A practical guide to “significance” beyond *p* values

Peter Nagy, Department of Biophysics and Cell Biology, University of Debrecen

Email: peter.v.nagy@gmail.com, nagyp@med.unideb.hu

Web page: peternagyweb.hu, peternagygroup.com

Biologists, medical doctors and students in these branches of science typically grapple with statistics, but finally many of them reach a basic level of understanding of “significance” and “p-values”. The bad news is that neither of these concepts characterizes sufficiently the reliability of our statistical decisions. This tutorial is intended for those practicing biologists and medical doctors who are interested in the limitations of simply reporting significance or *p* values. This text is not a comprehensive introduction to hypothesis testing, and basic level of understanding of its principles is assumed. Practical, how-to sections are written on a gray background. At the end of the tutorial, a brief summary without any theory is provided.

1. *How to interpret p values?*

Although the aim of this document is to show the limitations of *p* values, it is mandatory to understand clearly the meaning of *p* values so that one can comprehend its limitations. In statistical hypothesis testing, one typically compares a dataset to an assumption, and asks whether it is reasonable to assume that the data is in accordance with this assumption. This assumption is called the *null hypothesis*, typically denoted by H_0. Let us assume that an investigator would like to check if a drug has any effect on serum cholesterol. In any way of realizing this investigation (i.e. paired t-test or unpaired t-test), one will end up determining the difference between the serum cholesterol levels of treated vs. untreated persons. The null hypothesis in this case is set up in the following way if the test is two-sided: the serum cholesterol levels of treated and untreated individuals are identical. Afterwards it is numerically calculated how likely the calculated value of the test statistic is if the null hypothesis is indeed true. The previous statement is worth discussing further:

- What is exactly calculated and what kind of conclusion can we reach? In conventional statistical hypothesis testing, all calculations are based on the assumption that the null hypothesis is correct. Using this assumption (mean cholesterol in treated group=mean cholesterol in untreated group in our example) we can draw the probability distribution function of the calculated values of the test statistic. This curve shows the distribution of test statistic values if the null hypothesis is indeed true (Figure 1). If, according to this figure, the calculated value, or values similar to it, are obtained with a high probability, the null hypothesis is accepted. If, on the other hand, the probability of obtaining the calculated value according to the curve is low, we reason in the following way: although it is not impossible to obtain such a value for the test statistic if the null hypothesis is correct, but it is very unlikely. Therefore, it is more reasonable to assume that an alternative hypothesis is correct (cholesterol levels in the treated and untreated groups are unequal). It is matter of convention how unlikely the calculated value must be under the assumption of the correctness of the null hypothesis so that the alternative hypothesis is preferred. This threshold is called the level of significance, which is usually set to 5%, i.e. if the calculated value of the test statistic falls in the region harboring those extreme values that are obtained with a probability of 5% if the null hypothesis is true, we are going to prefer the alternative hypothesis. However, it must be kept in mind that even in the case of a true null hypothesis, the test statistic will assume values falling into the rejection area, i.e. the null hypothesis will be rejected although it is true. This scenario is called a type I error.

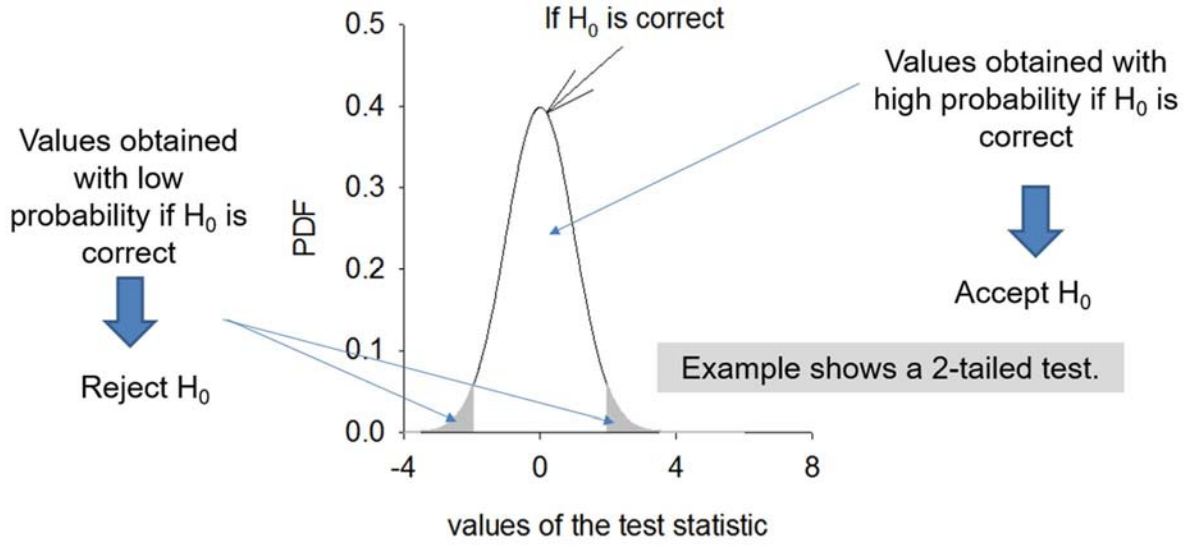


Figure 1. Acceptance or rejection of the null hypothesis. Using the distribution of values of the test statistic *according to the null hypothesis,* two kinds of regions are defined. The values obtained with high probability (white area) constitute the acceptance area. If the calculated value of the test statistic falls in this region, the null hypothesis is accepted (since such values are obtained with high probability if the null hypothesis is correct). Values in the gray region, constituting the rejection area, are obtained with low probability according to the null hypothesis, and consequently the null hypothesis will be rejected. The area of the gray regions, corresponding to the probability with which they are obtained according to the null hypothesis, is equal to the level of significance.

The *p* value gives us a numerical estimate for the likeliness of the calculated value according to the null hypothesis. It gives us the probability that such a value or values farther away from the assumed mean are obtained according to the null hypothesis (Figure 2). If this value is smaller than a threshold value (called the level of significance), the null hypothesis is rejected (Figure 2).

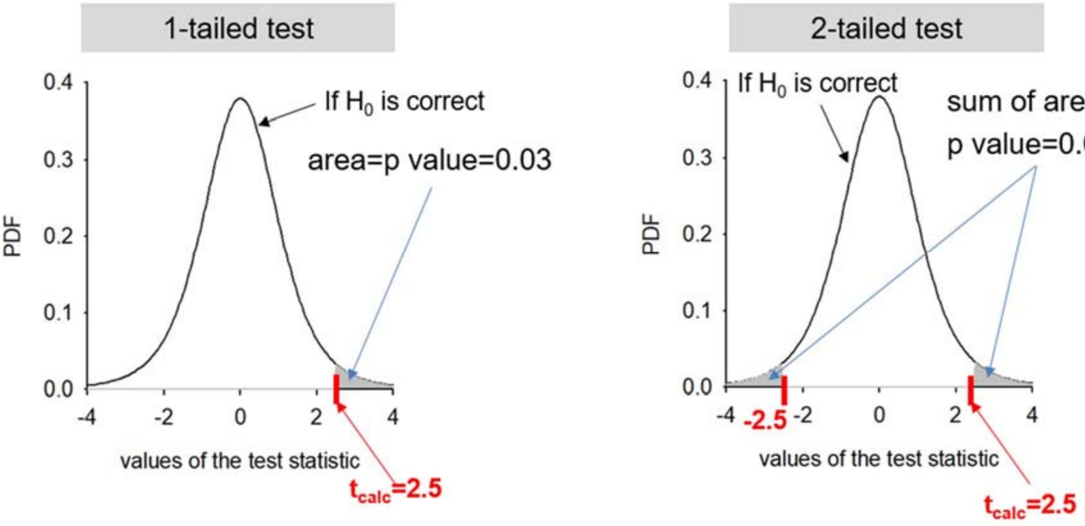


Figure 2. Interpretation of the *p* value. If the calculated value of the t-test is 2.5 (with the degree of freedom equal to 5), the probability that values equal to or more extreme than this value are obtained is 0.03 for a one-sided test. Due to the symmetry of the distribution of values, this probability is 2·0.03=0.06 for a two-sided test. These are the *p* values of the one-tailed and the two-tailed tests.

- Why is the probability of the calculated value analyzed with respect to the null hypothesis, and not the alternative hypothesis? In statistical hypothesis testing, we are always instructed to assume that the null hypothesis is correct. In most cases, this is equivalent to assuming that what we desire to show is not true, e.g. in our example of a drug lowering blood cholesterol, we are assuming that the drug does not work by setting up the null hypothesis the way we always do. Why do we do so? Are we too pessimistic to assume that the drug works? It turns out that we set up the null hypothesis in this way due to the fact that this is the only numerically testable hypothesis (Figure 3). If we assume that the drug does not work, then the expectation for the difference between the treated and untreated samples is a specific value, zero. Consequently, the distribution of statistic values according to this assumption can be drawn, and the calculated value can be analyzed according to this curve. What if we are brave enough to assume that the drug works? So that we can draw the distribution of statistic values according to this hypothesis, we would need to know the assumed difference between the treated and untreated populations. But what to assume? 1, 2, 10? Since we cannot specify any assumed value for the expectation for the difference between the two populations, the distribution of statistic values according to this hypothesis cannot be drawn. Consequently, this hypothesis cannot be numerically tested.

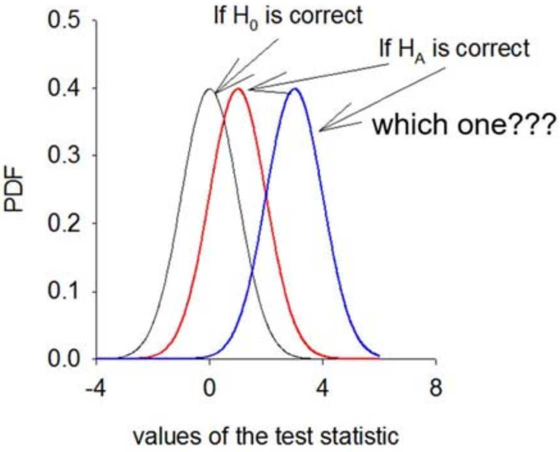


Figure 3. The null hypothesis is the only testable hypothesis. If the null hypothesis is correct, the mean of the difference between the treated and untreated groups is zero, allowing us to draw the distribution of statistic values according to this hypothesis (black curve). If we would prefer to assume that an alternative hypothesis is correct, we would have trouble drawing the distribution of statistic values according to this hypothesis since we cannot meaningfully specify any assumed value for the difference between the treated and untreated populations (blue and red curves).

- Do we have any reason to assume that the null hypothesis is true? In the previous section, it has been pointed out that the null hypothesis is assumed to be correct because it is the only testable hypothesis. However, the null hypothesis is not unique from among all possible hypotheses. It is not the most probable or most reasonable hypothesis. The statistical test is actually performed since we do not have any idea whether the null hypothesis is correct or how probable it is that the null hypothesis is true. As it will be discussed thoroughly in the following sections, this shortcoming has substantial consequences on the interpretation of the result of a statistical test.

2. *Interpretation and misinterpretation of the p value*

It has already been pointed out that the *p* value is equal to the probability with which a value of the test statistic equal to or more extreme than the actual calculated value is obtained *if the null hypothesis is true* (Figure 2). If a *p* value smaller than the level of significance prompts us to reject the null hypothesis, the *p* value can be correctly interpreted as the probability of falsely rejecting a *correct null hypothesis*. However, the *p* value is NOT the probability that rejecting the null hypothesis was a wrong decision. Although the two previous definitions may seem to be identical, restating them may help us see the difference:

<u>Correct interpretation:</u> The *p* value is the probability that the null hypothesis is wrongly rejected *assuming it is true* (if a *p* value smaller than the level of significance indeed leads us to reject it).

<u>Incorrect, but common interpretation:</u> The *p* value is the probability that the null hypothesis is wrongly rejected (if it is indeed rejected due to *p* smaller than the level of significance).

As it can be seen the only difference between the two definitions is the fact that the first one only considers cases when the null hypothesis is true, whereas the second one would apply to all cases. And herein lies the problem. Since the null hypothesis defines a specific, numerically definite distribution of statistic values, probabilities can be calculated with respect to this distribution. However, not even do we not know any parameter of the alternative hypothesis, we are not even in the position to know the probability that the null hypothesis is true (or the fraction of cases when the null hypothesis is true). *Therefore, the probability of incorrectly rejecting the null hypothesis in a statistical test cannot be calculated without further assumptions.* The following sections will address this issue further.

3. *False discoveries and the false discovery rate (FDR)*

The way null hypotheses are assumed invests rejection of the null hypothesis with a special role in biological research. Since it is assumed in the null hypothesis that the investigated drug or gene does not have any effect, rejection of this assumption is equivalent to finding a drug candidate or marker gene, i.e. a *discovery*. The rejection of a correct null hypothesis, i.e. a type I error, is called a *false discovery*. The probability of a false discovery is the *false discovery rate (FDR)*. It is of utmost importance to realize that <u>FDR is not equal to the *p* value</u> for reasons to be discussed in the next paragraphs.

Let us assume that a panel of drug candidates is analyzed for their efficacy to lower blood cholesterol. In statistical decision making neither parameters of the alternative hypothesis (e.g. the mean of blood cholesterol after treatment with an effective drug), nor the fraction of cases when the null hypothesis is true (i.e. the fraction of the drugs that are ineffective) is known. This circumstance prevents us from estimating the reliability of our decisions beyond the *p* value even though it would be desirable. In this section, two additional assumptions will be required for determining the false discovery rate: the power of the test (1-β, see later) and the fraction of correct null hypotheses, designated by *C*. While the assumption of any specific value for *C* is, at best, an educated guess, it will be demonstrated

- how the fraction of correct null hypothesis influences the difference between *p* and *FDR*
- depending on the fraction of correct null hypotheses *FDR* can be frighteningly high.

If the fraction of correct null hypotheses is *C*, a fraction of *C* of the drug candidates will be ineffective (Figure 4). The fraction of incorrect decisions for this group of drug candidates will be equal to α, the level of significance, i.e. a fraction of α of statistical tests investigating ineffective drugs will lead to rejection of the null hypothesis (type I error or false discovery). Such wrong decisions correspond to a fraction of $C\cdot\alpha$ of all statistical tests. Correct decisions in this group of drug candidates constitute a fraction $C\cdot(1-\alpha)$ of all statistical tests. The performance of statistical tests for incorrect null hypotheses (i.e. effective drugs) is determined by parameter β defined as the fraction of incorrect decisions for incorrect null hypotheses. The power of the statistical test, 1-β, is the fraction of correct decisions in such cases. A fraction of 1-β of statistical tests performed with incorrect null hypotheses will lead to a correct rejection of the null hypothesis (true discovery) constituting $(1-C)\cdot(1-\beta)$ of all statistical tests. Along similar lines, the fraction of incorrectly accepted null hypotheses (type II error) from among all statistical tests is $(1-C)\cdot\beta$.

