How do I check if Nielsen is right?

Fabio Campos*, Dino Lincoln**, Maria Neves***, Sergio Cavalcante****, Walter Correia*****

** UFPE-CESAR, fc2005@gmail.com*
*** UFPE, dinolincoln@gmail.com*
**** CESAR, marie@cesar.org.br*
***** CESAR, svc@cesar.org.br*
****** UFPE, wfmc10@gmail.com*

Abstract: This research presents a statistical model to scientifically validate the appropriate number of product reviewers that would better reflect the views of the target population to which the product is intended. In this paper we verified if the confidence intervals for the number of participants that according to Nielsen are enough to identify 70% of usability errors of a website. We also conducted a usability test of a website using a number of evaluators calculated by statistical methods and it was possible to obtain the double of the confidence level than with the procedure suggested by Nielsen. In conclusion, it appears that the definition of the number of users must follow a scientifically proven statistical procedure, as the one presented here, in order to achieve significant levels of degree of confidence.

Key words: Usability, Evaluation, Sample Space, Confidence Interval

1. Introduction

In the process of designing a new product, the evaluation is an important step. The objective of this phase is to collect accurately the perception of the evaluators, in order, if necessary, to make changes or corrections to the product before it is released to the market. Ideally, the perception of the evaluators should reflect the perception of the population to which the product is intended, so that its chance of success on the market is maximized.

One point of decision when conducting research with users to determine their perception of a product is how many users should be chosen to be part of the evaluation phase so that data collected from this sample will indeed reflect the views of the target population of users.

In the Design field, there is a controversial debate over the definition of this quantity, more specifically in the field of interface design for websites. According to Nielsen & Landauer [6], five trained users are enough to identify 70% of usability errors of a website. Some authors reaffirm this model of Nielsen, but others contest it [2].

An alternative approach to Nielsen´s would be to utilize a model to scientifically justify the amount of product reviewers [1,2], based on the use of a sample calculation model, widely used in statistics, to solve this problem. Through statistical procedures it is possible to verify how many evaluators should participate in the process and with what degree of confidence the opinion of the evaluators, or "sample" represents the opinion of the population of consumers of the product, or "sample space."

In Section 2 of this paper we simulated the procedure described by Nielsen, in his work on the evaluation of interfaces [4], and using statistics, we analyzed the confidence intervals of its results. In Section 3, we verified,

through a case study, how to determine, using a statistic sample calculation model, the number of participants that would represent the opinion of the population with a high degree of confidence. Finally, in Section 4, it was presented the conclusion and final considerations about the results of the previous sections.

2. Two Examples of How to Determine the Size of a Sample

This section will show how to determine, using statistic methods, the confidence degree relative to a number of participants that would represent a sample space. It will also analyze, using statistics theory, what it is the confidence intervals for the number of participants that according to Nielsen are enough to identify 70% of usability errors of a website.

2.1 Example A:

Prior to demonstrate how to determine the confidence degree of a sample, it is important to understand a few statistics concepts, such as the following equation, Figure 1, to calculate a minimum sample size [9].

$$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

where:

$$\therefore \begin{cases} n = \text{minimum sample size.} \\ Z_{\alpha/2} = \text{critical value for the confidence level desired.} \\ \sigma = \text{standard deviation of the sample.} \\ E = \text{maximum error of estimate.} \end{cases}$$

Figure 1. Formula for the minimum number of participants representing the sample space

It is also important to know that the critical value that represents the degree of confidence of the sample ($Z_{\alpha/2}$) is a constant interpreted according to the Table 1 [8]. For instance, if $Z_{\alpha/2}$ = 2.0, it means that the sample has 95.44% of confidence that will represent the population, or the "sample space" in question.

Table 1. Values corresponding to degrees of confidence

$Z_{\alpha/2}$	%	$Z_{\alpha/2}$	%	$Z_{\alpha/2}$	%	$Z_{\alpha/2}$	%
0,0	0,00	1,0	68,26	2,0	95,44	3,0	99,74
0,1	7,96	1,1	72,86	2,1	96,42	3,1	99,80
0,2	15,86	1,2	76,98	2,2	97,22	3,2	99,86
0,3	23,58	1,3	80,64	2,3	97,86	3,3	99,90
0,4	31,08	1,4	83,84	2,4	98,36	3,4	99,94
0,5	38,30	1,5	86,64	2,5	98,76	3,5	99,96
0,6	45,16	1,6	89,04	2,6	99,06	3,6	99,96
0,7	51,60	1,7	9108	2,7	99,30	3,7	99,98
0,8	57.62	1,8	92,82	2,8	99,48	3,8	99,98
0,9	63,18	1,9	94,26	2,9	99,62	3,9	100,0

As a rule of thumb, values above 92.5% are relevant to represent the views of the "sample space", since these values behave exponentially [7,8].

