Scientific Training Center for Plant Biotechnology
“Advanced level” training Session
“Power of a test, effect size, independence”

Pr. L. Gentzbittel, Pr. C. Ben

March 2022, Skoltech, Moscow, Russia

Contents

Before starting any experiment, careful planning needs to take place. For instance, how many samples are required for your experiment? This question is important for two reasons. First, an experiment with too few of samples may not be able to determine real differences between, say a control and experimental group. And second, too many samples may incur unnecessary costs. These questions center around the idea of a power analysis.

We will use the `pwr` package to realize preliminary computations to set-up experiments that may have sufficient power to see an effect – if it exists.

Several tricks are worth to know, without doing any power computations.

1 Allocating samples/ressources to a reference and new conditions, to evaluate with optimum accuracy.

Imagine we intend to compare the efficiency of new antibiotic molecules to the mean efficiency of the current reference molecule. We are primarily interested in **comparing to the reference** rather than **comparing the molecules with each other**. To test with the better precision, how much assays need to be allocated to the new molecules and to the reference molecule ?

In that particular example, 9 new antibiotics are to be compared to the reference molecule (e.g. streptomycin), by measuring growth of a bacterial strain after 48hours on media supplemented with the different antibiotics. We are allowed to use/we can perform 60 experimental units (say individual Petri dishes) for that experiment.

Let $\bar{x_i}$ the mean of growth on molecule i, σ^2 the variance of growth in experimental units (assumed to be the same for all molecules), and $\bar{x_0}$ the mean of growth on the reference molecule.

Recall : The variance of a mean, computed in a sample of n values is $\frac{\sigma^2}{n}$, with σ^2 the variance of the sample. This simply says that a mean is better estimated (its variance is reduced) using a large sample.

- We allocate the 10 molecules (9 new antibiotics and the reference streptomycin) in each of 6 units, for a total of 60 units. The variance of the difference between any of the average growth on a new antibiotics and the average growth on the reference antibiotics is given by :

$$Var(\bar{x_i} - \bar{x_0}) \approx Var(\bar{x_i}) + Var(\bar{x_0})$$

The greatest this variance, the more likely 0 will be in the confidence interval of the difference, implying that there is *no* difference between growth on a new antibiotics and the reference. In that particular design, the variance of the difference of the means is :

$$\left(\frac{1}{6} + \frac{1}{6}\right)\sigma^2 = \sigma^2/3 = 5\sigma^2/15$$

.

- Let use 5 experimental units (plates) to each of the 9 new molecules, and 15 plates to the reference streptomycin. The variance of the difference of the mean is :

$$\left(\frac{1}{5} + \frac{1}{15}\right)\sigma^2 = 4\sigma^2/15$$

i.e de 20% less than $\sigma^2/3$!

Case study: You want to quickly check for protein content in different cell lines as a response to particular stress and compare to a 'standard' response. Find the best repartition for 19 new cell lines and the 'standard' reference, given you will be able to perform 200 protein dosages at the same time. The variance of the response is expected to be the same among the 20 samples

```
200/20
```

```
[1] 10
```

- Let's use 10 samples per new cell line and 10 samples for the 'standard' reference. The variance of the difference of the mean of protein content between a new cell lines and the standard will be $\left(\frac{\sigma^2}{10} + \frac{\sigma^2}{10}\right) = \left(\frac{1}{10} + \frac{1}{10}\right)\sigma^2 = \frac{2}{10}\sigma^2 = \frac{522}{2610}\sigma^2$

```
200-9*19
```

```
[1] 29
```

- Let's use 9 samples per new cell lines and therefore 29 protein samples for the reference. The variance of the difference of the mean between a new accessions and the standard will be $\left(\frac{\sigma^2}{9} + \frac{\sigma^2}{29}\right) = \left(\frac{29}{261} + \frac{9}{261}\right)\sigma^2 = \frac{38}{261}\sigma^2 = \frac{380}{2610}\sigma^2$

The increase in precision is thus $\frac{522-380}{522} = 27\%$!!!!

- A major criteria to include in these computations is the minimum number of samples to evaluate to get an *accurate value* (low σ^2) of the mean phenotype (10 samples, 9 samples, ..., 6 samples ?).
- The general rule is thus : ***Allocate more samples to the reference than to the cases.***

2 The block/lot size shall be kept as small as possible.

We intend to estimate a disease incidence and we can use up to 100 plants. If we ***need*** to make different lots for practical reasons (not possible to handle 100 plants at the same time, not enough space in the greenhouse, heterogeneity within the greenhouse, etc...), what is the optimim size of the lots/batches : two lots of 50 plants each, 4 lots of 25 plants, ..., 20 lots of 5 plants each ?

