307; 2019) could promote bias. Irrefutable nonsense would rule.

More stringent thresholds of significance are needed for most fields, which currently assume statistical significance when *P* values are less than 0.05 (see, for example, D. J. Benjamin *et al. Nature Hum. Behav.* **2**, 6–10; 2018; J. P. A. Ioannidis *J. Am. Med. Assoc.* **319**, 1429–1430; 2018).

A company could, for example, claim that any results somehow support licensing of its product.

Careful thinking before a study starts should pick the best, fit-for-purpose statistical inference tool and pre-specify the rules of the game — whether frequentist, Bayesian or other. So, although the obstacle of statistical significance can be surmounted by trickery, removing it altogether is worse.

**John P. A. Ioannidis** *Stanford University, California, USA.*
*jioannid@stanford.edu*

**Retiring significance: raise the bar**

In my view, the proposal to retire statistical significance conflates two problems (V. Amrhein *et al. Nature* **567**, 305–307; 2019). These should be addressed separately.

One problem is the value of having a term that signifies whether an experiment provides evidence of an effect — that is, it achieves 'statistical significance'.

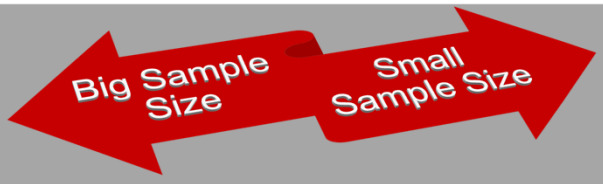
The second problem involves defining statistical significance as, say, $P<0.05$. Many scientists object to this threshold because it can prevent publication of experiments when $P>0.05$ (see, for example, D. Lakens *et al. Nature Hum. Behav.* **2**, 168–171; 2018). I have the opposite concern. Careful analysis of

**Retiring significance: keep hypothesis tests**

We agree that arbitrarily branding experimental findings as significant or non-significant generates a false sense of certainty (V. Amrhein *et al. Nature* **567**, 305–307; 2019). However, when done properly, hypothesis testing is an important precondition for estimating an effect size.

In his 1928 book *Statistical Methods for Research Workers*, British statistician Ronald Fisher remarked that "it is a useful preliminary before making a statistical estimate … to test if there is anything to justify estimation at all". And British polymath Harold Jeffreys declared in *Theory of Probability* in 1939 that "variation must be taken as random until there is positive evidence to the contrary". Hence, testing and estimation are complementary. Testing establishes whether there is an effect, and that helps to determine whether or not the magnitude needs to be estimated.

What happens when statistical testing is skipped and the null hypothesis is ignored? Well, noise would be interpreted as structural, and any differences between observations would be considered meaningful. Parameters would need to be estimated for all these differences, resulting in a "mere catalogue" of




- Choosing the sample size in a scientific experiment always involves balancing the increased information and precision that result from bigger samples and the reduced time and cost that result from smaller samples.
- Ethics and welfare requirements may further motivate reduction of the numbers of animals. We don't want unnecessarily many animals, but also need to avoid using too few so that the experiment is a waste of time. How do we choose the sample size? Using statistical & scientific approaches !
- However, specific testing techniques may require the use of certain sample size, e.g. in Toxicological testing as per OECD Guidelines

## Typical scenarios (too many ?)

***A animal science project requested for 500 day old chicks (DOC) to assess the effects of feeding 5 levels of probiotic on growth, meat quality and carcass composition.***

*Points to consider may include :*

- *Necessity/justification of the treatment levels used*
- *Distance/difference between the levels used (1 & 2 mg vs 100 & 200 mg) – **power analysis may be needed !***
- *Feasibility of handling 500 chicks,…. time taken for procedure, will the parameters of interest remain valid within the time difference between first and last animal ?*
- *Accepted practice for that field of expertise ?*

## Typical scenarios (too few ?)

***A theriogenology project requested for only 2 bulls, where semen will be sampled every two days for 30 days. Sperm will be used to investigate the efficacy of 3 new semen extenders under development***