Now, we are in a position to calculate the false discovery rate. The level of significance only determines the fraction of false discoveries within the group of correct null hypotheses. Similarly, the *p value only gives the probability of a false discovery for correct null hypotheses*. *The false discovery rate is the fraction of incorrectly rejected null hypotheses from among all rejected null hypotheses*:

$$FDR = \frac{\#\, of\ \ false\, discoveries}{\#\, of\ \ false\, discoveries + \#\, of\ true\, discoveries} = \frac{C\alpha}{C\alpha + (1-C)(1-\beta)} \qquad (1)$$

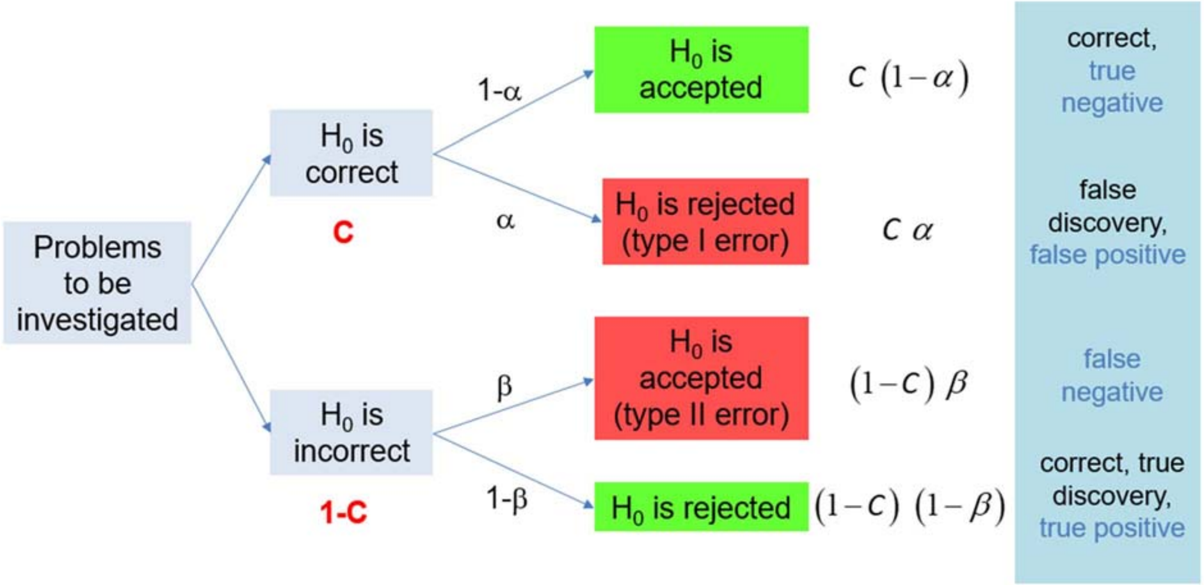


Figure 4. The flow chart of false discoveries. Statistical tests are applied to the results of certain studies, e.g. investigating a panel of drug candidates. In all the statistical tests, a null hypothesis will be assumed, e.g. H_0: the drug does not work. In an unknown fraction of cases, designated by *C*, the null hypothesis is indeed correct, i.e. the drug does not work. The performance of the statistical test in this group of drugs (or in this group of null hypotheses) is determined by the level of significance, α. In α% of these cases the null hypothesis will be falsely rejected corresponding to $C\cdot\alpha$ fraction of all tests. Performance of the statistical test for incorrect null hypotheses, corresponding to 1-*C* fraction of all tests, is determined by β, i.e. the fraction of incorrect null hypotheses failed to be rejected. The parameter, 1-β, is called the power of the test, which is the probability of rejecting an incorrect null hypothesis. The fraction of correctly rejected null hypotheses (true discoveries) from among all the statistical tests is given by (1-*C*)·(1-β). The blue text in the blue field on the right classifies decisions according to the terminology of medical diagnostic tests.

Implications of this equation are shown in Figure 5. When estimating the false discovery rate, α is usually assumed to be 0.05. β, as it will be discussed in the next sections, is determined by the unknown difference between the means according to the null and alternative hypotheses (i.e. the shift between the black and colored curves in Figure 3) and the sample size. β is usually assumed to be 0.1-0.2 in *FDR* calculations. Here, a β value of 0.2 will be used. If the fraction of correct null hypotheses is 0.8, the *FDR* is 20%, but it rises to 86% if the fraction of correct null hypotheses is 99%. Consequently, if a group of null hypotheses, in which most null hypotheses are correct, is tested (e.g. a panel of drug candidates in which most of them are useless; a panel of treatments in which most of them do not work) most discoveries, i.e. rejected null hypotheses, will be false. In conclusion, it can be stated that reaching true discoveries is exceedingly difficult if mostly correct null hypotheses are tested.

Calculation of the false discovery rate as described in this section can be calculated with an Excel workbook and a Matlab program to be described in more detail later (Figures 7 and 8).

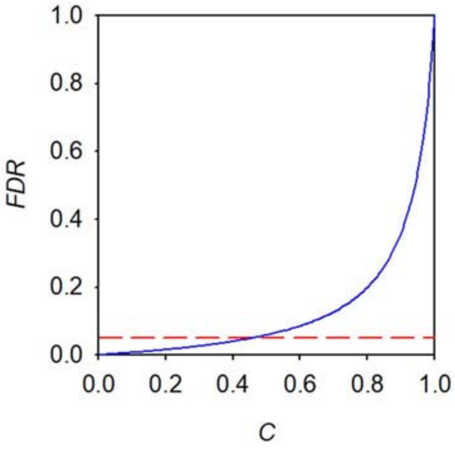


Figure 5. Dependence of the false discovery rate on the fraction of correct null hypotheses. The false discovery rate (*FDR*) was calculated according to equation (1) assuming α=0.05, β=0.2. The level of significance is shown by the horizontal, red, dashed line. The blue line shows that the *FDR* is approximately equal to the level of significance if the fraction of correct null hypotheses is low, but it steeply increases if *C* is beyond 50%.

4. Alternative estimation of the false discovery rate and the reverse Bayesian approach

At the end of section 2, it was pointed out that gaining further insight into the interpretation of statistical tests, especially into false discovery rates, requires assumptions besides the null hypothesis. In the previous section, the fraction of correct null hypotheses and the power of the test (1-β) was assumed allowing us to calculate the false discovery rate. In this section, a presumed distribution for the alternative hypothesis will be used instead of the power of the test. As it will be explained these two approaches lead to identical conclusions regarding the false discovery rate under certain circumstances. Before introducing this alternative approach, let us take a closer look at how to use both the null and alternative hypotheses to calculate the probabilities of correct and incorrect statistical decisions.

Using a two-sample t-test as an example, the correctness of the null hypothesis entails a zero expectation for the t-value calculated according to the following equation:

$$t=\frac{\overline{x}-\overline{y}}{\sqrt{\frac{SD_x^{\ 2}\left(N_x-1\right)+SD_y^{\ 2}\left(N_y-1\right)}{N_x+N_y-2}\left(\frac{1}{N_x}+\frac{1}{N_y}\right)}} \tag{2}$$

where $\overline{x}$ and $\overline{y}$ are the means of the two samples, SD_x and SD_y are the standard deviations of the samples, n_x and n_y are the corresponding sample sizes. The calculated t-values will be distributed according to Student's t-distribution (Figure 6). In conventional hypothesis testing the alternative hypothesis is the negation of the null hypothesis (e.g. H_0: no difference → H_A: there is a difference). However, this kind of formulation of the alternative hypothesis is not sufficient for doing calculations. Therefore, a specific form of the alternative hypothesis must be assumed instead (e.g. H_A: the difference is 1). Any specific form of the alternative hypothesis is merely speculative. If the alternative hypothesis is correct, calculated t-values will be distributed according to the noncentral t-distribution, which differs from the well-known Student's t-distribution in several respects:

- it is shifted with respect to Student's t-distribution by

$$\theta=\frac{E\left(\Delta x_{effect}\right)}{SD\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}}=\frac{d}{\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}} \tag{3}$$

  where $d=E\left(\Delta x_{effect}\right)/SD$ designates the effect size, i.e. the expected difference between the means according to the null and alternative hypotheses ($E(\Delta x_{\text{effect}})$) normalized by the standard deviation. The parameter defined by equation (3) is the noncentrality parameter (θ) of the noncentral t-distribution.
- it is an asymmetrical distribution that has a relatively long tail if the noncentrality parameter is large.

Now let us reinvestigate how we reach a decision in hypothesis testing. The null hypothesis is accepted or rejected based on the probability with which the calculated t-value occurs according to the null hypothesis (see Sections 1 and 2). However, independent of whether the null hypothesis is accepted or rejected, we may reach a wrong decision since any calculated t-value can be generated by the null or alternative hypothesis (blue or red curves in Figure 6). In this section, we are going to *compare explicitly the likelihoods with which a calculated t-value is produced if the null or the alternative hypothesis is correct*. There are two different ways such a comparison can be made. The principle is demonstrated in Figure 6 for a calculated t-value that is equal to the critical value (t_α):

- Conventional way based on cumulative probabilities (areas under the curve, **p-less-than approach**): In this approach we calculate cumulative probabilities of obtaining calculated t-values farther away from zero than the current one.
  - Step 1: estimate the probability that values at least this far from zero (i.e. values larger than t_α in Figure 6) are generated by the distribution according to the null hypothesis. The probability of this happening is the area under the blue curve corresponding to *t* values larger than t_α, i.e. the blue-shaded area, which is equal to the level of significance (α) since the calculated value is t_α. For an arbitrary calculated t-value, this probability is equal to the p-value.
  - Step 2: estimate the probability with which calculated t-values more extreme than the current one are produced if the alternative hypothesis is correct. This probability is equal to the unshaded area under the red curve in Figure 6, and it is called the *power of the test* (1-β).
- Approach 2 based on (non-cumulative) probabilities (**p-equals approach**): In this method, we attempt to estimate the likelihood with which values similar to the calculated one are obtained if the null or alternative hypothesis is correct. Basic statistical courses teach us that the probability with which a continuous random variable assumes values in a certain range is given by the area under the curve of the probability density function that can be determined by integration. Although the value of the probability density function at *x* is not a probability, it still provides a numerical estimate of how likely values close to *x* are produced. Therefore, the likelihood with which a calculated t-value is produced according to the null hypothesis is proportional to the value of the probability density function of Student's t-distribution (0.091 corresponding to t_α in Figure 6). Similarly, the likelihood with which the calculated t-value is produced if the alternative hypothesis is correct is proportional to the value of the probability density function of the noncentral t-distribution (0.34 in Figure 6). *Simulations showed that the p-equals approach provides a better approximation of the false discovery rate* (https://doi.org/10.1098/rsos.171085).

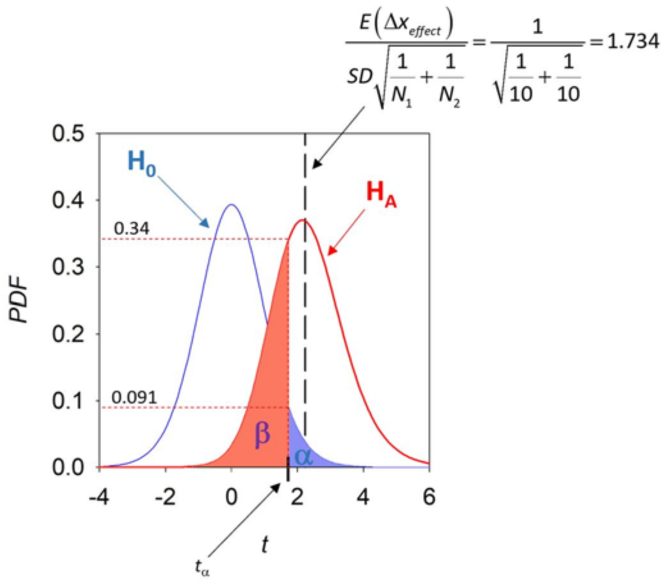


Figure 6. Calculation of the probabilities of type 1 and type 2 errors using the distributions according to the null and alternative hypotheses. If the null hypothesis is true (no difference between the means of the two samples), calculated values of the t-statistic are distributed according to Student's t-distribution (blue curve) with 18 degrees of freedom (both samples consist of 10 observations). If the null hypothesis is not true, calculated t-values are distributed according to an alternative hypothesis, whose parameters are unknown. The assumption in the figure was that the effect size, i.e. $E(\Delta x_{effect})/SD$, is one. This calculated t-values are distributed according to the noncentral t-distribution (red curve), whose peak is shifted with $E(\Delta x_{effect}) \Big/ \left(SD\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}\right)$ relative to the blue curve. The null hypothesis will be rejected if the calculated value falls in the rejection area (blue-shaded part; the graph corresponds to a one-sided test). Calculated t-values fall in this region with a probability of α, the level of significance. The unshaded part of the blue curve displays the acceptance region. The value separating the rejection area from the acceptance area is the critical value, designated by t_α, which is equal to 1.734 for a one-sided t-test with a degree of freedom of 18. Student's t-distribution assumes a value of 0.091 at 1.734. If the alternative hypothesis is true, but the calculated value falls in the acceptance region, a type 2 error is committed, whose probability is equal to β. The probability of obtaining t-values larger than t_α is equal to the unshaded area under the red curve, which is 0.694 in the current figure. The noncentral t-distribution (red curve) assumes a value of 0.34 at t_α.

The principles introduced in the previous paragraphs provide three ways to get a feeling of the reliability of our findings.

I. *Posterior and prior odds, likelihood ratio*. Although likelihood and probability are synonymous in most contexts, they have a slightly different meaning when discussing the results of hypothesis tests or regression. *Probability* is the plausibility of a certain model or assumption (e.g. the null or alternative hypothesis) given an observation or data. Using the terminology of probability theory, this is called the conditional probability of the model given the data: $P(model|data)$. According to the definition of conditional probability, it can be calculated according to the following equation:

$$P(model|data) = \frac{P(model \cap data)}{P(data)} \tag{4}$$

<u>*Likelihood*</u> follows a reverse logic in that it gives the probability of the data given a certain model (e.g. the null or alternative hypothesis):

$$P(data|model) = \frac{P(model \cap data)}{P(model)} \tag{5}$$

Using the definition of conditional probability according to equation (4), the *posterior probability* of the null (H_0) and alternative hypotheses (H_A) can be written as follows:

$$P(H_0|data) = \frac{P(H_0 \cap data)}{P(data)},\ P(H_A|data) = \frac{P(H_A \cap data)}{P(data)} \tag{6}$$

$P(H_0|data)$ and $P(H_A|data)$ are called posterior probabilities because we can only calculate them after the data has been collected. The likelihoods of the data given the null or the alternative hypothesis are given by the following equations:

$$P(data|H_0) = \frac{P(H_0 \cap data)}{P(H_0)},\ P(data|H_A) = \frac{P(H_A \cap data)}{P(H_A)} \tag{7}$$

Let us divide the two posterior probabilities with each other:

$$\frac{P(H_A|data)}{P(H_0|data)} = \frac{P(H_A \cap data)}{P(H_0 \cap data)} \tag{8}$$

Now, let us express the probabilities on the right side of the previous equation using the likelihoods defined in equation (7):

$$\underbrace{\frac{P(H_A|data)}{P(H_0|data)}}_{\text{posteirior odds}} = \underbrace{\frac{P(data|H_A)}{P(data|H_0)}}_{\text{likelihood ratio}} \underbrace{\frac{P(H_A)}{P(H_0)}}_{\text{prior odds}} \tag{9}$$

where $P(H_A)$ and $P(H_0)$ are called <u>*prior probabilities*</u> because they are (theoretically) defined already before collecting the data. The problem is that this piece of information is usually unknown in practice, and it has to be assumed. In the method described in Section 2, the fraction of correct null hypotheses was defined as *C*. Since the alternative and null hypotheses are mutually exclusive, the ratio of their probabilities can be expressed as follows:

$$P(H_0)+P(H_A)=1 \Rightarrow \frac{P(H_A)}{P(H_0)}=\underbrace{\frac{1-P(H_0)}{P(H_0)}}_{\text{odds of } H_A}=\frac{1-C}{C} \tag{10}$$

Since odds are defined as

$$odds=\frac{\text{probability that something is true}}{\text{probability that something is not true}} \tag{11}$$

$P(H_A)/P(H_0)$ in equations (9) and (10) is the *prior odds of the alternative hypothesis*, i.e. how many times the alternative hypothesis can be assumed to be more likely to be correct than false before collecting the data. Using the mutual exclusivity of the null and alternative hypotheses and the definition of odds, it can be shown that the left side of equation (9) is the *posterior odds of the null hypothesis*, i.e. how many times the alternative hypothesis is more probable to be correct than false after collecting the data.