Imagine we constitute p lots of n plants each. Let $\bar{X}$ the average value of the pn mesurements (ie the disease incidence we want to estimate), σ^2 the variance between plants within lots and σ^2_A the variance between lots (expected to be different from zero, otherwise why bother doing lots ?)

We have $Var(\bar{X}) = \sigma^2_X = \frac{\sigma^2_A}{p} + \frac{\sigma^2}{pn}$

- $p = 2$ lots of $n = 50$ mices $\bar{X}$ will be the average of the 2 means of each lot:

$$\sigma^2_{\bar{X}} = \frac{\sigma^2_A}{2} + \frac{\sigma^2}{100}$$

- $p = 4$ lots of $n = 25$ mices $\bar{X}$ will be the average of the 4 means of each lot:

$$\sigma^2_{\bar{X}} = \frac{\sigma^2_A}{4} + \frac{\sigma^2}{100}$$

- $p = 20$ lots of $n = 5$ mices $\bar{X}$ will be the average of the 20 means of each lot:

$$\sigma^2_{\bar{X}} = \frac{\sigma^2_A}{20} + \frac{\sigma^2}{100}$$

- In summary, keep block size / lot size ***as small as feasible***.

3 How much plants/independant samples ? Approximate computations.

It is possible to get rough sample size to compare two means.

Given the coefficient of variation ($V = \frac{\sigma}{\mu}$) and the difference between two means, δ, in percentage, then :

- for $\alpha = 0.05 \quad ; \quad 1 - \beta = 0.50$:

$$n \simeq \frac{8V^2}{\delta^2}$$

- $\alpha = 0.05 \quad ; \quad 1 - \beta = 0.90$:

$$n \simeq \frac{21V^2}{\delta^2}$$

Case study: Based on preliminary experiments, aerial biomass after infection of a partially resistant variety is as follows :

	Part. resistant
moyenne	100
variance	15^2

We deduce that $CV = \sigma/\mu = 0.15$

How much plants shall we score to see a 10% difference in means between this accession and another accession at $\alpha = 0.05$, in 90% of the experiments ?

Here $V = 0.15$ and $\delta = 0.10$, and the putative aerial biomass of the new accession after infection is expected around 110 $\quad (+10\%)$ or 90 $\quad (-10\%)$.

$$n \approx \frac{21V^2}{\delta^2} = \frac{21 \cdot (0.15)^2}{(0.10)^2} = 48$$

```r
(21*0.15^2)/(0.10^2)
```

```
[1] 47.25
```

4 How much measurements ? Introducing 'the effect size'.

The `pwr` package is the basis for power computations. It does not cover all experimental designs but still allows for very useful computations.

The case study is the following : we want to compute the number of samples to assess to decide if there is a significant decrease/increase in a trait after a treatment – or a significant difference in some trait value.

The following four quantities have an intimate relationship:

1. sample size (N)

2. effect size
3. significance level α = P(Type I error) = probability of finding an effect that is **not** there
4. power $1 - \beta$ = 1 - P(Type II error) = probability of **finding an effect that is there**

Given any three, we can determine the fourth.

There are thus four types of power analysis, defined by which one of these four parameters you wish to solve for:

- *a priori*: compute N, given α, power, effect size
- *post - hoc*: compute power, given α, N, effect size
- Criterion: compute α, given power, effect size, N
- Sensitivity: compute effect size, given α, power, N

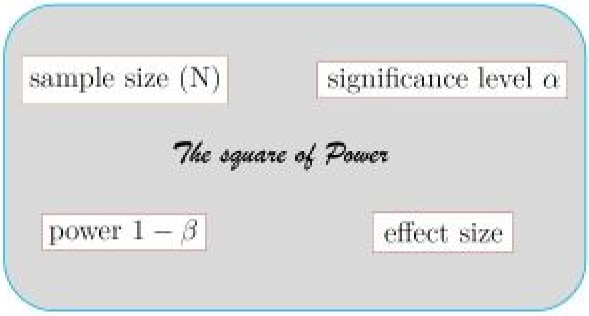
The ***a priori*** power analysis is what is usually done when designing a study. This tells you what sample size is needed to detect some level of effect with inferential statistics (i.e. with p-values). Funding agencies of course want to avoid chance findings, so an a priori power analysis is needed in all study proposals, except pilot studies.

A ***post hoc*** power analysis at the completion of a study is also wise, as your expected effect and actual effect may not align. This post -hoc power analysis tells you if you had sufficient plants to detect with inferential statistics the actual effect you found.

A **criterion** power analysis is seldom used by researchers.