*Points to consider may include :*

- *Reasons for using only 2 bulls ? Breeding bulls are expensive and hard to come by ?*
- *Issues of "pseudo replication" ? Sample size calculation may need to be based on **mixed model or repeated measure** assumptions.*
- *Will sperm quality change over the 30 day period ? length of spermatogenesis in bulls and their consistency ?*
- *Accepted practice for that field of expertise ?*

## Typical scenarios (sample size for control ?)

***A biomedical research requested 40 animals, in 4 equal groups of 10 animals each, to assess 2 active ingredients + 2 control groups under evaluation***

*Points to consider may include :*

- *Is it necessary for the control group to have the same sample size ? YES AND NO … Depends on "signal to noise" ratio….. **(see subsequent slides – slide 35)***
- *Is the "n" for control group justified in terms of mode(s) of action, and other grounds of comparison being considered.*
- *Sometimes, we may need to include MORE ANIMALS in the control group to improve the overall sensitivity of comparison (Bate and Karp, 2014 – published in Plos One, available at* http://dx.doi.org/10.1371/journal.pone.0114872 )...also from https://eda.nc3rs.org.uk/experimental-design-group

## Overview

*What is so great about statistics ? **NONE at ALL !***

*"Statistics never prove anything, they only indicate likelihood of the results of an investigation being the product of CHANCE"*

*It can be a really powerful SUMMARY TOOL linking all related variables/factors in a research to be considered together, to enable some conclusions to be made about them. Provided all experimental/statistical assumptions are CORRECT.*

## Overview

- The “n” is critical because :
  - determines the validity of a trial
  - economizing the research/time/labor cost
- But each statistical / experimental procedure has its own restrictions or rules on sample size ! No best approach to arrive at a viable sample size.
- In real life, our sample size is always restricted by

$$n = \frac{\text{RM available}}{\text{RM per sampling unit}}$$

## Overview

- Publications often do not tell you the success rate of a procedure !!
- Number of animals required may be HIGHER !
- E.g. Induction of Diabetes in Rats using STZ
  - At 40 mg/kg STZ, mortality rate can be 20 – 30 %, success rate about 75 %, effects may last only 6 weeks.
  - At 80 mg/kg STZ, mortality rate can be 50 – 60 %, success rate about > 85 %, longer effects but animal may get corneal opacity.
  - How should you calculate the STARTING sample size ?

## Overview

- Furthermore, ACUC forms often asks about

"What is the justification for the number of animals requested?"

- Justification must include :
  - ***a clear description of the experimental design*** *(not methods or procedures)*
  - ***specification of the number and species/strain of experimental/control animals per group/subgroup*** *in each experiment/procedure.*
  - ***The scientific/statistical rationale which supports*** *the size of the N in each group.*

This section will attempt to help you understand & recommend the best choice of sample size ……

## Sample Size

- Sample size depended on :
  - **Types of control used** (Own, Baseline, Parallel controls)
  - **Type of Experiments** (Pilot or Formal Testing of Hypothesis)
  - **Structure of the experimental design** (Groups and Replications)
  - **Magnitude of the observed effects** (smallest meaningful effect or minimum effect size)
  - **Precision of the measurement** methods and variation (verified by pilot trials)
  - **Statistical methods employed** (parametric vs non-parametric)
  - **Significance** (and confidence) level (95 % - 99%)
  - **Power of the tests** (or chance of correctly detecting a true treatment effect, usually 0.7 – 0.9 (70 - 90%) and above)
  - **Internationally agreed guides/standards (OECD guidelines) & conventions in generally accepted methodologies** (use published results from a similar experiments to back your claims)

## Sample Size : Types of Controls

- More groups = more animals
- Are those groups justified ?
- How critical is positive control, how critical are negative control, baseline, age controls……
- Using each animal as its own control (for before and after measurements), will cut down sample size but is this justifiable ?
- *Specific requirement(s) for control and types of control are sometimes specified for each respective categories of testing as per internationally agreed conventions (e.g. OECD guidelines).*

## Sample Size : Types of Experiment

- Examples for experiment based on success or failure of an expected outcome….
- Example, production of transgenic animals by gene insertion into fertilized eggs or embryonic stem cells.
- Why ?

