R tutorial, annex to session 3

H. Seitz (IGH)
(herve.seitz@igh.cnrs.fr)

October 8, 2014

Contents

1 Why do we need statistics?

The main problem, in experimental science, stems from the irreproducibility of experimental measurements. If the experimenter performs twice the same measurement on identical samples, the numerical output (displayed by the measurement device) is never exactly the same: the two numbers will never be perfectly equal, including all the decimals.

Such lack of reproducibility is due to two causes: first, measurements are never performed exactly identically (two pipettings will never have the exact same volume, even if the pipette is set to the same value; two incubations will never last exactly for the same time, second for second; *etc.*). Secondly, measured samples are also never really identical (two wild-type mice are not genetically identical: they are not clones; and anyway, they didn't eat exactly the same food, they didn't experience the same bacterial infections, *etc.*; two bacterial cultures will never have grown in the exact same conditions of oxygenation and temperature, depending on their localization in the incubator; *etc.*). These phenomena are obvious to anyone, and they have an unavoidable consequence: if I compare two samples (let's say, a wild-type and a mutant), I will certainly get different values (we saw that two *a priori* identical samples will never give the exact same value, with every decimal being the same; clearly, two samples which are meant to be different will also give different results). The whole question is to be able to tell whether the measured difference is due to the phenomenon of interest (here: whether it is due to the mutation that distinguishes the wild-type from the mutant), or whether it is due to the mere fact that I performed two measurements (and, every time I perform two measurements, the results are always, at least a little bit, different).

It is the experimenter's responsibility to make sure that the measurements were performed in conditions as similar as possible (it is better to load the wild-type and mutant samples on the same electrophoresis gel, rather than on two different gels; it is better to pipet reasonnable volumes: you don't pipet 0.5 μL with a 200 μL pipette; *etc.*), but even when the experimenter takes good care of this, the two measurements will unavoidably give results that will be, at least a little bit, different, and it will have to be determined whether the difference is due to the studied phenomenon, or to the unavoidable irreproducibility of experimental measurements.

That is exactly where statistics will help us: estimating the variability within a series of measurements on wild-type samples, the variability within a series of measurements on mutant samples, and comparing that to the observed difference between the two groups. If the wild-types are more similar to each other (and the mutants are more similar to each other) than wild-types are similar to mutants, then the observed difference will be mostly due to the studied phenomenon (here: the mutation of interest). Otherwise, it will be impossible to conclude: the potential effect of the mutation will be obscured by the large variability in measurements, and the experimenter won't be able to tell whether the mutation indeed had an effect on the measurement.

Clearly, for statistical tests to give us that sort of information, you will need to provide them with several replicates of the wild-type measurement, and several replicates of the mutant measurement. The more replicates you have, the more precise the test (it will be able to estimate more precisely the variability within each group, as well as the difference between groups). According to some sort of a long-standing tradition, biologists perform their experiments in three replicates: there is no theoretical justification for that number, and if the experimenter can perform more replicates, he'd better do! We will see later (in section 4) that a given statistical test cannot detect significant differences (with a cutoff of 0.05 on the p-value) if you only provide three replicates of the wild-type and three replicates of the mutant. It is of course

not always possible to prepare dozens of replicates for each measurement (each replicate takes time and uses reagents), but whenever it is possible, the experimenter will be able to detect much smaller differences than with just three replicates. His analysis will be more precise. Another point needs to be emphasized: replicates have to be independent (*i.e.*: each one has to be a full experiment). For example, if the experiment is a measurement of the abundance of an mRNA in mouse liver by qRT-PCR: the experimenter may want to dissect a single mouse, extract RNA from that liver, then perform several RT-PCR reactions on that single RNA sample. These replicates aren't independent: they differ only by the last step (the RT-PCR step). Comparing these replicates will only give us information on the reproducibility of this last step, and it won't assess the reproducibility of RNA extraction from mouse liver, nor the variability that may exist between two mice. This is important, because when he compares the wild-type liver to a mutant liver, he will obviously compare two different mice. In order to capture every possible source of irreproducibility, it is necessary to make sure that the replicates differ as much (between each other) than the two groups of measurements (wild-type and mutant). They have to come from distinct dissections, RNA extractions and RT-PCRs.

