Power analysis is extremely important in statistics since it allows us to calculate how many chances we have of obtaining realistic results. Sometimes researchers tend to underestimate this aspect and they are just interested in obtaining significant p-values. The problem with this is that a significance level of 0.05 does not necessarily mean that what you are observing is real.
In the book "Statistics Done Wrong" by Alex Reinhart (which you can read for free here: https://www.statisticsdonewrong.com/) this problem is discussed with an example where we can clearly see that a significance of 0.05 does not mean that we have 5% chances of getting it wrong, but actually we have closer to 30% chances of obtaining unrealistic results. This is because there are two types of errors in which we can incur (for example when performing an ANOVA), the type I (i.e. rejecting a null hypothesis when it is actually true) and type II (i.e. accepting a null hypothesis when it is actually false).

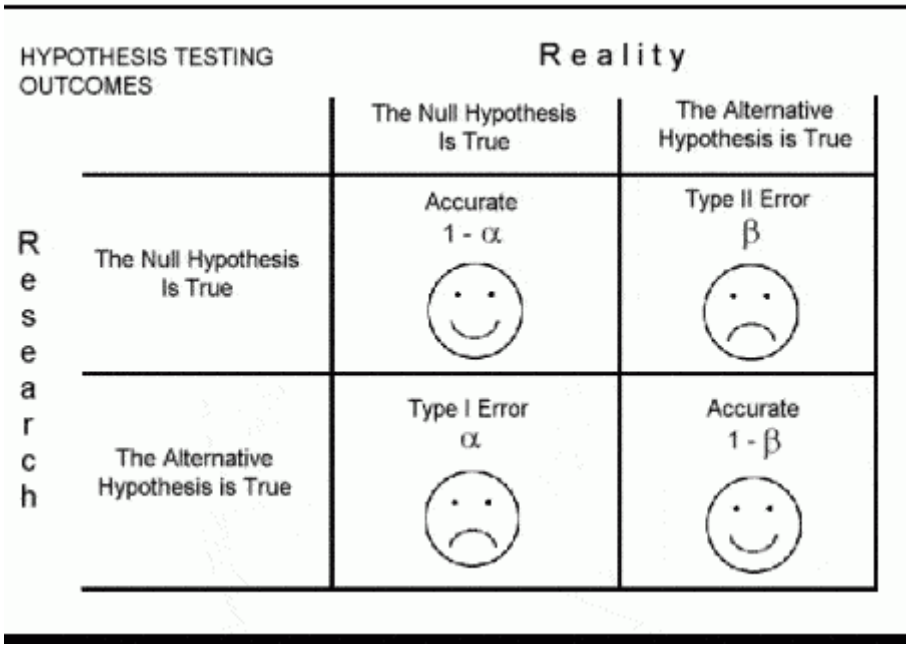

Image taken from: https://datasciencedojo.com/

The probability of incurring in a type I error is indicated by α (or significance level) and usually takes a value of 5%; this means that we are happy to consider a scenario where we have 5% chances of rejecting the null hypothesis when it is actually true. If we are not happy with this, we can further decrease this probability by decreasing the value of α (for example to 1%). On the contrary the probability of incurring in a type II error is expressed by β, which usually takes a value of 20% (meaning a power of 80%). This means we are happy to work assuming that we have a 20% chance of accepting the null hypothesis when it is actually false.
If our experiment is not designed properly we cannot be sure whether we actually incurred in one of these two errors. In other words, if we run a bad experiment and we obtain a insignificant p-value it may be that we incurred in a type II error, meaning that in reality our treatment works but its effect cannot be detected by our experiment. However, it may also be that we obtained a significant p-value but we incurred in a type I error, and if we repeat the experiment we will find different results.
The only way we can be sure to run a good experiment is by running a power analysis. By definition power is the probability of obtaining statistical significance (not necessarily a small p-value, but at least a realistic outcome). Power analysis can be used before an experiment to test whether our design has good chances of succeeding (*a priori*) or after to test whether the results we obtained were realistic.

Effect Size

A simple and effective definition of effect size is provided in the book "Power Analysis for Experimental Research" by Bausell & Li. They say:

"effect size is nothing more than a standardized measure of the size of the mean difference(s) among the study's groups or of the strength of the relationship(s) among its variables".

Despite its simple definition the calculation of the effect size is not always straightforward and many indexes have been proposed over the years. Bausell & Li propose the following definition, in line with what proposed by Cohen in his "Statistical Power Analysis for the Behavioral Sciences":

$$ES = d = \frac{\bar{y}_B - \bar{y}_A}{SD_{pooled}}$$

where ES is the effect size (in Cohen this is referred as d). In this equation, Ya is the mean of the measures for treatment A, and Yb is the mean for treatment B. The denominator is the pooled standard deviation, which is computed as follows:

$$SD_{pooled} = \sqrt{\frac{(SD_B^2 + SD_A^2)}{2}}$$

where SD are the standard deviation for treatments B and A.