A **sensitivity** power analysis is used when the sample size is predetermined by study constraints. For example, if there are only 20 evaluations/measurements you can afford, determining how many you need is less relevant. Instead, one determines what level of effect you could find with the subjects you have. This is referred to as the minimal detectable effect (MDE).

This is the ‘Square of Power’ !



4.1 The (often unknown) difference between the two alternative hypothesis is the ‘effect size’.

- When comparing two means (using t-test):

$$d = \frac{|\mu_1 - \mu_2|}{\sigma}$$

It is suggested that d values of 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes respectively. **Your own subject matter experience should be brought to bear.**

- When comparing **multiple** means (using one-way ANOVA):

$$f = \sqrt{\frac{\sum_{i=1}^{k} p_i \times (\mu_i - \mu)^2}{\sigma^2}}$$

avec $p_i = n_i/N$, n_i = number of observations in group i (k groups), N = total number of observations, μ_i = mean of group i, μ = grand mean, σ^2 = residual variance.

It is suggested that f values of 0.1, 0.25, and 0.4 represent small, medium, and large effect sizes respectively. **Your own subject matter experience should be brought to bear.**

- When comparing two proportions:

$$h = 2\arcsin(\sqrt{p_1}) - 2\arcsin(\sqrt{p_2})$$

with $p_1 \geq p_2$.

It is suggested that h values of 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes respectively. **Your own subject matter experience should be brought to bear.**

- for linear models such as multiple regression.

$$f^2 = \frac{R^2}{1 - R^2}$$

where R^2 is the multiple correlation coefficient.

It is suggested that f^2 values of 0.02, 0.15, and 0.35 represent small, medium, and large effect sizes.

4.2 Let's plan our experiments.

```
library(pwr)
```

4.2.1 Percentage of diseased/dead indivuals and disease incidence.

- An attempt to NOT blindly/stupidly plan an experiment: I suspect that the disease incidence in a population is close to 50%. I would like to compare it with the reference (susceptible) population that has a disease incidence index of 80%, by testing 15 individuals within each population. I'm a serious scientist and intend to use a test level $\alpha = 0.05$.

In this example, the effect size is :

```
( EffectSize <- 2*asin(sqrt(0.80)) - 2*asin(sqrt(0.50)) )
```

```
[1] 0.6435011
```

```
# pwr.2p.test(h = , n = , sig.level =, power = )
pwr.2p.test(h = EffectSize, n = 15, sig.level = 0.05)
```

```
     Difference of proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.6435011
              n = 15
      sig.level = 0.05
          power = 0.4217528
    alternative = two.sided

NOTE: same sample sizes
```

meaning that in $\approx$ 58% of the cases, I wil NOT detect the difference between "50% disease incidence" and "80% disease incidence" at $\alpha = 0.05$. Given that result, what is the probability that three (3) replicates of the experiment will *all* detect the difference? "THREE" (3) is the 'golden magic number' of replicates in some trendy papers to be able to publish.

```
# (this is the probability of a binomial law with p = 0.42, a sample size of 3,
# and 3 successes).
dbinom(3, 3, 0.42)
```

```
[1] 0.074088
```

In other words (more simply), because the replicates are (expected to be) independent, the probability that all replicates detect the difference is $0.42 \times 0.42 \times 0.42 = 0.074$

Well, not much ! OK, How much samples do I need to detect that difference in 90% of the experiments, keeping the same level of the test:

```
# pwr.2p.test(h = , n = , sig.level =, power = )
pwr.2p.test(h = EffectSize, sig.level = 0.05, power = 0.9)
```

```
     Difference of proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.6435011
              n = 50.749
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: same sample sizes
```

And given that result, what is the probability that three replicates of the experiment will *all* give three positive results (i.e. detect the difference) ? This is $0.90 \times 0.90 \times 0.90 = 72.9\%$. Not very nice.

- OK, let's say I will 90% chances that all three replicates detect the effect. In that case, the power of a unitary experiment is $\sqrt[3]{0.90}$ ie 0.965. How much individuals to sample per experiment are requested ?

```r
# pwr.2p.test(h = , n = , sig.level =, power = )
pwr.2p.test(h = EffectSize, sig.level = 0.05, power = 0.965)
```

```
     Difference of proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.6435011
              n = 68.7141
      sig.level = 0.05
          power = 0.965
    alternative = two.sided

NOTE: same sample sizes
```

- ***In summary***, we need ≈ 69 individual in both population lines, in each of the three replicates (thus a total of 207 individuals) to detect the difference between a population that has 50% disease incidence and a population that has 80% disease incidence, at $\alpha = 0.05$ in all three replicates.
- And what if I would like to use a significance level of $\alpha = 0.01$ and a power of $1 - \beta = 0.965$ – to be able to have 90% chances of getting three replicates (over three) that detect the effect ?

