Sample size: how many participants do I need in my research?*

Jeovany Martínez-Mesa[1] David Alejandro González-Chica[2] João Luiz Bastos[2]
Renan Rangel Bonamigo[3] Rodrigo Pereira Duquia[3]

DOI: http://dx.doi.org/10.1590/abd1806-4841.20143705

Abstract: The importance of estimating sample sizes is rarely understood by researchers, when planning a study. This paper aims to highlight the centrality of sample size estimations in health research. Examples that help in understanding the basic concepts involved in their calculation are presented. The scenarios covered are based more on the epidemiological reasoning and less on mathematical formulae. Proper calculation of the number of participants in a study diminishes the likelihood of errors, which are often associated with adverse consequences in terms of economic, ethical and health aspects.
Keywords: Cross-sectional studies; Dermatology; Epidemiology; Prevalence; Risk factors; Sampling studies

INTRODUCTION

Investigations in the health field are oriented by research problems or questions, which should be clearly defined in the study project. Sample size calculation is an essential item to be included in the project to reduce the probability of error, respect ethical standards, define the logistics of the study and, last but not least, improve its success rates, when evaluated by funding agencies.

Let us imagine that a group of investigators decides to study the frequency of sunscreen use and how the use of this product is distributed in the "population". In order to carry out this task, the authors define two research questions, each of which involving a distinct sample size calculation: **1)** What is the proportion of people that use sunscreen in the population?; and, **2)** Are there differences in the use of sunscreen between men and women, or between individuals that are white or of another skin color group, or between the wealthiest and the poorest, or between people with more and less years of schooling? Before doing the calculations, it will be necessary to review a few fundamental concepts and identify which are the required parameters to determine them.

WHAT DO WE MEAN, WHEN WE TALK ABOUT POPULATIONS?

First of all, we must define what is a *population*. Population is the group of individuals restricted to a geographical region (neighborhood, city, state, country, continent etc.), or certain institutions (hospitals, schools, health centers etc.), that is, a set of individuals that have at least one characteristic in common. The *target population* corresponds to a portion of the previously mentioned population, about which one intends to draw conclusions, that is to say, it is a part of the population whose characteristics are an object of interest of the investigator. Finally, *study population* is that which will actually be part of the study, which will be evaluated and will allow conclusions to be drawn about the target population, as long as it is representative of the latter. Figure 1 demonstrates how these concepts are interrelated.

We will now separately consider the required parameters for sample size calculation in studies that aim at estimating the frequency of events (prevalence of health outcomes or behaviors, for example), to test associations between risk/protective factors and dichotomous health conditions (yes/no), as well as with health outcomes measured in numerical scales.[1]

Received on 17.05.2014.
Approved by the Advisory Board and accepted for publication on 24.05.2014.
* Work carried out at the Latin American Cooperative Oncology Group (LACOG), Universidade Federal de Santa Catarina (UFSC), and Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Brazil.
Financial Support: None
Conflict of Interest: None

[1] Latin American Cooperative Oncology Group – Porto Alegre (RS), Brazil.
[2] Universidade Federal de Santa Catarina (UFSC) – Florianópolis (SC), Brazil.
[3] Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA) – Porto Alegre (RS), Brazil.

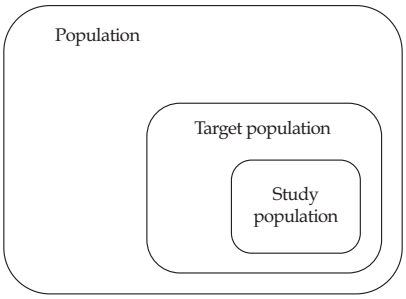

Figure 1: Graphic representation of the concepts of population, target population and study population

The formulas used for these calculations may be obtained from different sources – we recommend using the free online software OpenEpi (www.openepi.com).[2]

WHICH PARAMETERS DOES SAMPLE SIZE CALCULATION DEPEND UPON FOR A STUDY THAT AIMS AT ESTIMATING THE FREQUENCY OF HEALTH OUTCOMES, BEHAVIORS OR CONDITIONS?

