

Designing for power

Class 7 in MATH 615: Approaches to Quantitative Analysis in the Life Sciences

Effect size, errors and power. Experimental design. Sampling strategies — blocks and strata. Variations on ANOVA — nested effects, split plots, repeated measures.

Setup

Why do experiments?

Minimize error term.

Minimize number of predictor variables.

Control range of predictors.

Significance and effect size

Effect size measures the *strength* of a relationship between variables (as opposed to a significance test which determines whether the relationship likely occurred by chance or not). Remember, a relationship may be significant, but not necessarily interesting if the effect size is small.

One common measure of effect size in regression-type data is one we already know: R^2 , a.k.a. the **coefficient of determination** (or the proportion of variance explained). This is always positive.

A related measure is R , the **correlation coefficient**, which ranges from -1 to 1 , and thus incorporates the sign of the relationship.

Neither of these measures incorporates the absolute magnitude of the slope term, but rather its size relative to the remaining error in the response.

For ANOVA-type data with two groups, the most common statistic is Cohen's d :

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

where \bar{x}_1 and \bar{x}_2 are the means of the two groups, and s is the standard deviation of groups around their means (assumed to be the same). There are many others.

'Type' errors revisited and power

The concept of 'type' errors applies whenever you decide that something is or is not 'significant.' They are therefore central to the hypothesis-testing approach, which codifies the concept of significance, but are also relevant to the other approaches in which a yes-no significance decision is not built in to the method, but which may nevertheless be applied when an investigator inter-

prets the results. Here we will discuss them in the hypothesis-testing context.

Type I error: rejecting the null hypothesis when it is true (a false positive significance)

Type II error: accepting the null hypothesis when it is false (a false negative significance)

In hypothesis testing, Type I is determined by the α criterion for assigning significance to a p value. Remember, p is the probability of getting a statistic, like F , as large or larger than your observed value *by chance* (i.e., by accident). If you set $\alpha = 0.05$, then in that fraction of experiments you will assign significance when none exists — your Type I error rate. So it make sense that this error rate is usually represented as α .

What about Type II? This is the probability of failing to detect a relationship that does, in fact, exist, and is generally represented as β . We usually think in terms of the complement of the Type II error: the ability to detect a true relationship. This is called **power**.

$$\text{Power} = 1 - \beta$$

β , and therefore power, is determined by *three* things: by the Type I error rate, by your sample size, and by the effect size. It can be quite difficult to calculate.

Type I error: α determines whether you accept a measured statistic as significant or not. The larger it is, the more likely you are to assign significance. This increases your Type I error rate, but it also *decreases* your Type II error rate and therefore increases your power. (Anything that increases the rate at which you assign significance increases the rate at which you do so correctly, as well as incorrectly.)

Sample size: The more data you have, the less likely you are to miss a true relationship. (In a linear model, more data means a larger model sum-of-squares with the same degrees of freedom, which means a larger model mean square, which means a larger F statistic, which is more likely to be significant.)

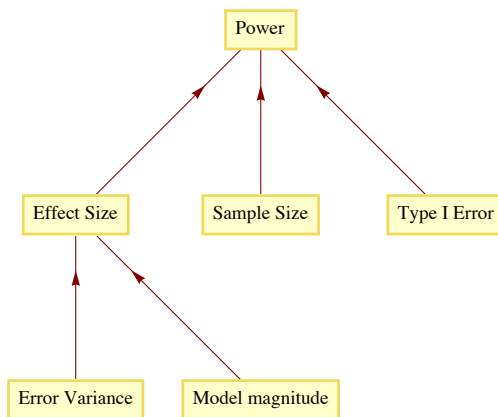
Effect size: The less error variance relative to the model variance (i.e., the greater the effect size), the easier it is to detect the influence of the predictor(s). (If the error variance is very small, the predictors are almost the only thing left determining the response.)