2 What is the t-test?

We saw that the main question is to estimate how reliable the measured difference is between wild-type and mutant samples (and we know for sure that there will be some difference). This it the usage of a statistical test named "t-test": you have to provide it with the two series of measurements (the series of wild-type replicates and the series of mutant replicates) and it will calculate a value named "p-value". Most biologists know that the "p-value" measures the reliability of the measured difference, and the smaller the p-value, the more "significant" the difference (which means that it will have a signification from the experimenter's point of view; in our example: a low p-value means that the observed difference is mostly due to the mutation that distinguishes the wild-type from the mutant, more than to the intrinsic irreproducibility of experimental measurements). Yet it would be useful to know exactly the meaning of that "p-value". Here it is: the t-test p-value is the probability that the two sampled populations have the same mean.

It is very easy to misunderstand that definition! You could think that the p-value is the probability that the series of wild-type replicates, and the series of mutant replicates, have the same mean. That would be a trivial question! As the measurements have been done, it is easy to calculate their means. So it would be easy to compare the mean of the wild-type values to the mean of the mutant values. If they are equal, then the probability that they are equal is 1, and if they are different, the probability that they are equal is 0 (hence there could be no intermediate p-value between 0 and 1). And as two measurements always give results which are, at least a little bit, different, then we know for sure (even before performing the experiment !) that the two means will be different. So the p-value would always be 0 (it would not be informative).

Clearly, that was not the definition of the p-value; the difference is due to a word which seems a little bit out of place in the middle of the definition: the word "sampled". Here is how to interpret that definition: the experimenter (knowing that he has to do several replicates of the measurement) can prepare 10 replicates (of his Western blot, of his qPCR, ...). But why would he not prepare 10 more replicates? Or 20 more? Or even, an infinity of replicates? Of course, for all sorts of practical reasons, it won't be possible to prepare an infinite number of replicates (each replicate takes some time; each Western blot replicate uses a little bit of antibody, and there is not an infinite amount of antibody on Earth; ...), but you can imagine

that, in an ideal world, it would be possible to perform an infinite number of replicates. There is, somewhere in an ideal universe, a list of the possible measurements of all these replicates. An infinite list of values, which would contain all the possible results of the measurement, during all the life of the Universe, and even more ... And you can interpret, with a new point of view, the experiment of a biologist performing 10 replicates of his measurement: doing 10 replicates of the measurement, that's like picking randomly 10 values among the infinity of possible values, resulting from an infinity of replicates ...

Well, this ideal population, infinitely large, where the experimenter picks values when he performs his experiment, that is the "sampled population". It is "sampled", because the experimenter samples his 10 (or 20, or 30, ...) replicates when he performs the experiment.

Then the definition of the p-value takes a whole new meaning: the p-value is the probability that the two ideal populations (infinitely large, from where the wild-type and mutant replicates are sampled) have the same mean. See how powerful the t-test is ! It allows you to compare infinite populations of replicates, which will, of course, always be inaccessible to a real experiment. It allows you to reason on a perfect experiment (the one which has an infinite number of replicates, both for the wild-type and the mutant), without having to actually do it.

It is now clear, why the lower the p-value, the more significant the difference: when you provide it with a list of replicates of the wild-type and of the mutant, the t-test will estimate what the two ideal populations (from which the two datasets were sampled) look like, and it will estimate the probability that their means are equal. If it is very unlikely that their means are equal (p-value much smaller than 1), then you will have to conclude that a perfect experiment (with an infinity of replicates in each dataset) would certainly detect a difference between the means of the two series – in other words: when the number of replicates is infinite (which is the best way of eliminating technical sources of irreproducibility in the measurement, because such fluctuations will affect identically each series), there will still be a difference between the series of wild-types, and the series of mutants.

Figure 1 illustrates that process: let's consider two series of values (dataset 1 and dataset 2). These two series are centered on average values close to 15 (top panel) and therefore, the ideal populations from which they were sampled probably have a mean close to 15 (ideal, sampled populations are represented in blue in the middle panel). In that particular example, the t-test p-value equals $\approx$0.74. Now, if one of the two datasets (dataset 2) is shifted by two units towards the right (bottom right panel), the probability that ideal populations have the same mean is much smaller (6.9×10^{-8}).

It is important to realize that no mathematical relationship can allow you to calculate the p-value from the difference between the means of the two datasets: the p-value also depends on other parameters:

- The shape of the distribution of each dataset: it distributions are very tight (in each dataset, values are very close to each other, and the histogram shows a tall and narrow peak), then the t-test will declare, with much more certainty, that the ideal populations have different means (it will be able to estimate more reliably that the two ideal populations are separated from each other, because each of them has been better resolved). Reciprocally, if the two datasets are very wide, it will be harder for the t-test to say that the means of ideal populations are different: the p-value will be larger.