Now, we are going to use equation (9) to calculate the posterior odds of the alternative hypothesis. The prior odds of the alternative hypothesis can be determined according to equation (10). Since this calculation involves the assumed fraction of true null hypotheses (*C*), the prior odds are also an assumption. The likelihood ratio can be calculated in two principally different ways (p-less-than approach and p-equals approach). In Figure 6, likelihood of the data given the alternative hypothesis ($P(data|H_A)$) according to the p-less-than approach is the area of under the curve of the alternative hypothesis right of the calculated t-value. Since in Figure 6, the calculated t-value is t_α, this happens to be the power of the test ($1-\beta=0.694$ in Figure 6). According to the p-equals approach, $P(data|H_A)$ is proportional to the value of the probability density function of the alternative hypothesis at the calculated value (0.34). Similarly, the likelihood of the data if the null hypothesis is true ($P(data|H_0)$) according to the p-less-than approach is 0.05 in Figure 6, and it is proportional to 0.091 according to the p-equals approach. Accordingly, we have two different estimates for the likelihood ratio:

$$\begin{aligned}&\text{p-equals approach}: \frac{P\left(data \mid H_A\right)}{P\left(data \mid H_0\right)} = \frac{0.34}{0.091}\\ &\text{p-less-than approach}: \frac{P\left(data \mid H_A\right)}{P\left(data \mid H_0\right)} = \frac{0.694}{0.05}\end{aligned} \tag{12}$$

Assuming $t_{\text{calc}}=t_{\alpha}$ an effect size of 1, sample sizes of 10 and C=0.9, the posterior odds of the alternative hypothesis can be calculated according to equation (9) using the p-equals and p-less-than approaches:

$$\begin{aligned}&\text{p-equals approach}: \frac{P\left(H_A \mid data\right)}{P\left(H_0 \mid data\right)} = \frac{P\left(data \mid H_A\right)}{P\left(data \mid H_0\right)} \frac{P\left(H_A\right)}{P\left(H_0\right)} = \frac{0.34}{0.091} \frac{1-0.9}{0.9} = 0.415\\ &\text{p-less-than approach}: \frac{P\left(H_A \mid data\right)}{P\left(H_0 \mid data\right)} = \frac{P\left(data \mid H_A\right)}{P\left(data \mid H_0\right)} \frac{P\left(H_A\right)}{P\left(H_0\right)} = \frac{0.694}{0.05} \frac{1-0.9}{0.9} = 1.542\end{aligned} \tag{13}$$

How to interpret this result? Before collecting the data and performing the statistical test, the odds of the alternative hypothesis were (according to the assumed value of 0.9 for C) $(1-0.9)/0.9=0.111$, i.e. the alternative hypothesis was 9-times more probable to be false than true. After the statistical test, the odds of the alternative hypothesis increased to 0.415 and 1.542 according to the p-equals and p-less-than approaches, respectively. According to the p-equals approach, the alternative hypothesis is still less likely to be true (since its odds are less than one) given the results of the test. According to the p-less-than approach, the alternative hypothesis is 1.542× more likely to be true than false after the statistical test. The charm of this approach is the fact that it shows that neither the null, nor the alternative hypothesis can be selected with 100% certainty to be true even after calculating the statistical test. This is in striking contrast with the misleading conclusion of accepting or rejecting the null hypothesis in conventional hypothesis testing that allures many investigators to believe that rejecting the null hypothesis with a p-value of 0.05 proves that the null hypothesis is wrong.

Three different ways of performing these calculations in practice are described here.

1. I have created an Excel workbook, in which Excel functions and macros perform statistical calculations. The program is available at the following web address: https://peternagyweb.hu/Excel/Peter_ManyStatProbes_with_Excel.xlsm

The sheet named “FDR” calculates the posterior odds of the alternative hypothesis (Figure 7). More detailed description is available in the legends to the figure.

2. Alternatively, a custom-written, graphical user interface-controlled Matlab program, fdrEstimation.m, performs the same calculations (Figure 8). The Matlab file can be downloaded at the following web address: https://peternagyweb.hu/Statistics/fdrEstimation.m
3. An online application available at http://fpr-calc.ucl.ac.uk/ or http://shiny.ieis.tue.nl/fpr_calc/ (Figure 9).

II. *Alternative determination of the false discovery rate:* In section 3, the false discovery rate was calculated using an assumed value for the power of the test, 1-β. Based on the concepts introduced in this section, 1-β can be calculated if an assumption is made for the effect size, i.e. E(Δx_{effect})/SD. According to the principles described in the current section, the power of the test is the likelihood $P\left(data|H_A\right)$ defined by equation (7). Similarly, the probability of committing a type I error (α) is the likelihood $P\left(data|H_0\right)$, also defined by equation (7). Both of these likelihoods can be calculated according to the p-equals and the p-less-than approach as described previously. Using these two pieces of information let us rewrite equation (1):

$$FDR=\frac{C\alpha}{C\alpha+(1-C)(1-\beta)}=\frac{C\,P\left(data|H_0\right)}{C\,P\left(data|H_0\right)+(1-C)\,P\left(data|H_A\right)} \tag{14}$$

The advantage of this way of calculation is the fact that the likelihoods $P\left(data|H_0\right)$ and $P\left(data|H_A\right)$ can easily be computed not only for the critical value, but for any calculated value of the t-test enabling us to calculate the false discovery rate corresponding to any of our statistical tests. According to Figure 6 and the likelihoods determined in equation (12), the false discovery rate can be calculated as follows:

$$\text{p-equals approach}:\frac{C\,P\left(data|H_0\right)}{C\,P\left(data|H_0\right)+(1-C)\,P\left(data|H_A\right)}=\frac{0.9\cdot 0.091}{0.9\cdot 0.091+0.1\cdot 0.34}=0.707$$

$$\text{p-less-than approach}:\frac{C\,P\left(data|H_0\right)}{C\,P\left(data|H_0\right)+(1-C)\,P\left(data|H_A\right)}=\frac{0.9\cdot 0.05}{0.9\cdot 0.05+0.1\cdot 0.694}=0.393$$

(15)

When applying equation (14) in practice, the question of what to assume for the effect, i.e. E(Δx_{effect}), needs to be answered. This assumption has basically nothing to do with statistics, it is determined by how large an effect is biologically or medically relevant, or what the expected effect is based on previous experiences ("an educated guess"). It is also important to realize that the FDR estimation described in Section 3 and in this current section provide identical results for the p-less-than case if

- a calculated t-value equaling the critical value (t_α) is used. The critical value is influenced by the degree of freedom, the level of significance and whether the test is one- or two-sided.
- β, required for the method in Section 3, is calculated in accordance with the level of significance, the sidedness of the test, the sample sizes, E(Δx_{effect}) and the SD, parameters required for this current approach.

By providing the effect size, the sample sizes, the calculated t-value, the fraction of true null hypotheses and whether the test is one- or two-sided the Excel workbook described previously and the Matlab program fdrEstimation determine the false discovery rate (Figures 7 and 8). The online application performs the same kind of calculations for two-sided tests (http://fpr-calc.ucl.ac.uk/ or http://shiny.ieis.tue.nl/fpr_calc/) (Figure 9).

III. *Reverse Bayesian approach:* Although it is sometimes straightforward to choose the fraction of true null hypotheses for the calculations described in the preceding sections, it might be very subjective in other cases. The approach described in this section provides yet another method especially suited for such cases. Let us rewrite equation (9) for the posterior odds of the alternative hypothesis by eliminating P(H$_0$) using the equality P(H$_0$)+P(H$_A$)=1:

$$\frac{P(H_A|data)}{P(H_0|data)}=\frac{P(data|H_A)P(H_A)}{P(data|H_0)P(H_0)}=\frac{P(data|H_A)P(H_A)}{P(data|H_0)(1-P(H_A))} \qquad (16)$$

Solution of the previous equation for P(H$_A$) yields the following expression:

$$P(H_A)=\frac{P(H_A|data)P(data|H_0)}{P(data|H_A)P(H_0|data)+P(H_A|data)P(data|H_0)} \qquad (17)$$

Assuming the data allowed us to reject the null hypothesis $P\left(H_0\middle|data\right)$ gives that fraction of those null hypothesis rejections when the null hypothesis was true, i.e. $P\left(H_0\middle|data\right)$ is the false discovery rate. Since $P\left(H_0\middle|data\right)+P\left(H_A\middle|data\right)=1$, the previous equation can be rewritten to obtain its final form:

$$P\left(H_A\right)=\frac{\left(1-FDR\right)P\left(data\middle|H_0\right)}{P\left(data\middle|H_A\right)FDR+\left(1-FPR\right)P\left(data\middle|H_0\right)} \qquad (18)$$

FDR on the right side of the previous equation is the desired (maximum) false discovery rate to be achieved. By substituting the likelihoods $P\left(data\middle|H_0\right)$ and $P\left(data\middle|H_A\right)$, *the minimum prior probability of the alternative hypothesis (P(H_A)), required for the desired maximum false discovery rate, can be determined*.

The likelihoods can be calculated using the p-equals and p-less-than approaches as described previously. Let us stipulate that the false discovery rate must not be more than 0.05:

$$\text{p-equals approach: } P\left(H_A\right)=\frac{\left(1-FDR\right)P\left(data\middle|H_0\right)}{P\left(data\middle|H_A\right)FDR+\left(1-FPR\right)P\left(data\middle|H_0\right)}=\frac{0.95\cdot 0.091}{0.34\cdot 0.05+0.95\cdot 0.091}=0.836$$

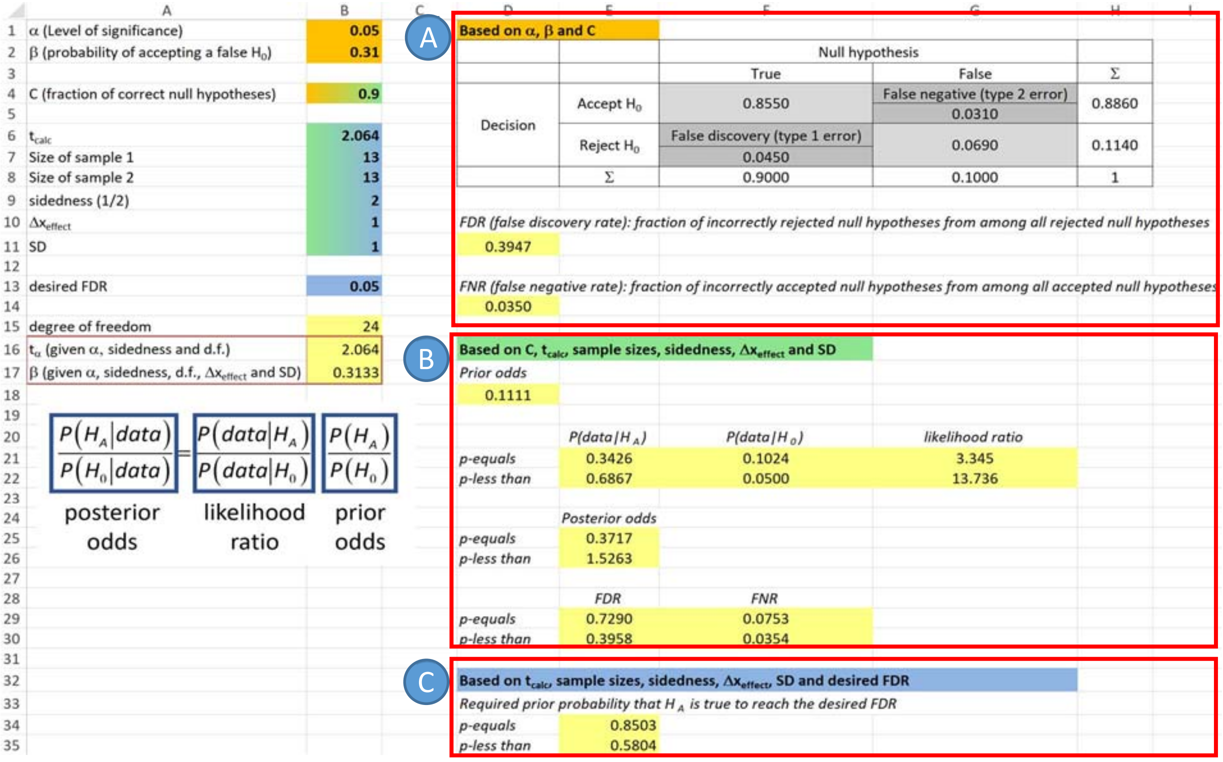
$$\text{p-less-than approach: } P\left(H_A\right)=\frac{\left(1-FDR\right)P\left(data\middle|H_0\right)}{P\left(data\middle|H_A\right)FDR+\left(1-FPR\right)P\left(data\middle|H_0\right)}=\frac{0.95\cdot 0.05}{0.694\cdot 0.05+0.95\cdot 0.05}=0.578$$

(19)

These results imply that we have to be 83.6% (according to the p-equals approach) or 57.8% (according to the p-less-than approach) certain that the alternative hypothesis is correct in order to achieve the desired maximum false discovery rate of 5%. We have to ask ourselves if it is reasonable to assume that the alternative hypothesis is more probable to be correct than the null hypothesis. If such an assumption is not reasonable, the false discovery rate can be higher than the desired value of 5% in this example. As stated previously the results provided by the p-equals approach is more in accordance with reality.

The three applications described in the previous sections, i.e. the “FDR” sheet of the Excel workbook (Figure 7), the fdrEstimation Matlab program (Figure 8) and the online application (http://fpr-calc.ucl.ac.uk/ or http://shiny.ieis.tue.nl/fpr_calc/, Figure 9) can all calculate the required prior probability of the alternative hypothesis according to equation (18).

Due to the flexibility of the methods described in this section (i.e. the capability to use them for any arbitrary calculated t-value), they are preferred to the one introduced in Section 3. The principles described in this section are mostly based on the paper “The reproducibility of research and the misinterpretation of p-values” by Colquhoun, D. (*R Soc Open Sci* **4**, 171085 (2017); https://doi.org/10.1098/rsos.171085).



A	B
α (Level of significance)	0.05
β (probability of accepting a false H_0)	0.31
C (fraction of correct null hypotheses)	0.9
t_{calc}	2.064
Size of sample 1	13
Size of sample 2	13
sidedness (1/2)	2
Δx_{effect}	1
SD	1
desired FDR	0.05
degree of freedom	24
t_α (given α, sidedness and d.f.)	2.064
β (given α, sidedness, d.f., Δx_{effect} and SD)	0.3133

$$\frac{P(H_A|data)}{P(H_0|data)} = \frac{P(data|H_A)}{P(data|H_0)} \frac{P(H_A)}{P(H_0)}$$

posterior odds   likelihood ratio   prior odds

A

Based on α, β and C

		Null hypothesis		
		True	False	Σ
Decision	Accept H_0	0.8550	False negative (type 2 error) 0.0310	0.8860
	Reject H_0	False discovery (type 1 error) 0.0450	0.0690	0.1140
	Σ	0.9000	0.1000	1

FDR (false discovery rate): fraction of incorrectly rejected null hypotheses from among all rejected null hypotheses

0.3947

FNR (false negative rate): fraction of incorrectly accepted null hypotheses from among all accepted null hypotheses

0.0350

B

Based on C, t_{calc}, sample sizes, sidedness, Δx_{effect} and SD

Prior odds

0.1111

	P(data|H_A)	*P(data|H_0)*	*likelihood ratio*
p-equals	0.3426	0.1024	3.345
p-less than	0.6867	0.0500	13.736

	Posterior odds
p-equals	0.3717
p-less than	1.5263

	FDR	*FNR*
p-equals	0.7290	0.0753
p-less than	0.3958	0.0354

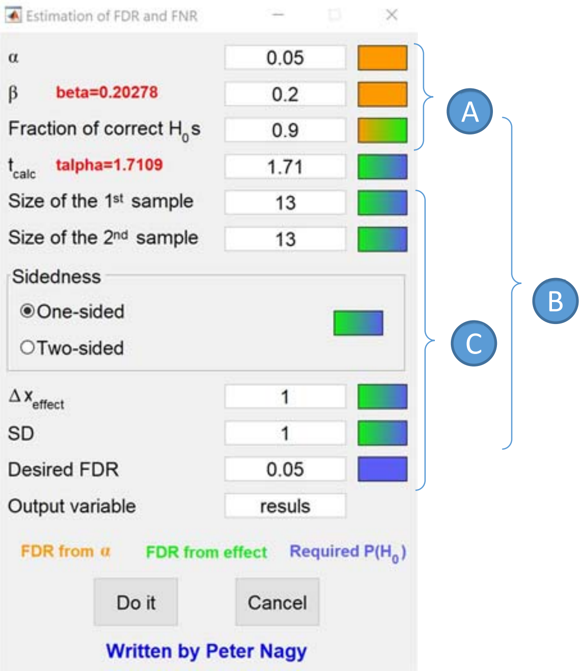
C

Based on t_{calc}, sample sizes, sidedness, Δx_{effect}, SD and desired FDR

Required prior probability that H_A is true to reach the desired FDR

p-equals	0.8503
p-less than	0.5804

Figure 7. Determination of the reliability of discoveries with Excel. The figure shows the “FDR” sheet of the Excel workbook downloadable at https://peternagyweb.hu/Excel/Peter_ManyStatProbes_with_Excel.xlsm. This Excel sheet calculates all measures of the reliability of discoveries described in Sections 3 and 4. Part A contains the estimation of the false discovery rate and the false negative rate based on the fraction of true null hypotheses, α and β (described in Section 3). This part is labeled by orange, and consequently the completely and partially orange input cells in the upper left part of the sheet must be filled in for this calculation. Part B contains the results of methods II and III of Section 4. The completely or partially green-colored input cells are required for these calculations. So that the two different kinds of determination of the false discovery rate (parts A and B of the sheet) provide identical results, two conditions must be met: (i) t_{calc} (cell B6) must be equal to t_α (i.e. the critical value corresponding to the level of significance, provided in cell B1). This t_α value is shown in the yellow cell of B16. (ii) In addition, the β value, to be entered in cell B2, must be equal to what is calculated in cell B17 based on the effect size, degree of freedom, level of significance and sidedness. Part C of the sheet calculates the required prior probability of alternative hypothesis calculated according to approach III of Section 4. The completely or partially blue-colored input cells must be filled in for this calculation.