As you can image, power is not simple to calculate exactly. The math required depends in large part on the statistical design — it is different for regressions, ANOVAs, etc. But it's worth doing, and what's more, it is worth doing as part of the process of experimental design, or sampling design if you are collecting data from an unplanned experiment (a.k.a. the real world). Here's why:

You know the Type I error rate up front, because it is up to the investigator to set it.

Given that, if you know any *two* out of the power, the sample size and the effect size, you can calculate the third. Generally, in designing an experiment, you want to know how big a sample to collect — too small, and you won't see the effect you hypothesize even if its there; too big and you are just wasting time and effort, and perhaps wasting money or even killing animals. (Institutional animal welfare supervisors generally insist that you experiment on the smallest number of animals possible consistent with having a good chance of testing your hypothesis, so this kind of analysis is mandatory.)

So, for example, you could specify the effect size you want to be able to detect, and the probability you want to have of detecting that, and calculate the sample size necessary to give you that probability. In practice, the effect size is a combination of an interesting thing (the effect of the predictors) and an uninteresting thing (the error variance), and scientists would like to be able to detect a certain level of the effect of the predictor (e.g., the slope of a regression), so you need an estimate of the error variance in order to convert that to an effect size. This can often be obtained from a simple pilot experiment — you don't need to do the whole thing.



■ Experimental design power analysis steps

1. **Decide on the Type I error you will use.** This depends on the nature of your research. It is typically 0.05 for ‘pure science’ but might be smaller if making a “Type 1” mistake could be very costly, as in medical research.
2. **Decide on the minimum magnitude of the effect you want to be able to detect.** This is where biological significance comes in. What will you (and your peers) consider important enough to take notice of? Note that picking a small magnitude will lead to you needing a larger sample size (and therefore effort) *while at the same time* perhaps making people less excited about a positive result if the magnitude is indeed small.
3. **Decide on the power, the probability you want to have of being able to detect the effect.** This comes down to logistics and personality! For example, if setting up an experiment is much harder than actually collecting data once it’s running, then it makes sense to be more sure of detecting your hypothesised effect from any given experiment (rather than running the risk of having to repeat it). So you would pick a large value, such as 0.95, requiring a larger sample size. If the experiment requires very little set-up, you might take a bigger risk of missing your effect, knowing that you could always re-do the experiment (i.e., collect more sample) later.
4. **Estimate the error variance.** This might involve a simple experiment collecting some data under the same general conditions as the experiment, using the same methods and equipment, but without actually applying any treatments (so all measurements have the same expected value). A fairly small sample (~30) will give you a decent estimate of the variance around that expectation. This is where your skill as an experimenter comes in — a careful experiment in controlled conditions will have a small error variance.
5. **Choose your treatment values (if necessary).** If your treatments are purely qualitative, then this step may not be relevant. If they are quantitative, then you need to think about the range of values you will use. Generally, the larger the range, the larger your effect size. But there may be good reasons not to use extreme values. Usually there is some practical application of any result of your experiment, and this can provide a guideline as to sensible values. If you are testing the effect of fertilizer treatment on plant growth, it makes sense to use treatments in the ‘normal’ range (e.g., what farmers typically use).
6. **Calculate the required sample size.** See below for methods.
7. **Re-evaluate your experiment (or your life)!** The first time you do the process above, you may find that you need ten million samples, taking 13 years to collect. At that point, you must re-evaluate something. Are you looking for too small a magnitude of the effect? If you make it larger, you will need less samples (but might miss a small but real effect). Does your entire career depend in getting a positive result? If not, reduce your power. How bad would a false positive be? If not too bad, you might increase your acceptable Type I error rate. Do you really enjoy doing experiments/collecting data? Do you/others really care about the hypothesis? If not, perhaps choose something easier to study! Are you testing a cure for cancer, or just trying to graduate in a reasonable amount of time?
8. **Repeat 1–7 until an acceptable compromise is reached.**

■ Power analysis via simulations

Simple experimental designs have known formulas for calculating power, and there are online tools that do it for you. For more complicated designs the calculations become very complicated very quickly. Now that we have computers, a more conceptually

straightforward way to calculate power is to *simulate your experiment* based on your choices for steps 1 through 3 and 5 above, and your data from step 4. Let's do a simple example.