  *- considerable variation in the proportion of successful gene or DNA incorporation into the cell's genome.*

  *- variability in the implantation of the transferred cell.*

  *- DNA integrates randomly into the genome.*

  *- Expression varies widely as a function of the integration site and transgene copy number.*

## Sample Size : Types of Experiment

- Sample size can be estimated if the success or failure rate can be predicted.
- Sample size, n, is given by

$$n = \frac{\log \beta}{\log p'}$$

Dell, R.B., Holleran, S.and Ramakrishnan, R. (2002). Sample Size Determination. ILAR Journal. 43:207 – 213

- Where beta = 0.05 to 0.20 and (1-beta) is the chosen power of the test (or chance of correctly detecting there is a treatment effect), and p' is the proportion of animal that did not have the desired goal (or not transgenic in this case).

## Sample Size : Types of Experiment

- Assuming the success rate of producing transgenic animal is 5 % and investigator only have about 90 % chance of getting the transgenics
  - *Sample size = log (1-0.90) / log (1.0-0.05)*
    *= log 0.1 / log 0.95*
    *= 44.89 (45 animals)*
- Assuming 30 % of the animal is infected and the investigator need to have 95 % chance of detecting the disease
  - *Sample size = log (1-0.95) / log (1.0-0.3)*
    *= log 0.05 / log 0.7*
    *= 8.4 (9 animals)*

## Sample Size : Types of Experiment

- Further examples for sample size calculation in experiments based on success or failures of goals.
- Examples of this type of experiment is the likelihood of developing a reaction to a compound, and then show it symptomatically.
  - *E.g. 80 % of the exposed animals showed serological/cellular changes to a novel compound, but only 1 in 3 animals showed visible allergic reactions.*
  - *E.g. All animals showed levels of teratogens within regulatory limits after the exposure, of this 50 % of the exposed animals demonstrated the presence of higher levels of teratogens but within the regulatory guidelines, however 5 % foetus will develop mild degree of deformity*
  - *The end point and goals of the experiment will determine the base number of sample size.*

**! SUPPLEMENTARY MATERIALS !**

## Sample Size : Types of Experiment

- Assuming 80 % of the exposed animals showed cellular changes to a novel compound, but only 1 in 3 animals showed visible allergic reactions (which is our desired endpoint, TG404:Acute dermal irritation, recommended n is 1+2)
  - *Sample size = log (1-0.80) / log (1.0-0.33)*
    *= log 0.2 / log 0.66 (-0.6989/-0.1804)*
    *= 3.87 (4 animals)*
- *It means that we need a minimum of 4 animals in order to demonstrate an allergic reaction …how does this compare to OECD TG404 guidelines which state that up to 2 additional animals can be used if the first animal showed reaction ? How do we overcome this ? Hint : strains, animals, environmental & procedural differences*

! SUPPLEMENTARY MATERIALS !

## Sample Size : Types of Experiment

- Conversely, if 80 % of the exposed animals showed cellular changes to a novel compound, and this experiment had 95 % level of confidence that what we are seeing is true (standard level of statistical confidence anyway) (which is our desired endpoint, TG442A:Skin sensitization, recommended n is 4 per dose)
  - *Sample size = log (1-0.95) / log (1.0-0.80)*
    *= log 0.05 / log 0.2 (-1.3010/-0.6989)*
    *= 1.86 (2 animals)*
- *What does this mean ? How does this compare to the OECD TG442A guidelines ?*

## Sample Size : Types of Experiment

- Experiments designed for the formal testing of hypothesis usually require more elaborate knowledge of the sampling population, variables measured, experimental design employed, effect size, test power & confidence and estimates of variability & precision.
- We can determine the sample size by taking into account all the determinants in the above and use :
  - *A). Specific formulas (those of Cochran & Snedecor, 1989; Fleiss, 1981; Dell et al., 2002).*
  - *B). Altman Normogram*
  - *C). Statistical softwares to estimate sample size (G*Power, SPSS Sample Power, nQuery Advisor, Sample Size macro in Minitab) & online calculators*

## Sample Size : Types of Experiment

G*Power 3 is available from

http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/download-and-register :

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods,* 39, 175-191.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

G*Power allows for both “A priori” & “Compromise” types of estimation.