α 0.05

β beta=0.20278 0.2

Fraction of correct H_0s 0.9

t_{calc} talpha=1.7109 1.71

Size of the 1st sample 13

Size of the 2nd sample 13

Sidedness

● One-sided

○ Two-sided

Δx_{effect} 1

SD 1

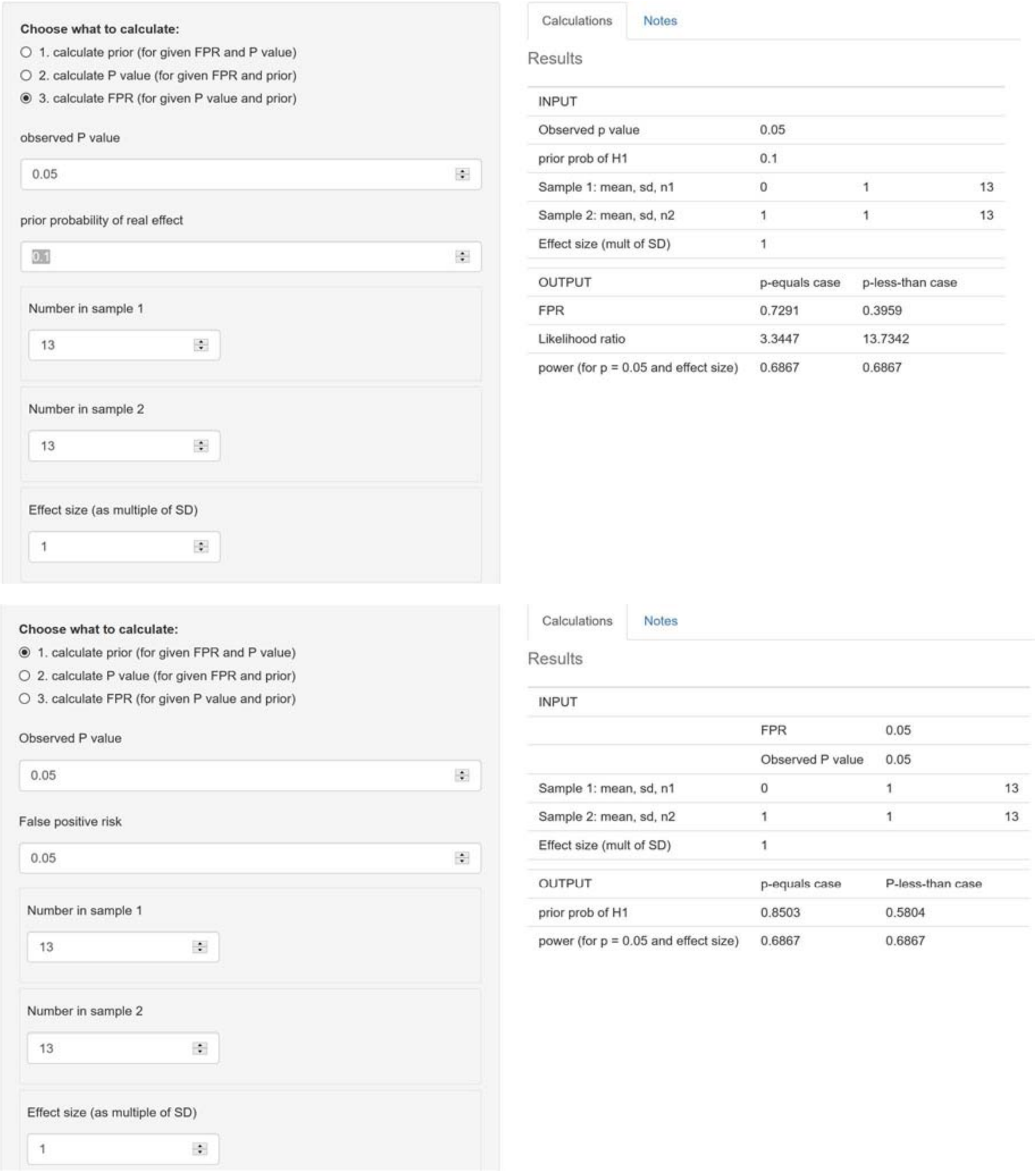
Desired FDR 0.05

Output variable resuls

Figure 8. Graphical user interface of the Matlab program fdrEstimation (https://peternagyweb.hu/Statistics/fdrEstimation.m). The program calculates all the characteristics of the reliability of statistical conclusions discussed in Sections 3 and 4. Three different groups of outputs are generated (FDR from α or effect, Required $P(H_0)$). These are shown with color-codes above the buttons in the graphical user interface. The results will be saved in a structure variable with eight fields. The names of fields holding the different kinds of results are shown in parentheses below in the legend. The name of this output variable is specified in the lowest input field. In order to calculate the false discovery rate (field: *fdr*), the false negative rate (field: *fnr*) and the probabilities of different statistical decisions (field: *fdrTable*) from α, β and the fraction of true null hypotheses, the top three fields (labeled by orange or partially orange colors) must be filled in (A). The prior odds (field: *oddsPrior*), the likelihood ratios (field: *lrTable*) and the posterior odds (field: *oddsPosteriorTable*) are calculated if the input cells labeled by green or partially green colors are filled in (B). The minimum required probability of the alternative hypothesis for reaching a certain, desired false discovery rate (field: *minHATable*) is calculated if the input fields labeled by blue or partially blue colors are filled in (C). The last two groups of calculations (green and blue) are calculated using the p-equals and p-less-than approaches as well. The red text next to β and t_{calc} displays those values of these parameters at which the two different calculations of the false discovery rate provide identical results for the p-less-than approach. The program can be used in two different ways: (i) entering "fdrEstimation" without any arguments will run the program in GUI (graphical user interface) mode; (ii) entering the following command will execute the program without displaying the GUI:

r=fdrEstimation(alpha,beta,fractionCorrect,tcalc,N1,N2,sidedness,dxEffect,sd,desiredFDR);

alpha, beta – probability of rejecting a true H_0 or accepting a false H_A, respectively;
fractionCorrect – the fraction of correct null hypotheses;
tcalc – the calculated value of the t-test;
N1,N2 – the sample sizes;
sidedness – 1 or 2 corresponding to a one-sided or two-sided test, respectively;
dxEffect – $E(\Delta x_{effect})$; sd – standard deviation
desiredFDR – the desired false discovery rate to be achieved.



Choose what to calculate:

○ 1. calculate prior (for given FPR and P value)

○ 2. calculate P value (for given FPR and prior)

● 3. calculate FPR (for given P value and prior)

observed P value: 0.05

prior probability of real effect: 0.1

Number in sample 1: 13

Number in sample 2: 13

Effect size (as multiple of SD): 1

Results

INPUT			
Observed p value	0.05		
prior prob of H1	0.1		
Sample 1: mean, sd, n1	0	1	13
Sample 2: mean, sd, n2	1	1	13
Effect size (mult of SD)	1		

OUTPUT	p-equals case	p-less-than case
FPR	0.7291	0.3959
Likelihood ratio	3.3447	13.7342
power (for p = 0.05 and effect size)	0.6867	0.6867

Choose what to calculate:

● 1. calculate prior (for given FPR and P value)

○ 2. calculate P value (for given FPR and prior)

○ 3. calculate FPR (for given P value and prior)

Observed P value: 0.05

False positive risk: 0.05

Number in sample 1: 13

Number in sample 2: 13

Effect size (as multiple of SD): 1

Results

INPUT			
	FPR	0.05	
	Observed P value	0.05	
Sample 1: mean, sd, n1	0	1	13
Sample 2: mean, sd, n2	1	1	13
Effect size (mult of SD)	1		

OUTPUT	p-equals case	P-less-than case
prior prob of H1	0.8503	0.5804
power (for p = 0.05 and effect size)	0.6867	0.6867

Figure 9. A web-based application for estimating the false discovery rate and the required prior probability of the alternative hypothesis. The application is available at http://fpr-calc.ucl.ac.uk/, direct link to the application: http://shiny.ieis.tue.nl/fpr_calc/. If number 3 is chosen in the "Choose what to calculate" part, the false discovery rate and the likelihood ratio will be calculated. The calculations are performed according to part II of Section 4 ("Alternative determination of the false discovery rate"). The results are identical to those provided by the Excel sheet and the 'fdrEstimation' Matlab program if the calculations are performed for a two-sided test in the latter ones. If number 1 is chosen from among the "Choose what to calculate" radio buttons, the required prior probability of the alternative hypothesis will be calculated according to part III of Section 4. These calculations are also performed for a two-sided test by the application.

5. Sample size estimation

Besides interpreting the implications of the result of a statistical test, it is also meaningful to calculate the sample size, which is required for the investigation to have the power to reject the null hypothesis assuming a certain effect size. For a two-sample t-test, the effect size is the difference between the two sample means divided by the SD. While such calculations do not reduce the risk of false discoveries, they shed light on false negative results, i.e. they reveal situations when there is little chance of discoveries (rejecting the null hypothesis).

This section will use a two-sample t-test as an example to demonstrate the principle. The well-known formula for the two-sample t-test was provided in a previous section (equation (2)). Assuming the two SDs are identical and provided the two sample sizes are $n_x=n$, $n_y=r \cdot n$, where r is the ratio of the two sample sizes, this formula takes the following form:

$$t=\frac{\overline{x}-\overline{y}}{\sqrt{\frac{SD^2\left(N+r\,N-2\right)}{N+r\,N-2}\left(\frac{1}{N}+\frac{1}{r\,N}\right)}}=\frac{\overline{x}-\overline{y}}{SD\sqrt{\frac{1}{N}+\frac{1}{r\,N}}} \tag{20}$$

The sample size is determined with the following two assumed parameters:

(i) a certain level of significance, designated by α

(ii) a certain power, $1-\beta$, to reject the null hypothesis if it is indeed false.

Let us first write an equation corresponding to the first point. If the null hypothesis is true, the expected value of the numerator of equation (20) will be zero, i.e. the blue curve in Figure 10 is centered at zero. If the null hypothesis is true, equation (20) is distributed according to Student's t-distribution with a degree of freedom of $N+N\cdot r-2$. According to condition (i), we would like to have α% of the calculated t values to fall in the rejection area. This requirement is met if the expression is equal to t_α, the value above which the area under the blue curve in Figure 10 is equal to the level of significance:

$$\frac{\Delta x_{calc}}{SD\sqrt{\frac{1}{N}+\frac{1}{r\,N}}}=t_\alpha \tag{21}$$

Now, let us write an equation corresponding to the second condition, i.e. the power to reject the null hypothesis if the alternative hypothesis is true. If the alternative hypothesis is true, t values calculated according to equation (20) are distributed according to the noncentral t-distribution with a noncentrality parameter given by equation (3), which takes the following form using *N* and *r N* as the sample sizes:

$$\theta = \frac{E\left(\Delta x_{effect}\right)}{SD\sqrt{\frac{1}{N}+\frac{1}{r\ N}}} \tag{22}$$

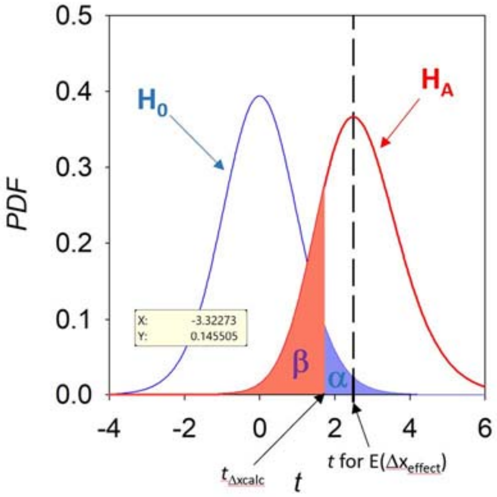


Figure 10. Principle of calculating the sample size to reach a certain power to reject the null hypothesis. Distribution of calculated values of the t-test, if the null hypothesis is true, is shown by the blue curve labeled by H_0. Distribution of test statistic values according to the alternative hypothesis (red curve labeled by H_A) is shifted relative to the blue curve by the assumed value of the effect, i.e. the t-value corresponding to E(Δx_{effect}). The sample size required to reach a certain power is computed by determining the calculated t-value ($t_{\Delta xcalc}$) that provides the specified level of significance (α=type 1 error rate, blue-shaded area) and power (1-β, red-shaded area, β - type 2 error rate). The blue-shaded area corresponds to type 1 errors since these values are generated by the distribution according to the null hypothesis, but are extreme enough to be larger than the critical value. The red-shaded area corresponds to type 2 errors since these calculated *t* values are not extreme enough to reject the null hypothesis (are in the acceptance region), but still belong to the distribution according to the alternative hypothesis. If the sample size is equal to the minimum sample size meeting these requirements, Δx_{calc} values determined by the two methods (equations (21) and (23)) are equal to each other.

This distribution is different from the relatively well-known Student's t-distribution. According to condition (ii), we would like to have β% of this curve to be left of the calculated *t* value. This requirement is expressed by the following equation:

$$\frac{\Delta x_{calc}}{SD\sqrt{\frac{1}{N}+\frac{1}{r\ N}}} = tnc_{\theta,\beta} \tag{23}$$

where $tnc_{\theta,\beta}$ is the value compared to which β% of calculated t-values, distributed according to a noncentral t-distribution with degree of freedom of $N+N\cdot r-2$ and a noncentrality parameter of θ, are smaller. If the sample size is just large enough to meet the requirements specified by points (i) and (ii) above, Δx_{calc} values calculated according to equations (21) and (23) are equal:

$$t_\alpha = tnc_{\theta,\beta} \tag{24}$$

Solving equation (24) for N provides the minimal sample size required to reach a certain level of significance and power. The sample size given by the above equation is the size of one of the samples, whereas the other sample has a size of $r\cdot N$. Since both t_α and $tnc_{\theta,\beta}$ depend on N, the above equation does not have a closed form solution. Many textbooks provide a simplified approach to the problem by approximating the t-test with a z-test. Equations (21) and (23) take the following form in that case:

$$\frac{\Delta x_{calc}}{SD\sqrt{\frac{1}{N}+\frac{1}{r\,N}}} = z_\alpha\,,\ \frac{\Delta x_{calc} - E\left(\Delta x_{effect}\right)}{SD\sqrt{\frac{1}{N}+\frac{1}{r\,N}}} = z_\beta \ \Rightarrow\ z_\alpha\ SD\sqrt{\frac{1}{N}+\frac{1}{r\,N}} = E\left(\Delta x_{effect}\right) - z_\beta\ SD\sqrt{\frac{1}{N}+\frac{1}{r\,N}}$$

(25)

The right-hand side of the expression is obtained by expressing Δx_{calc} from both equations. The closed-form solution of equation (25) for N is:

$$N = \frac{(1+r)}{r}\left(\frac{SD\left(z_\beta - z_\alpha\right)}{\Delta x_{effect}}\right)^2 \text{,if } r\text{=1} \Rightarrow 2\left(\frac{SD\left(z_\beta - z_\alpha\right)}{\Delta x_{effect}}\right)^2 \tag{26}$$

The effect of the sample size on the power of the statistical test is due to the inverse relationship between the standard error of the mean (SEM) and the square root of the sample size:

$$SEM = \frac{SD}{\sqrt{N}} \tag{27}$$

The standard error of the mean estimates the precision with which the sample mean approximates the mean of the population. Consequently, smaller differences between the two sample means can be found statistically significant by increasing sample sizes (and decreasing SEMs). This principle is demonstrated by Figure 11.