Suppose your experimental design is a one-way ANOVA with three qualitative treatments A, B and C. The hypothesis is that the *group* means are different, i.e., that they have some spread, rather than being a single value, so you have to decide on a minimum meaningful measure of this spread. A simple measure of the spread is a standard deviation. Going through the steps, you choose $\alpha = 0.05$, a group mean standard deviation that you would like to be able to detect of 1.5 units (of whatever units you are measuring), and a power to detect this difference of $\beta = 0.8$. You also do a short experiment to measure some values from your experimental set-up with just one of the treatments, and find that the standard deviation of those measurements is 3.2 units, and that they are approximately Normally distributed.

Now you simulate data from this setup, with varying within-treatment sample size, fit your proposed statistical analysis to the results, and assign significance

```

sim[numGroups_, groupSD_, dataSD_,  $\alpha$ _, sampleSize_] :=
Module[{treatments, groupMeans, responses, data, nom, p},
  treatments = Flatten[Table[Table[i, {sampleSize}], {i, 1, numGroups}]];
  groupMeans = RandomReal[NormalDistribution[0, groupSD], {numGroups}];
  responses = Flatten[
    Map[RandomReal[NormalDistribution[#, dataSD], {sampleSize}] &, groupMeans]];
  data = Transpose[{treatments, responses}];
  nom = LinearModelFit[data, {Group}, {Group}, NominalVariables  $\rightarrow$  Group];
  p = nom["ANOVATablePValues"][[1]];
  If[p <  $\alpha$ , 1, 0]
]

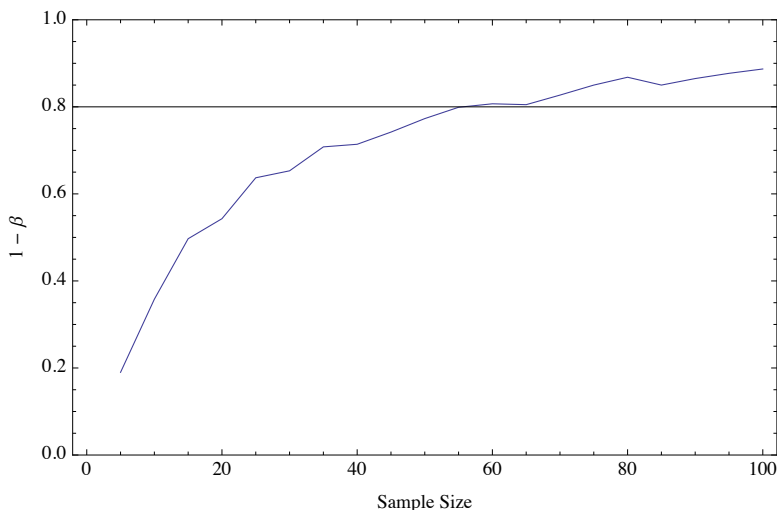
sim[3, 1.5, 3.2, 0.05, 50]

0

power[numGroups_, groupSD_, dataSD_,  $\alpha$ _, sampleSize_] :=
Total[Table[sim[numGroups, groupSD, dataSD,  $\alpha$ , sampleSize], {1000}]] / 1000.

Timing[powerGraph = Table[
  {sampleSize, power[3, 1.5, 3.2, 0.05, sampleSize]}, {sampleSize, 5, 100, 5}];]
{199.284, Null}

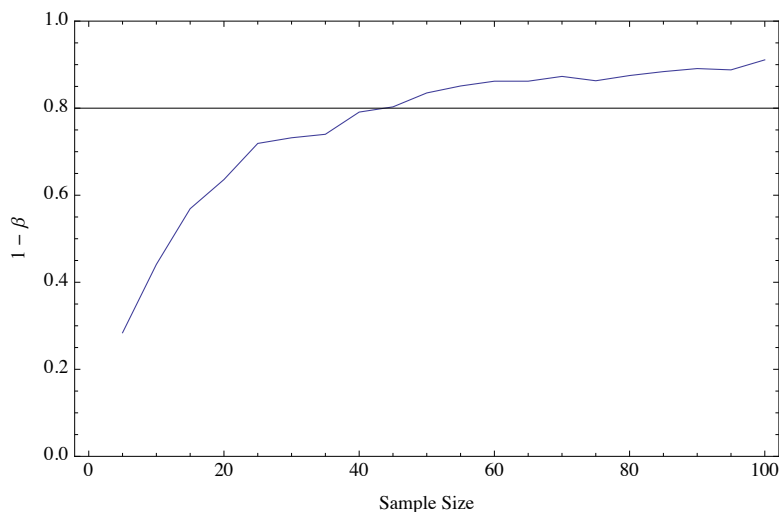
```



