Acta of Bioengineering and Biomechanics
Vol. 2, No. 1, 2000

The importance of sample size and statistical power in experimental research. A comparative study

LUCA CRISTOFOLINI

DIEM, Engineering Faculty, University of Bologna,
Viale Risorgimento 2, 40136 Bologna, Italy; email: cristofolini@tecno.ior.it
tel: +39-051-6366864, fax: +39-051-6366863

MICAELA TESTONI

Laboratorio di Tecnologia Medica, Rizzoli Orthopaedic Institutes,
Via di Barbiano 1/10, 40136, Bologna, Italy

The aim of this paper is to stress the importance of a proper statistical determination of the sample size in experimental research, and to underline the possible effect of an experiment with an inadequate number of cases on the results. The first part of the paper introduces the statistical concepts needed for the sample size calculation. Type I and II errors are defined and the associated probabilities are presented. Statistical power of a test is explained and its correlation with the sample size and experimental variability is defined. In a second section, the criteria for the calculation of the sample size are described. In the third part of the article, a statistical comparison between a real experiment and a numerical simulation is shown to highlight the consequences of the selection of different sample sizes. The risk of drawing mistaken conclusions caused by an inadequate sample size calculation is thus calculated.

Key words: sample size, statistical power, α and β errors, null hypothesis H_0

Glossary of symbols used

H_0 – Null hypothesis. Any difference between samples is caused by random variations.

H_A – Alternative hypothesis. The difference observed between samples is caused by the presence of a systematic factor.

Z_α – Abscise corresponding to the α value in the standard normal distribution. It is the threshold that separates the no-rejection area from the rejection ones. It is possible to define this score for each statistic distribution (T_α for Student's distribution, F_α for Fisher's distribution, etc.).

Z_β – Abscise corresponding to the β value in the standard normal distribution under H_A. It is possible to define this score for each statistic distribution (T_β for Student's distribution, F_β for Fisher's distribution, etc.).

σ – Population standard deviation.
s – Sample standard deviation.
α – Probability of rejecting H_0 when H_0 is actually true. *Type I error.*
β – Probability of failing to reject H_0 when H_0 is actually false. *Type II error.*
$1-\beta$ – Statistical power. Probability of detecting an existing difference between samples.
δ – Relevant difference between populations.
n – Sample size.
θ – Non-centrality parameter.

1. Introduction

The relevance and reliability of experimental findings are severely affected by the planning of the experiment. The sample size calculation (i.e. the number of specimens to be tested) is one of the most important tools of a correct planning [12]. Often a researcher is interested in detecting the difference between two or more samples (for instance in assessing whether a new prosthesis offers more or less stress shielding than an existing one, or in assessing whether a material wears more or less than another one, or in assessing whether a drug is effective in reducing the occurrence of osteoporosis). Hence, it is necessary to define strictly such a sample size that allows detecting the required difference.

In fact, the first researcher's question should be: how many cases guarantee relevant results?

Not always can this problem be easily solved, because due to resource limitations (time and money) the choice is very limited. Increasing the scale of experiment makes it generally expensive and time-consuming. On the other hand, in many cases patients and resources are "wasted" in a study that is not conclusive because of a too small sample size. This causes poor efficacy of the test employed and it may reduce any chance of revealing aspects of clinical importance. It is possible that no significant difference between groups is observed as a consequence of an inadequate number of cases, whereas a larger sample would have detected the same difference as significant. An investigator should then seriously consider whether it is worth performing the investigation, when the number of cases that can be afforded is too limited. An answer to this question can be given by considering the statistical power.

The aim of this article is to point out the importance of a correct sample size determination.

This paper is divided into three sections. The first section presents the tools needed for a complete understanding of the topics related to statistical testing and the sample size determination. The second section defines the criteria for determining the sample size. In the third section, a numerical simulation is presented in order to show the results that are obtained for different sample sizes. The simulation is based on the data gathered in real experiment with the aim of evaluating the mechanical features of bone cement.