This is the main definition but then every software or functions tend to use indexes correlated to this but not identical. We will see each way of calculating the effect size case by case.

One-Way ANOVA

Sample size

For simple models the power calculating can be performed with the package pwr:

```
library(pwr)
```

In the previous post (Linear Models) we worked on a dataset where we tested the impact on yield of 6 levels of nitrogen. Let's assume that we need to run a similar experiment and we would like to know how many samples we should collect (or how many plants we should use in the glass house) for each level of nitrogen. To calculate this we need to do a power analysis.

To compute the sample size required to reach good power we can run the following line of code:

```
pwr.anova.test(k=6, f=0.25, sig.level=0.05, power=0.8)
```

Let's start describing the options from the end. We have the option power, to specify the power you require for your experiment. In general, this can be set to 0.8, as mentioned above. The significance level is alpha and usually we are happy to accept a significance of 5%. Another option is k, which is the number of groups in our experiment, in this case we have 6 groups. Clearly if we were considering two treatments, the first with 6 levels and the second with 3, k would have been 6*3=18.
Finally we have the option f, which is the effect size. As I mentioned above, there are many indexes to express the effect size and f is one of them.
According to Cohen, f can be expressed as:

$$f = \frac{\sigma_m}{\sigma}$$

where the numerator is the is the standard deviation of the effects that we want to test and the denominator is the common standard deviation. For two means, as in the equation we have seen above, f is simply equal to:

$$f = \frac{1}{2}d$$

Clearly, before running the experiment we do not really know what the effect size would be. In some case we may have an idea, for example from previous experiments or a pilot study. However, most of the times we do not have a clue. In such cases we can use the classification suggested by Cohen, who considered the following values for f:

Size of effect	*f*
small	.1
medium	.25
large	.4

The general rule is that if we do not know anything about our experiment we should use a medium effect size, so in this case 0.25. This was suggested in the book Bausell & Li and it is based on a review of 302 studies in the social and behavioral sciences. for this reason it may well be that the effect size of your experiment would be different. However, if you do not have any additional information this is the only thing the literature suggest.

The function above returns the following output:

```
> pwr.anova.test(k=6, f=0.25, sig.level=0.05, power=0.8)
   Balanced one-way analysis of variance power calculation
        k = 6
        n = 35.14095
        f = 0.25
   sig.level = 0.05
     power = 0.8
NOTE: n is number in each group
```

In this example we would need 36 samples for each nitrogen level to achieve a power of 80% with a significance of 5%.

Power Calculation

As I mentioned above, sometimes we have a dataset we collected assuming we could reach good power but we are not actually sure if that is the case. In those instances what we can do is the *a posteriori* power analysis, where we basically compute the power for a model we already fitted.

As you remember is the previous post about linear models, we fitted the following:

```
mod1 = aov(yield ~ nf, data=dat)
```

To compute the power we achieved here we first need to calculate the effect size. As discussed above we have several options: d, f and another index called partial eta squared.
Let's start from d, which can be simply calculated using means and standard deviation of two groups, for example N0 (control) and N5:

```
 numerator = (mean(dat[dat$nf=="N5","yield"])-
mean(dat[dat$nf=="N0","yield"]))
 denominator =
sqrt((sd(dat[dat$nf=="N5","yield"])^2+sd(dat[dat$nf=="N0","yield"])^2)/2)
 d = numerator/denominator
```

This code simply computes the numerator (difference in means) and the denominator (pooled standard deviation) and then computes the Cohen's d, which results in 0.38.

Again Cohen provides some values for the d, so that we can determine how large is our effects, which are presented below:

Size of effect	*d*
small	.2
medium	.5
large	.8

From this table we can see that our effect size is actually low, and not medium as we assumed for the *a priori* analysis. This is important because if we run the experiment with 36 samples per group we may end up with unrealistic results simply due to low power. For this reason it is my opinion that we should always be a bit more conservative and maybe include some additional replicates or blocks, just to account for potential unforeseen differences between our assumptions and reality.

The function to compute power is again pwr.anova.test, in which the effect size is expressed as f. We have two ways of doing that, the first is by using the d values we just calculated and halve it, so in this case f = 0.38/2 = 0.19. However, this will tell you the specific effects size for the relation between N0 and N5, and not for the full set of treatments.