## Sample Size Determination

• To use the formulas, normograms and statistical softwares, we need to understand the concept of effect size.

•Effect size (ES) is a name given to a family of indices that measure the magnitude of a treatment effect.

• Unlike significance tests, these indices are independent of sample size.

• In general, ES can be measured in two ways:
a) as the standardized difference between two means, or
b) as the correlation between the independent variable classification and the individual scores on the dependent variable. This correlation is called the "effect size correlation"

## Sample Size Determination

- Why do we need effect size ??
- Because any given experiment can be statistically significant if given large enough sample size !
- Consider this one-way ANOVA example :

•ANOVA is normally used to analyze effects of treatment on groups of subjects. It measures the ratio of Treatment Effects / Experimental Error corrected for total number of samples and number of groups, this will give you the F value… defined as :

$$\frac{\text{Treatment Effect}}{\text{Experimental Error}} \times \frac{(\text{Total sample} - \text{number of groups})}{(\text{Number of groups} - 1)}$$

We normally wanted large F values (typically > 3) for our statistical analysis to be significant…. *Can you imagine what happen if there are 63 subjects in 3 groups ? What about 100…200 subjects…..*

## Sample Size Determination

- Why do we need effect size ?? The problem with ANOVA….

## Sample Size Determination

- How do we calculate Effect Size (ES) ?

Effect size indicate the magnitude of treatment effects

Effect size, ES = $\frac{\text{mean difference}}{\text{standard deviation of the means}}$

ES = (Mean 1 – Mean 2) / $\sqrt{}$ [[(SD1)$^2$ + (SD2)$^2$]/2]

So, using the following example :

**Effect size is calculated differently for different kinds of statistical tests… refer to sample size calculation software for details**

***Experiment to examine the effects of Drug A on a liver metabolite level (versus untreated subjects) in rats. Group A (n=8) was treated with Drug A and Group CTRL (n=8) was the control.***

The results :

| | |
|---|---|
| *Group A* | *15,18,12,19,17,18,16,19* |
| *Group CTRL* | *11,13,15,17,18,10,15,12* |

## Sample Size Determination

- How do we calculate Effect Size (ES) ?

***Experiment to examine the effects of Drug A on a liver metabolite level (versus untreated subjects) in rats. Group A (n=8) was treated with Drug A and Group CTRL (n=8) was the control.***

The results :

| | |
|---|---|
| *Group A* | *15,18,12,19,17,18,16,19* |
| *Group CTRL* | *11,13,15,17,18,10,15,12* |

Therefore ES = (16.8-13.9) / $\sqrt{}$ [[(2.4)$^2$ + (2.9)$^2$]/2] = 1.1

- To estimate the sample size, simply fit the ES into the sample estimation software or the Altman Normogram……
- Next Question : What does a reasonably “good” ES look like ?

## Sample Size Determination

Generally, we need the ES to be as reasonably big as possible. Therefore

*We normally use 0.2 or 0.35 as a start … ES=0.35 is also known as Cohen's criterion*

Small ES, ES < 0.2

Medium ES ES = 0.2 to 0.5

Large ES ES>0.6

Extremely large ES. ES > 4.0 is considered perfect …. You need less than 1 animal in an experiment to prove there is a treatment effect… *in other words no experiment is necessary !*