2. Statistical power in hypothesis testing

A strong correlation exists between sample size, statistical power and variance. This issue involves statistical concepts such as the null hypothesis H_0 and alternative hypothesis H_A [4]. Under the null hypothesis H_0 every difference between the samples is caused by random variations, whereas under the alternative hypothesis H_A the differences observed are caused by the presence of a systematic factor.

2.1. The α and β values

Let $\bar{x}_A$ and $\bar{x}_B$ be the means of two random samples (e.g., x is the amount of wear debris produced by two prosthesis designs with different coatings, A and B). We want to evaluate if a systematic difference exists between A and B (e.g., we want to understand whether A wears to a greater or lesser degree than B). We may think of $\bar{x}_A$ and $\bar{x}_B$ as random variables. In fact, $\bar{x}_A$ and $\bar{x}_B$ represent the average of one of the samples that can be extracted from the two respective normal populations. If A and B are identical populations, the distribution of all possible values $\bar{x}_A - \bar{x}_B$ is a bell distribution centred in 0 (figure 1; H_0 distribution) [1]. If A and B are two populations with different mean values (μ_A is the mean value for the population A; μ_B is the mean value for the population B, where $\mu_A \neq \mu_B$), the distribution of all possible values $\bar{x}_{A,i} - \bar{x}_{B,j}$ is a bell distribution centred in $\delta = \mu_A - \mu_B$ (figure 1; H_A distribution).

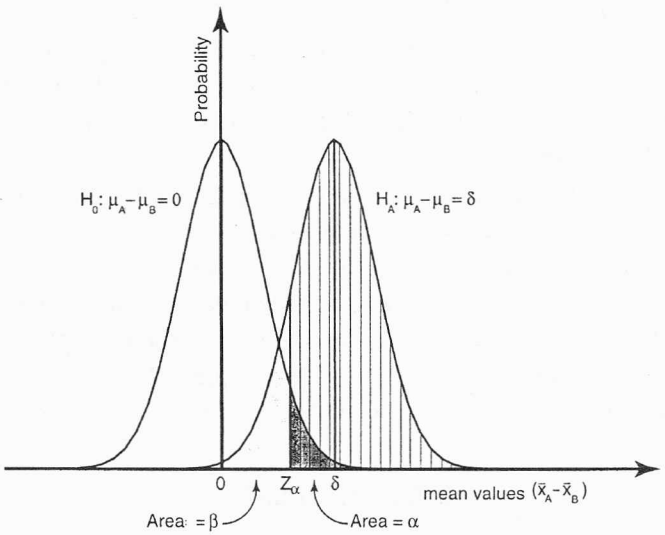

Fig. 1. H_0 is the distribution generated from the difference of all possible samples under the null hypothesis: the samples analysed are not different. H_A is the distribution generated from the difference of all possible samples under the alternative hypothesis: the samples analysed are actually different and the difference is equal to δ

Hence, if we reject the null hypothesis H_0 we conclude that the two populations (e.g., the two prosthetic designs) are significantly different. Conversely, if we do not reject H_0 we conclude that our data do not allow us to point out any difference between A and B. These results are probabilistic conclusions: then in both cases we may commit an error:

- The first kind of error (*type I error*, table 1) consists in rejecting the hypothesis H_0 when H_0 is actually true [3]. The probability of committing such an error is indicated by α (significance level).
- The second kind of error (*type II error*, table 1) consists in failing to reject the hypothesis H_0 when H_0 is false [3]. The probability of committing such an error is indicated by β.

Table 1. Statistical errors

Null hypothesis	Null hypothesis	
	True	False
Not rejected	Correct decision ($1-\alpha$)	*Type II error* (β)
Rejected	*Type I error* (α)	Correct decision ($1-\beta$)

In figure 1, the tails of the distributions H_0 and H_A represent the α and β errors, respectively. Z_α is the abscissa corresponding to the α ordinate under the distribution H_0:

- Area equal to α is the tail of the distribution H_0: the samples on the right side of the threshold Z_α are identified as the false-positive (figure 1). For these cases we mistakenly reject H_0 and the probability of this error is equal to α.
- Area equal to β is the tail of the distribution H_A: the samples on the left side of threshold Z_α are identified as the false-negative (figure 1). β is the probability of failing to detect a real difference between the groups.