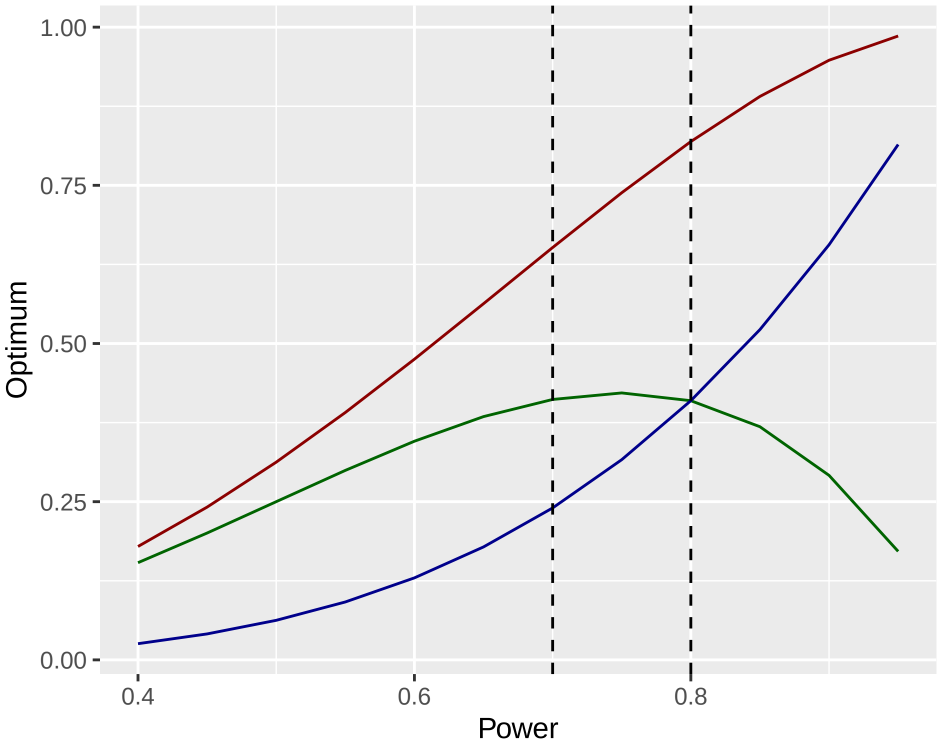
```r
# pwr.2p.test(h = , n = , sig.level =, power = )
pwr.2p.test(h = EffectSize, sig.level = 0.01, power = 0.965)
```

```
     Difference of proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.6435011
              n = 92.98506
      sig.level = 0.01
          power = 0.965
    alternative = two.sided

NOTE: same sample sizes
```

Wow ! A total of 279 individuals per population is required (three experiments of 93 individual per population each).

- Well, why not doing 4 replicates with a reasonable power, and simply present the three replicates that are OK?

```r
# what unitary power gives a reasonable probability to get at least 3 successful replicates over 4 assa
power <- seq(from = 0.4, to = 0.95, by = 0.05)
Optimize <- data.frame(Power = power)
Optimize <- Optimize %>%
            mutate(Trois = dbinom(3, 4, Power)) %>%
            mutate(Quatre = dbinom(4, 4, Power)) %>% rowwise() %>%
            mutate(Optimum = sum(Trois, Quatre))

Optim <- ggplot( Optimize, aes( x = Power, y = Optimum)) +
         geom_line(col = "darkred") +
         geom_line(data = Optimize, aes( x = Power, y = Trois), col = "darkgreen") +
         geom_line(data = Optimize, aes( x = Power, y = Quatre), col = "darkblue") +
         geom_vline(xintercept = c(0.7, 0.8), linetype = 2, col = "black")
print(Optim) # to visualize
```



```
# if we need to plot the legend, it is required to reorganise the data from wide to long format
# and map to a colour aesthetic. Optimize is NOT in a well-organised data format.
```

So, a power ≈ 0.75 maximises the probability of getting 3 or 4 positive results over 4 experiments. How much samples to reach that power ?

```
# pwr.2p.test(h = , n = , sig.level =, power = )
pwr.2p.test(h = EffectSize, sig.level = 0.01, power = 0.75)
```

```
     Difference of proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.6435011
              n = 51.02505
      sig.level = 0.01
          power = 0.75
    alternative = two.sided

NOTE: same sample sizes
```

Well, by the way : What is your opinion about the magical 'three replicates' that are required/mandatory to publish in scientific journals ?