- The number of replicates in each dataset: the more replicates there are, the easier it will be for the t-test to estimate precisely what the ideal, infinite population looks like – so the easier it will be for the t-test to detect a difference, even if it is small, between the means

of the ideal populations. That is the reason why (in the bottom panel of figure 1) the p-value is so low (with just 3 or 4 replicates, with a ≈10% difference between the means of the two datasets, for datasets as widely distributed as here, the t-test would certainly output a much larger p-value). Here, the large number of replicates (200 in dataset 1, 100 in dataset 2) allows the t-test to estimate precisely the ideal populations, hence it can be more affirmative about the difference between their means.

These considerations show that the p-value is not the ultimate criterion: a p-value can be very low, even when the difference between two datasets is small (if their distributions are narrow enough and if the number of replicates is large enough). It is actually possible to decrease the p-value as much as you want, simply by increasing the number of replicates (the required number of replicates to reach any given p-value will depend on the shape of the two distributions). You should not consider that the p-value measures the amplitude of the difference (the amplitude can be measured by calculating the ratio or the difference between the means or the medians, or other estimators, of the two datasets): the p-value only measures how reliable the measured amplitude is.

In sum, each of these two values (amplitude of the difference, and p-value) allows you to measure a characteristic of the difference between the two datasets; each of them is necessary to characterize it, and none is sufficient by itself.

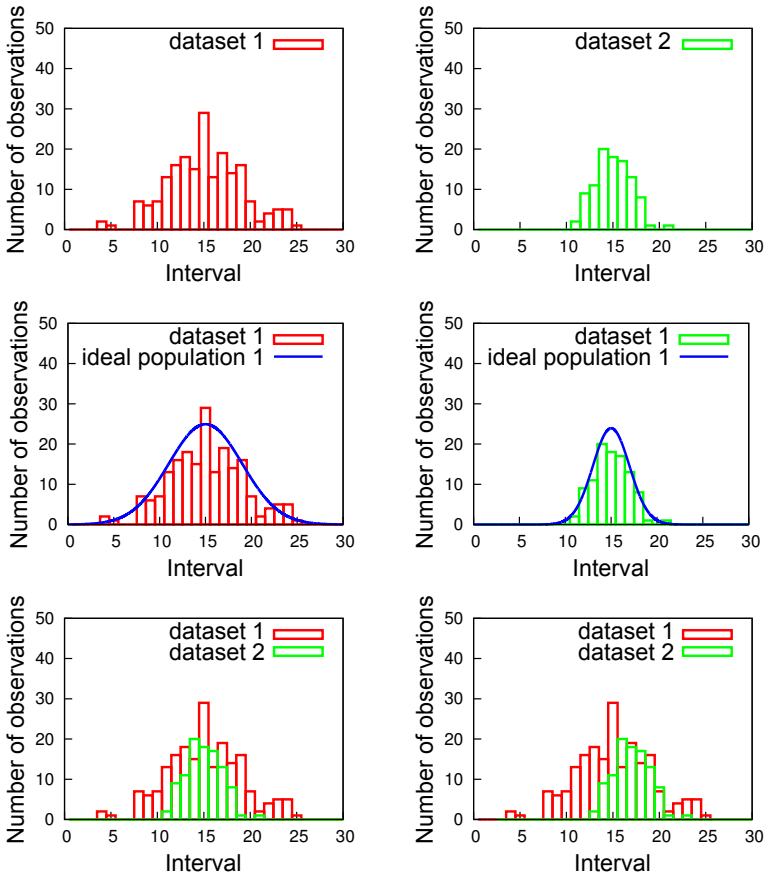

Figure 1: Let's consider two datasets: dataset 1 (composed of 200 replicates, centered on a value close to 15; top left panel) and dataset 2 (composed of 100 replicates, also centered on a value close to 15; top right panel). These histograms represent the number of observations falling in each interval of width=1 (for example, 18 observations of dataset 1 fall between 12.5 and 13.5, so the bar centered on 13 has a height of 18). Here, the two datasets were generated computationnally, they were purposedly sampled from ideal, infinite populations having a mean of 15: these ideal populations are represented in blue on the middle panel. The t-test will estimate the probability that the ideal populations (which are usually unknown) from which the datasets were sampled have the same mean: that probability equals 0.74, indicating that, according to the t-test, it is very likely that these two datasets were sampled from infinite populations that have the same mean (which is indeed the case here). If one of the two datasets is shifted by two units towards the right (bottom right panel; dataset 2 is now centered on a value close to 17), the t-test now considers it very unlikely that the sampled, infinite populations have the same mean (probability=6.9×10^{-8}).