No incontrovertible rules exist to define the values that should be used for the α and β errors. Their values are fixed by each researcher and they depend on the application field and on details needed for the study [2]. For example, in *in vitro* test, usually the α values between 0.1% and 1% are accepted. As far as the β value is concerned, this is frequently fixed in the range of 10%–25%.

2.2. Delta value (δ)

All the statistics described above, as well as the determination of statistical power, are related to the decision of not rejection/rejection of the null hypothesis H_0. Not rejecting/rejecting H_0 is equivalent to establishing whether a difference between groups is significant. An investigator must then indicate the magnitude of the differ-

ence (δ) between the population means that is considered as relevant to investigation. In other words, the investigator assumes that the differences smaller than δ are of no practical interest (no matter if they are real or not).

As an example, let us consider a comparison between the micromotion of different hip implants. Let us consider two samples: the control group and a novel design. A 0.02% improvement might be of scant interest in this case, where differences between stems easily reach 50% of the values measured. An investigator might decide, for instance, that only an improvement greater than 10% is relevant to his task. Hence, the value of δ is fixed on 10%. The value of δ is therefore determined on the basis of the relevant difference the investigator wants to detect.

2.3. Experimental variability

An indicator of the experimental variability is required in order to compare it with the experimental difference between the samples. The standard deviation of the population (σ) or that of the sample (s) are assumed as indicators of the experimental variability. Several times it may be difficult to indicate *a priori* the value of σ that characterizes the experiment. In this case, the possible solutions are as follows:

a) to develop a pilot study in order to estimate the variability;

b) to use appropriate values extrapolated from literature;

c) to fix a value of the difference δ, expressed as a fraction of σ, for example $\delta = 1.2$ times σ, in this way the problem of determining σ is overcome.

2.4. Statistical power ($1 - \beta$)

The probability ($1 - \beta$) of correctly drawing a true-positive conclusion is complementary to the *type II error* (β). This value is called the *statistical power* [3] and is the measure of the probability of rejecting (correctly) the null hypothesis H_0 when H_0 is false (i.e., it indicates the likelihood of detecting a significant difference by a test). The statistical power of the test is defined by the samples on the right side of the threshold Z_α under H_A (vertically shaded area, figure 1).

Moving Z_α, we can modify the two error values: shifting Z_α to the left causes an increase of β and a decrease of α, whereas shifting Z_α to the right results in a decrease of α and an increase of β. As will be shown in the next sections, a possible way to decrease both errors simultaneously is to increase the sample size.

The correlation between the number of cases (n) and statistical power ($1 - \beta$) can be described by a curve that converges to 1 with increasing n [11]. The convergence rate of the curve depends on the variability of the data.

All data in this paper are generated using a software implemented with SAS 6.11 (SAS Institute Inc., Cary, North Carolina, USA). Figures 2 and 3 present the results of two fictitious experiments concerning the comparison between the means of two sam-

ples (unpaired *t*-test). In figure 2, the power curves defined on the basis of the alternative hypothesis with a delta value (δ) equal to 1.8 are represented. In figure 3 the value of δ is fixed equal to 2.