- Assuming a significance level of 0.01 and a common sample size of 30 for each proportion, what effect size can be detected with a power of .75?

```
# pwr.2p.test(h = , n = , sig.level =, power = )
pwr.2p.test(n = 30, sig.level = 0.01, power = 0.75)
```

```
     Difference of proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.8392269
              n = 30
      sig.level = 0.01
          power = 0.75
```

```
    alternative = two.sided

NOTE: same sample sizes
```

```
# translating in percentages :
# let p2 = 3% (resistant line, with low disease incidence).
# looking for the upper limit (less resistant lines) ie p1
(p1 <- (sin( asin(sqrt(0.03)) + (0.84/2) ))^2 )
```

```
[1] 0.3133191
```

```
2*asin(sqrt(p1)) - 2*asin(sqrt(0.03))  # OK
```

```
[1] 0.84
```

```
# check the other way :
# pwr.2p.test(h = , n = , sig.level =, power = )
pwr.2p.test(h = 0.84, n = 30, sig.level = 0.01)
```

```
     Difference of proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.84
              n = 30
      sig.level = 0.01
          power = 0.7509482
    alternative = two.sided

NOTE: same sample sizes
```

In the above design (30 samples assessed per case), the population exhibiting an average disease incidence less than 31% will be considered, in 75% of the experiments, as resistant as the population exhibiting a disease incidence level of 3%.

- **Be careful**, percentages are not symetrical: When infected with $Pm8$, a line has a disease incidence of 60%. Using 30 plants and $\alpha = 0.01$, what will be the more susceptible or more resistant lines that we will be able to detect ?

```
# let p1 = 60% (partially resistant line, with high disease incidence)
( p2 <- (sin(asin(sqrt(0.60)) - 0.84/2))^2 )
```

```
[1] 0.2019471
```

```
2*asin(sqrt(0.6)) - 2*asin(sqrt(p2))
```

```
[1] 0.84
```

```
# let p2 = 60% (partially resistant line, with high disease incidence)
( p1 <- (sin(asin(sqrt(0.60)) + 0.84/2))^2 )
```

```
[1] 0.9315454
```

```
2*asin(sqrt(p1)) - 2*asin(sqrt(0.6))
```

```
[1] 0.84
```

In that case, using 30 plants allows a power of 75% when comparing disease incidence of 60% compared to 20% at $\alpha = 0.01$; or when comparing disease incidence of 60% to 93% at $\alpha = 0.01$. Remember that in the remaining 25% of the experiments, this difference will *NOT* be declared significant at $\alpha = 0.01$.

4.2.2 Biomass or other quantitative traits, using t-tests.

The method is similar to that used for percentages, and the `pwr` package uses an unified syntax.
The main difference with the previous computation on percentage is the definition of the effect size, that involves not only the difference between the means but also the variance.

- What is the power of a one-tailed t-test, with a significance level of 0.01, 25 data points in each group, and an effect size equal to 0.75?

```r
# pwr.t.test(n = , d = , sig.level = , power = , alternative = )
pwr.t.test(n = 25, d = 0.75, sig.level = 0.01, alternative = "greater")
```

```
     Two-sample t test power calculation 

              n = 25
              d = 0.75
      sig.level = 0.01
          power = 0.5988572
    alternative = greater

NOTE: n is number in *each* group
```

- A reference tissue exhibits a trait $= 2.5$ with $\sigma^2 = 0.73$. How much samples to use to differentiate another tissue with expected trait $= 4$ and $\sigma^2 = 0.8$, at $\alpha = 0.01$ and with a power of 90% ?
  In a first attempt, we will use a pooled estimate of the variance : $\sigma^2_{pooled} = \frac{sigma^2_{s1}+sigma^2_{s2}}{2}$. In the above case, it is 0.765.

  The effect size $d = \frac{|\mu_1-\mu_2|}{\sigma_{pooled}} = 1.7149859$

```r
# pwr.t.test(n = , d = , sig.level = , power = , alternative = )
pwr.t.test(d = 1.71, sig.level = 0.01, alternative = "greater", power = 0.90)
```

```
     Two-sample t test power calculation 

              n = 10.38185
              d = 1.71
      sig.level = 0.01
          power = 0.9
    alternative = greater

NOTE: n is number in *each* group
```

So approx 11 samples are required.