The required sample size for having a 0.8 chance of detecting group differences assuming a group standard deviation of 1.5 is approximately 60. Note that this is a minimum. If your sample size is larger, you will be more likely to detect the same relation-

ship and/or able to detect a smaller relationship. If you accepted a lower value for β , obviously you would need a smaller sample size. Not so obviously, you could also be more generous about Type I errors by making α larger:

```
powerGraph =
  Table[{sampleSize, power[3, 1.5, 3.2, 0.10, sampleSize]}, {sampleSize, 5, 100, 5}];
ListLinePlot[powerGraph, AxesOrigin -> {0, 0.8}, Frame -> True,
  FrameLabel -> {"Sample Size", "1 -  $\beta$ "}, PlotRange -> {All, {0, 1.00001}}]
```



In this case the required sample size is approximately 45.

Let's do a regression example, with four *quantitative* levels of treatment, 0, 10, 20, 30. The hypothesis is that the response varies linearly with the treatment, and the treatment values are specified by the experimenter. The slope term that you would like to be able to detect is 0.1. You choose $\alpha = 0.05$, and a power to detect a significant slope of $\beta = 0.9$. You also do a short experiment to measure some values from your experimental set-up with just one of the treatment values, and find that the standard deviation of those measurements is 2.1 units, and that they are approximately Normally distributed.

Now you simulate data from this setup, with varying within-treatment sample size, fit your proposed statistical analysis to the results, and assign significance

```
sim2[groupMeans_, slope_, dataSD_,  $\alpha$ _, sampleSize_] :=
  Module[{treatments, responses, data, nom, p},
    treatments = Flatten[Table[Table[i, {sampleSize}], {i, groupMeans}]];
    responses =
      Flatten[Map[RandomReal[NormalDistribution[# * slope, dataSD], {sampleSize}] &,
        groupMeans]];
    data = Transpose[{treatments, responses}];
    nom = LinearModelFit[data, {Group}, {Group}];
    p = nom["ANOVATablePValues"][[1]];
    If[p <  $\alpha$ , 1, 0]
  ]

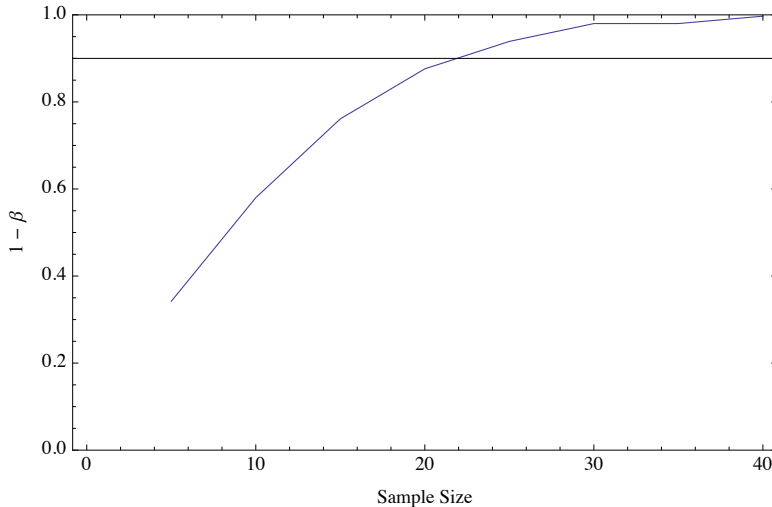
sim2[{0, 10, 20, 30}, 0.1, 3.2, 0.05, 10]

1

power2[treatmentMeans_, slope_, dataSD_,  $\alpha$ _, sampleSize_] :=
  Total[Table[sim2[treatmentMeans, slope, dataSD,  $\alpha$ , sampleSize], {1000}]] / 1000.
```

```
powerGraph = Table[{sampleSize, power2[{0, 10, 20, 30}, 0.1, 3.2, 0.05, sampleSize]},
  {sampleSize, 5, 40, 5}];
```