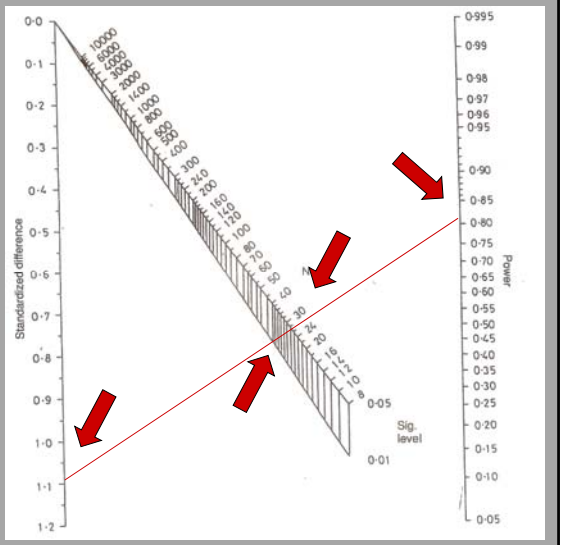
ES = 0 means that there is no difference between treatment and untreated group…. *You are wasting your time with animal experiment….*

## Sample Size Determination

• Using the Altman's Normogram

Note that Standardized difference is actually ES.

*Assuming that we have ES of 1.1 and the experiment has a test power of 80 % or 0.80, total number of samples required would be 26…..or 13 per group if we are testing at 95 % confidence level !*

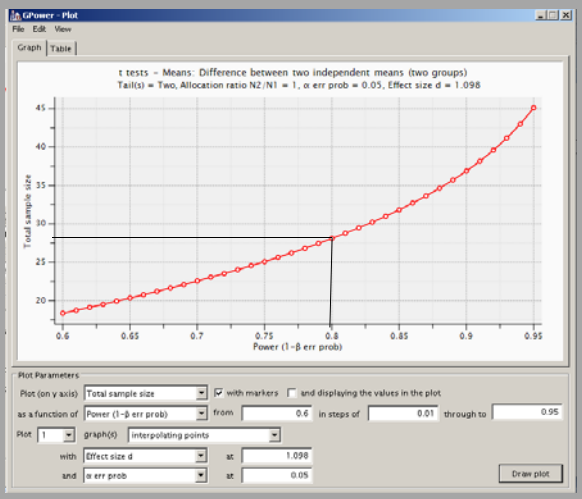



Altman, D.G. (1991). Practical statistics for medical research. Chapman and Hall, London.

## Sample Size Determination

- Using the G*Power software

*Assuming that we have ES of 1.1 and the experiment has a test power of 80 % or 0.80, total number of samples required would be 28…..or 14 per group if we are testing at 95 % confidence level ! Similar to what have been reported by Altman's Normogram*




## Sample Size Determination

- Using specific formula for studies comparing 2 groups of means

*Formula is given by :*

$$n = 1 + 2C\left(\frac{s}{d}\right)^2,$$

Snedecor, G.W. and Cochran, W.G. (1989). Statistical Methods. 8th Ed. Ames: Iowa State Press.

*Where n = sample size for ONE group*

*s = standard deviation of the entire test population, or pooled standard deviation*

*d = the magnitude of difference that we postulated/hoped to detect*

*C = a constant given by the following table*

Sig. level

| Test power 1-β | α = 0.05 | 0.01 |
|---|---|---|
| 0.8 | 7.85 | 11.68 |
| 0.9 | 10.51 | 14.88. |

## Sample Size Determination

- Using the values from the example,

$s = \sqrt{[(2.4)^2 + (2.9)^2]/2} = 2.66$

*d = difference between Mean Group A – Mean Group Control*
*= 1.68 – 13.9 = 2.9*

*C = 7.85, from the table assuming that we wanted 80 % test power at 95 % confidence level.*