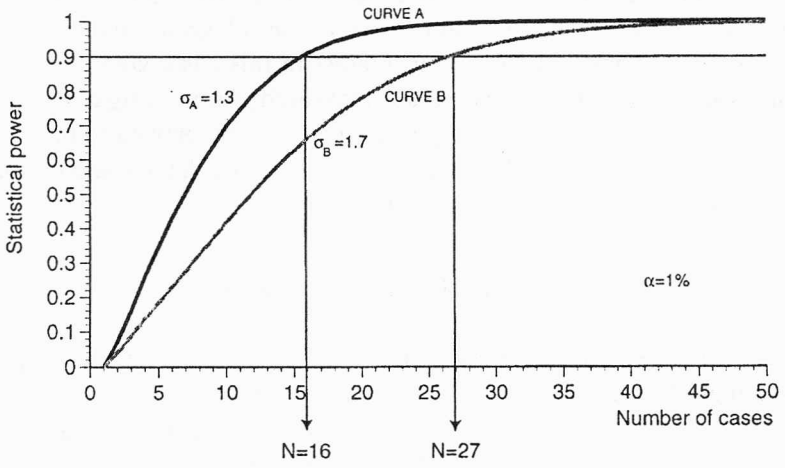

Fig. 2. Power curves for different σ values

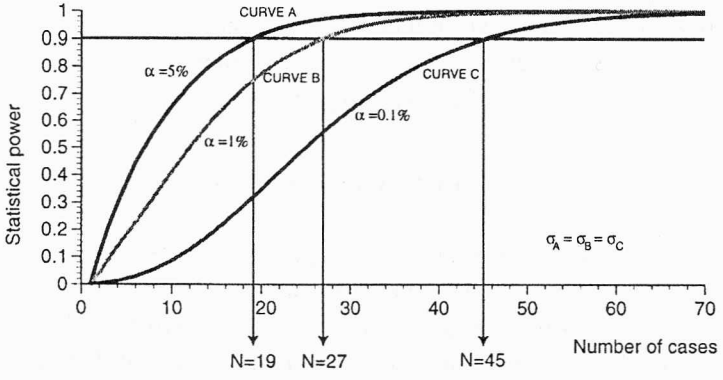

Fig. 3. Power curves for different α values (σ constant)

In figure 2, representing the two data groups, *A* and *B*, the values of the feature σ are equal to 1.3 and 1.7, respectively. Hence, *A* is characterized by a smaller variability than *B*. Both curves are defined for an α value equal to 1%. If we want to achieve a power equal to 90% in the first experimental situation (curve *A*), we need 16 cases, whereas for the second situation (curve *B*), we need at least 27 observations per sample. This example proves that the larger the experimental variability, the higher the number of cases required to achieve the same statistical power.

In figure 3, the authors determined the power curves for different values of α (curve *A*, α = 5%; curve *B*, α = 1%, curve *C*, α = 0.1%), whereas the value of σ is

kept constant ($\sigma_A = \sigma_B = \sigma_C = 1.9$). In other words, the same experimental data are analysed with different thresholds for α. If we are ready to accept a probability of committing a *type I error* (α) equal to 5%, the test will achieve a power of about 90% with only 19 cases. If we assume, however, $\alpha = 1\%$ and $\alpha = 0.1\%$, we will achieve the same statistical power as above (90%) considering at least 27 and 45 cases, respectively.

Both the examples presented above show that it is possible to achieve more powerful statistical results increasing the sample size. The sample size required is highly correlated with experimental variability. The knowledge of the experimental variability (σ) is a condition necessary for determination of the sample size [13].

3. Calculating the sample size required

In the first section of this paper, the concepts needed to determine the sample size were introduced. In this second section, the procedure of determining the sample size required is specified. The steps that have to be taken prior to determining the sample size are as follows:

1. Definition of the null hypothesis H_0.
2. Selection of an appropriate α value (*type I error*).
3. Selection of an appropriate β value (*type II error*).
4. Determination of the value of δ.
5. Determination of a variability statistical indicator such as the standard deviation (σ).

Several books and papers supply formulas, tables and figures to calculate the sample size in all experimental designs and for all statistical tests. In table 2, there is presented a list of references providing us with the tables and figures which are necessary for calculating the sample size in some most common experimental situations. Moreover, several software packages that calculate the sample size for the most common experimental design (comparison between two samples, comparison between proportions, analysis of variance, etc.) are available on the market.