4.2.3 Biomass or other quantitative traits, using ANOVA.

- For a one-way ANOVA comparing 5 groups, calculate the sample size needed in each group to obtain a power of 0.80, when the effect size is moderate (0.25) and a significance level of 0.05 is employed.

```r
pwr.anova.test(k = 5, f = 0.25, sig.level = 0.05, power = 0.8)
```

```
     Balanced one-way analysis of variance power calculation 

              k = 5
              n = 39.1534
              f = 0.25
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

- It is also possible to draw a *power curve*, to optimize the design. Let's keep *the same design* and compute the power as a function of sample size. This will be also a nice example of *programming with R*.

```r
N <- seq(from = 5, to = 200, by = 5)

# carefully observe the structure of the result when applying the function.
# This is a *list* and we can access each individual result:
(Compute <- pwr.anova.test(k = 5, n = 5, f = 0.25, sig.level = 0.05))
```

```
     Balanced one-way analysis of variance power calculation 

              k = 5
              n = 5
```

```
                f = 0.25
        sig.level = 0.05
            power = 0.121178

NOTE: n is number in each group
```

```r
# get n:
Compute[[1]]
```

```
[1] 5
```

```r
# get power :
Compute[[5]]
```

```
[1] 0.121178
```

```r
# Now, apply the pwr.anova.test function for each value of N,
# and get the 5th result of each computation.
# we will use the lapply() function:

ListOftests <- lapply( N ,
                       function(x) pwr.anova.test(k = 5, n = x, f = 0.25, sig.level = 0.05)
                       # this (l)applies the pwr.anova.test for each value of the vector N
                       )
# ListOftests is a list, following R nomenclature. A list handle data of differnt types.
# e.g.: results of pwr.anova.test for n= 195
ListOftests[[39]]
```

```
     Balanced one-way analysis of variance power calculation

                k = 5
                n = 195
                f = 0.25
        sig.level = 0.05
            power = 0.9999997

NOTE: n is number in each group
```

```r
# e.g.: value of the power for n = 200
ListOftests[[40]][5]
```

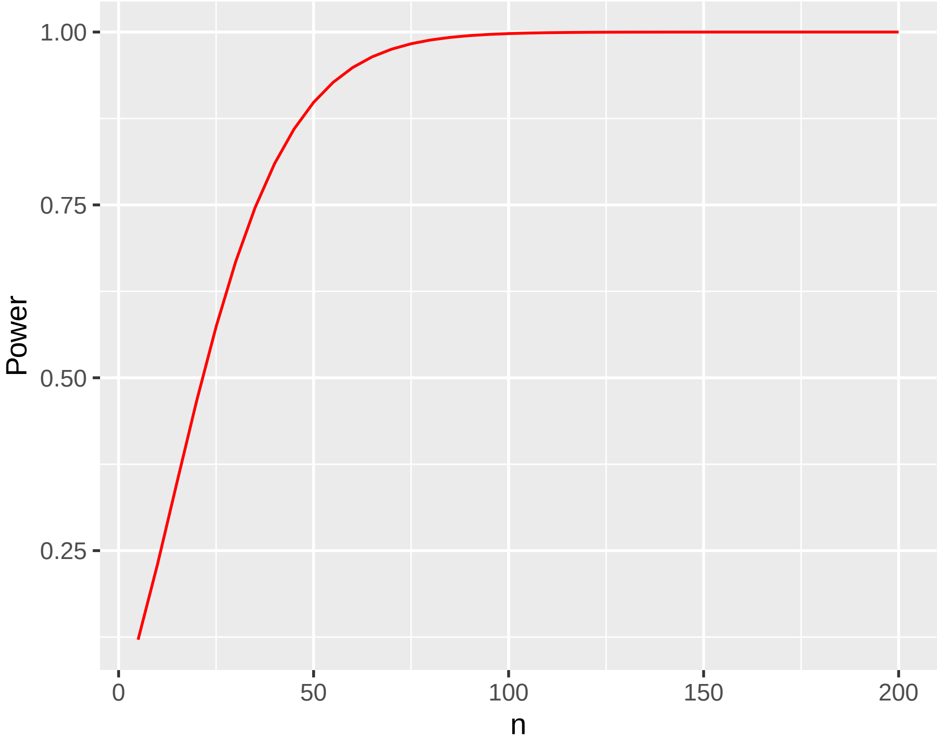
```
$power
[1] 0.9999998
```

```r
Powers0 <- sapply(ListOftests,
                       "[[", 5)  # this (s)applies the [[ function to extract the 5th element
# sapply returns a vector ; lapply returns a list.
head(Powers0, 20)
```

```
 [1] 0.1211780 0.2306936 0.3494298 0.4663312 0.5738000 0.6675964 0.7461564
 [8] 0.8097710 0.8598435 0.8983113 0.9272466 0.9486106 0.9641256 0.9752270
[15] 0.9830642 0.9885298 0.9922990 0.9948716 0.9966108 0.9977762
```

