Sample size and significance – somewhere between statistical power and judgment prostration

Cezary Watała

Department of Haemostatic Disorders, Medical University of Lodz, Poland

Submitted: 14 November 2006
Accepted: 17 December 2006

Arch Med Sci 2007; 3, 1: 5-13

Corresponding author:
Prof. Cezary Watała
Department of Haemostatic Disorders
Medical University of Lodz
Medical University Hospital No. 2
113 Zeromskiego Street
90-549 Lodz, Poland
Phone: +48 42 6393471
Fax: +48 42 6787567
E-mail: cwatala@csk.umed.lodz.pl

Abstract

When performing scientific research we are so "embraced" to use the tool of inductive logic in our reasoning that we often express more generalized opinions on the population of interest based on relatively small sample(s) of a general population. What we take care about in such situations is that chosen segments are representative for a whole set of elements in the general population. To cope with such a demand we always want to know how large our selected subpopulation should be to enable us to detect the experimental effect of interest not only at a certain level of significance, but also with the highest possible power of statistical reasoning. Thus, when designing our experiment, we have to compromise between a sample size not too small to ensure that our sample is sufficiently representative, and not too large to benefit from the sampling procedure at all. The tools for the estimation of minimum required sample size and the analysis of power, which help us to make quick decisions on how to compromise reasonably between significance, statistical power and sample size, are discussed in this paper.

Key words: sample size, statistical significance, statistical power, hypothesis testing, experimental design.

Usually we do not have access to the entire population of interest, or we simply do not want to investigate all elements of such a population. The reasons for the latter are usually very trivial: the entire population may be too large to be exhaustively measured, or it is too expensive and too time-consuming to allow more than some small fragment of the population to be monitored. Therefore, we often express our opinions on the population of interest based on a relatively small sample (s) of the general population. What we take care about in such situations is that chosen segments are representative for a whole set of elements of the general population. Random selection of these elements is an absolute 'must' in all respected procedures, and in most experimental approaches we also care that our data are independent of each other. The significance of these two basic requirements in sampling elements from the general population are discussed in considerable detail in another paper in this journal [1].

Sampling – theoretical background

On the basis of a small amount of sample data we try to compute the so-called statistic in order to evaluate some characteristic of a population

called a parameter. Mean and standard deviation are typical parameters characterizing a population under study. How close these parameters are to the real ones, representing the entire population, reflects how good a representation of the whole population our sample data are. For example, we want to estimate the mean weight of boys in the age range of 15-17 years, inhabiting a certain city with a total population of, say, 100,000. Such a mean value, μ, is a parameter of the general population of boys in the age range of 15-17 inhabiting the city. Assuming that the population of our interest numbers 17,500 boys, we will not investigate all of them but rather will draw a random sample of 100 boys. It is obvious that the number of boys to be pooled (n) will be quite small relative to the size of the total city population of these boys (N=17,500). Once the sample is selected we can estimate the mean weight of 100 boys, $\bar{x}$, which is called a statistic of the sample population. Of course, $\bar{x}$ will never be identical to μ; it will always deviate at least a little by a value which may be called the sampling error or imprecision of measurement. This is so simply because $\bar{x}$ involves what may be referred to as "the luck of the draw". This means that each single draw may result in slightly different $\bar{x}$ values, desirably close to, but not identical to the real μ. Obviously, in any analyzed sample, we may be absolutely sure that there will be some sampling error concerning the variable of interest. The main problem is that we are never certain how large this error is. If we knew how large this error was, we would actually know the exact value of the measured parameter, so we would never need to do any approximation.

Based on theoretical considerations we are of course not able to estimate what will happen in any particular experiment, but rather what will tend to happen in a larger population of a given size that we examined. Of course, the distribution of a measured parameter is given by some statistics over repeated measurements, with the estimated measure of a central tendency, the sample mean $\bar{x}$, getting closer to the population mean μ with increasing sample size (N). Due to natural variability and experimental error, in our sample of data there will always be a small percentage of values that are greater or less than μ. The distribution of values around μ reflects that our $\bar{x}$ is simply an imperfect indicator of μ, and shows how big is the "noise" around the "signal" when monitoring a given parameter of interest. Recalling the equation for SEM

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

we may notice that the experimental error gets smaller (and the accuracy increases) with increasing N. We have clearly shown that in the example discussed above. At large enough N we can be very certain that our estimated $\bar{x}$ will get very close to the population μ; we may say that large N leads us to nearly perfect accuracy.