Many texts provide the data necessary to calculate the sample size in a form of figure, called *operating characteristic curve* (O.C. curve) [14]. In order to explain the O.C. curves, one has to define a non-centrality parameter (θ), intrinsic to the relevant statistic distribution under the alternative hypothesis, e.g.

- for the comparison between two samples (unpaired *t*-test, Student's distribution) $\theta = \delta$;
- for the analysis of the variance θ the non-centrality parameter of the Fisher distribution is used.

The θ values are arranged in theoretical tables [19] and it is possible to calculate the non-centrality parameter θ based on the information introduced above. The O.C. curve is then a plot of the *type II error*, β, versus the non-centrality parameter θ.

Table 2. References for sample size calculation

Statistic distribution	Typical applications	References
Z-test	Comparison between proportions	FLEISS [5], 33–48, ARMITAGE [1], 200–202, SAHAI [20]
Exact test	Exact test for 2×2×2 tables	HABER [6]
Z-test	Comparison between two independent samples; σ is known	MONTGOMERY [15], 31–32, FISHER [4], 158–159
t-test	Comparison between two independent samples; σ is unknown	MONTGOMERY [15], 31–32, FISHER [4], 158–159
F-test	Single-factor experiments (ANOVA)	MONTGOMERY [15], 110–114, KASTENBAUM [7]
F-test	Two-factor experiments (ANOVA)	ODEH [18], 7–11
F-test	Randomised block design	MONTGOMERY [15], 143–145, KASTENBAUM [8]
F-test	Latin squares design	ODEH [18], 19–22
	Simple linear and quadratic regression	ODEH [18], 22–25
	Clinical trials	LACHIN [9], LACHIN [10]
	ROC analysis	OBUCHOWSKI [17], [16]

For other statistical tests such as survival analysis, logistic regression and comparison between proportions, other parameters are required (in place of θ) to determine the sample size.

4. Application: what happens if the wrong sample size is used?

This third section presents the statistical power analysis for an *in vitro* experiment and shows a statistical comparison between a real experiment and a simulation involving different numerical sample sizes. The likelihood of drawing wrong conclusion because of an inadequate sample size is calculated.

4.1. Reference experiment: materials and methods

The reference experiment was actually performed in our laboratory [21]. The experiment consists in determining the flexural Young's modulus for two commercial types of bone cement (Cemex HV and Cemex UHV, Tecres, Sommacampagna, Verona, Italy), as determined in the four point bending test. The test was performed following the procedure defined in the ISO 5833 standard, Section F. The purpose of the investigation was to assess whether there was a significant difference between the values of the Young modulus of two cements. An unpaired *t*-test analysis was applied.

4.2. Statistical power calculation

A statistical power software is used to calculate the sample size needed for proving statistically that there exists the relevant difference between the samples. The software was developed by the authors using SAS.

Following the steps listed in previous paragraphs, the following parameters were identified:

1. Null hypothesis H_0: the values of the Young modulus of two cements do not differ from each other.

2. The α value (*type I error*): an α error probability equal to 0.1% is considered acceptable for the experiment performed.

3. The β value: the target is a statistical power $(1 - \beta)$ of at least 85%. This corresponds to a β value equal or smaller than 15%.

4. The δ value: the researcher is interested in investigating a difference between cement mean values equal to 100 MPa, $\delta = 100$. This corresponds to about 3–4% of the mean expected. Smaller differences are of no practical interest.