```r
# -----------------
# OK, all in a single step using pipes %>% :
Powers <- N %>% lapply( # note : the first argument (here N) is NOT specified using %>%
                       function(x) pwr.anova.test(k = 5, n = x, f = 0.25, sig.level = 0.05)
                       ) %>%
                       sapply( "[[", 5 )  # note : the first agrument (here the list)
                                          # is NOT specified using %>%
# draw power curve:
(graphe <- N %>% data.frame( n = . )  %>%
          mutate( "Power" = Powers) %>%
          ggplot( aes(x = n, y = Power)) +
                 geom_line(col = "red")
```

```
    )
```



4.3 Reporting a Power Analysis

Writing a power analysis summary in a grant proposal need not be complex. A simple paragraph or two that reflects your study plans and addresses possible contingencies for additional variables, attrition, etc. usually is satisfactory. Most funding agencies are interested in study feasibility with a well thought out study plan, which a power analysis is a component of.

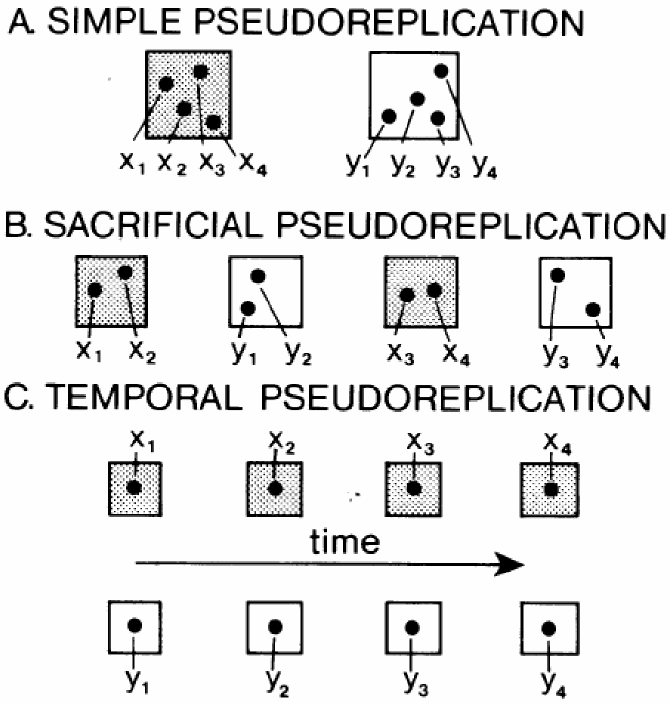
Below is a sample power analysis paragraph for a simple design, which can be modified easily to reflect the study specifics:

Sample size estimation:

A statistical power analysis was performed for sample size estimation, based on data from pilot study/published study X (N=. . .), comparing . . . . to . . . .. The effect size (ES) in this study was . . . ., considered to be extremely large/large/medium/small using Cohen's (1988) criteria. With an $\alpha = \ldots$ and power $= \ldots$, the projected sample size needed with this effect size (`pwr` package of **R**) is approximately N = ... for this simplest between/within group comparison. Thus, our proposed sample size of ..N+.. will be more than adequate for the main objective of this study and should also allow for expected attrition and our additional objectives of controlling for possible mediating/moderating factors/subgroup analysis, *etc. . . , etc. . .*

5 Data independence: "fight pseudo-replication" !

Ideas in this points are mostly based on `Hurlbert, S.H., 1984. Pseudoreplication and the Design of Ecological Field Experiments. Ecological Monographs 54, 187-211. https://doi.org/10.2307/1942661`



A. SIMPLE PSEUDOREPLICATION
x1 x2 x3 x4
y1 y2 y3 y4
B. SACRIFICIAL PSEUDOREPLICATION
x1 x2
y1 y2
x3 x4
y3 y4
C. TEMPORAL PSEUDOREPLICATION
x1