The general rule governing the relation between sample size and sampling error is that the larger the sample size (n), the smaller the sampling error. Therefore, to increase the accuracy of our estimations we would desire to have our sample size large enough in order to make sampling error as small as possible. However, making a sample large enough to minimize sampling error and produce a reasonable accuracy of measurements inevitably means waste of time and money. So, there is a point in diminishing sample size, although not too much, as data would tend to be not precise and thus not of much use. Then we have to compromise between sample size not too small to ensure sufficient representativeness of our sample, and not to large to benefit from the sampling procedure at all.

The tools for the estimation of minimum required sample size and the analysis of power are thought to help us in making quick decisions on how to compromise reasonably between significance, statistical power and sample size.

The logic behind hypothesis testing

Undoubtedly, the most often used statistical technique in experimentation is testing of statistical hypotheses. First, before presenting the logic behind this strategy, we need to distinguish between two often misused terms: research hypothesis and statistical hypothesis. A research hypothesis is a general statement describing some natural phenomenon, association, difference, mechanism, likelihood of a given process, etc. We may refer to a research hypothesis as the hypothetical scenario of an examined phenomenon (or phenomena).

In contrast, a statistical hypothesis is a kind of a mathematical equation, precisely defining what we are comparing, linking or neglecting. As such, we may consider a statistical hypothesis as a (smaller) part of a more general and complex research hypothesis. Behind such a relation it stands that any research hypothesis may be "decomposed" into a few or several statistical hypotheses, each of which verifies a single equation or association.

Hence, as we can easily guess, the logic of how one can build up the statistical hypotheses is a kind of firm estimation – we are not absolutely free in how to state the null hypothesis and the alternative hypothesis. This originates from the fact that we are merely able to reject the null hypothesis at a certain likelihood, and never to accept it (to prove it is true).

It is habitual that statistical hypotheses are stated as logically compounding theses. They are coupled in such a way that the null hypothesis (the basic one) assumes equality and lack of differences ($\mu_1=\mu_2$), while, in contrast, the opposing alternative

hypothesis assumes the occurrence of differences ($\mu_1 \neq \mu_2$). Such a parity of statistical hypotheses naturally implies that these two opposite statements must by mutually exclusive.

The principle underlying the statistical evaluation showing that a hypothetical inequality $\mu_1 \neq \mu_2$ becomes true is the calculation of the so-called test statistic, based on our experimental data. Further, the outcome of the calculated test statistic equal to zero identifies two (or more) identical means. Relevant to that, the more the calculated value of the test statistic differs from zero, the higher the likelihood that the compared means are statistically (i.e. not by pure chance) different. In other words, the higher the estimated test statistic, the lower the chances that the null hypothesis (stating an equality) is true. Accordingly, a higher test statistic assures us that the calculated difference between means should be considered regular and not accidental. The important conclusion of the above is that we are only able to classify statistical hypotheses as true or false with a given likelihood, more or less differing from 1, and never with absolute certainty. If we are not able to deny the null hypothesis, it cannot be rejected at a given stage, but it certainly does not mean that it is true. It inevitably means that we are aware of committing an error in one of two circumstances:

- when we erroneously reject the null hypothesis in the case when we have no evidence that in fact it is false, or
- when we accept the null hypothesis when in fact it is not true.