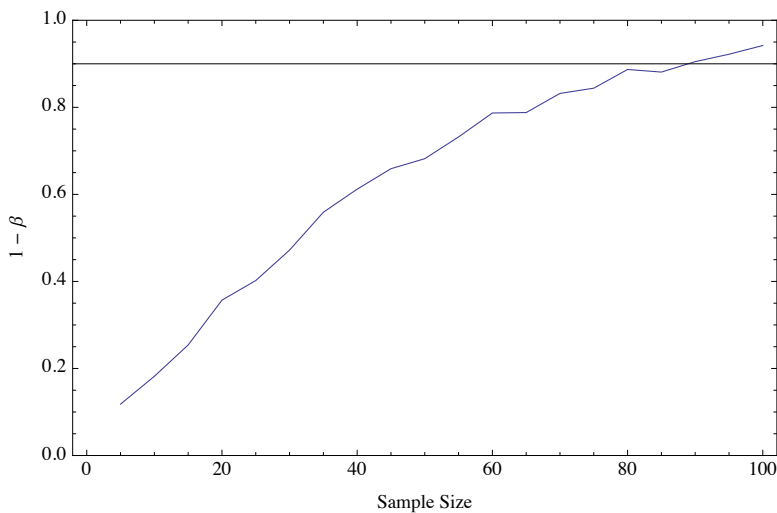
```
ListLinePlot[powerGraph, AxesOrigin -> {0, 0.9}, Frame -> True,
  FrameLabel -> {"Sample Size", "1 -  $\beta$ "}, PlotRange -> {All, {0, 1.00001}}]
```



The required sample size is about 22 observations per treatment. If we halve the minimum detectable slope to 0.05, then the required sample size jumps to about 80 observations.

```
powerGraph = Table[{sampleSize, power2[{0, 10, 20, 30}, 0.05, 3.2, 0.05, sampleSize]},
  {sampleSize, 5, 100, 5}];
```