The follow example will analyze statistically the model proposed by Nielsen [4] where the first 5 users found approximately 75% of usability errors that affected the interaction with a website. To this end, it is considered the illustrative simulation shown on Table 2.

Table 2. Sample "A" simulation

Participant	Percentage of usability errors found
A_1	30 %
A_2	55 %
A_3	20 %
A_4	80 %
A_5	70%

To calculate the standard deviation, as shown on equation explained in Figure 1, it was used the formula [7] and arranged notations shown on figures 2 and 3 below.

$$\sigma = \sqrt{\frac{(A_1 - M_a)^2 + (A_2 - M_a)^2 + (A_3 - M_a)^2 + (A_4 - M_a)^2 + (A_5 - M_a)^2}{M}}$$

Figure 2. Formula for the standard deviation of five scores

where:

$$\therefore \begin{cases} \sigma = \text{standard deviation.} \\ M_a = \text{arithmetic mean of research results.} \\ A_1, A_2, A_3, A_4, A_5 = \text{Scores associated with each member of the search.} \\ M = \text{number of users participating in the research.} \end{cases}$$

Figure 3 - Notations used for the standard deviation formula

To obtain the standard deviation is therefore necessary to extract the arithmetic mean of the sample presented on Table 2, as shown on Figure 4.

$$M_a = \frac{30 + 55 + 20 + 80 + 70}{5} \rightarrow M_a = \frac{255}{5} \rightarrow M_a = 51$$

Figure 4. Arithmetic mean of the sample "A"

Proceeding with verification, we obtain the following standard deviation (σ_A) , shown on Figure 5.

$$\sigma_A = \sqrt{\frac{2620}{5}} \rightarrow \sigma_A = \sqrt{524} \rightarrow \sigma_A = 22{,}8910$$

Figure 5. Standard deviation of the sample "A"

Using this standard deviation and the considering a sample of 5 users (n_A = 5) performing the test, we can obtain the following confidence level for the sample "A", shown on Figure 6.

$$n_A = \left(\frac{Z_{\alpha/2} \,.\, \sigma}{E}\right)^2 \rightarrow 5 = \left(\frac{Z_{\alpha/2} \,.\, 22{,}891}{5}\right)^2 \rightarrow Z_{\alpha/2} = \sqrt{\frac{125}{523{,}998}} \rightarrow Z_{\alpha/2} = 0{,}4884$$

Figure 6. Obtaining the confidence level of the sample "A"

According to the scale of Table 1, the confidence level for $Z_{\alpha/2}$ = 0.4884 is of approximately 40%. This means that, in this experiment, there is only 40% of probability that the 5 users would reflect the sample space.

2.2 Example B:

Still for illustrative purposes, consider the same method of Nielsen, but assuming that the users that participated in the evaluation phase were extremely well qualified designers in detecting usability problems. Let's also suppose that their percentage of usability errors found in the website would consists of the following shown on Table 3.

Table 3. Sample "B" simulation

Participant	Percentage of usability errors found
B_1	100 %
B_2	100 %
B_3	80 %
B_4	60 %
B_5	60%

Similarly, by repeating the procedure showed in the previous example, we obtained again the standard deviation (σ_B= 17.8885) and the degree of confidence of the example, as explained in Figure 7.

$$n = \left(\frac{Z_{\alpha/2} \,.\, \sigma}{E}\right)^2 \rightarrow 5 = \left(\frac{Z_{\alpha/2} \,.\, 17{,}8885}{5}\right)^2 \rightarrow Z_{\alpha/2} = \sqrt{\frac{5 \,.\, 25}{319{,}9984}} \rightarrow Z_{\alpha/2} = 0{,}625$$

Figure 7. Degree of confidence of the "B" sample

By examining Table 1, it is possible to find that for $Z_{\alpha/2}$= 0.625, the confidence level of the "B" sample is of 45%. This value represents that there is only 45% of probability that the 5 users would reflect the sample space.

3. Case study

Based on the results obtained by the simulations done on the previous section, the number of 5 evaluators represents a low degree of confidence, less than 50% percentage, of probability to represent the sample space, thus, considering the aforementioned rule of thumb, 5 evaluators would not be enough. The following case study aims to demonstrate how to find a number of evaluators that would represent, with a high degree of confidence (that is, higher than 92.5%), the sample space.