The risk of such a misleading decision is defined as the likelihood of committing one of two statistical errors (incidentally, there are two types of statistical errors simply because there is parity in the assigning of two opposing and logically compounding hypotheses) (Table I). If we erroneously reject a "true" null hypothesis, we commit statistical error type I (error α). If we do not reject a "false" null hypothesis, we make statistical error type II (error β). In other words, the significance of a statistical test is nothing else but the risk of the error α. In turn, the likelihood of rejection of a "false" null hypothesis is known as the power of statistical testing (Figure 1).

		real world: H_0 is true	real world: H_0 is false
decision	reject H_0	**type I error** (probability = significance)	**decision correct** (probability = test power)
	accept H_0	**decision correct** (probability = 1 – significance)	**type II error** (probability = 1 – test power)

Figure 1. Principle of hypothesis testing and logical outcomes for committing type I and type II statistical error

This clearly explains why we always try to use tests of the highest statistical power of testing: to minimize the risk of not rejecting a null hypothesis which is not true. Powerful tests lead us to more reliable rejection of untrue null hypotheses, as far as the tested difference really occurs. The conventions are much more rigid with respect to α than β: while it is commonly accepted that α should be kept at or below 0.05, β is required not to exceed 0.2, which of course means that the statistical power should be at least 80% to detect a reasonable difference from what is stated in the null hypothesis (see also below).

Let us suppose that we are interested in showing that boys aged 15-17 years are on average taller than girls at the relevant age. We can state it as an equation: $\bar{x}_{boys} > \bar{x}_{girls}$. We intuitively believe it is true, but we need to prove it statistically. To perform the analysis, we need to arrange two opposite statistical hypotheses prior to collecting the experimental data. We remember that a null hypothesis (H_0) may only be rejected (and never accepted) and we know it is something logically opposite to the alternative hypothesis (H_A), which we believe is true. Now, we need to gather data and, using the statistical theory behind the hypothesis testing, we have to show from our data that it is likely that H_0 is false, and should be rejected. Thus, by rejecting the null hypothesis, we actually support what we believe. This kind of statistical reasoning is often called

Table I. Comparison of type I and type II statistical errors

type I error	type II error
designation: α	designation: β
definition: incorrect rejection of true H_0	definition: incorrect acceptance of false H_0
set in advance	dependent on other input parameters
not affected by sample size when set in advance	strongly depends on sample size and significance
increases with the number of tests or end points (correction for multiple testing needed)	may be estimated only as a function of the true population effect
	becomes smaller as the sample size gets larger
	becomes smaller as the number of tests or end points increases

reject-support (RS) testing, because while rejecting H_0 we support our experimental theory. In practice, an RS type experiment usually concerns the comparison of two means of a control and experimental group, when the experimenter believes that the tested treatment has an effect and tries to confirm it using a significance test that allows the null hypothesis to be rejected. In the case of RS testing the commitment of α (type I) error is relevant to a false-positive outcome for the experimenter's theory. From the researcher's point of view an α error is extremely undesirable, because it means a waste of time and energy, particularly when such a false-positive outcome is interesting from a theoretical standpoint. It stimulates further research, which obviously will not replicate the results of the original work, which was incorrect. It leads to confusion and frustration originating from inability to confirm the primary results. On the other hand, not a lesser tragedy is committing a β (type II) error in RS testing, because a "true" theory is rejected. Thus, we may not gain a benefit from a better therapy (which is incorrectly not shown as interesting compared to the control) and we lose a worthwhile procedure and discount an interesting idea of a researcher. Ultimately, we benefit if both errors are kept reasonably low, although in practice, especially at low sample sizes, there is often a trade-off between α and β errors.

Clearly the opposite logic occurs in so-called accept-support (AS) testing. In AS testing it is H_0 which we believe and intend to accept. Such a situation is routine e.g. in the pharmaceutical sciences, when proving that some preparation or analysis is not worse than another one (usually earlier performed). By accepting the null hypothesis in AS testing the researcher's theory is supported. Therefore, under such conditions an α (type I) error is a false negative for our theory (we deny a true H_0), whereas a β (type II) error is a false positive (we accept a wrong H_0). Consequently, favouring a very low type I error in RS testing is relevant to maximizing the belief in the researcher's theory in AS testing.

Considering the above, we are always challenged to compromise in how to get enough power of statistical reasoning with the lowest possible sample size. Small samples imply of course low power, but too high power may also be an obstacle. In RS testing even trivial differences between means in very large groups lead us unreservedly to reject the null hypothesis regardless of the real difference between groups. It is even worse in AS testing, since very high sample size often makes a researcher decide against a theory, even if the theory corresponds to the data nearly perfectly. In this particular case, high precision is "against" the researcher.