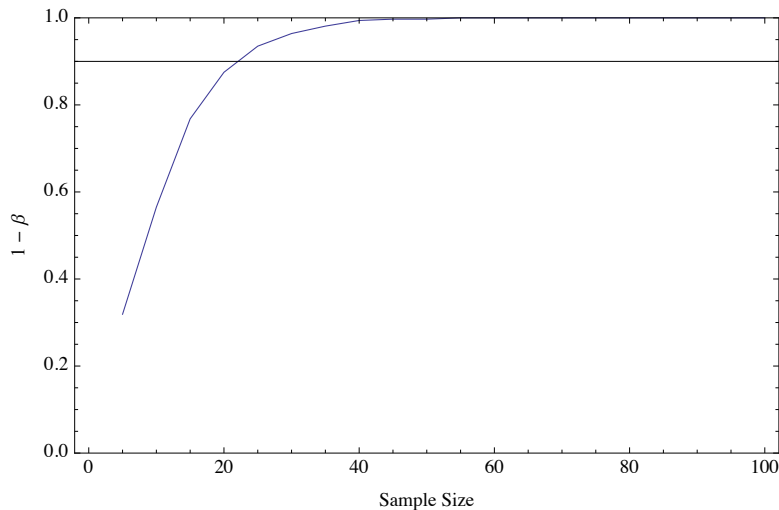
```
ListLinePlot[powerGraph, AxesOrigin -> {0, 0.9}, Frame -> True,
  FrameLabel -> {"Sample Size", "1 -  $\beta$ "}, PlotRange -> {All, {0, 1.00001}}]
```



Finally, let's double the range of our predictor (assuming that we are still within a sensible range of values).

```
powerGraph = Table[{sampleSize, power2[{0, 20, 40, 60}, 0.05, 3.2, 0.05, sampleSize]},
  {sampleSize, 5, 100, 5}];
```

```
ListLinePlot[powerGraph, AxesOrigin -> {0, 0.9}, Frame -> True,
FrameLabel -> {"Sample Size", "1 -  $\beta$ "}, PlotRange -> {All, {0, 1.00001}}]
```



Now we are back to needing only about 22 observations per treatment.

■ When you don't control the predictors

What if you are not doing a completely designed experiment, but rather designing a sampling strategy for the real world? In that case, you may not be able to control the values of your predictor variables, but you need to know approximately what those values will be in order to do a power analysis. (The greater the variance in predictors, the easier it is to find a relationship.) So you might need another preliminary round of observations of your predictor values alone, followed by fitting a distribution to them from which you can draw in your simulation. (Alternatively, in the simulation you could simply draw from your observed values.)

■ Bottom line

Power analysis can be a pain, but *you can't do a serious experiment or sampling protocol without it*. Who wants to collect far more data than they need? Or collect a lot, only to find that it's not nearly enough for the question they wanted to answer? As well as giving you a quantitative sample size goal, the analysis forces you to think carefully about your larger goals, such as 'what result would be useful or interesting?' It also forces you to figure out the analysis method first, and even set it up (if you use a simulation approach, you actually *do* the analysis on your simulated data). So once you have collected your data, you just plug it in and you're done.

Don't be the student who collects data without a clear plan, and only then, armed with a spreadsheet of numbers, thinks about an analysis method, only to find that the data are not appropriate or adequate for the question they were (vaguely) thinking about. ("I kind of wanted to show that *this* is related to *this*.") These students often find their way to me...

This should also sell you on the usefulness of learning a programming tool, whether it be a statistics-oriented package like R or a general purpose package like *Mathematica* or MATLAB. Some other dedicated statistics have experimental design modules that cover the 'standard' types of design (ANOVA, regression, etc.).

Error rates, multiple tests and Bonferroni corrections

Recap of last week.

Sampling

There are special kinds of sampling protocol which can improve power.

Key goal: sample should be *representative* of the population.

Classic approach: sample should be *large* and *random*.

If it is small, it might (by chance) not be representative.

■ Replication ('within' sample size)

Replication is the sample size *within treatment combinations*. Replicates are independent observations from the same combination of treatments. Generally, the more treatments you have, the more combinations you have, the more samples you need overall to disentangle them all. Also, greater power is achieved when sample size within treatments is equal.

■ Warning: pseudoreplication!