For this case study a total of 40 design students attending the discipline of Human-Machine Interface from two Brazilian universities were presented the 10 usability heuristics of Nielsen [5]. Subsequently, these students evaluated the usability of the website a university [1] shown on Figure 8.

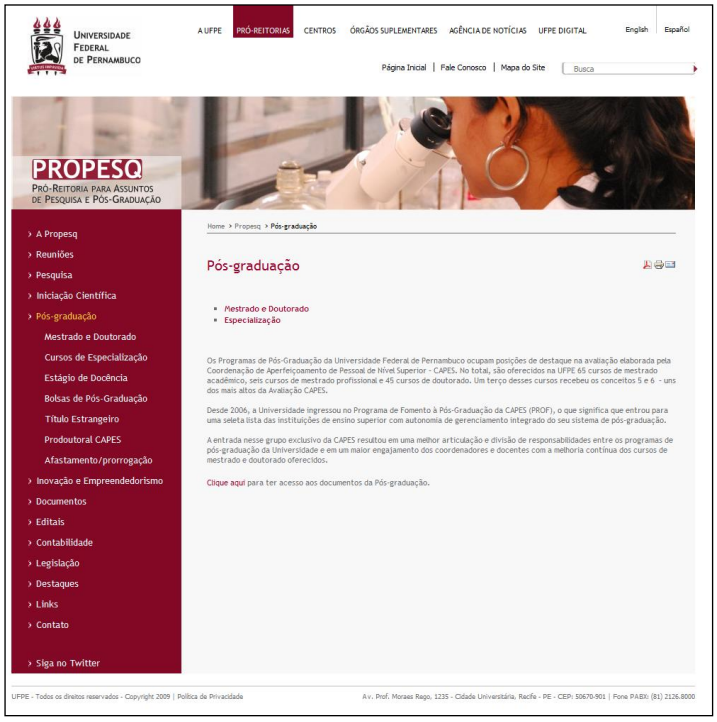

Figure 8. Interface from the website evaluated

To make a parallel with the method used by Nielsen, these 40 evaluators were organized in groups of 5 evaluators each, as presented on Table 4. This table also shows the number of errors found by each user, along with the percentage relative to the "maximum number of errors to be found", and the standard deviation and confidence level for each group.

Table 4. Usability errors reported by users

Group	Evaluator ID	Errors (Errors Percentage)	Standard deviation (SD), and confidence level ($Z_{\alpha/2}$)	Group	Evaluator ID	Errors (Errors Percentage	Standard deviation (SD), and confidence level ($Z_{\alpha/2}$)
A	1	50 (*98%*)	SD:	E	21	11 (*22%*)	SD:
	2	51 (*100%*)	28,673		22	8 (*16%*)	24,352
	3	21 (*41%*)	$Z_{\alpha/2}$:		23	23 (*45%*)	$Z_{\alpha/2}$:
	4	20 (*39%*)	0,78		24	43 (*84%*)	0,92
	5	21 (*41%*)	(56%)		25	17 (*33%*)	(64%)
B	6	27 (*53%*)	SD:	F	26	26 (*51%*)	SD:
	7	10 (*20%*)	27,702		27	25 (*49%*)	16,849
	8	50 (*98%*)	$Z_{\alpha/2}$:		28	25 (*49%*)	$Z_{\alpha/2}$:
	9	25 (*49%*)	0,81		29	6 (*12%*)	1,33
	10	43 (*84%*)	(58%)		30	10 (*20%*)	(81%)
C	11	22 (*43%*)	SD:	G	31	14 (*27%*)	SD:
	12	23 (*45%*)	13,832		32	25 (*49%*)	14,161
	13	11 (*22%*)	$Z_{\alpha/2}$:		33	29 (*57%*)	$Z_{\alpha/2}$:
	14	25 (*49%*)	1,62		34	9 (*18%*)	1,58
	15	33 (*65%*)	(90%)		35	19 (*37%*)	(88%)
D	16	50 (*98%*)	SD:	H	36	3 (*6%*)	SD:
	17	15 (*29%*)	28,479		37	19 (*37%*)	15,005
	18	21 (*41%*)	$Z_{\alpha/2}$:		38	25 (*49%*)	$Z_{\alpha/2}$:
	19	50 (*98%*)	0,79		39	21 (*41%*)	1,49
	20	30 (*59%*)	(57%)		40	21 (*41%*)	(86%)

The group "A", from Table 4, was composed by students considered by their teachers as “outstanding” in the area of usability. Therefore, the group “A” was chosen to be the control (reference) group for the other groups. The maximum number of errors found by the group “A”, was 51 errors; thus, since this is the “reference” group, the number 51 will represent the maximum number of errors to be potentially found by other evaluators in the analysis of the website.