5. Statistical indicator of the variability: σ is obtained from previous experiments ($\sigma = 50$ MPa).

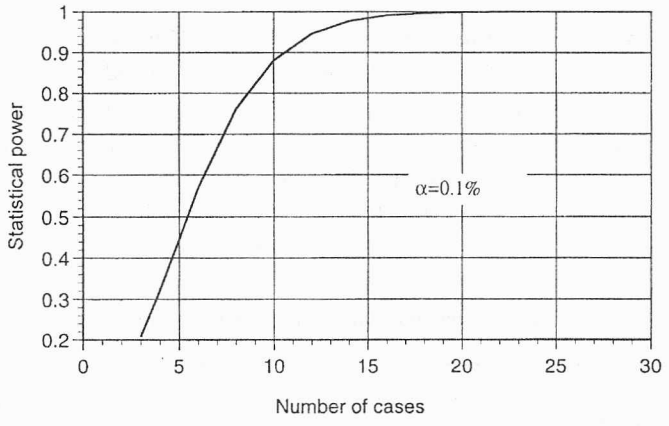

Fig. 4. Power curve determined for the reference experiment

With this information, it was possible to determine the number of specimens required for detecting a statistical difference between the samples of about $\delta/\sigma = 2$ times the standard deviation. Figure 4 presents the power curve for this experiment (implemented with the statistical power software described above). The formula used for the sample size determination is the following [4]:

$$n > 2\left[\frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)\sigma}{\delta}\right]^2.$$

The analysis performed is presented in table 3. In order to detect a significant difference between the samples with a statistical power $1 - \beta = 85\%$ (α is fixed equal to 0.1%), it is necessary to consider at least nine specimens per cement type (figure 4). The researchers actually used a sample of nine specimens for each of the two cement types under investigation.

Table 3. Statistical power analysis

Samples size	$\delta \pm \sigma$	*Type I error* α	*Type II error* β	Power test $(1-\beta)$
3	100 ± 50	0.1%	79%	21%
4	100 ± 50	0.1%	68%	32%
6	100 ± 50	0.1%	43%	57%
7	100 ± 50	0.1%	30%	70%
8	100 ± 50	0.1%	24%	76%
9	100 ± 50	0.1%	15%	85%
10	100 ± 50	0.1%	12%	88%
12	100 ± 50	0.1%	5%	95%

4.3. Reference experiment: results and discussion

The reference experiment has shown a significant difference between the values of the flexural Young modulus of the two cement types (unpaired *t*-test; $p < 0.01\%$). The null hypothesis H_0 must then be rejected. The probability α (originally fixed to less than 0.1%) of committing an error when rejecting the null hypothesis was actually less than 0.01% (table 4). The statistical power $(1 - \beta)$ achieved by the study was equal to 85% (table 3). It means that the failure probability in detecting an existing difference between the samples is equal to $\beta = 100\% - 85\% = 15\%$.

Table 4. Experimental results from the reference experiment

Cement type	Sample size	Mean value (MPa)	Standard deviation (MPa)	*p* value (α)
HV	9	2491	45	
UHV	9	2321	56	<0.0001
$H_0: \mu_{HV} = \mu_{UHV}$				

This result represents the reference point for the results produced by the numerical simulations described below.

4.4. Numerical experiment: materials and methods

The aim of this numerical experiment was to determine the statistical results that would be obtained if a smaller sample size were used. For this purpose, four numeri-

cal simulations corresponding to sample sizes of 3, 4, 6 and 8 specimens were implemented. All simulations were based on the data extracted from experimental measurements taken during the reference experiment. The analyses were developed using SAS.

The first simulation consists in:

- extracting all possible sub-samples of three elements from each of the two populations of the nine specimens (a total of 84 sub-samples were generated from each population);
- making all possible combinations of two sub-samples (one from each of the two populations).

Hence, a total of 7056 sub-samples consisting of three elements from each of the two cement types are obtained. Similarly, for the other three simulations all possible combinations of respectively four, six and eight elements were extracted from the same population of nine specimens (table 5).

Table 5. Number of simulation made for numerical experiment

Sample size	Number of couples of sub-samples extracted
3	7056
4	15876
6	7056
8	81
Total number of simulations	30069

An unpaired *t*-test was applied to each pair of sub-samples generated, so as to compare the values of the flexural Young modulus (similar to the comparison that was made on the nine specimens in the reference experiment). Thus, for each sub-sample a result in terms of significance or insignificance of the difference was obtained. The fraction of tests indicating a significant difference was determined for the size three, size four, six, and eight simulations. These results were compared with those produced by the experiment with nine specimens (reference experiment) which indicated a significant difference ($p < 0.01\%$).