Pseudoreplication is when you treat observations as independent replicates when they are not. Suppose you have square plots receiving different fertilizer treatments, and your response is plant biomass. The treatments are applied to the plots, and so the biomass of a single plot is the correct observation. Proper replicates would be multiple plots with replicated treatments. If you instead measured the biomass of each plant within your plot and treated *those* as replicates, that would be **pseudoreplication**. The plant measurements are not independent because all the plants in one plot share the *exact same* treatment, and also share any localized environmental variation that contributes to the error term. Pseudoreplication leads to an underestimated error variance, and therefore a greater chance of finding significance, and therefore an increased Type I error rate.

See Hurlbert, S. (2004) On misinterpretations of pseudoreplication and related matters: a reply to Oksanen. *Oikos* 104(3): 591–597. DOI 0.1111/j.0030-1299.2004.12752.x

■ Randomization

Treatments should be allocated to experimental units in a random manner. I hope I convinced you in the first class that humans don't do 'random' very well — we tend to be too systematic, or have unconscious bias. So use a random number generator or something similar to do the assignments.

■ Blocking

In good lab experiments, conditions not related to the treatments are carefully controlled so as to add the minimum amount to the error variance. Sometimes this isn't possible, and in field experiments, it's almost always not possible. One way to minimize the problem is with **blocking**. If there is an obvious uncontrolled environmental factor, then you choose different levels of the factor and, as much as possible, replicate all treatments within each level. The uncontrolled factor then becomes part of the experimental design. Analysis will test for an effect of the new factor, but you can test for treatment effects *independent* of that factor (i.e., as if it wasn't there), especially if you were able to replicate all treatment combinations within level of that factor.

■ Stratified sampling

Sometimes a population is divided into obvious sub-populations, often with different proportions. (An example would a human population divided into ethnic groups.) A truly random sample might fail to include members of the rarer sub-populations. In **stratified sampling**, samples are random *within sub-populations*, and the size of the sample taken from each sub-population is usually based in its proportion in the overall population (proportionate stratified random sampling). If you want to make group

comparisons, this method is essential to ensure that all groups are represented. In fact, if group comparisons were your main goal, you might use disproportionate stratified random sampling and take the same sample size from each group. Even if you only want to estimate whole-population parameters, proportionate stratified random sampling has more power than fully random sampling *if the subpopulations really are homogenous and have different group means of the variable of interest.*

Crossed effects, nested effects and split-plot designs

■ Crossed effects

Of you have more than one factor (type of treatment), you usually design an experiment so that they are *crossed*, meaning that you have all possible treatment *combinations*. The example we discussed before was treating soil with nitrogen- and or phosphorus-based fertilizer. There are two factors, with two levels each, giving four treatment combinations.

<input type="checkbox"/>	No N	N
No P	<input type="checkbox"/>	<input type="checkbox"/>
P	<input type="checkbox"/>	<input type="checkbox"/>

A crossed design allows you to test for interaction effects, when the effect of a combination of treatments is different from the sum of their individual effects.

In the case of blocking, the block factor becomes another crossed effect. Suppose your plots were in a field, which had a wet half and a dry half, and you replicated all your treatments a number of times in each half of the field. You would now have a three-way design with $2 \times 2 \times 2 = 8$ treatment combinations.

Here's another version of blocking: Your plants are in 196 small pots, spread across four environmental chambers that hold 48 pots each. The chambers are of different design, and different ages, and although they provide identical conditions in theory, you can't be sure. You would make sure that each chamber held 12 pots of each of the four fertilizer treatment combinations, and then make 'chamber' a factor (with four levels), giving you $2 \times 2 \times 4 = 16$ treatment combinations. That way you could test for chamber differences, and get a more powerful test of fertilizer treatment effects by factoring out any chamber effects (which would otherwise add noise).