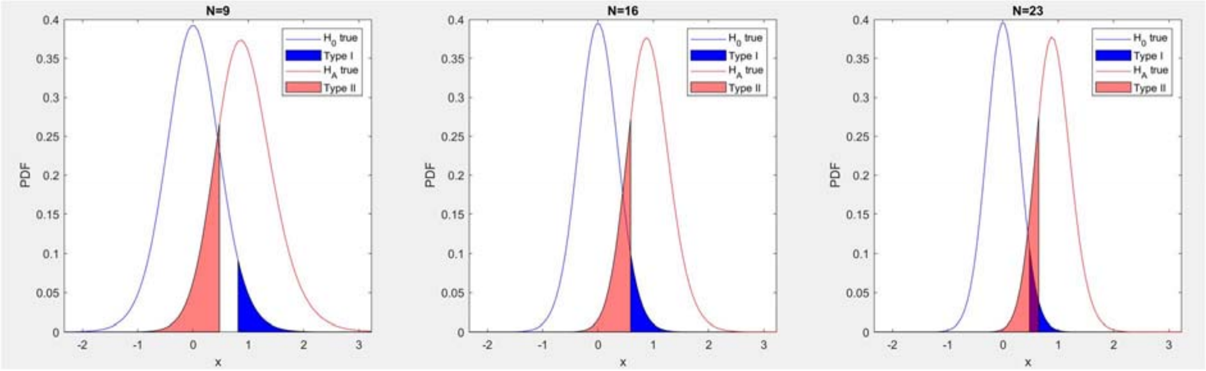


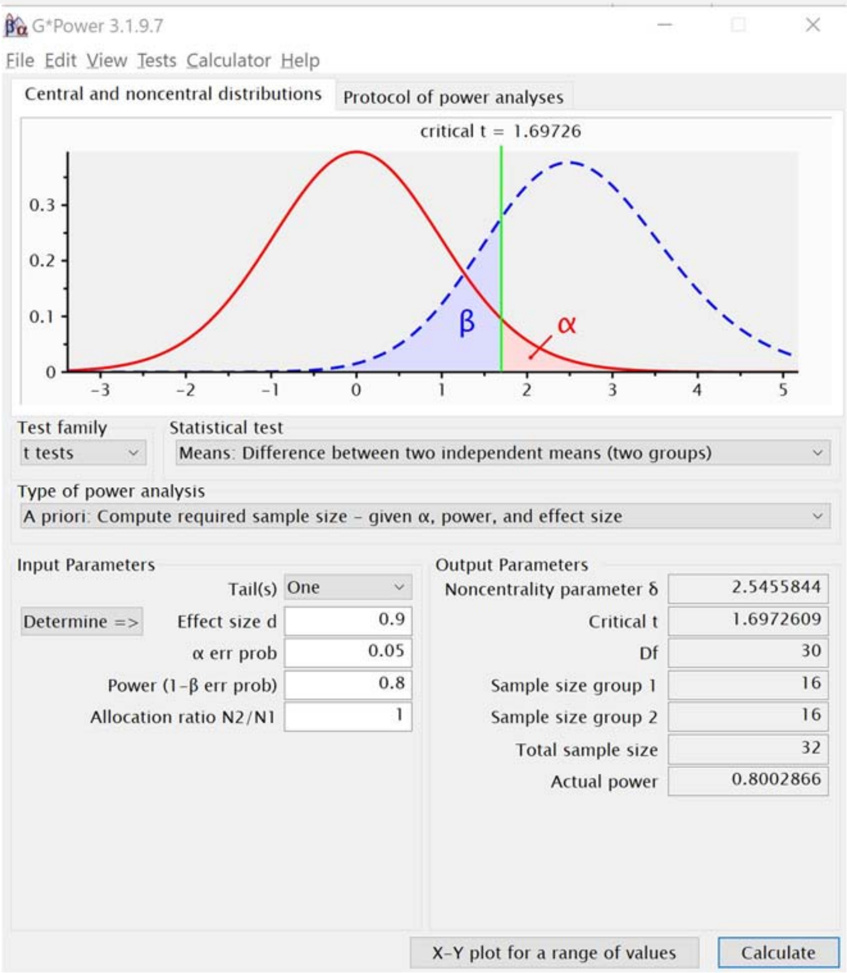
Figure 11. Demonstration of the effect of different sample sizes on the power to reject the null hypothesis. Distribution of sample means according to the null and alternative hypotheses are shown by the blue and red curves, respectively. The blue curves are centered at 0, whereas the red ones are centered at $E(\Delta x_{\mathrm{effect}})$ in all panels, but their width decreases with increasing sample size. The blue- and red-shaded areas correspond to type 1 and type 2 errors, respectively. The SD of the measurement was assumed to be 2, and the effect was 90% of the SD. A level of significance of 5% and a required power of 80% was used for generating the figures. Consequently, the blue and red areas are 5% and 20%, respectively. The required sample size to meet these conditions is *N*=16. If the sample size is smaller than this threshold, more than 20% of the distribution according to the null hypothesis is in the acceptance area of the blue curve, i.e. less than 80% of tests will be rejected if the alternative hypothesis is correct. If the sample size is exactly 16, 80% of the distribution is above the critical value according to the null hypothesis, i.e. the left boundary of the blue-shaded area. If the sample size is larger than 16, the false negative rate is less than 20% if the alternative hypothesis is correct.

Such estimations of the required sample size for reaching a certain statistical power to reject the null hypothesis assuming a certain effect size are increasingly required for manuscript submissions, by granting agencies and institutional review boards, especially for animal experiments and clinical trials. The importance of calculating such minimal sample sizes lies in the fact that *samples smaller than those estimated as described above basically preclude or significantly reduce the chance of rejecting the null hypothesis, i.e. finding a discovery, at the assumed effect size*. Investigations based on too small samples are doomed for failure and are thought not to be worth to be performed. Therefore, sample size estimations are usually carried out *a priori*, i.e. before the actual investigation.

The following section describes five different ways of estimating the required sample size:

1. G*Power (available at www.gpower.hhu.de): An intuitive, easy-to-use, graphical user interface-controlled application, in which the settings for determining the sample size for

a two-sample t-test are shown in Figure 12. The program is flexible, and is capable of sample size estimations for many other kinds of statistical problems.


G*Power 3.1.9.7
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 1.69726
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
5
β
α
Determine =>
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

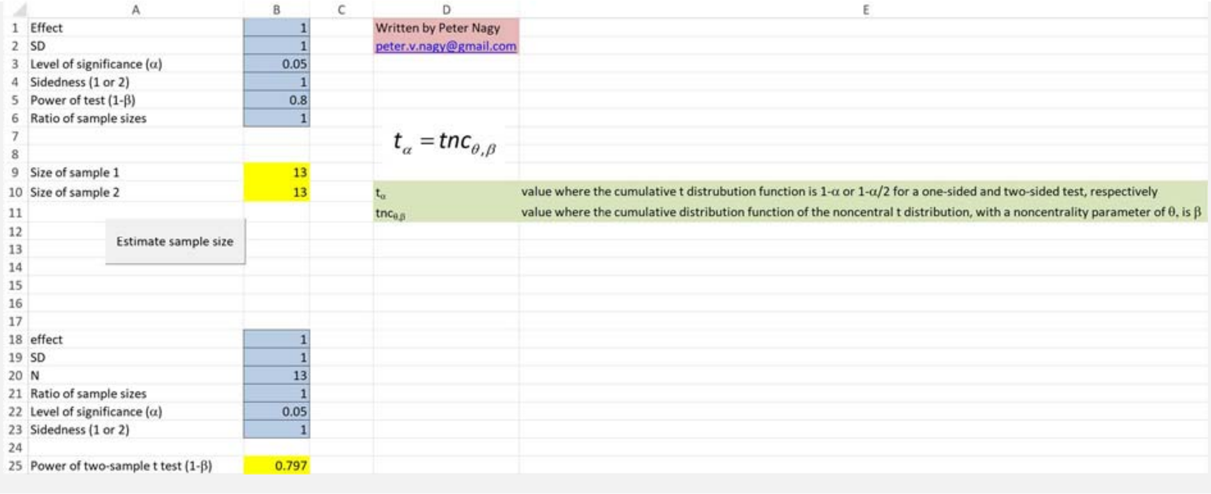
Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters	
Tail(s)	One
Effect size d	0.9
α err prob	0.05
Power (1-β err prob)	0.8
Allocation ratio N2/N1	1

Output Parameters	
Noncentrality parameter δ	2.5455844
Critical t	1.6972609
Df	30
Sample size group 1	16
Sample size group 2	16
Total sample size	32
Actual power	0.8002866

Figure 12. Graphical user interface of G*Power for calculating the required sample size for a 2-samle t-test. As opposed to Figure 11 and the graphical output of the Matlab program sampleSizeForTtest, which display the sample means on the horizontal axis, this application plots the calculated t-values on the x axis.

2. Excel: The sheet named “Sample size & power” in the Excel workbook, introduced in Figure 7, determines the sample size required to reach a certain statistical power in a two-sample t-test given a certain effect size. In addition, the same sheet also estimates the statistical power of the test given the effect size and the sample size (Figure 13).



	A	B	C	D	E
1	Effect	1		Written by Peter Nagy	
2	SD	1		peter.v.nagy@gmail.com	
3	Level of significance (α)	0.05			
4	Sidedness (1 or 2)	1			
5	Power of test (1-β)	0.8			
6	Ratio of sample sizes	1			
7				$t_\alpha = tnc_{\theta,\beta}$	
8					
9	Size of sample 1	13			
10	Size of sample 2	13		t_α	value where the cumulative t distrubution function is 1-α or 1-α/2 for a one-sided and two-sided test, respectively
11				$tnc_{\theta,\beta}$	value where the cumulative distribution function of the noncentral t distribution, with a noncentrality parameter of θ, is β
12	Estimate sample size				
13					
14					
15					
16					
17					
18	effect	1			
19	SD	1			
20	N	13			
21	Ratio of sample sizes	1			
22	Level of significance (α)	0.05			
23	Sidedness (1 or 2)	1			
24					
25	Power of two-sample t test (1-β)	0.797			

Figure 13. Determining the required sample size and the power of a statistical test in Excel. The Excel workbook is available for download from https://peternagyweb.hu/Statistics. The sheet named "Sample size & power" is shown in the figure. If the sample size for a certain power is to be calculated, the top part is to be used. The user must fill in the blue cells, and the sample size will be shown in the yellow cells after clicking on the button "Estimate sample size". The bottom part of the sheet calculates the statistical power of a test given a certain sample and effect size. After filling in the blue cells, the statistical power will be immediately shown in the yellow field.

3. Built-in function in Matlab: the following command calculates the required sample size to reject the null hypothesis using a two-sample t-test (hence 't2'):

- assuming a mean and SD of 10 and 1, respectively, according to the null hypothesis
- a mean of 10.9 according to the alternative hypothesis
- a power of 80%
- right-tailed statistical test
- a level of significance of 5%
- a ratio of sample sizes of 2:

```
[N1,N2]=sampsizepwr('t2',[10 1],10.9,0.8,[],'tail','right','alpha',0.05,'ratio',2)
```

The function requires the Statistics and Machine Learning Toolbox in Matlab.

4. Second method in Matlab. A custom-written function sampleSizeForTtest.m, available at https://peternagyweb.hu/Statistics/sampleSizeForTtest.m will do the same kind of calculation for a two-sample t-test. This function also requires the Statistics and Machine Learning Toolbox in Matlab. The program can be simply run by entering "sampleSizeForTtest" at the Matlab command prompt and the data can be entered in a graphical user interface (Figure 14). Alternatively, the program can also be executed according to the following syntax:

```
Nsample= sampleSizeForTtest(dxEffect,sd,alpha,beta,Nratio,sidedness,visual);
```

- dxEffect – the difference between the means according to the null and alternative hypotheses
- sd – standard deviation
- alpha, beta – level and significance and the probability of a type 2 error given the correctness of the alternative hypothesis (1-β = power of the test), respectively
- Nratio – ratio of the sample sizes
- sidedness – 1 or 2 specifying a one-sided and two-sided test, respectively
- visual – if 1, the program will generate a demo figure like the one shown in Figure 11.

5. Performing the calculation in R.

- Install the ‘pwr’ package: Packages > Install packages. Select a mirror, and then the ‘pwr’ package. Alternatively, type “install.packages(‘pwr’) at the command prompt.
- Load the package by typing the following command: library(pwr)
- The following command will do the calculation for a two-sample, one-sided t-test, with a level of significance of 5%, a power of 90% and an effect size of 1:

```
pwr.t.test(d=1, sig.level=0.05, power = 0.9, type = 'two.sample', alternative = 'greater')
```

If a priori sample size estimation has not been carried out or there is any doubt that the required statistical power has been reached, it is possible and advisable to estimate the statistical power of the test according to Figure 10, i.e. to calculate the red-shaded area relative to the total area under the curve corresponding to the alternative hypothesis.

This can be accomplished in several different ways:

1. The Excel sheet introduced in Figure 13.
2. Built-in function in Matlab: the following command calculates the power of a two-sample t-test with the following parameters:

- assuming a mean and SD of 10 and 1, respectively, according to the null hypothesis
- a mean of 10.9 according to the alternative hypothesis
- a sample size of 16 for one of the samples
- right-tailed statistical test
- a level of significance of 5%
- a ratio of sample sizes of 2:

```
power=sampsizepwr('t2',[10 1],10.9,[],16,'tail','right','alpha',0.05,'ratio',2);
```

The function requires the Statistics and Machine Learning Toolbox in Matlab.

3. A custom-written Matlab function 'determinePowerTtest'. If the program is executed without input arguments, a graphical user interface is displayed (Figure 14). Alternatively, the program can be run according to the following syntax:

   power=determinePowerTtest(dxEffect,sd,N,Nratio,alpha,sidedness)

   dxEffect – the difference between the means according to the null and alternative hypotheses

   sd – standard deviation

   N – sample size

   Nratio – ratio of the sample sizes

   alpha – level of significance

   sidedness – 1 or 2 specifying a one-sided and two-sided test, respectively

   This function is available at

   https://peternagyweb.hu/Statistics/determinePowerTtest.m

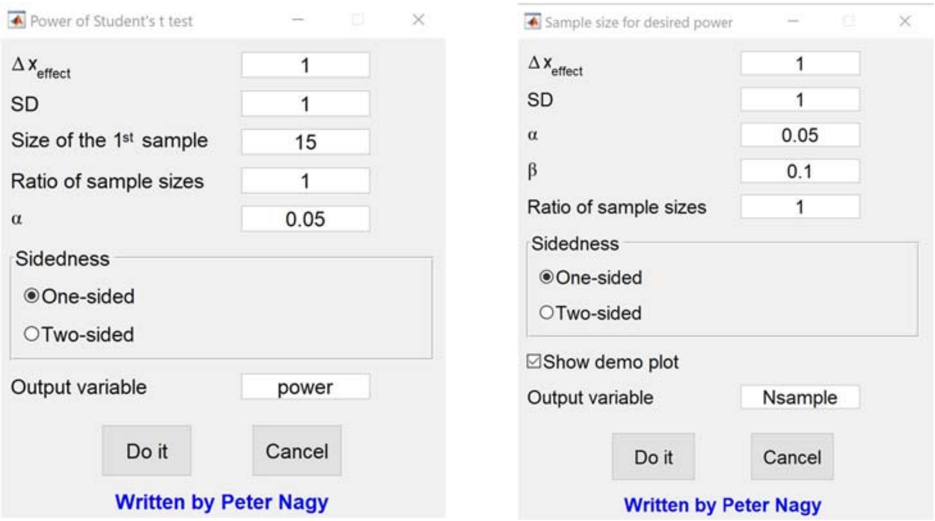
   The function requires the Statistics and Machine Learning Toolbox in Matlab.