In summary, when we test the hypotheses in reject-support research we are intuitively interested in rejecting the null hypothesis, while there is a demand to keep the risk of type I error (α, significance) very low. High sample size works for the researcher; therefore the estimation of minimal sample size always pays off. We should be very concerned about the risk of type II (β) error and the appropriate statistical power. We have to keep in mind, however, that too much power is against us, as it makes trivial differences inappropriately become highly significant. In AS testing, we want to accept H_0. We are required to control type II (β) error and avoid accepting a false H_0, but we should also pay a lot of attention to minimizing type I (α) error and not to reject a true null hypothesis too hastily. Paradoxically, large sample sizes work against us, because if there is too much power, our theory may be "rejected" based on even very trivial fluctuations in our data.

Statistical significance

Conventionally, the term "significance" is used to describe our belief that we do not reject the true null hypothesis. Significance level refers to the risk of committing a type I error in such a meaning that higher significance denotes numerically a lower risk (Table 1). The latter is designated by the Greek alpha, whereas "p" or "P" is used to denote the significance level: low and very low "p" (or α) values are relevant to high or very high significance levels. There is a bit of a mix-up in the literature as regards use of either "α" or "p", and some researchers, less familiar with statistical terminology, are often confused as to which notation should be used to describe statistical significance. It has been accepted that "α" "should be reserved for a pre-chosen probability at the stage of experiment planning, while the term "p value" should be used to indicate a probability *a posteriori*, i.e. the one calculated after performing our study. Importantly, we can consider a significance level either a binary or exploratory measure. In the first case, we simply accept *a priori* a certain level of significance (α) and use it for hypothesis testing. Say, we decided to test our hypotheses at the significance of $\alpha < 0.001$. This means that if we reject the null hypothesis, there will be less than 1 in 1000 chance of being wrong (i.e. that we rejected a true H_0). Then we of course calculate the test statistic based on our experimental data, and compare the estimated value with the theoretical one, tabular, for $\alpha = 0.001$ at a given number of degrees of freedom. If our calculated "experimental" value of test statistic is higher than the tabular value, we reject the null hypothesis and accept the alternative one. Otherwise, we have no right to reject H_0. Overall, this approach is qualitative, because our choice is one of two possibilities and the employed test gives a one-bit outcome (1 or 0, YES or NO). When we reject H_0 and our calculated test statistic is much greater than the tabular one, it is quite likely that our rough description "less than 1 in 1000" may even approach the value of "1 in 5000" or "1 in 10,000".

However, it is usually of no further interest for us: we are satisfied that H_0 was rejected at the minimal, satisfactory for us, level of significance. There is a quite opposite approach when thinking about significance in exploratory terms. When using statistical software we often get the result expressed as the exact value of *a posteriori* significance, evaluated from the exact estimated value of the test statistic calculated from our experimental data. Using this analogue, quantitative approach, we do not need to use any convention in saying that 1% may be accepted as sufficiently significant, while 5% may not. The exploratory manner of viewing significance is particularly readily employed in multivariate statistical methods, when we are often at "the balance" of significance when juggling with various regression models including a variety of parameters.

Statistical power and sample size

These two techniques are milestones in the process of designing an experiment. They allow us to decide a) how large a sample is needed to enable statistical judgments that are accurate and reliable and b) how likely our statistical test will be to detect effects of a given size in a particular situation. Both sample size estimation and the analysis of statistical power are so important because without these calculations we may risk the possibility that sample size will be too high or too low. In the first case, if sample size is too low, the experiment will lack the precision to provide reliable answers to the questions raised by the investigator. On the other hand, if sample size is too large, you risk that you waste your time and resources, and you gain a minimal benefit. Hence, both seem crucial to perform a study in a cost-effective and scientifically useful manner.

Statistical power

When designing our experiment we should take care to level up the statistical power of our reasoning high enough to be able to detect a reasonable falsification of our null hypothesis. We remember that in powerful statistical procedures (tests) the risk of committing type II statistical error is minimal. This means that H_0 is rejected always when it is false, the researcher gets the support for what (s)he believes in, and therefore an experiment is worth doing at all. Numerous factors may have a considerable impact on statistical power. The most important include a) sample size, b) expected size of experimental effect, c) variation of data gathered in the experiment, and d) the type of test used in the calculus.