■ Nested effects

But now consider the following: you have a single treatment of light intensity, and this is provided by a series of smaller growth chambers with controllable lighting. Each chamber holds only eight pots, and you have four chambers with each of three intensity levels, for a total of 12 chambers and 96 pots. If you measure biomass in individual pots, you potentially have pseudoreplication, because all eight pots in a chamber share a single treatment application, and the chambers within a single treatment level might vary in light intensity despite your best efforts. But there is way to test for this by adding chamber as a factor. However, in this case, chambers are not crossed with, but rather **nested** within, the light treatment. Each chamber is associated with only one treatment level. In tabular form, the design can be represented like this:

High light	Chamber 1	<input type="checkbox"/>
High light	Chamber 2	<input type="checkbox"/>
High light	Chamber 3	<input type="checkbox"/>
High light	Chamber 4	<input type="checkbox"/>
Medium light	Chamber 5	<input type="checkbox"/>
Medium light	Chamber 6	<input type="checkbox"/>
Medium light	Chamber 7	<input type="checkbox"/>
Medium light	Chamber 8	<input type="checkbox"/>
Low light	Chamber 9	<input type="checkbox"/>
Low light	Chamber 10	<input type="checkbox"/>
Low light	Chamber 11	<input type="checkbox"/>
Low light	Chamber 12	<input type="checkbox"/>

When you set up an ANOVA-type analysis, you can specify when you have nested effects. If they are present, and you don't incorporate them, they add to the error variance within your groups, reducing your power. If they are present and substantial, you can factor them out and recover some of the power.

■ Split-plot

Split-plot designs are a sort of hybrid between crossed and nested designs. They are used when the application of one of the factors in a design must, for some reason be applied at a different scale than the other. One example might be fertilizer and seed type treatments in an agricultural setting. Fertilizer treatments are often applied most easily at the whole-field level, whereas different seed types can be planted in smaller plots. So you might have a few fields with the fertilizer treatments, and multiple plots in each field with different seed types. The design is crossed in the sense that all combinations of fertilizer and seed type occur, but it is nested in the sense of spatial scale and replication. The 'whole plot' factor is replicated less times than the 'sub-plot' factor, and there is therefore less power to detect differences.

Split-plot designs look the same as crossed designs once you get the data into your spreadsheet — you have to know how the experiment was conducted to analyse it correctly. Most packages have split-plot option that allows you select specify which data columns represent whole-plot and sub-plot factors.

Fixed and random effects

Statistics packages also will allow you to designate factors as either fixed or random. A factor is fixed when the levels under study are the only levels of interest. A factor is random when the levels under study are a random sample from a larger population. A rule of thumb is that if you specified the level of a treatment as an investigator, it's fixed. If the variation was due to things you couldn't control, then it's random. Nested factors such as the unplanned chamber variation in the previous example are generally random (as one would not make a nested factor by choice). Also regression predictors from field observations are typically random. The exact methods of calculating mean squares, and therefore the probability of getting a significant result, vary depending on whether factors are fixed or random.

Number of factors in experiments

The more factors, the more tests (especially if interactions are permitted), so the more likely a false positive or two.

Repeated measures

In repeated measures, individual subjects receive multiple treatments or treatment combinations (ideally, all of them). Thus individual subjects are crossed with treatments, which allows between-subject differences to be factored out, and therefore increases power for detecting treatment differences — a strong advantage.

□	Person 1	Person 2	Person 3	Person 4
Treatment 1	□	□	□	□
Treatment 2	□	□	□	□
Treatment 3	□	□	□	□

However, regular ANOVA is *not* ok because repeated measures (the treatments applied to each subject) are not independent of each other. That is because the unique idiosyncrasies of each subject affects their response to all the treatments they receive. There are special options for specifying repeated measures in most analysis tools.

Obviously this can only be done if repeated treatment is possible. It is a common technique in longitudinal clinical studies — age is the repeated factor.

There are disadvantages though, including ‘practice’ effects (subjects change in their response to treatments over time because of adaptations to previous treatments) and differential carry-over effects (similar to practice effect, but the sequential treatments interact more directly). Both of these can be minimized *somewhat* by randomizing the order of treatments between subjects.