4. Performing the calculations in R. Install and load the 'pwr' package as described previously for sample size determination. The following command will calculate the power of a two-sample, one-sided t-test at a level of significance of 5% assuming the size of a sample is 18:

   pwr.t.test(d=1, sig.level=.05, n = 18, type = 'two.sample', alternative = 'greater')

If the required statistical power is not reached due to the low sample size, the risk of false negative decisions (type 2 errors) increases. The false negative rate (FNR), i.e. the probability of type 2 errors, can be calculated according to the principle introduced in Figure 4:

$$FNR=\frac{(1-C)\beta}{(1-C)\beta+C(1-\alpha)} \tag{28}$$



Power of Student's t test

Δx_{effect} 1

SD 1

Size of the 1st sample 15

Ratio of sample sizes 1

α 0.05

Sidedness

● One-sided

○ Two-sided

Output variable power

Sample size for desired power

Δx_{effect} 1

SD 1

α 0.05

β 0.1

Ratio of sample sizes 1

Sidedness

● One-sided

○ Two-sided

☑ Show demo plot

Output variable Nsample

Figure 14. Graphical user interfaces of the Matlab programs *determinePowerTtest* (left) and *sampleSizeForTtest* (right).

6. *Controlling the false discovery rate*

In the previous sections, an overview has been given about procedures for estimating the false discovery rate *in a single statistical test*. When multiple comparisons are performed, the situation is further complicated by the fact that a false discovery can be made in any of the comparisons. The **family-wise error rate** is the probability of committing at least one type I error when performing multiple hypothesis tests. If no measures are taken, the family-wise error rate can be frighteningly high. If the probability of a type I error in a single test is α, the probability of not committing a type I error is a single comparison is 1-α. If *k* hypothesis tests are performed, $(1-\alpha)^k$ is the probability that no type I error is committed in all the tests. Consequently, the probability that at least one type I error is committed, i.e. the family-wise error rate (*FWER*), is given by the following equation:

$$FWER = 1-(1-\alpha)^k \tag{29}$$

A plot of this equation as a function of the number of comparisons shows that the family-wise error rate approaches 1 when the number of tests is higher than 50 (Figure 15).

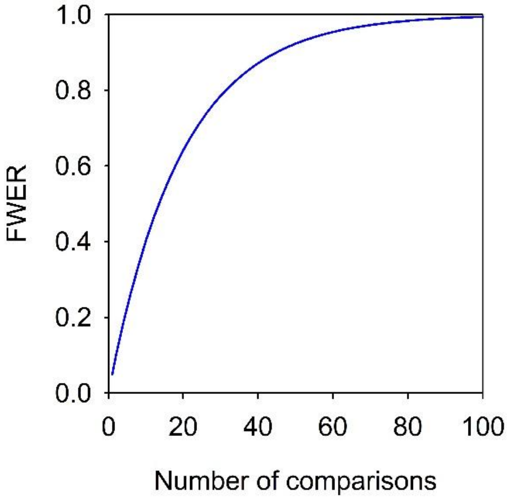


Figure 15. The family-wise error rate (FWER), i.e. the probability of rejecting at least one correct null hypothesis, as a function of the number of tests performed if the level of significance is 5%.

While the approaches described in the previous sections are useful for controlling the false discovery rate in a single hypothesis test, different methods are used in multiple comparisons. Procedures commonly used for controlling the false discovery rate or family-wise error rate in multiple, post-hoc comparisons after ANOVA (e.g. the Bonferroni correction, Sheffé's method, Tukey's HSD test) are appropriate for a couple of tests, and they are not discussed in this tutorial. Two of the approaches required for controlling the false discovery rate if tens or hundreds of hypothesis tests are performed will be described in this section.

Both of these methods are based on the following principle. If the null hypothesis is true and a statistical test is performed a large number of times, the obtained p-values will be homogeneously distributed between 0 and 1, which is the consequence of the probability integral transform (proof is available here). While the proof may be a bit complicated, qualitative understanding is somewhat easier. If the null hypothesis is true, the test statistic will generate random values according to its distribution, i.e. the p-values of these test statistics will not be preferentially small or large (Figure 16A). If a certain fraction of null hypotheses are false, then the statistical test carried out for these null hypotheses will most likely generate small p-values. Consequently, when a mixture of true and false null hypotheses are tested, the distribution of p-values will have a peak close to zero (Figure 16B). If a horizontal line is fitted to the flat part of the histogram of p-values, the p-values under and above this line correspond to those cases when the null hypothesis is true and false, respectively. If the horizontal scale is divided into two parts at the p-value equal to the level of significance, false negative, false positive, true negative and true positive decisions can be graphically represented, and from these values the false discovery rate can be calculated according to equation (1). In the context of multiple comparisons, the false discovery rate calculated according to Figure 16 is often called the **q-value**. According to a formal definition of the q-value, **it is the probability of making a false discovery if a p-value and all p-values smaller than the current one are considered to be "significant" (i.e. discoveries)**.

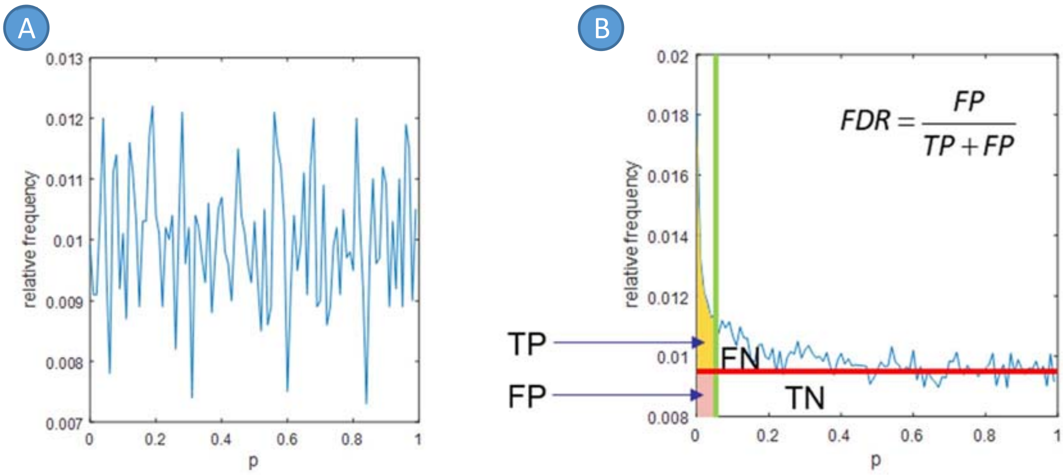


Figure 16. Distribution of p-values. If the null hypothesis is true for all statistical tests carried out, the distribution of p-values will be uniform (A). If, on the other hand, a certain fraction of null hypotheses are wrong, the p-values corresponding to statistical tests carried out with these null hypotheses will produce a peak close to zero (B). Assuming the null hypothesis is rejected if the p-value is smaller than the level of significance (α=0.05 in the figure), p-values left of the vertical green line correspond to positive results ("discoveries"), i.e. when the null hypothesis is rejected. p-values right of the green line correspond to acceptance of the null hypothesis. If a horizontal line is fitted to the flat part of the distribution of p-values (red line), false and correct decisions can be identified, since p-values corresponding to true null hypotheses are under this horizontal line. The yellow area corresponds to true positive results (TP, true discoveries), since these p-values are smaller than the level of significance (hence the null hypothesis is rejected) and they are above the red line (hence they were not produced by tests carried out with correct null hypotheses). The pink area corresponds to false positives (FP, false discoveries), since these small p-values were generated by statistical tests carried out with correct null hypotheses. Using similar reasoning, false negative (FN) and true negative (TN) decisions can also be identified. The false discovery rate (*FDR*) is the fraction of false positive results from among all positive results. Graphically, *FDR* corresponds to the pink area divided by the sum of the pink and yellow areas.

Two methods will be introduced that can be used for controlling the false discovery rate when a large number of hypothesis tests have been carried out. Both approaches are based on principles described in Figure 16 in that they determine q-values and they only require p-values as an input.

6.1. The Benjamini-Hochberg method

Let us assume that there is an array of p-values, and let us arrange them from smallest to largest. The position of the p-value in the list, its rank, is denoted by *i*. The false discovery rate that should not be surpassed, must be specified. Let us designate this desired false discovery rate by *Q*. For each p-value, the Benjamini-Hochberg critical value (BH_{crit}) is calculated according to the following equation:

$$BH_{crit} = \frac{i}{m}Q \tag{30}$$

where *m* is the total number of statistical tests. The largest p-value that is smaller than its corresponding Benjamini-Hochberg critical value, and all p-values smaller than it, will be considered to be significant, positive findings (Table 1). In order to determine the q-value, which is sometimes called the "Benjamini-Hochberg adjusted p-value", the following calculation is carried out for all p-values:

$$q_{initial} = p\frac{m}{i} \tag{31}$$

The q-value corresponding to a certain p-value is the smallest q_{initial} calculated for the p-value under consideration or for any of the larger p-values.

Sorted p-values	Description	Rank	BH_{crit}	Significant	q-values
0.001	Fat	1	0.002	Yes	0.025
0.002	Sugar	2	0.004	Yes	0.025
0.013	Sex	3	0.006	No	0.1083
0.045	Weight	4	0.008	No	0.2679
…	…	…	…	…	…

Table 1. Result of the Benjamini-Hochberg procedure. p-values are arranged in ascending order and ranks are assigned to each of them. The Benjamini-Hochberg critical value (BH_{crit}) is calculated according to equation (30). The false discovery rate (*Q*) was specified to be 0.05. If the p-value is smaller than the corresponding BH_{crit}, the corresponding result, and any other result with a smaller p-value, is deemed "significant", i.e. a discovery. The probability of making a false discovery if a certain p-value and all smaller p-values are considered significant is given by the q-value. E.g. for the 3rd row, **the q-value of 0.1083 means that if the third statistical test and all others with smaller p-values (i.e. altogether the top three rows) would be considered significant, the false discovery rate would be 10.83% for the property "Sex".**

How does the Benjamini-Hochberg method ensure that the false discovery rate is smaller than *Q* and how is it related to Figure 16? According to Figure 17, every test for which the blue curve (p-value) is under the orange one (BH_{crit}) is considered significant. If BH_{crit} is plotted against the rank (*i*), the resulting line has a slope of Q/m. The largest p-value that is considered significant is p_R, and its rank is *R*. Consequently, the number of rejected null hypotheses is *R*. Now, let us determine how many true null hypotheses are among these *R* rejected ones. According to Figure 16A, the distribution of p-values is uniform if the null hypothesis is true. If the number of true null hypotheses is m_0, the number of such null hypotheses having p-values smaller than p_R is $m_0\, p_R$. Consequently, the false discovery rate is:

$$FDR = \frac{m_0\ p_R}{R} \tag{32}$$

According to the orange line:

$$\frac{p_R}{R} = \frac{Q}{m} \ \Rightarrow\ p_R = \frac{R\ Q}{m} \tag{33}$$

Substituting this result into equation (32)

$$FDR = \frac{m_0\ p_R}{R} = \frac{m_0\ R\ Q}{m\ R} = \frac{m_0\ Q}{m} \leq Q \tag{34}$$

The last inequality follows from the fact that $m_0/m \leq 1$. This result proves that rejecting null hypotheses for which the Benjamini-Hochberg critical value is larger than the p-value ensures that the false discovery rate is smaller than *Q*.

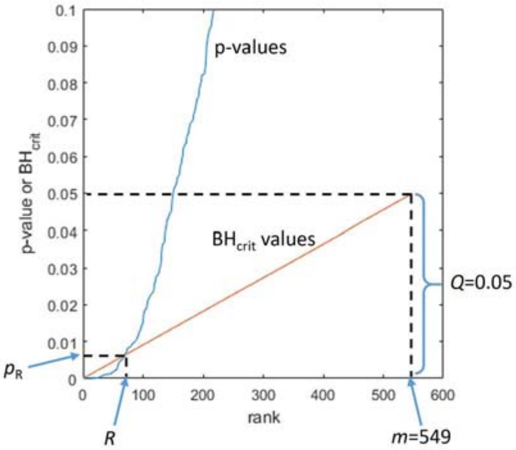
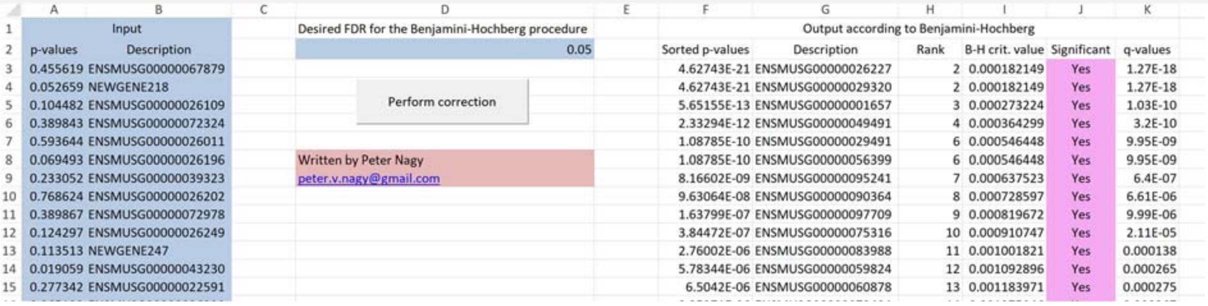


Figure 17. How does the Benjamini-Hochberg procedure control the false discovery rate? The blue and orange curves show the p-values and the Benjamini-Hochberg critical values plotted against the ranks. The largest significant p-value is p_R with a rank of *R*. The rest of the explanation is available in the text.

Four ways of performing the Benjamini-Hochberg procedure are described below:

1. The “Controlling FDR” sheet of the Excel file available at the following web address: https://peternagyweb.hu/Excel/Peter_ManyStatProbes_with_Excel.xlsm (Figure 18).



	A	B	C	D	E	F	G	H	I	J	K
1	Input			Desired FDR for the Benjamini-Hochberg procedure		Output according to Benjamini-Hochberg					
2	p-values	Description		0.05		Sorted p-values	Description	Rank	B-H crit. value	Significant	q-values
3	0.455619	ENSMUSG00000067879				4.62743E-21	ENSMUSG00000026227	2	0.000182149	Yes	1.27E-18
4	0.052659	NEWGENE218		Perform correction		4.62743E-21	ENSMUSG00000029320	2	0.000182149	Yes	1.27E-18
5	0.104482	ENSMUSG00000026109				5.65155E-13	ENSMUSG00000001657	3	0.000273224	Yes	1.03E-10
6	0.389843	ENSMUSG00000072324				2.33294E-12	ENSMUSG00000049491	4	0.000364299	Yes	3.2E-10
7	0.593644	ENSMUSG00000026011				1.08785E-10	ENSMUSG00000029491	6	0.000546448	Yes	9.95E-09
8	0.069493	ENSMUSG00000026196		Written by Peter Nagy		1.08785E-10	ENSMUSG00000056399	6	0.000546448	Yes	9.95E-09
9	0.233052	ENSMUSG00000039323		peter.v.nagy@gmail.com		8.16602E-09	ENSMUSG00000095241	7	0.000637523	Yes	6.4E-07
10	0.768624	ENSMUSG00000026202				9.63064E-08	ENSMUSG00000090364	8	0.000728597	Yes	6.61E-06
11	0.389867	ENSMUSG00000072978				1.63799E-07	ENSMUSG00000097709	9	0.000819672	Yes	9.99E-06
12	0.124297	ENSMUSG00000026249				3.84472E-07	ENSMUSG00000075316	10	0.000910747	Yes	2.11E-05
13	0.113513	NEWGENE247				2.76002E-06	ENSMUSG00000083988	11	0.001001821	Yes	0.000138
14	0.019059	ENSMUSG00000043230				5.78344E-06	ENSMUSG00000059824	12	0.001092896	Yes	0.000265
15	0.277342	ENSMUSG00000022591				6.5042E-06	ENSMUSG00000060878	13	0.001183971	Yes	0.000275

Figure 18. The “Controlling FDR” sheet of the Excel workbook. Information must be entered into the blue cells. The p-values and corresponding descriptions are in the first two columns, while the desired false discovery rate is in cell D2. A click on the “Perform correction” button runs both the Benjamini-Hochberg and Storey procedures (the latter one is to be described later). p-values are ordered, and they will be displayed in column F along with their corresponding ranks, the Benjamini-Hochberg critical values, a statement whether the p-value is significant and the q-values (columns F-K). A “Yes” in column J indicates a discovery (rejection of the null hypothesis). In this case, the corresponding cell in column J is colored purple.