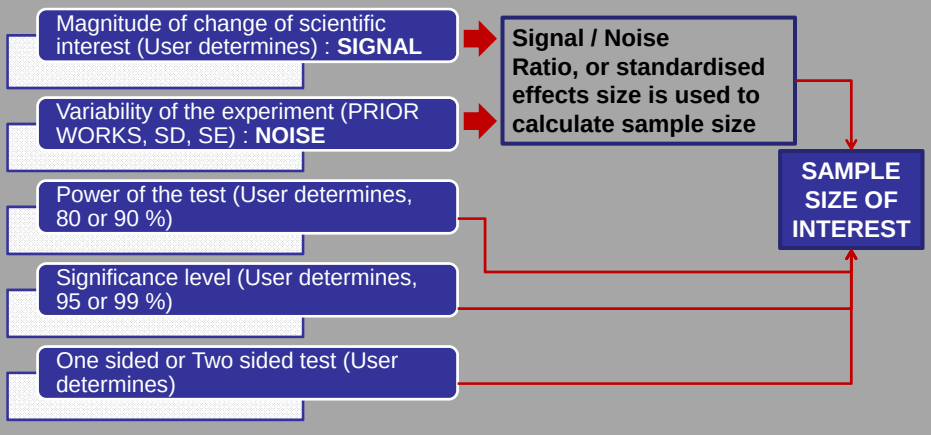
*Therefore number of samples per group, n = 13.11 or 13 animals*
*Thus, total sample size = 13 x 2 groups = 26 !*

***Similar to what have been reported by Altman's Normogram !***

*Please refer to Dell et al., (2002) for further discussion on sample size calculations for proportions…..*

## Sample Size & Power Analysis

In general, the relationship between factors that determined sample size, and power of an experiment can be visualised here :



## Sample Size & Power Analysis

So, when the CALCULATED group sample size for the control group for a given S/N ratio is less than the REAL LIFE group sample size for the control group….**You can use lesser number of animals in the control group.**

*E.g. S/N ratio calculation for single factor, TWO GROUPS comparison ONLY at 90 % or 80 % power, at 95 % confidence level. THIS TABLE IS NOT SUITABLE FOR 3 groups or more, multi factorial comparisons.*

*The corresponding table can be generated using G*Power 3….*

*(from http://www.3rs-reduction.co.uk/html/6__power_and_sample_size.html)*

Group size as a function of S/N ratio (5% sig., 2-sided)

| SN ratio | 90% power | 80% power |
|---|---|---|
| 0.2 | 526 | 393 |
| 0.4 | 132 | 99 |
| 0.6 | 59 | 45 |
| 0.8 | 34 | 26 |
| 1.0 | 22 | 17 |
| 1.2 | 16 | 12 |
| 1.4 | 12 | 9 |
| 1.6 | 9 | 7 |
| 1.8 | 8 | 6 |
| 2.0 | 6 | 5 |
| 2.2 | 6 | 4 |
| 2.4 | 5 | 4 |
| 2.6 | 4 | 4 |
| 2.8 | 4 | 3 |
| 3.0 | 4 | 3 |

## Sample Size & Precision of Measurements

- More precise = less animals
- More precise = easily observable experimental effects of interest
- Effect size (ES) is a name given to a family of indices that measure the magnitude of a treatment effect.
- Bigger ES = Easier to detect experimental difference
- See example in the next slide

## Sample Size & Precision of Measurements

- Methods have 10 x difference in precision

| Measurement precision | Group 1 | Group 2 | Difference | ES calculated | Sample size per group |
|---|---|---|---|---|---|
| ± 0.1 | 91.5 | 97.4 | 5.9 | 0.536363 | 22 |
| ± 1.0 | 92 | 97 | 5 | 0.454545 | 31 |

Assumptions :
SD = 11
Test Power = 80 %
Significance level = 5 %

## Sample Size & Experimental Design

• Types of experimental design will affect the efficiency of a research method in detecting treatment effects – increase or reduce groups