It is worth noting that even with a high rate of errors found; the sampling group "A" returned a low degree of confidence, 56%.

The Table 5 shows a summary of the maximum percentage of errors found by the evaluators of each group.

Table 5. Change in percentage of usability errors found

Group	A	B	C	D	E	F	G	H
Errors found	100%	98%	65%	98%	84%	51%	57%	49%
Evaluator ID	2	8	15	16	24	26	33	38

Based in the statistical model of "infinite sample space" and "sampling groups" [8], it is possible to deduce the number of evaluators needed to get a high degree of confidence [7]. For this example, this number would be of 32.8, as it is shown on Figure 9.

$$nA = \left(\frac{Z_{\alpha/2} \,.\, \sigma A}{E}\right)^2$$

$$nA = \left(\frac{2 \,.\, 28{,}673}{10}\right)^2$$

$$nA = 32{,}8$$

Figure 9. Statistical calculation for obtaining the number of evaluators

To check the consistency of this information, we can use, for example, the data from the users from the groups "B", "C", "D", "E", "F", "G" and "H", taking the first 33 (round up of 32,8) answers from the 35 available from these groups. The standard deviation of this sample would be 24.542 and the confidence level would be of 97,86%, based on the calculations shown at Figure 10.

$$n_A = \left(\frac{Z_{\alpha/2} \,.\, \sigma_A}{E}\right)^2$$

$$33 = \left(\frac{Z_{\alpha/2} . \, 24{,}542}{10}\right)^2$$

$$Z_{\alpha/2} = \frac{10\sqrt{33}}{24{,}542}$$

$$Z_{\alpha/2} = 2{,}34$$

Figure 10. Calculating the confidence level of the sample with 33 evaluators

Based on these figures we can say that if there were 33 users evaluating the website then it would mean a sample that reflects the sample space of users of this product with a statistical confidence level of roughly 98%.

4. Final Considerations

By analyzing the procedure described by Nielsen in one of his works on the evaluation of interfaces [5], was possible to verify, based on basic statistical theory, that the degree of confidence of the results found by using his procedure was very low.

The results of the case studies done indicates that it is possible to obtain a much higher degree of confidence of the results of a usability test of a website, if ones makes use of a number of evaluators calculated by statistical methods.

The statistical approach presented in this work do not intend to be exhaustive, in the sense that it represents only one of several statistical tools available to assist the designer in this task.

In short, despite the remarkable work done by Nielsen [5], creating a set of heuristics to evaluate usability of websites, it is not useful to use the prescribed "5 users technique" described in his works [4], at least if one needs a level of confidence adequate to represent a population. It appears that the definition of the number of users must follow a scientifically proven statistical procedure, as the one presented here, in order to achieve significant levels of degree of confidence.

6. References

[1] Federal University of Pernambuco. *Website* . Available at <http://www.ufpe.br/propesq/index.php?option=com_content&view=article&id=69&Itemid=137 > [Accessed 12 March 2013].

[2] Figueiroa, D. L., Campos, F. (2012*): A avaliação de artefatos em Design e os problemas decorrentes da aleatoriedade (in Portuguese).* PhD Thesis. Universidade Federal de Pernambuco, Departamento de Design. Recife.

[3] Levin, J. , Fox, J. (1987) *Statistics for social sciences*. 2nd Ed. Editora Harbra Ltda. São Paulo.

[4] Nielsen, J. (2006) *Quantitative Studies: How Many Users to Test?*. Available at <http://www.nngroup.com/articles/quantitative-studies-how-many-users/ > [Accessed 12 March 2013].

[5] Nielsen, J. (1993) *Usability Engineering*. Academic Press, Boston.

[6] Nielsen, J.; Landauer, T.K. *A mathematical model of the finding of usability problems*. In: Proceedings of ACM INTERCHI'93 Conference 1993, ACM Press, pp. 206-213

[7] Pocinho, M.; Figueiredo, J.P. (2004) *Estatística e Bioestatística*. Lisboa, Madeira.

[8] Stevens,J.: Appied multivariate statistics for the social sciences. University of Cincinnati.

[9] Triola, M. F. (2008) *Essentials of Statistics*. 3rd Ed. Prentice Hall. Boston