When approaching the first research question defined at the beginning of this article (What is the proportion of people that use sunscreen?), the investigators need to conduct a prevalence study. In order to do this, some parameters must be defined to calculate the sample size, as demonstrated in chart 1.

Chart 2 presents some sample size simulations, according to the outcome prevalence, sample error and the type of target population investigated. The same basic question was used in this table (prevalence

Chart 1: Description of different parameters to be considered in the calculation of sample size for a study aiming at estimating the frequency of health ouctomes, behaviors or conditions

Parameter	Description	Remark
Population size	Total population size from which the sample will be drawn and about which researchers will draw conclusions (target population)	Information regarding population size may be obtained based on secondary data from hospitals, health centers, census surveys (population, schools etc.). The smaller the target population (for example, less than 100 individuals), the larger the sample size will proportionally be.
Expected prevalence of outcome or event of interest	The study outcome must be a percentage, that is, a number that varies from 0% to 100%.	Information regarding expected prevalence rates should be obtained from the literature or by carrying out a pilot-study. When this information is not available in the literature or a pilot-study cannot be carried out, the value that maximizes sample size is used (50% for a fixed value of sample error).
Sample error for estimate	The value we are willing to accept as error in the estimate obtained by the study.	The smaller the sample error, the larger the sample size and the greater the precision. In health studies, values between two and five percentage points are usually recommended.
Significance level	It is the probability that the expected prevalence will be within the error margin being established.	The higher the confidence level (greater expected precision), the larger will be the sample size. This parameter is usually fixed as 95%.
Design effect	It is necessary when the study participants are chosen by cluster selection procedures. This means that, instead of the participants being individually selected (simple, systematic or stratified sampling), they are first divided and randomly selected in groups (census tracts, neighborhood, households, days of the week, etc.) and later the individuals are selected within these groups. Thus, greater similarity is expected among the respondents within a group than in the general population. This generates loss of precision, which needs to be compensated by a sample size adjustment (increase).	The principle is that the total estimated variance may have been reduced as a consequence of cluster selection. The value of the design effect may be obtained from the literature. When not available, a value between 1.5 and 2.0 may be determined and the investigators should evaluate, after the study is completed, the actual design effect and report it in their publications. The greater the homogeneity within each group (the more similar the respondents are within each cluster), the greater the design effect will be and the larger the sample size required to increase precision. In studies that do not use cluster selection procedures (simple, systematic or stratified sampling), the design effect is considered as null or 1.0.

Chart 2: Sample size calculation to estimate the frequency (prevalence) of sunscreen use in the population, considering different scenarios but keeping the significance level (95%) and the design effect (1.0) constant

Target population	Prevalence (p) of outcome					
	Sunscreen use at work p=10%		Sunscreen use in sports p=35%		Sunscreen use at the beach p=50%	
	Acceptable Error 2 p.p.	Acceptable Error 5 p.p.	Acceptable Error 2 p.p.	Acceptable Error 5 p.p.	Acceptable Error 2 p.p.	Acceptable Error 5 p.p.
	Sample					
Health center users investigated in a single day (population = 100)	90	59	96	78	97	80
All users in the area covered by a health center (population size = 1,000)	464	122	687	260	707	278
All users from the areas covered by all health centers in a city (population size = 10,000)	796	137	1794	338	1937	370
The entire city population (N = 40.000)	847	138	2072	347	2265	381

p.p.= percentage points

of sunscreen use), but considering three different situations (at work, while doing sports or at the beach), as in the study by Duquia et al. conducted in the city of Pelotas, state of Rio Grande do Sul, in 2005.[3]

The calculations show that, by holding the sample error and the significance level constant, the higher the expected prevalence, the larger will be the required sample size. However, when the expected prevalence surpasses 50%, the required sample size progressively diminishes – the sample size for an expected prevalence of 10% is the same as that for an expected prevalence of 90%.