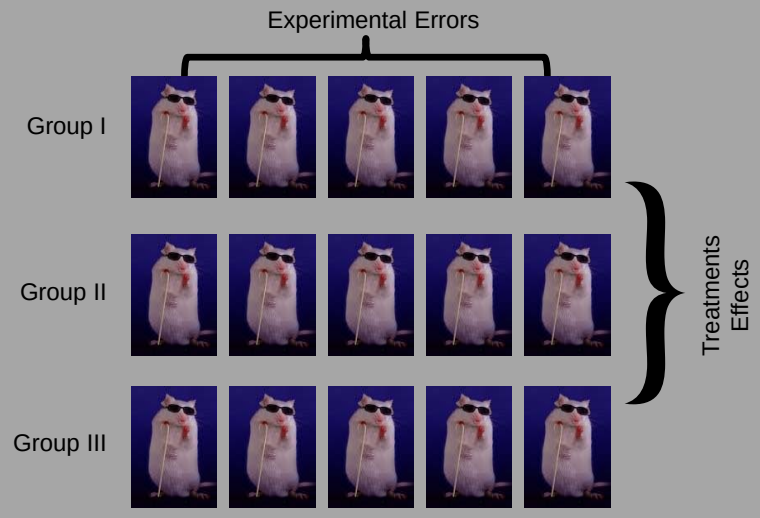
• Common experimental designs include :

**A.Completely randomised design (CRD)**
*- when investigating effect(s) of ONE FACTOR*
*- one direction of change*
*- data matrix (arrangement) follows 1-way-ANOVA*

**B.Randomised Complete Block Design (RCBD)**
*- when investigating effects of ONE FACTOR while ruling out the effect of the other Factor (known as confounding factor).*
*- two directions of change*
*- data matrix follows 2-way-ANOVA*

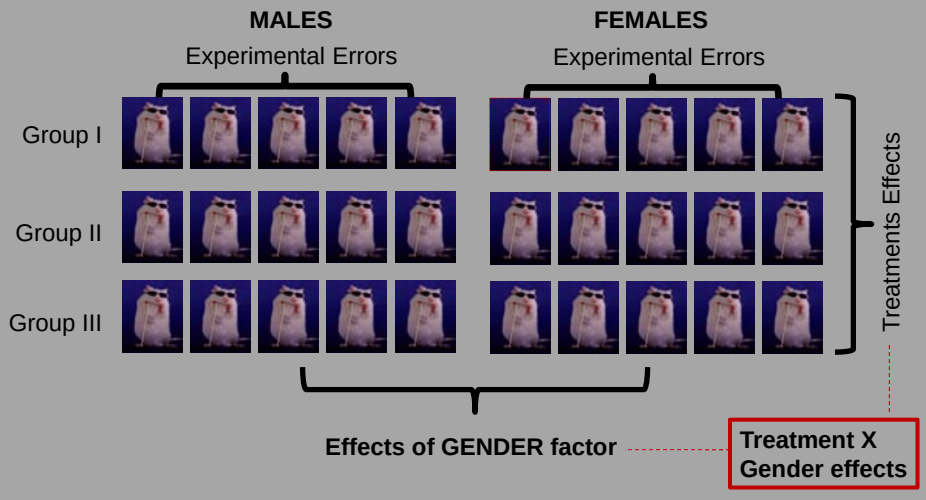
## Sample Size & Experimental Design

THE “SIMPLE” CRD Design




## Sample Size & Experimental Design

THE RCBD Design that takes into account effects of factors



## Sample Size & Experimental Design

- Common experimental design included :

**C.Factorial design**

*- when investigating effect(s) of many FACTORS (more than 1)*
*- multiple directions of change*
*- data matrix (arrangement) follows 1,2,3,4,5…..-way-ANOVA depending on the number of factors.*

**D.Latin Square**

*- when investigating effects of TWO FACTORS that are minimally interacting with each other.*
*- allow researchers to have more sample than usual – but watch out for "wash-out periods"*
*- two (crossed) directions of change*
*- data matrix follows 2-way-ANOVA*

## Sample Size & Experimental Design

Effects of experimental design on sample size & test efficiency.