At this link there is an Excel file that you can use to convert between indexes of effect size:
http://www.stat-help.com/spreadsheets/Converting%20effect%20sizes%202012-06-19.xls

Another way to get a fuller picture is by using the partial Eta Squared, which can be calculated using the sum of squares:

$$\eta_p^2 = \frac{SS_{treatment}}{SS_{treatment} + SS_{residuals}}$$

This will tell us the average effect size for all the treatments we applied, so not only for N5 compared to N0, but for all of them.
To compute the partial eta squared we first need to access the anova table, with the function anova:

```
> anova(mod1)
Analysis of Variance Table
Response: yield
          Df Sum Sq Mean Sq F value    Pr(>F)
nf         5  23987  4797.4  12.396 6.075e-12 ***
Residuals 3437 1330110  387.0
```

```
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this table we can extract the sum of squares for the treatment (i.e. nf) and the sum of squares of the residuals and then solve the equation above:

```
> EtaSQ = 23987/(23987+1330110)
> print(EtaSQ)
[1] 0.01771439
```

As for the other indexes, eta squares also has its table of interpretation:

Size of effect	η^2
small	.01
medium	.09
large	.25

The relation between f and eta squared is the following:

$$f = \sqrt{\frac{\mu^2}{(1-\mu^2)}}$$

so to compute the f related to the full treatment we can simply do the following:

```
> f = sqrt(EtaSQ / (1-EtaSQ))
> print(f)
[1] 0.1342902
```

So now we have everything we need to calculate the power of our model:

```
> pwr.anova.test(k=6, n=571, f=f, sig.level=0.05)
   Balanced one-way analysis of variance power calculation
        k = 6
        n = 571
        f = 0.1342902
   sig.level = 0.05
      power = 1
NOTE: n is number in each group
```

To compute the power we need to run again the function pwr.anova.test, but this time without specifying the option power, but replacing it with the option n, which is the number of samples per group.
As you remember from the previous post this was an unbalanced design, so the number of samples per group is not constant. We could either use a vector as input for n, with all the samples per each group. In that case the function will return a power for each group. However, what I did here is putting the lowest number, so that we are sure to reach good power for the lowest sample size.

As you can see even with the small effect size we are still able to reach a power of 1, meaning 100%. This is because the sample size is more than adequate to catch even such a small effect. You could try to run again the sample size calculation to actually see what would be the minimum sample requirement for the observed effect size.

Linear Model

The method we have seen above is only valid for one-way ANOVAs. For more complex model, which may simply be ANOVA with two treatments we should use the function specific for linear models.

Sample Size

To calculate the sample size for this analysis we can refer once again to the package pwr, but now use the function pwr.f2.test.
Using this function is slightly more complex because here we start reasoning in terms of degrees of freedom for the F ratio, which can be obtained using the following equation:

$$F = \frac{\mathrm{MS}_{\mathrm{between}}}{\mathrm{MS}_{\mathrm{within}}}$$

From: https://cnx.org/contents/crKQFJtQ@14/F-Distribution-and-One-Way-ANO

where MS between is the mean square variance between groups and MS within is the mean square variance within each group.
These two terms have the following equations (again from: https://cnx.org/contents/crKQFJtQ@14/F-Distribution-and-One-Way-ANO) :

$$\mathrm{MS}_{\mathrm{between}} = \frac{\mathrm{SS}_{\mathrm{between}}}{\mathrm{df}_{\mathrm{between}}} = \frac{\mathrm{SS}_{\mathrm{between}}}{k-1}$$

$$\mathrm{MS}_{\mathrm{within}} = \frac{\mathrm{SS}_{\mathrm{within}}}{\mathrm{df}_{\mathrm{within}}} = \frac{\mathrm{SS}_{\mathrm{within}}}{n-\mathrm{k}}$$

The degrees of freedom we need to consider are the denominators of the last two equations. For an *a priori* power analysis we need to input the option u, with the degrees of freedom of the numerator of the F ratio, thus MS between. As you can see this can be computed as k-1, for a one-way ANOVA.
For more complex model we need to calculate the degrees of freedom ourselves. This is not difficult because we can generate dummy datasets in R with the specific treatment structure we require, so that R will compute the degrees of freedom for us.
We can generate dummy dataset very easily with the function expand.grid:

```
> data = expand.grid(rep=1:3, FC1=c("A","B","C"), FC2=c("TR1","TR2"))
> data
  rep FC1 FC2
1   1   A TR1
2   2   A TR1
3   3   A TR1
4   1   B TR1
5   2   B TR1
6   3   B TR1
```

```
7   1   C TR1
8   2   C TR1
9   3   C TR1
10  1   A TR2
11  2   A TR2
12  3   A TR2
13  1   B TR2
14  2   B TR2
15  3   B TR2
16  1   C TR2
17  2   C TR2
18  3   C TR2
```

Working with expand.grid is very simple. We just need to specify the level for each treatment and the number of replicates (or blocks) and the function will generate a dataset with every combination.
Now we just need to add the dependent variable, which we can generate randomly from a normal distribution:

```
data$Y = rnorm(nrow(data))
```

Now our dataset is ready so we can fit a linear model to it and generate the ANOVA table:

```
> mod.pilot = lm(Y ~ FC1*FC2, data=data)
> anova(mod.pilot)
Analysis of Variance Table
Response: Y
      Df Sum Sq Mean Sq F value Pr(>F)
FC1    2 0.8627 0.4314 0.3586 0.7059
FC2    1 3.3515 3.3515 2.7859 0.1210
FC1:FC2  2 1.8915 0.9458 0.7862 0.4777
Residuals 12 14.4359 1.2030
```

Since this is a dummy dataset all the sum of squares and the other values are meaningless. We are only interested in looking at the degrees of freedom.
To calculate the sample size for this analysis we can refer once again to the package pwr, but now use the function pwr.f2.test, as follows:

```
pwr.f2.test(u = 2, f2 = 0.25, sig.level = 0.05, power=0.8)
```

The first option in the function is u, which represents the degrees of freedom of the numerator of the F ratio. This is related to the degrees of freedom of the component we want to focus on. As you probably noticed from the model, we are trying to see if there is an interaction between two treatments. From the ANOVA table above we can see that the degrees of freedom of the interaction are equal to 2, so that it what we include as u.
Other options are again power and significance level, which we already discussed. Moreover, in this function the effect size is f2, which is again different from the f we've seen before. F2 again has its own table:

Size of effect	f^2
small	.02
medium	.25
large	.4

Since we assume we have no idea about the real effect size we use a medium value for the *a priori* testing.

The function returns the following table:

```
> pwr.f2.test(u = 2, f2 = 0.25, sig.level = 0.05, power=0.8)
   Multiple regression power calculation
        u = 2
        v = 38.68562
        f2 = 0.25
   sig.level = 0.05
     power = 0.8
```

As you can see what the function is actually providing us is the value of the degrees of freedom for the denominator of the F test (with v), which results in 38.68, so 39 since we always round it by excess.
If we look to the equation to compute MS withing we can see that the degrees of freedom is given by n-k, meaning that to transform the degrees of freedom into a sample size we need to add what we calculated before for the option u. The sample size is then equal to n = v + u + 1, so in this case the sample size is equal 39 + 2 + 1 = 42

This is not the number of samples per group but it is the **total number of samples**.

Another way of looking at the problem would be to compute the total power of our model, and not just how much power we have to discriminate between levels of one of the treatments (as we saw above). To do so we can still use the function pwr.f2.test, but with some differences. The first is that we need to compute u using all elements in the model, so basically sum the decrees of freedom of the ANOVA table, or sum all the coefficients in the model minus the intercept:

```
u = length(coef(mod3))-1
```

Another difference is in how we compute the effects size f2. Before we used its relation with partial eta square, now we can use its relation with the R2 of the model:

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

With these additional element we can compute the power of the model.

Power Calculation

Now we look at estimating the power for a model we've already fitted, which can be done with the same function.

We will work with one of the models we used in the post about Linear Models:

```
mod3 = lm(yield ~ nf + bv, data=dat)
```

Once again we first need to calculate the observed effect size as the eta squared, using again the sum of squares:

```
> Anova(mod3, type="III")
Anova Table (Type III tests)
Response: yield
             Sum Sq  Df F value  Pr(>F)
(Intercept) 747872   1 2877.809 < 2.2e-16 ***
nf           24111   5  18.555 < 2.2e-16 ***
bv          437177   1 1682.256 < 2.2e-16 ***
Residuals   892933 3436
```

```
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this example, I used the function Anova (with option type="III") in the package car just to remind you that if you have an unbalanced design, like in this case, you should use the type III sum of squares.
From this table we can obtain the sum of squares we need to compute the eta squared, for example for nf we will use the following code:

```
> EtaSQ = 24111/(24111+892933)
> EtaSQ
[1] 0.02629209
```

Then we need to transform this into f2 (of f squared), which is what the pwr.f2.test function uses:

```
> f2 = EtaSQ / (1-EtaSQ)
> f2
[1] 0.02700203
```

The only thing we need to do now is calculating the value of v, i.e. the denominator degrees of freedom. This is equal to the n (number of samples) – u – 1, but a quick way of obtaining this number is looking at the anova table above and take the degrees of freedom of the residuals, i.e. 3436.

Now we have everything we need to obtain the observed power:

```
> pwr.f2.test(u = 5, v = 3436, f2 = f2, sig.level = 0.05)
   Multiple regression power calculation
        u = 5
        v = 3436
        f2 = 0.02700203
   sig.level = 0.05
     power = 1
```

which again returns a very high power, since we have a lot of samples.

Source:

https://www.r-bloggers.com/power-analysis-and-sample-size-calculation-for-agriculture/