There are several variants of the t-test, and you will have to choose the best one depending on the situation. It is also important to know that the t-test is not applicable in every circumstance: you need to be able to identify these circumstances and to know what alternatives should be chosen. This is what we will discuss now.

3 t-test variants

When the experimenter tries to use the t-test, regardless of the computational program he is using (R, Excel, ...), he has to choose among several options of the test. Here is how to choose the most appropriate ones:

1. "Equal variances" or "different variances" ? The variance measures data dispersion in each dataset (in figure 1, the variance of dataset 1 is larger than that of dataset 2: values are spread further away from their mean). In the initial version of the t-test, invented in the early 20th century (that version is called "Student's t-test"), variances of the two datasets had to be equal for the test to be applicable (in fact, they are never perfectly equal, as always: the two numbers won't be equal, with all the same decimal digits; they just have to be similar, which is usually called "homogeneous" variances). If they are very different, you should not use the t-test's original version, rather a later version (named "Welch's t-test"), which tolerates variance heterogeneity. But if variances are similar ("homogeneous"), it is better to use the test's original version ("t-test with equal variances" or "t-test with homogeneous variances" or "Student's t-test"), which is more precise than Welch's t-test.

2. "Paired" or "unpaired" data? You can use the t-test with paired data when there is an evident logical link between each replicate of one dataset, and one (only one) replicate of the other dataset, and which could possibly affect the result of the measurement. That logical link could be: the two measurements were performed on the same object (*e.g.*, if the experiment is a measurement of body weight performed on patients before and after a diet, there exists an evident logical link between each replicate of the first dataset and one given replicate of the second dataset: they were performed on the same individual). Pairing data gives more accuracy to the test: in that example, it allows the test to ignore heterogeneities in body weight across patients (rather than comparing, as two whole groups, the two datasets, the test will compare each patient's body weight before the treatment, to his body weight after treatment). It is thus preferrable to pair data whenever they can be paired: eliminating heterogeneity between measurement replicates, you will give the test the possibility to detect smaller differences.

3. "One-tailed" or "two-tailed" ? Everything we have seen so far deals with the "two-tailed t-test", which aims at determining whether the two datasets are significantly different. The "one-tailed t-test" will test whether values in one of the two datasets tend to be larger than values in the other dataset (hence it tests only one sense of variation). You should use that variant of the t-test only when you can exclude *a priori* one of the two senses of variation (for example: if the experiment consists in measuring the length of a wood stick, then cutting a piece of it, then measuring it again: it is undisputable that the lengths measured in the second series of measurements have to be smaller than the lengths measured in the first series; each time a wood stick appears to be longer in the second measurement, we know for sure that it is a measurement error). As you can imagine,

that sort of situation is very rare in our job (it is rare to know in advance the sense of variation of a value; in all my career, I never had to use the one-tailed t-test). Not only its usage is very rare, but the one-tailed t-test is dangerous: it divides p-values by two, so it can make unsignificant differences appear significant. Using the one-tailed t-test in an illegitimate case is scientific fraud: so you really have to double-check that there is indeed an undisputable reason to exclude *a priori* a sense of variation before using the one-tailed t-test (and if any of you ever encounters that sort of situation, please let me know: I would be curious to see an example of one-tailed t-test usage in experimental biology !).

4 When can I use the t-test ?

We have seen how useful the t-test can be in biology; unfortunately that test is not applicable in every situation, and before running the test you should verify that it is applicable. The main limitation to t-test applicability deals with the shape of data distribution: for the t-test to be usable, each of the two datasets has to follow a normal distribution (also called "Gaussian distribution", it is the bell-shaped curve of the ideal data distributions shown in blue in figure 1). The normal distribution is described by a precise mathematical equation (every bell-shaped curve does not follow a normal distribution !); of course, a dataset coming from the real world won't perfectly follow the normal distribution (see how the histograms of datasets 1 and 2 do not fit exactly with the blue curves on figure 1). So you don't need the datasets to stick perfectly to a normal distribution – you just need them not to deviate too much from it.

There are statistical tests that can assess whether a dataset follows a normal distribution (the Shapiro-Wilk test, as well as the Kolmogorov-Smirnov test, which will be a little less precise). You feed these tests with a dataset, and they give a p-value, which is the probability that the dataset was sampled from an ideal population, infinitely large, that follows a normal distribution (so you will have to run that test on each of the two datasets, and only use the t-test if each of the two p-values was large – in general, people consider that they have to be larger than 0.05; but they tend to be easily larger than 0.05 whenever the number of replicates is small: the more replicates you have, the more you can trust the result of these tests).