4.5. Numerical experiment: results

Table 6 summarises the results of the numerical simulation. For each of the sample sizes simulated, it presents the fraction of times the unpaired *t*-test determined a statistically significant result, considering the α values of 0.1%, 1% and 5%.

With the same α value chosen as for the reference experiment ($\alpha = 0.1\%$), the difference appears insignificant in 94 and 68 times out of 100, with respectively 3 and 4

specimens per sample. The results of no significance decrease to 50.4% and 23.3% of the cases, respectively, if $\alpha = 1\%$ is accepted. With three specimens per sample and accepting a higher α value (5%), the comparisons are still insignificant 17 times out of 100.

Table 6. Numerical experiment results

Sample size	$\alpha = 5\%$		$\alpha = 1\%$		$\alpha = 0.1\%$	
	n.s.	signif.	n.s.	signif.	n.s.	signif.
3	17.5%	82.5%	50.4%	49.6%	94.2%	5.8%
4	0%	100%	23.3%	76.7%	68.3%	31.7%
6	0%	100%	0%	100%	2.1%	97.9%
8	0%	100%	0%	100%	0%	100%

n.s. – no significant difference between the samples; signif. – significant difference between the samples.

4.6. Numerical experiment: discussion

The unpaired *t*-test was applied to sample data extracted from the data obtained in the reference experiment. At first, let us consider an α value equal to that assumed in the reference experiment (0.1%). In 94 out of 100 cases of the first numerical simulation ($n = 3$), the difference appears to be insignificant (this is assumed as a "wrong" result). That means that the result obtained with a sample of 3 specimens is in agreement with the reference experiment only 6 times out of 100 (table 6). In fact, the statistical power of the unpaired *t*-test with 3 elements per sample was equal to 21% (table 3). In other words, the failure probability in detecting an existing difference between the data using 3 elements per sample was about 79%.

With four specimens per sample the percentage of "wrong" results decreased to 68%. Considering six specimens, the test would have failed to detect a significant difference only 2 times out of 100. However, the statistical power for a sample with 6 elements (table 3) is not high yet.

The number of "wrong" results (no significant difference detected) decreases if a larger *type I error* is accepted.

5. Conclusions

In scientific literature statistical methods are frequently used and a researcher has now familiarized himself with probability indicators such as the α value. A second probability indicator, or, better, a second type of error (β) has to be considered when analysing the data. In fact, the condition of "insignificant difference" could hide an

experiment with an inadequate number of cases. If a significant difference is detected, an investigator may be satisfied with the information obtained. Conversely, a result of insignificance does not add any information, as we are not able to establish whether the lack of significance is due to a poor sample size or to the fact that the two populations do not differ from each other. This possible flare is checked by verifying the statistical power ($1 - \beta$) of the test. If the statistical power is low ($1-\beta < 75$–80%) the experiment looses statistical relevance, because the observed case number is too small to detect real differences between the data.

Correcting an experiment after it has been performed is definitely a harder task than planning it carefully in advance. Therefore, the sample size should be appropriately chosen in the first steps of the experiment so as to avoid loss of time and resources.

The present paper illustrates this situation analysing the Young modulus of two cement types, by comparing real experiment and numerical simulation. When the result of numerical experiment was insignificant a researcher would have failed to reject the null hypothesis. He would not have succeed to point out a (real) difference between the two cement types. But if an investigator had calculated the statistical power prior to running the experiment, he would have found that he needed at least eight specimens to reduce the probability of the *type II error* (the failure probability in detecting an existing difference between the samples) to an acceptable value of $\beta = 24\%$ (the test statistical power, $1 - \beta = 76\%$). The study proves that a sample size of less than eight specimens is inadequate for the analysis of the reference experimental situation.