Obviously, the larger the sample size, the greater the power. We are aware, however, that increasing sample size needs investment of more energy and effort, longer time and higher costs of an experiment. That is why it is much more logical to make sample size large enough, but not unreasonably and wastefully large. Further, expecting high experimental effect we assume that our null hypothesis will be false to a substantial extent. The larger the expected difference, the higher the power of our reasoning. Remember, however, that the magnitude of experimental effect should be significant practically (clinically), and not merely statistically. High variability in the monitored parameter(s) means lower power. Consequently, better precision and higher consistency of experimental data improve statistical power. Finally, not all tests are equal with respect to statistical power: some of them are more powerful than others. Therefore we should choose the test wisely, according to its applicability to our data and our benefit to minimize type II error.

Sample size

Prior to running any experiment we are always confronted with several questions:

- how many measurements/readings should we gather to reason on the significance of an experimental effect with adequate power?
- can we be sure that having investigated a large group of elements we will be able to reason on the occurrence or lack of significant differences?
- what is our estimation of sample size based on? in other words, what do we need to input to calculate a sample size?
- is it reasonable to bother with the estimation of sample size? maybe it would be enough to continue our study as long as we still have financial support and there are still available data to be gathered,
- should the estimation of sample size be performed *a priori* (before running an experiment) or *a posteriori* (after the collection of data)? in other words, do we need to know how many to collect in advance of our experiment? or do we estimate the power, which comes off the already performed study?

The answer given by the leading statistical advisors of contemporary scientific research is definite and clear. The estimation of sample size **must** be performed before running any experiment to ensure *a priori* that our statistical testing will be done with adequate power. Leaving the above statistical argument aside, we may give two practical reasons why the estimation of sample size makes sense in research. We do the estimation to avoid collecting wastefully large number of data under circumstances when:

- we are able to see at first sight (even after collecting very few data) that our experimental effect is evident,
- there is no real experimental effect and we have no chance to show it even after collecting large samples (maybe there is a need to employ another methodology to perform measurements).

The logic of the estimation of sample size in these two situations is the same: in both cases we only risk wasting time, effort and money when we unreasonably continue our experiment, without considerably improved impact on significance and statistical power.

On the other hand, we should not stop our experiment too early. With too small sample size we risk inadequate power and erroneous acceptance of a false null hypothesis. Even if our experimental effect is obvious and we are "too impatient" to finish the study appropriately from a statistical point of view, we may not detect the expected effect, which makes our experiment hardly worth doing.

Let us analyze the following example. We are to investigate the effectiveness of two medical treatments, denoted A and B. Our initial assumption is that treatment A is more effective than treatment B in at least 70% of examined patients. This means of course that in 7 of 10 patients treatment A will turn out to be more effective than B, for which the proportion will be 5 of 10 patients. We are interested in showing a significant difference between these two treatments with significance of at least 5%, and we would like to detect the difference with the power of at least 90%. Our question is: how many patients should we enrol in the study?