What should you do if at least one of the two datasets deviates too much from a normal distribution (p-value$<$0.05 after a Shapiro-Wilk test or a Kolmogorov-Smirnov test) ? There are two possibilities:

1. Use a test that is more robust than the t-test when datasets are not normally distributed. That test is called the "Wilcoxon test" (or: "Mann-Whitney test"). You can use it just like the t-test (you have to provide it with the two datasets, and it will calculate the probability that they were sampled from infinite populations having the same mean), but that one is also applicable when datasets do not follow a normal distribution. Such robustness has a cost: the Wilcoxon test is much less powerful than the t-test (*i.e.*: for the same two datasets, it will detect significant differences less easily). That difference in statistical power is a real hurdle when the two datasets contain few replicates (in particular, the Wilcoxon test p-value can never be lower than 0.1 when each dataset contains three replicates).

2. Use a mathematical transformation. When a dataset does not follow a normal distribution, it is common that the logarithm of its values follow a normal distribution (logarithm

is a monotonously increasing function, that increases very slowly, so it tends to squeeze differences, hence it will bring values closer to each other; aberrant values, which are often responsible for deviations to normality, then have less influence on the general shape of the distribution). You can then apply the t-test on the logarithm of your data, rather than on the data themselves (and if the logarithms are significantly different, then the untransformed values are also significantly different; but, thanks to the log-transformation, you will get a more precise estimation of the p-value for that difference). Other mathematical transformations can have the same effect (square root, arcsine, ...) and their usage sometimes looks like some sort of a random search for some unjustified mathematical transformation, that will distort the dataset in a way that pleases the experimenter ... It is thus preferable to use a mathematical transformation which has a physical sense (the logarithm of a reagent concentration, for example, has a physical meaning: it can be used to calculate pH's and pKa's; it is less logical to calculate logarithms for values that are not multiplicative, but additive).

The procedure to follow is shown in figure 2. Most statistical tests cited in that figure (Shapiro-Wilk test, Kolmogorov-Smirnov test, Levene test, Wilcoxon test) are not available in Excel. Free software like R (`http://www.r-project.org/`) or PAST (`http://folk.uio.no/ohammer/past/`) provide these tests.

Another limitation of the t-test is evident, but often ignored: that test allows you to compare the reliability of a measured difference between two datasets, nothing more. It doesn't allow you to compare two data distributions, for example.

Let's imagine the experiment is a comparison of the distribution of cells between the cell cycle phases (mitosis, G1 phase, S phase, G2 phase). We have to compare two series of counting values (one in each experimental condition; for example, a wild-type and a mutant). As always, we know we won't find exactly the same distribution between cell cycle phases in both conditions -- – and we would like to know whether the difference is significant. The experimenter could be tempted to use the t-test to compare, one by one, the cell counts in each phase of cell cycle, in both conditions. But two problems are to be expected:

1. As discussed earlier, the t-test needs several replicates of each measurement. Here, the measurement is a cell counting (probably several hundreds of cells were counted): in a sense, the experiment has already been performed many times (in each category – here: in each cell cycle phase – the cell count is the result of the analysis of hundreds of cells). It seems strange to be forced to do several replicates of an experiment ... which already contained hundreds of countings.

2. Assuming that the first problem has been ignored (the experimenter indeed performed several independent countings, of hundreds of cells each), the t-test could give aberrant results: for example, it may show that the number of cells in the G1 phase is significantly different between the two experimental conditions, without detecting significant changes in the other three phases of the cycle (which is possible, if the difference compensating for the difference in G1 phase spreads rather uniformly among the other three phases). How, then could you interpret that result: is the distribution of cells among cell cycle phases affected by the tested conditions ?

That second potential problem illustrates the nature of the experimenter's error: here, the aim was not to compare (category by category) the number of cells in both conditions. The aim was to compare the whole distributions to each other.