A mistake of this kind may cause problems regarding the waste of material and human resources employed in a study. Moreover, in some cases also ethical considerations are involved (e.g., if a drug is tested, whose effectiveness and/or toxicity are unknown). An inadequate choice of the sample size may have a relevant impact from a clinical point of view.

Summarising, the correct definition of sample size is one of the most important tools of the study. If this is not taken into account, the risk of drawing wrong conclusions and wasting time and resources can dramatically increase.

References

[1] ARMITAGE P., BERRY G., *Sampling*, [In:] *Statistical methods in medical research*, 3rd ed., Oxford, Blackwell Scientific Publications, 1994, pp. 78–92.

[2] BETHED R.M., DURAN B.S., BOULLION T.L., *Statistical inference: hypothesis testing*, [In:] *Statistical methods for engineers and scientists*, New York, Marcel Dekker, Inc., 1995.

[3] COLTON T., *Statistics in medicine,* Little, Brown & Co., 1974.

[4] FISHER L.D., VAN BELLE G., *Statistical inference: populations and samples*, [In:] *Biostatistics: a methodology for the health sciences*, New York, John Wiley & Sons, Inc., 1993, pp. 75–137.

[5] FLEISS J.L., *Determining sample size needed to detect a difference between two proportions*, [In:] *Statistical methods for rates and proportions*, New York, John Wiley & Sons, Inc., 1981, pp. 33–49.

[6] HABER M., *Sample sizes for the exact test of "no interaction" in 2×2×2 tables*, Biometrics, 1983, 39:493–498.

[7] KASTENBAUM M.A., HOEL D.G., *Sample size requirements: one-way analysis of variance*, Biometrika, 1970, 57:421–430.

[8] KASTENBAUM M.A., HOEL D.G., *Sample size requirements: randomized block designs*, Biometrika, 1970, 57:573–577.

[9] LACHIN J.M., *Introduction to sample size determination and power analysis for clinical trials*, Biometrics, 1981, 2:93–113.

[10] LACHIN J.M., FOULKES M.A., *Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance and stratification*, Biometrics, 1986, 42:507–519.

[11] LIEBER R., *Statistical significance and statistical power in hypothesis testing*, Journal of Orthorpaedic Research, 1990, 8:304–309.

[12] MANLY B.F.J., *Synthesis: carrying out a research study*, [In:] *The design and analysis of research studies*, Cambridge, Cambridge University Press, 1992, pp. 321–325.

[13] MATTHEWS D.E., FAREWELL V.T., *The question of sample size*, [In:] *Using and understanding medical statistics*, 2nd ed., London, Karger, 1988, pp. 197–205.

[14] MONTGOMERY D.C., *More about single-factor experiments*, [In:] *Design and analysis of experiments*, 3rd ed., New York, John Wiley & Sons, 1991, pp. 95–133.

[15] MONTGOMERY D.C., *Design and analysis of experiments*, New York, John Wiley & Sons, 1991.

[16] OBUCHOWSKI N.A., *Computing sample size for receiver operating characteristic studies*, Investigative Radiology, 1994, 29:238–243.

[17] OBUCHOWSKI N.A., MCCLISH D.K., *Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices*, Statistic in medicine, 1997, 16:1529–1542.

[18] ODEH R.E., FAX M., *Sample size choice*, New York, Marcel Dekker, 1991.

[19] PEARSON E.S., HARTLEY H.D., *Biometrika: tables for statisticians*, Cambridge, Biometrika Trustees at the University Press, 1972.

[20] SAHAI H., KHURSHID A., *Formulae and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two-sample design: a review*, Statistics in medicine, 1996, 15:1–21.

[21] VESCHI M., CRISTOFOLINI L., BALEANI M., TONI A., *Progetto dell'esperimento per la caratterizzazione a flessione a 4 punti di cemento acrilico per uso chirurgico*, E. Leone Publ. Catania, 1997; 203 p., XXVI Meeting of the Associazione Italiana Analisi Sollecitazioni.