x2
x3
x4
time
y1
y2
y3
y4

FIG. 5. Schematic representation of the three most common types of pseudoreplication. Shaded and unshaded boxes represent experimental units receiving different treatments. Each dot represents a sample or measurement. Pseudoreplication is a consequence, in each example, of statistically testing for a treatment effect by means of procedures (e.g., t test, U test) which assume, implicitly, that the four data for each treatment have come from four independent experimental units (=treatment replicates).

Replication, randomization, and independence.— Replication and randomization both have two functions in an experiment: they improve estimation and they permit testing. Only their roles in estimation are implied in Table 1. Replication reduces the effects of "noise" or random variation or error, thereby increasing the *precision* of an estimate of, e.g., the mean of a treatment or the difference between two treatments. Randomization eliminates possible bias on the part of the experimenter, thereby increasing the *accuracy* of such estimates.

In operational terms, a lack of independence of errors prohibits us from knowing α, the probability of a type I error. In going through the mechanics of a significance test, we may specify, for example, that $\alpha = 0.05$ and look up the corresponding critical value of the appropriate test criterion (e.g., t or F). However, if errors are not independent, then true α is probably higher or lower than 0.05, but in any case unknown. Thus interpretation of the statistical analysis becomes rather subjective.

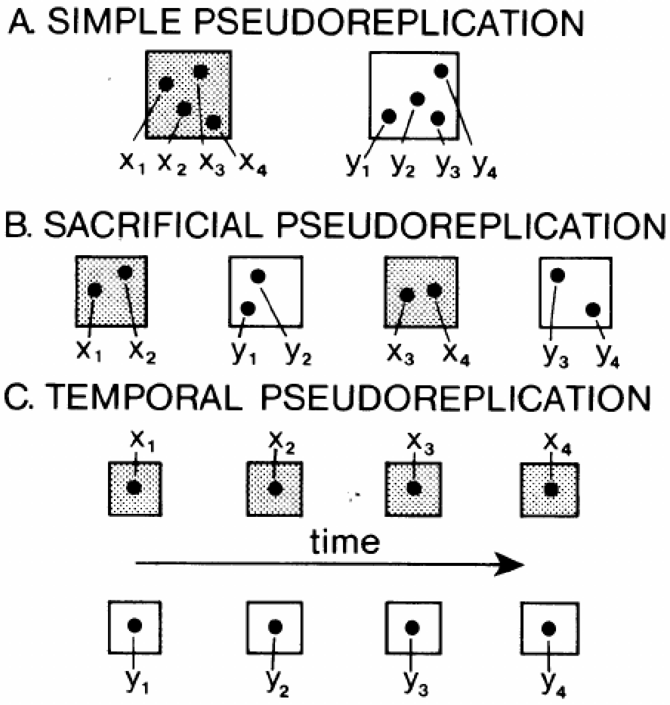


FIG. 5. Schematic representation of the three most common types of pseudoreplication. Shaded and unshaded boxes represent experimental units receiving different treatments. Each dot represents a sample or measurement. Pseudoreplication is a consequence, in each example, of statistically testing for a treatment effect by means of procedures (e.g., *t* test, *U* test) which assume, implicitly, that the four data for each treatment have come from four independent experimental units (=treatment replicates).

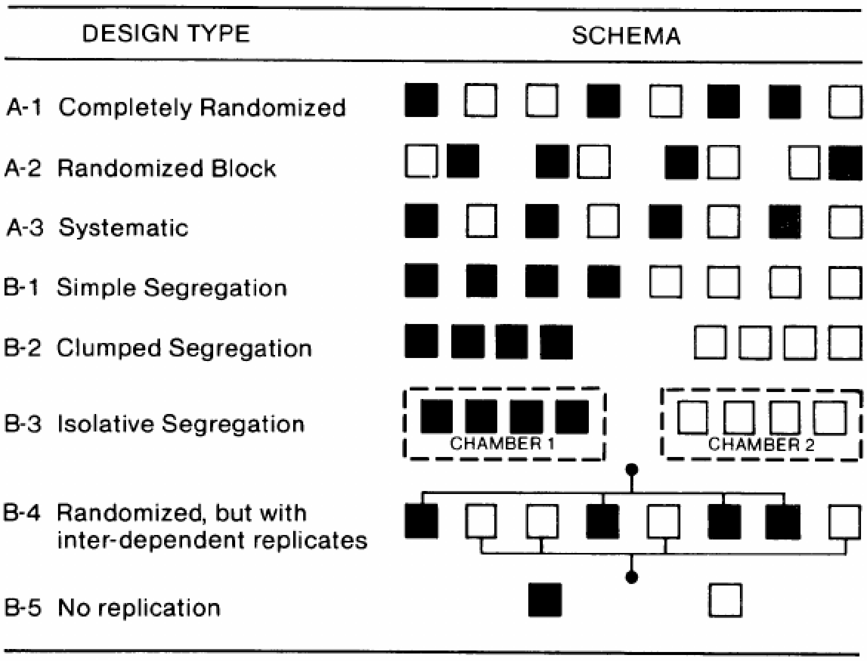


FIG. 1. Schematic representation of various acceptable modes (A) of interspersing the replicates (boxes) of two treatments (shaded, unshaded) and various ways (B) in which the principle of interspersion can be violated.