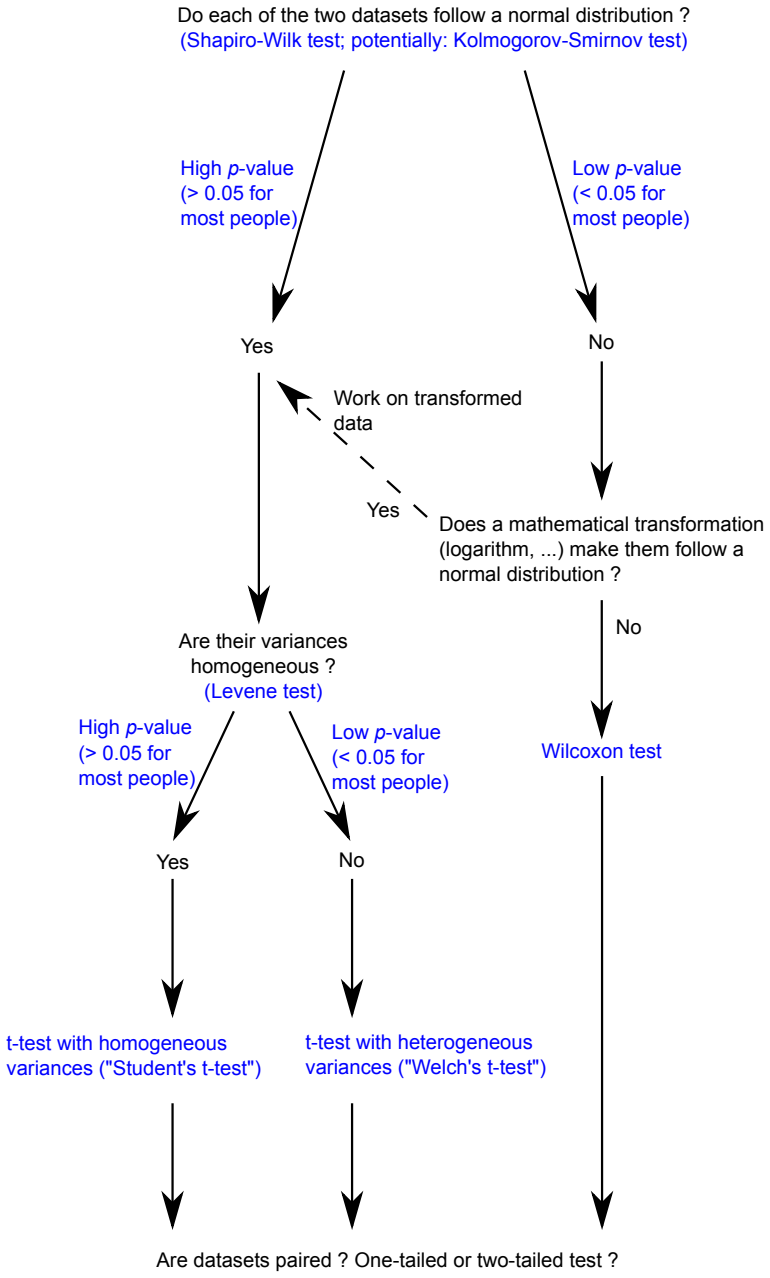

Figure 2: This figure summarizes the procedure to follow to compare two datasets. Note that several tests are available for most questions (normality can be assessed either with the Shapiro-Wilk test or with the Kolomogorov-Smirnov test; both tests generally agree, but if they disagree you should trust the Shapiro-Wilk test, which is more precise; variance homogeneity can also be assessed with other tests than the Levene test: they generally give the same answer).

That sort of problem is inaccessible to the t-test. Another test is needed here: either Pearson's χ^2 test (if cell counts are large enough; in general, one considers that the number of observations in each category should be larger than 5); or Fisher's exact test (which doesn't have such a limitation, but which takes more computational time to run; this is usually not a problem any more, now that computers have become tremendously fast). You have to provide these tests with the series of observation counts, in each experimental condition, and the calculated p-value

is the probability that the two datasets were sampled from ideal, infinite populations that have the same distribution.

It is particularly important to remember that these two tests should only be applied to observations counts, raw counting results, which have not been normalized (for example, you don't run these tests on percentages, whose sum has been arbitrarily normalized to 100 by multiplying every count by a normalization factor). This is because these two tests take the number of observations into account in order to measure how reliable the measured differences are (they consider that 500 is very different from 450, whereas 50 is not very different from 45): such sensitivity to raw count values is a useful feature, and it is lost if you increase or decrease arbitrarily every value by multiplying it by a normalization factor.

5 Conclusion

The notions and illustrations presented in this document do not aim at covering every possible use of statistics in biology; their main purpose is to explain concepts which are much more general, and to warn you against some frequent mistakes in experimental biology publications. They should help the reader understanding the readings necessary to handle more complex problems than the ones cited here.