2. correctFDR.m Matlab program (https://peternagyweb.hu/Statistics/correctFDR.m, Figure 19). The program performs both the Benjamini-Hochberg and the Storey procedures. It can be run from the command prompt using input arguments. Syntax:

   [BHtable,storeyTable,storeyData]=correctFDR(pTable,desiredFDR);

   pTable – a two-column table with the first and second columns containing the p-values and descriptions, respectively.

   desiredFDR – the false discovery rate to be achieved in the Benjamini-Hochberg procedure.

   BHtable – a table output containing results of the Benjamini-Hochberg procedure.

   The other two output arguments will be described in 6.2.

   If no input arguments are provided, a graphical user interface is displayed (Figure 19). The program requires the Bioinformatics Toolbox in Matlab.

3. A built-in Matlab function, mafdr, can also perform the Benjamini-Hochberg correction:

   qValues=mafdr(pValues, 'bhfdr', true);

   pValues – an array of p-values.

   'bhfdr', true – specifies that the Benjamini-Hochberg procedure is to be used.

   qValues – an array of q-values corresponding to each p-value in the input.

This built-in Matlab function provides fewer types of output than the correctFDR, custom-written program. mafdr requires the Bioinformatics Toolbox of Matlab.

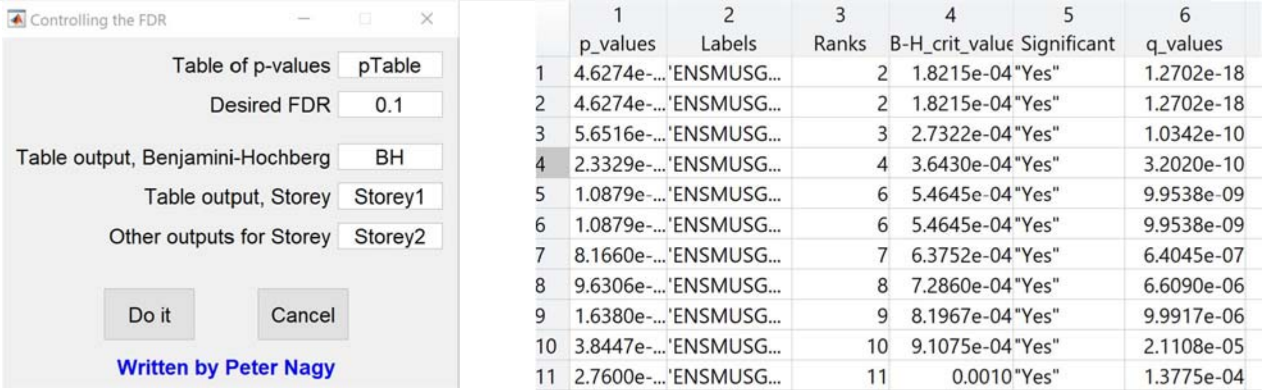


Table of p-values: pTable
Desired FDR: 0.1
Table output, Benjamini-Hochberg: BH
Table output, Storey: Storey1
Other outputs for Storey: Storey2

Written by Peter Nagy

	1	2	3	4	5	6
	p_values	Labels	Ranks	B-H_crit_value	Significant	q_values
1	4.6274e-...	'ENSMUSG...	2	1.8215e-04	"Yes"	1.2702e-18
2	4.6274e-...	'ENSMUSG...	2	1.8215e-04	"Yes"	1.2702e-18
3	5.6516e-...	'ENSMUSG...	3	2.7322e-04	"Yes"	1.0342e-10
4	2.3329e-...	'ENSMUSG...	4	3.6430e-04	"Yes"	3.2020e-10
5	1.0879e-...	'ENSMUSG...	6	5.4645e-04	"Yes"	9.9538e-09
6	1.0879e-...	'ENSMUSG...	6	5.4645e-04	"Yes"	9.9538e-09
7	8.1660e-...	'ENSMUSG...	7	6.3752e-04	"Yes"	6.4045e-07
8	9.6306e-...	'ENSMUSG...	8	7.2860e-04	"Yes"	6.6090e-06
9	1.6380e-...	'ENSMUSG...	9	8.1967e-04	"Yes"	9.9917e-06
10	3.8447e-...	'ENSMUSG...	10	9.1075e-04	"Yes"	2.1108e-05
11	2.7600e-...	'ENSMUSG...	11	0.0010	"Yes"	1.3775e-04

Figure 19. The graphical user interface of correctFDR (left) and its output table containing results of the Benjamini-Hochberg procedure (right). The columns of the table are identical to columns F-K of the Excel sheet (Figure 18). The variables defined by "Table output, Storey" and "Other outputs for Storey" will be described in Figure 22.

4. The following lines of code perform the Benjamini-Hochberg correction in R:

```
p_values=read.table(file = "clipboard", sep = "\t")
p_values=as.numeric(unlist(p_values))
q=p.adjust(p_values,method="BH")
```

The first line reads a list of p-values from the clipboard to which they have been copied from Excel. The second line converts the p-values to a numeric data type, and the third line performs the Benjamini-Hochberg correction returning the q-values for every p-value. Data can also be read directly from a file instead of the clipboard using the following syntax:

```
p_values = read.xlsx("D:/pvalues.xlsx",1)
```

The above code reads the first sheet from the file "D:/pvalues.xlsx". The read.xlsx command requires the "xlsx" package, which must be installed and loaded:

```
install.packages("xlsx")
library(xlsx)
```

6.2. The method of Storey

The Storey approach is more closely based on Figure 16. The aim of this method is to explicitly determine the fraction of correct null hypotheses, which is usually designated by π_0. If the total number of tests carried out in the investigation is *m*, the number of correct null hypotheses is $m \cdot \pi_0$. These null hypotheses are under the red line in Figure 20A (and also in

Figure 16). The principle of estimating π_0 is to find the horizontal part of the histogram of p-values (i.e. the blue curve in Figure 20A). Below, a step-by-step protocol is briefly described for finding π_0:

a. Let us choose an arbitrary threshold, designated by λ, and let us assume that all p-values above λ indeed correspond to true, non-significant null hypotheses. In this case, the number of true null hypotheses with p-values above λ (designated by $\#\{p_i>\lambda\}$) is given by the following equation (since the p-values of non-significant, true null hypotheses are uniformly distributed between p=0 and p=1, i.e. they are in the rectangle under the red line in Figure 20):

$$\#\{p_i > \lambda\} = (1-\lambda)\pi_0 m \tag{35}$$

π_0 can be expressed from the above equation as follows:

$$\pi_0 = \frac{\#\{p_i > \lambda\}}{(1-\lambda)m} \tag{36}$$

b. Let us determine π_0 for a range of λ values between 0 and 1, and plot π_0 as a function of λ (Figure 20B).

c. Find the minimum of this curve, which will be used as an estimate for π_0.

Once the fraction of correct null hypotheses has been determined as described above, estimate the false discovery rate for every p-value according to the equation and principles described in Figure 16, which is put into practice by the following equation for a p-value of p_x (Figure 20C):

$$FDR(p_x) = \frac{\Pi_0 m\, p_x}{\#\{p_i \le p_x\}} \tag{37}$$

The numerator of this equation corresponds to the number of false positive findings and the denominator is the total number of positive results if all null hypotheses with p-values smaller than or equal to the current one (p_x) are rejected. (The term $\#\{p_i \le p_x\}$ instructs us to count the number of p-values (p_i) smaller than or equal to p_x.)

Finally, the q-value must be determined for every p-value. The q-value corresponding to a p-value is either the false discovery rate corresponding to the same p-value or the q-value of the next larger p-value, whichever is smaller.

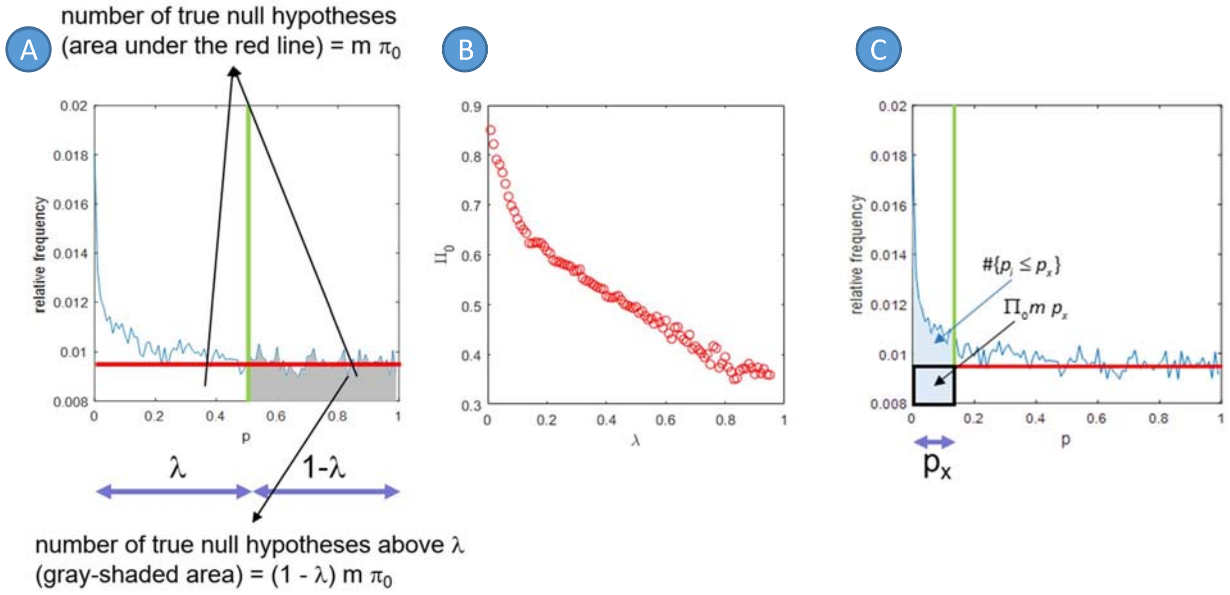
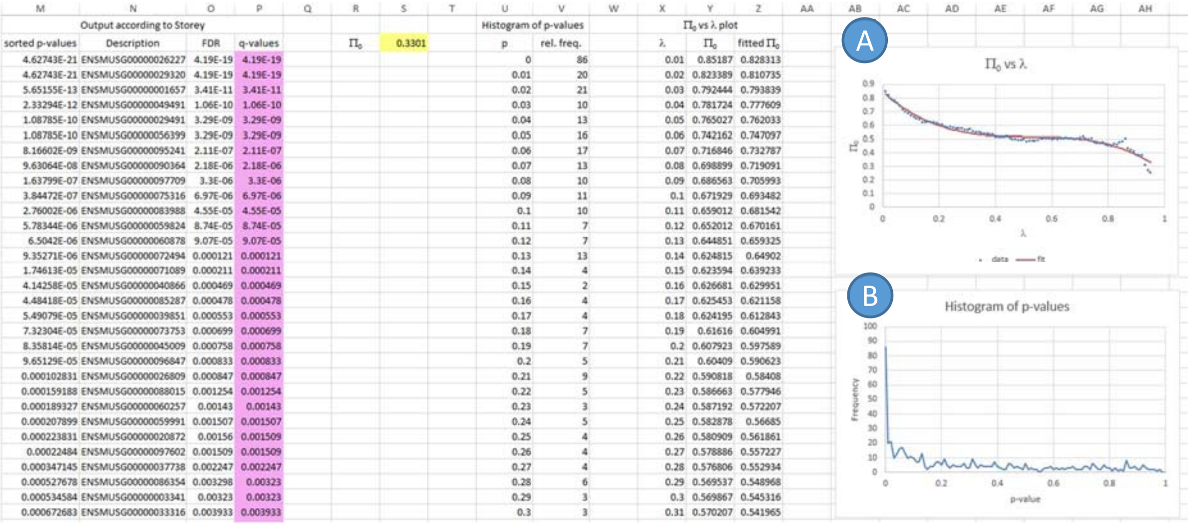


Figure 20. The principle of determining the false discovery rate according to Storey. A histogram of p-values is generated (A). The p-values of true null hypotheses are uniformly distributed between 0 and 1. Since most large p-values correspond to these true null hypotheses, fitting a horizontal line (red line in A) to the horizontal part of the histogram at large p-values will identify the p-values corresponding to true null hypotheses. They are located in the rectangle under the red line. If the total number of tests and the fraction of true null hypotheses are designated by *m* and π_0, respectively, the number of true null hypotheses, proportional to the area under the red line, is $m \cdot \pi_0$. If an arbitrary value, designated by λ, is chosen on the horizontal axis, the number of true null hypotheses above λ is $(1-\lambda)m \cdot \pi_0$. π_0 is determined according to equation (36) for an array of λ values, and a plot of π_0 vs. λ is generated (B), from which the minimum π_0 is determined. This lowest π_0 will be used as an estimate for the fraction of correct null hypotheses. Using this π_0 value, the false discovery rate is determined for every p-value. The following logic is used (C). If we reject a null hypothesis with a p-value of p_x and all other null hypotheses with smaller p-values, then the number of discoveries (rejected null hypotheses including true and false discoveries) is equal to the number of tests with p-values smaller than or equal to p_x. This is designated by $\#\{p_i \leq p_x\}$ in the figure, and it is proportional to the shaded-area left of the green line. From among these rejected null hypotheses, those whose p-values are under the red line correspond to false discoveries (true null hypotheses in the black square). Their number is the total number of correct null hypotheses ($\pi_0 \cdot$m) multiplied by p_x, since this is the area of the black square. Consequently, the false discovery rate corresponding to a certain p-value can be determined according to equation (37).

Four different ways of performing the correction according to Storey are introduced here.

1. The “Controlling FDR” sheet of the Excel file available at the following web address: https://peternagyweb.hu/Excel/Peter_ManyStatProbes_with_Excel.xlsm (Figure 18). The “Perform correction” button executes both the Benjamini-Hochberg and the Storey procedures, and the results of the two procedures are displayed side-by-side (Figure 21).