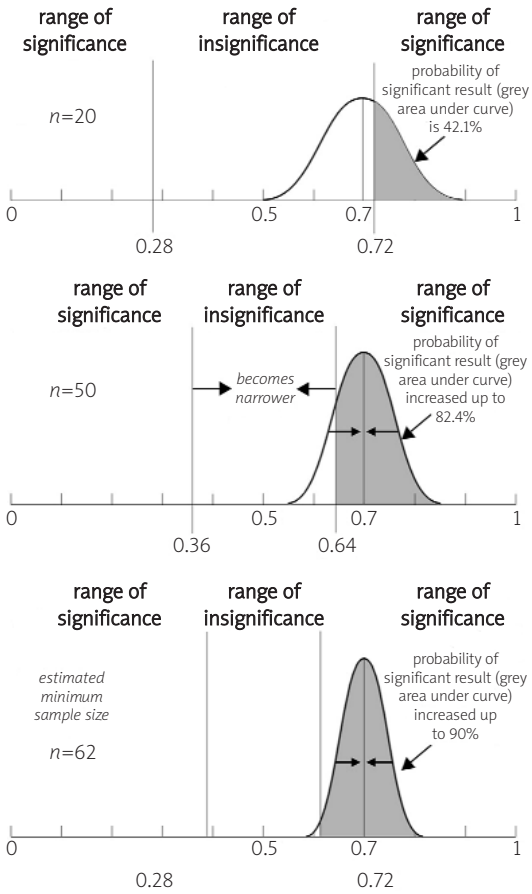

Figure 2. Probability of detecting an experimental effect (statistical power) at the significance of 5% and various sample sizes, when testing the hypothesis assuming the predominance of medical treatment A over B in at least 70% of cases (p (H_A)=0.7, while H_0 assumes the equality of treatments, p (H_0)=0.5)

First, let us assume that we will perform the study with 20 patients. We know that our result will reach statistical significance of at least 5% if it departs from a central value by at least 1.96-fold of the value of the standard error of the mean. Our null hypothesis says that the proportion of patients equals p=0.5 for both treatments (A and B), which clearly means that in one half of the patients treatment A will be effective, whereas in the other half it will not. Further, we expect a minimum experimental effect of 0.7–0.5=0.2. The standard error for the conditions of our null hypothesis will be:

$$SE = \sqrt{(0.5 \times 0.5)/20} = 0.1118$$

and our confidence intervals will be respectively:

$$+95\%CI = p + 1.96 \times SE = 0.5 + 1.96 \times 0.1118 = 0.72,$$
and
$$-95\%CI = p - 1.96 \times SE = 0.5 - 1.96 \times 0.1118 = 0.28$$

This means that a value significantly different (at α=0.05) from p=0.5 should be either lower than 0.28 or higher than 0.72. Keeping in mind that the real, expected proportion will be at least 0.7, we may ask: what is the probability that our observations will provide a result lying above 0.72 significant at α=0.05? This probability is relevant to the shaded area under the normal distribution curve in Figure 2 (upper plot), for which the mean value equals p=0.7 and $SE = \sqrt{(0.7 \times 0.3)/20} = 0.1025$. The estimated z value of the normal distribution equals:

$$\frac{0.72 - 0.7}{0.1025} = 0.1951$$

and the cumulative distribution (the area under the normal curve for $x \geq 0.72$) is 0.421. In other words, in a sample of 20 patients we have only a chance of 42.1% to detect a difference of at least 0.2 between treatments A and B (or proportion greater than the "cut-off" value of 0.72 required to reject H_0). If we increase the sample to 50 patients we will have:

$$SE = \sqrt{(0.5 \times 0.5)/50} = 0.0707$$
$$+95\%CI = p + 1.96 \times SE = 0.5 + 1.96 \times 0.0707 = 0.64,$$
and
$$-95\%CI = p - 1.96 \times SE = 0.5 - 1.96 \times 0.0707 = 0.36.$$

For $SE=\sqrt{(0.7 \times 0.3)/50}=0.06481$ the calculated z value of the normal distribution is

(0.64–0.7)/0.06481=–0.926, and the respective cumulative distribution equals 1–0.1762=0.8238. Hence, in the sample of 50 patients there is a probability of 82.4% to show that treatment A is more effective in at least 70% of cases (Figure 2, middle).

The power of our reasoning has increased from about 42% to over 82% and we may easily notice that it is clearly due to a lower SE and narrower distribution curve. However, we are still a little bit away from the demanded 90% power of detecting the departure from the null hypothesis stating that p=0.5. To reach the threshold values, we need to increase the sample size a little bit more. How much? We know that our outcome will remain significant (at α=0.05) below the value 0.5–1.96 × SE or above the value 0.5 + 1.96 × SE. Thus, we need a sample large enough to ensure that 90% of the area of our distribution lies above the "cut-off" point of 0.5 + 1.96 × SE (Figure 2, lower). The *z* statistic of a normal distribution corresponding to the cumulative distribution of 90% equals –1.28 and is relevant to the observed value of

$$0.7–1.28 \times SE = 0.7–0.28 \times \sqrt{(0.7 \times 0.3)/n}$$

Therefore, the estimated sample size should be large enough to ensure that:

$$0.7–1.28 \times \sqrt{(0.7 \times 0.3)/n} > 0.5 + 1.96 \times \sqrt{(0.5 \times 0.5)/n},$$