- When a test is inefficient, sample size tend to INCREASE
- **We have an index known as Design Effect Factor (DEF) which serves to compare the efficiency of a selected experimental design AGAINST Completely Randomized Design (CRD)**
- **DEF = Pooled Variance of the data using a selected experimental design / Pooled Variance of the data when analyzed as CRD**
- **Important DEFF values**
- ***DEF < 1.0, the new experimental design is better compared to CRD***
- ***DEF = 1.0, the new experimental design is comparable to CRD***
- ***DEF > 1.0, the new experimental design is not as good as CRD***

## Sample Size & Experimental Design

- Consider the following example :

***You suspect that your pilot data for a study attempting to compare the effects of Compound X on blood pressure in rats is showing noticeable gender effect. You proceed to compare the efficiency of a CRD model vs RCBD model where you block the subjects by sex.***

***CRD design, BP in rats post treatment :***

***120,116, 137,135,145, 118,125, 109, 138, 127***

***RCBD design, BP in rats by gender :***

***Males : 137, 135, 145, 125, 138***

***Females : 120, 116, 118, 109, 127***

## Sample Size & Experimental Design

***CRD design, BP in rats post treatment :***

***120,116,137,135,145,118,125,109,138,127***

***$Variance_{CRD}$ = 52.0***

***RCBD design, BP in rats by gender :***

***Males : 137, 135, 145, 125, 138 Variance = 52.0***

***Females : 120, 116, 118, 109, 127 Variance = 42.0***

***Therefore, Pooled Variance for Males & Females***

***$Variance_{RCBD} = \sqrt{[[(52)^2 + (42)^2]/2]} = 47.26$***

***DEFF = 47.26 / 52***

***= 0.91 so what is your verdict ???? What can you do to the calculated sample size ???***

## Sample Size & Statistical Tests

• Statistical tests can be grouped according to their variable types

**A) Parametric data**
- continuous only, absolute
- follows a normal distribution.
- 1,2.3, 4.567, 9, 10.1……..

**B) Non-parametric data**
- discrete, ratio, percentage, converted numbers
- does not follow any type of distribution (non-normal).
- not really absolute !

*Differentiating parametric data from non-parametric data can be done using Shapiro-Wilk's Test or even a simple probability (O/E) plot.. Kolmogorov Smirnov Test is not really recommended at this point*

## Sample Size & Statistical Tests

• Statistical tests can be grouped according to their variable types

| | **Parametric** | **Non-parametric** |
|---|---|---|
| *Measurement* | Mean | Median,Mean,Mode |
| *Comparing* | Independent t-test | Mann-Whitney U Test |
| *2 groups* | Paired t-test | Wilcoxon Sign Ranked Test |
| *Comparing > 2 groups* | 1-way-ANOVA | Kruskal-Wallis H Test |
| | 2-way-ANOVA (no replicates) | Friedman's Test |
| | 2-way-ANOVA (with replicates) | Transform data and do as 2-way-ANOVA |

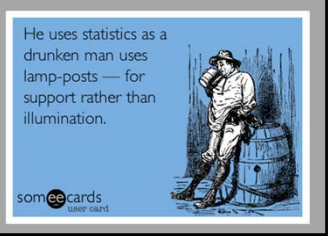
## Sample Size & Statistical Tests

- Statistical tests can be grouped according to their variable types

| | Parametric | Non-parametric |
|---|---|---|
| *Correlations* | Pearson's r<br><br>C.O.D can be used | Spearman's Rank Correlation r (rho)<br>C.O.D. is useless |
| *Regression* | Parametric regression models | Logistic regression & non-parametric regression models |
| *Others* | *No parametric test generally require less sample size to work, but suffers from lack of sensitivity to a comparable parametric test !* | Chi-Square or G-Test Test of Independence, Goodness of fit, Test of equality of proportions |

## Concluding Remarks

- There is no absolute and fool proof approach for sample size estimation. Available statistical tools merely advises us on the best option/estimate available.
- Important to refer to prior works and follow conventionally accepted sample size and methodologies.
- Avoid unnecessary suffering and wastage of experimental subjects.
- Justification for the sample size should be scientifically and statistically sound while following the 3R approach, Refine, Reduce, Replace…




Thank you
Aerial view of UPM's main entrance