M	N	O	P	R	S	U	V	X	Y	Z
	Output according to Storey					Histogram of p-values		Π_0 vs λ plot		
sorted p-values	Description	FDR	q-values	Π_0	0.3301	p	rel. freq.	λ	Π_0	fitted Π_0
4.62743E-21	ENSMUSG00000026227	4.19E-19	4.19E-19			0	86	0.01	0.85187	0.828313
4.62743E-21	ENSMUSG00000029320	4.19E-19	4.19E-19			0.01	20	0.02	0.823389	0.810735
5.65155E-13	ENSMUSG00000001657	3.41E-11	3.41E-11			0.02	21	0.03	0.792444	0.793839
2.33294E-12	ENSMUSG00000049491	1.06E-10	1.06E-10			0.03	10	0.04	0.781724	0.777609
1.08785E-10	ENSMUSG00000029491	3.29E-09	3.29E-09			0.04	13	0.05	0.765027	0.762033
1.08785E-10	ENSMUSG00000056399	3.29E-09	3.29E-09			0.05	16	0.06	0.742162	0.747097
8.16602E-09	ENSMUSG00000095241	2.11E-07	2.11E-07			0.06	17	0.07	0.716846	0.732787
9.63064E-08	ENSMUSG00000090364	2.18E-06	2.18E-06			0.07	13	0.08	0.698899	0.719091
1.63799E-07	ENSMUSG00000097709	3.3E-06	3.3E-06			0.08	10	0.09	0.686563	0.705993
3.84472E-07	ENSMUSG00000075316	6.97E-06	6.97E-06			0.09	11	0.1	0.671929	0.693482
2.76002E-06	ENSMUSG00000083988	4.55E-05	4.55E-05			0.1	10	0.11	0.659012	0.681542
5.78344E-06	ENSMUSG00000059824	8.74E-05	8.74E-05			0.11	7	0.12	0.652012	0.670161
6.5042E-06	ENSMUSG00000060878	9.07E-05	9.07E-05			0.12	7	0.13	0.644851	0.659325
9.35271E-06	ENSMUSG00000072494	0.000121	0.000121			0.13	13	0.14	0.624815	0.64902
1.74613E-05	ENSMUSG00000071089	0.000211	0.000211			0.14	4	0.15	0.623594	0.639233
4.14258E-05	ENSMUSG00000040866	0.000469	0.000469			0.15	2	0.16	0.626681	0.629951
4.48418E-05	ENSMUSG00000085287	0.000478	0.000478			0.16	4	0.17	0.625453	0.621158
5.49079E-05	ENSMUSG00000039851	0.000553	0.000553			0.17	4	0.18	0.624195	0.612843
7.32304E-05	ENSMUSG00000073753	0.000699	0.000699			0.18	7	0.19	0.61616	0.604991
8.35814E-05	ENSMUSG00000045009	0.000758	0.000758			0.19	7	0.2	0.607923	0.597589
9.65129E-05	ENSMUSG00000096847	0.000833	0.000833			0.2	5	0.21	0.60409	0.590623
0.000102831	ENSMUSG00000026809	0.000847	0.000847			0.21	9	0.22	0.590818	0.58408
0.000159188	ENSMUSG00000088015	0.001254	0.001254			0.22	5	0.23	0.586663	0.577946
0.000189327	ENSMUSG00000060257	0.00143	0.00143			0.23	3	0.24	0.587192	0.572207
0.000207899	ENSMUSG00000059991	0.001507	0.001507			0.24	5	0.25	0.582878	0.56685
0.000223831	ENSMUSG00000020872	0.00156	0.001509			0.25	4	0.26	0.580909	0.561861
0.00022484	ENSMUSG00000097602	0.001509	0.001509			0.26	4	0.27	0.578886	0.557227
0.000347145	ENSMUSG00000037738	0.002247	0.002247			0.27	4	0.28	0.576806	0.552934
0.000527678	ENSMUSG00000086354	0.003298	0.00323			0.28	6	0.29	0.569537	0.548968
0.000534584	ENSMUSG00000003341	0.00323	0.00323			0.29	3	0.3	0.569867	0.545316
0.000672683	ENSMUSG00000033316	0.003933	0.003933			0.3	3	0.31	0.570207	0.541965

Figure 21. Output of Storey procedure in the Excel workbook. p-values are sorted and the corresponding descriptions, the false discovery rates and the q-values are displayed (columns M-P). Significant p-values whose q-value is smaller than the false discovery rate specified in cell D2 (Figure 18) are colored purple. The fraction of correct null hypotheses (π_0) is determined for an array of λ values between 0 and 1 according to equation (36), and the plot is displayed (A). The minimum π_0 value is determined in the following way. A cubic polynomial fit is performed, and the lowest π_0 is determined from the fit. The estimated π_0 value is shown in the yellow cell (S2). The data corresponding to the graph is in columns X-Z. A histogram of p-values (B) and the corresponding data (columns U-V) are also displayed.

2. correctFDR.m Matlab program (https://peternagyweb.hu/Statistics/correctFDR.m, Figure 19 and Figure 22). The program performs both the Benjamini-Hochberg and the Storey procedures. It can be run from the command prompt using input arguments. Syntax:

   [BHtable,storeyTable,storeyData]=correctFDR(pTable,desiredFDR);

   pTable – a two-column table with the first and second columns containing the p-values and descriptions, respectively.

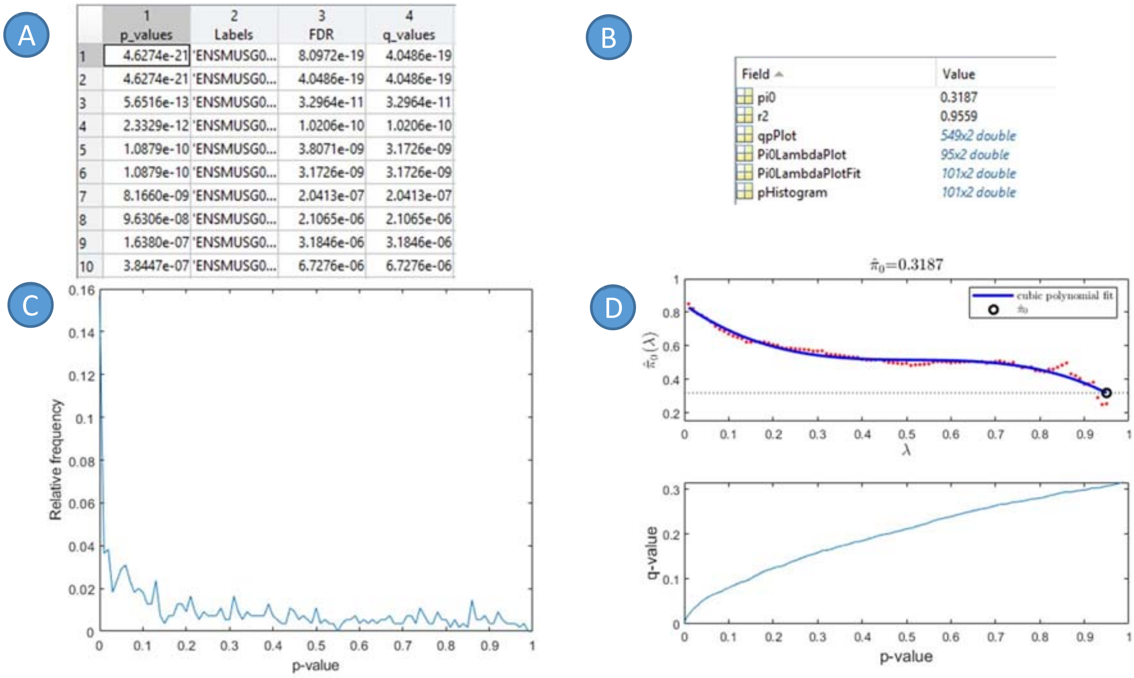
   desiredFDR – the false discovery rate to be achieved in the Benjamini-Hochberg procedure.

   BHtable – a table output containing results of the Benjamini-Hochberg procedure.

   storeyTable – a Matlab table containing results of the Storey procedure (Figure 22A).

   storeyData – a Matlab structure variable containing various other results of the Storey procedure (Figure 22B).

   The program can also be used without any input arguments. In this case, a graphical user interface is displayed (Figure 19). The program requires the Bioinformatics Toolbox in Matlab.



	1 p_values	2 Labels	3 FDR	4 q_values
1	4.6274e-21	'ENSMUSG0...	8.0972e-19	4.0486e-19
2	4.6274e-21	'ENSMUSG0...	4.0486e-19	4.0486e-19
3	5.6516e-13	'ENSMUSG0...	3.2964e-11	3.2964e-11
4	2.3329e-12	'ENSMUSG0...	1.0206e-10	1.0206e-10
5	1.0879e-10	'ENSMUSG0...	3.8071e-09	3.1726e-09
6	1.0879e-10	'ENSMUSG0...	3.1726e-09	3.1726e-09
7	8.1660e-09	'ENSMUSG0...	2.0413e-07	2.0413e-07
8	9.6306e-08	'ENSMUSG0...	2.1065e-06	2.1065e-06
9	1.6380e-07	'ENSMUSG0...	3.1846e-06	3.1846e-06
10	3.8447e-07	'ENSMUSG0...	6.7276e-06	6.7276e-06

Field ▲	Value
pi0	0.3187
r2	0.9559
qpPlot	549x2 double
Pi0LambdaPlot	95x2 double
Pi0LambdaPlotFit	101x2 double
pHistogram	101x2 double

Figure 22. Outputs of the correctFDR.m Matlab program. Sorted p-values along with their descriptions, the corresponding false discovery rates (equation (37)) and q-values are displayed in a Matlab table (A). A Matlab structure variable is also generated as an output with the following fields (B): π_0 – the determined fraction of true null hypotheses; r2 – square of the correlation coefficient of the polynomial fit of the π_0 vs. λ plot; qpPlot – a two-column array storing the data of the q vs. p plot (shown in part D); Pi0LambdaPlot – a two-column array storing the data of the π_0 vs. λ plot (shown in part D); Pi0LambdaPlotFit – a two-column array storing the data of the fit to the π_0 vs. λ plot; pHistogram – a two-column array storing the data of the histogram of p-values (shown in part C). Matlab designates the fraction of true null hypotheses by $\hbar_0$ instead of π_0 (D).

3. The mafdr command in Matlab can perform the Storey procedure. Syntax:

   ```
   [fdr,qValues,pi0]=mafdr(pValues);
   ```

   mafdr performs the Storey correction if *'bhfdr', true* is omitted from the argument list. If *'bhfdr', true* is given in the argument list, the Benjamini-Hochberg correction is performed (see above in section 6.1 about the Benjamini-Hochberg procedure).

   pValues – array of p-values.

   fdr, q – arrays containing the false discovery rates and the q-values, respectively, corresponding to each p-value.

   pi0 – the estimated fraction of true null hypotheses.

   Optionally, the method of estimating the fraction of correct null hypotheses can also be specified. By default, a bootstrapping procedure is used. The following syntax specifies that π_0 is to be determined by polynomial fitting:

```
[fdr,qValues,pi0]=mafdr(pValues,’method’,’polynomial’);
```

The following syntax specifies that the plot shown in Figure 22D is to be generated:

```
[fdr,qValues,pi0]=mafdr(pValues,’showplot’,true);
```

More detailed help about the function is available on the following web page:

https://www.mathworks.com/help/bioinfo/ref/mafdr.html

4. The Storey procedure can be carried out in R by typing the following lines of code at the command prompt:

```
install.packages("BiocManager")
BiocManager::install("qvalue")
library(qvalue)
p_values=read.table(file = "clipboard", sep = "\t")
p_values=as.numeric(unlist(p_values))
storey=qvalue(p_values)
names(storey)
install.packages("xlsx")
library(xlsx)
write.xlsx(storey$qvalues,"d:/myqvalues.xlsx")
```

The first three lines install and load the libraries required for running the qvalue function. Lines 4-5 load an array of p-values that has been copied to the clipboard. Line 6 performs the Storey procedure and saves the results into a variable named “storey”. Line 7 lists the fields of the variable or object. One of the fields is called “qvalues”. This field can be referenced by the syntax storey$qvalues. It will be saved to an Excel file named “myqvalues.xlsx” in line 10. In order to run the write.xlsx command, the “xlsx” library must be installed and loaded in lines 8-9. Data can also be read directly from a file instead of the clipboard using the following syntax:

```
p_values = read.xlsx("D:/pvalues.xlsx",1)
```

The above code reads the first sheet from the file "D:/pvalues.xlsx". read.xlsx also requires installation and loading of the “xlsx” library.

7. *A very brief summary*

The last section of this tutorial summarizes the punchlines of the previous chapters without any theory. Those statements are presented here that are essential from a practical point of view, and the most important equations and figures are referenced.

p-values and the false discovery rate

- p-value is the probability of rejecting a true null hypothesis. It is not equal to the probability that rejecting the null hypothesis was wrong. This latter probability is the false discovery rate.
- If the fraction of true null hypotheses is low, the p-value is approximately equal to the false discovery rate. However, when the null hypothesis is very likely to be true, the false discovery rate is much larger than the p-value (Figure 5).
- The false discovery rate can only be estimated if assumptions are made regarding the alternative hypothesis.
    - One of these assumptions is the power of the test, i.e. the probability that a true alternative hypothesis is accepted. The false discovery rate can be estimated according to equation (1) and Figure 4 in such a case.
    - Another assumption enabling calculation of the false discovery rate is the effect size, i.e. the difference between the expected means of the two samples according to the null and alternative hypotheses normalized by the SD. The false discovery rate can be calculated according to equation (14) and Figure 6 in this case.

  Estimation of the false discovery rate according to both approaches is carried out by the Excel workbook introduced here (Figure 7) and the fdrEstimation Matlab program (Figure 8). An online application performs these calculations only according to the second approach (Figure 9).

- According to an alternative interpretation of statistical tests, we have some sort of prior knowledge about the probability of the correctness of the alternative hypothesis. This is expressed by the prior odds of the alternative hypothesis. The statistical test essentially refines the prior odds and gives us the posterior odds of the alternative hypothesis, i.e. we will know how many times we can be more certain about the correctness of the alternative hypothesis (equation (9)). The Excel workbook, the fdrEstimation Matlab program and the online application can all perform these calculations (Figures 7-9).

- Another approach for characterizing the strength or reliability of a discovery achieved by rejecting the null hypothesis is the required prior probability of the alternative hypothesis so that the false discovery rate remains under a stipulated threshold (equation (18)). The higher this prior probability is, the less we can trust our discovery. The Excel workbook, the fdrEstimation Matlab program and the online application can all perform these calculations (Figures 7-9).

Sample sizes and the power of a statistical test

- Statistical confirmation of an effect requires a minimum sample size that depends on how large the effect is and how certain we would like to be to detect a real effect (power of the test). The smaller an effect is, the larger the sample needed is for its detection by a statistical test. This negative correlation is the consequence of the inverse relationship between the standard error of the mean and the sample size, i.e. larger samples allow us to estimate the means of the populations, from which they were taken, more accurately (equation (27) and Figure 11). The Excel workbook (Figure 13), a Matlab program (sampleSizeForTtest, Figure 14) and the G*Power program (Figure 12) can calculate the minimum sample size required for detecting a certain effect.

Controlling the false discovery rate for a large number of statistical tests

- When multiple hypothesis tests are performed, the probability of making at least one false discovery, called the family-wise error rate, increases with the number of tests according to equation (29) and Figure 15.
- Conventional approaches, like the ones used in ANOVA, are inappropriate for a large number of comparisons.
- If true null hypotheses are tested multiple times, the distribution of p-values will be uniform between 0 and 1 (Figure 16A). Both the Benjamini-Hochberg and the Storey procedures use this principle for controlling the false discovery rate.
- Both the Benjamini-Hochberg and the Storey procedures only require an array of p-values, and not the calculated statistical tests.
- Both approaches provide the q-value for every p-value. Although there is a slight difference between the false discovery rate and q-values (discussed on page 32 under

equation (31)), the q-value roughly corresponds to the false discovery rate if all p-values smaller than or equal to the current one are considered to be significant (if the corresponding null hypotheses are rejected).

- For the Benjamini-Hochberg procedure, one needs to specify a desired false discovery rate (*Q*), from which a Benjamini-Hochberg critical value is calculated for every p-value using the p-value and its rank (*i*) according to equation (30). If the p-value is smaller than the Benjamini-Hochberg critical value, the corresponding null hypothesis is rejected.
- The Benjamini-Hochberg procedure can be carried out with the “Controlling FDR” sheet of the Excel workbook introduced in this tutorial (Figure 18) or the correctFDR Matlab function (Figure 19).
- The Storey procedure is based explicitly on the uniform distribution of p-values when testing true null hypotheses. A horizontal straight line is fitted on the flat part of the p-value histogram providing the fraction of true null hypotheses (Figure 20A). p-values in the rectangle under this horizontal line correspond to true null hypotheses.
- The q-value in the Storey procedure is calculated according to equation (37), which basically represents the fraction of p-values in the black box (false positives) compared to those in the shaded area (false and true positives) in Figure 20C.
- The Storey procedure can be carried out with the “Controlling FDR” sheet of the Excel workbook introduced in this tutorial (Figures 18 and 21) or the correctFDR Matlab function (Figures 19 and 22).