which gives

$$0.7–0.5 > \frac{1.96 \times \sqrt{(0.5 \times 0.5)} + 1.28 \times \sqrt{(0.7 \times 0.3)}}{\sqrt{n}}$$

$$n > \frac{[1.96 \times \sqrt{(0.5 \times 0.5)} + 1.28 \times \sqrt{(0.7 \times 0.3)}]^2}{(0.2)^2}$$

$$n > 61.36$$

Our sample should include at least 62 patients to be able to detect with the power of 90% the difference between two treatments at the significance of 5%, if we assume that treatment A should be more effective than treatment B in at least 70% of observations.

Overall, in planning a study we must estimate what constitutes the reasonable minimum effect that we wish to detect, the minimum power to detect that effect, and the sample size that will achieve that desired level of power. In an approach to estimate the minimum sample size we have to know:

- within-group variability (i.e. SD, SEM; what the level of experimental error is, what the accuracy/precision of the monitoring technique we use is),
- size of experimental effect (e.g. expected discrimination between groups being compared, how far we expect the compared groups to be different from each other),
- significance of a monitored experimental effect (risk of committing type I statistical error), and
- power of detecting the effect (risk of committing type II error).

It is worth emphasizing that all the above parameters (possibly except for the first under some circumstances) are set by the researcher. We decide on the values of these parameters even without doing any research, just based on the appropriate design of what we are supposed to study. Thus, we create the final "image" of our experiment ourselves, prior to collecting data. This may be hard to believe and we further may ask: how can we figure out e.g. the parameter variability or the size of experimental effect without calculating these parameters for the collected data? I can assure you, we are able to do it. If we think we are not, it simply means that we are not yet prepared mentally well enough to perform the experiment and we need more time to think it over. Let us briefly discuss the nature of each of the input parameters.

- The within-group variability of a parameter may be a derivative of either experimental error (originating e.g. from imprecision of a given detection technique), natural biodiversity within a population, or both, depending on the nature of the investigated issue. Only in extremely rare situations we have no idea how large such variation is. It could happen when we investigate parameter(s) or use a methodology which has never been examined or tested before, i.e. if our approach is absolutely novel in respect of what we monitor and how we monitor. Usually we are able to extrapolate the information on such variability from other studies performed earlier. Thus, it is just a matter of good and efficient literature search in order to learn more about the parameter we are supposed to study.
- Also the size of experimental effect is what we decide and not what we accept. Or at least, it should be so. It should be our decision what extent of the effect is "satisfactory" for us to consider e.g. that a given medical treatment is effective. Thus, we are not satisfied with any significant effect, but only with that which seems reasonable in terms of a clinically important impact. It seems reasonable to expect an effect visible in clinical practice and not any recorded change. What the satisfactory size of experimental effect should be depends of course on the particular study; however, effects exceeding 20% are usually considered clinically important.
- We remember that the level of significance means for us the risk of committing type I statistical error (rejection of non-false null hypothesis). In more frequently performed RS testing not too high significance (moderately low α) means simply not accepting the researcher's theory too hastily. With extremely high significance we assure ourselves

that accepting our theory will almost never be wrong. In the clinical sciences it reflects a reliable diagnosis and trustful prognosis. When deciding about the significance level appropriate for our experiment, we forget for a while about its exploratory capabilities. We look at the value of significance not in terms of quantity (what is the exact likelihood that a rejected H_0 may be true?), but rather in terms of quality (is it true or false?). The latter requires us to accept some convention on what may be referred to as "true", and what should be referred to as "false". Such a convention is nothing but an arbitrarily chosen threshold value of significance level used for rejecting or not rejecting H_0. As discussed above, the exploratory (quantitative) approach is much more informative. Here, we employ a binary (qualitative, 0/1) approach, but we have to decide ourselves where the threshold lies and what is the significance value that justifies acceptance of our theory. In the social sciences, and sometimes also in the biological sciences, the arbitrary chosen significance is usually 5% (which means a value of α=0.05), serving as a kind of universal threshold in a binary approach. Less often it may be 1% or 0.1% (meaning α=0.01 or α=0.001) for designating so-called "high significance". What does it mean that α=0.05? It means that for every 100 statements, conclusions, decisions, etc. you are wrong in 5 cases. We may say that we are secure of being right in 95 per 100 cases. Is it much or little? If we state the diagnosis in 100 patients and we are wrong in 5 of them, is this few enough to be acceptable, or is it so many that we are disqualified? This clearly shows that in clinical research, further extrapolated to clinical practice, there is no such thing as one commonly accepted boundary significance value. The tolerable significance level depends on a particular problem in a given clinical study. Usually it is set much lower than 0.05, but it is we, the researchers, who decide on this value based on our knowledge of what we investigate, what we are supposed to show, and what we would like to evidence. Therefore, we might say that any arbitrarily chosen significance level, in disjunction with a given scientific problem that needs resolving, deludes us with a false sense of security.

- Power of detecting the searched effect is the last parameter which we need to determine for successful experiment design. We remember that statistical power has be specified wisely, not too low to enable reliable detection, but also not too high to avoid showing unreal effects. Commonly, it is set within the range of 80-90% (see the discussion above).

There are plenty of available statistical packages, either professional [2, 3] or public domain [4-7], which enable us to estimate sample size and/or establish statistical power based on the input parameters of the minimum expected difference (experimental effect) and significance. More professional statistical programs offer detailed analysis of graphs showing the relationships of power vs. sample size for different extents of discrimination and different levels of experimental errors. The sample size can then be deduced by analyzing such graphs. In addition, some software packages can directly calculate the desired sample size for the input values of the user's choice. The second approach, though faster and apparently more convenient, is not good for inexperienced researchers. If you have the possibility to analyze the graphs, do it. It will always provide you with more information than some raw calculations. For instance, for a given statistical power, the plot of sample size vs. level of experimental error or expected experimental effect can show us how sensitive the estimated sample size is to the actual variability in our sample or the amount of difference we would like to see [2]. By playing a little bit with such estimators you will see for example that to reliably detect a small difference from the null hypothesis, we would need much greater *N* than for detection of a larger discrepancy. Look at the comparison given below [3]:

Sample size for a paired or single sample Student's t test

Significance (α)	0.05	0.05
Power (1-β)	0.9	0.9
Difference of mean from zero	5	10
Standard deviation	10	10
Estimated minimum sample size	45 pairs	13 pairs
Degrees of freedom	44	12

Concluding remarks

In general, it is much better if we are aware of the overall properties of a statistical test under different circumstances before we run an experiment instead of being confronted with unexpected difficulties after the fact. First of all, we have to keep in mind that even minor departures from the expected values of an experimental error and experimental effect may require a huge increase in sample size. We may not always be prepared to accommodate our research appropriately, e.g. due to budget limitations. Therefore, it is reasonable to play a little with different variants of our input parameters used to calculate a sample size, and not to accept the most optimistic variant. Certainly, we should do it at the stage of experimental design, before running our experiment, in order to create a wider "window of opportunities" for proper adjusting of sample size in our research.

References

1. Watala C. How to plan an experiment? I. Randomization: current fad or (ever) lasting fashion? Arch Med Sci 2006; 2: 58-65.
2. STATISTICA for Windows [Computer program manual]. Tulsa, OK: StatSoft, Inc., 2300 East 14th Street, Tulsa, OK 74104; email: info@statsoftinc.com, 2000.

3. StatsDirect statistical software, version 2.3.7. StatsDirect Ltd., 2004.
4. Simple Interactive Statistical Analysis (SISA) [Computer software]. Retrieved 13-11-2006 from http://home.clara.net/sisa/power.htm.
5. UCLA's Dept. of Statistics Power Calculator [Computer software]. Retrieved 13-11-2006 from http://stat.ubc.ca/~rollin/stats/ssize/.
6. Lenth R.V. (2006) Java Applets for Power and Sample Size [Computer software]. Retrieved 13-11-2006, from http://www.stat.uiowa.edu/~rlenth/Power
7. Schoenfeld DA. Statistical considerations for clinical trials and scientific experiments [Computer software]. Retrieved 13-11-2006 from http://hedwig.mgh.harvard.edu/sample_size/size.html.