The investigator should also define beforehand the precision level to be accepted for the investigated event (sample error) and the confidence level of this result (usually 95%). Chart 2 demonstrates that, holding the expected prevalence constant, the higher the precision (smaller sample error) and the higher the confidence level (in this case, 95% was considered for all calculations), the larger also will be the required sample size.

Chart 2 also demonstrates that there is a direct relationship between the target population size and the number of individuals to be included in the sample. Nevertheless, when the target population size is sufficiently large, that is, surpasses an arbitrary value (for example, one million individuals), the resulting sample size tends to stabilize. The smaller the target population, the larger the sample will be; in some cases, the sample may even correspond to the total number of individuals from the target population – in these cases, it may be more convenient to study the entire target population, carrying out a census survey, rather than a study based on a sample of the population.

SAMPLE CALCULATION TO TEST THE ASSOCIATION BETWEEN TWO VARIABLES: HYPOTHESES AND TYPES OF ERROR

When the study objective is to investigate whether there are differences in sunscreen use according to sociodemographic characteristics (such as, for example, between men and women), the existence of association between explanatory variables (exposure or independent variables, in this case sociodemographic variables) and a dependent or outcome variable (use of sunscreen) is what is under consideration.

In these cases, we need first to understand what the hypotheses are, as well as the types of error that may result from their acceptance or refutation. A hypothesis is a "supposition arrived at from observation or reflection, that leads to refutable predictions".[4] In other words, it is a statement that may be questioned or tested and that may be falsified in scientific studies.

In scientific studies, there are two types of hypothesis: the null hypothesis (H_0) or original supposition that we assume to be true for a given situation, and the alternative hypothesis (H_A) or additional explanation for the same situation, which we believe may replace the original supposition. In the health field, H_0 is frequently defined as the equality or absence of difference in the outcome of interest between the studied groups (for example, sunscreen use is equal in men and women). On the other hand, H_A assumes the existence of difference between groups. H_A is called *two-tailed* when it is expected that the difference between the groups will occur in any direction (men using more sunscreen than women or vice-versa). However, if the investigator expects to find that a specific group uses more sunscreen than the other, he will be testing a one-tailed H_A.

In the sample investigated by Duquia et al., the frequency of sunscreen use at the beach was greater in men (32.7%) than in women (26.2%).[3] Although this what was observed in the sample, that is, men do wear more sunscreen than women, the investigators must decide whether they refute or accept H_0 in the target population (which contends that there is no difference in sunscreen use according to sex). Given that the entire target population is hardly ever investigated to confirm or refute the difference observed in the sample, the authors have to be aware that, independently from their decision (accepting or refuting H_0), their conclusion may be wrong, as can be seen in figure 2.

In case the investigators conclude that both in the target population and in the sample sunscreen use is also different between men and women (rejecting H_0), they may be making a type I or Alpha error, which is the probability of rejecting H_0 based on sample results when, in the target population, H_0 is true (the difference between men and women regarding sunscreen use found in the sample is not observed in the target population). If the authors conclude that there are no differences between the groups (accepting H_0), the investigators may be making a type II or Beta error, which is the probability of accepting H_0 when, in the target population, H_0 is false (that is, H_A is true) or, in other words, the probability of stating that the frequency of sunscreen use is equal between the sexes, when it is different in the same groups of the target population.

		Result in the target population	
		There is no difference in sunscreen use between the sexes (H_0 true)	There is difference in sunscreen use between the sexes (H_A true)
Results in the sample	There is no difference in sunscreen use between the sexes (Accepted H0)	CORRECT	Error type II (Beta)
	There is difference in sunscreen use between the sexes (Rejected H0)	Error type I (Alpha)	CORRECT

Figure 2: Types of possible results when performing a hypothesis test

In order to accept or refute H_0, the investigators need to previously define which is the maximum probability of type I and II errors that they are willing to incorporate into their results. In general, the type I error is fixed at a maximum value of 5% (0.05 or confidence level of 95%), since the consequences originated from this type of error are considered more harmful. For example, to state that an exposure/intervention affects a health condition, when this does not happen in the target population may bring about behaviors or actions (therapeutic changes, implementation of intervention programs etc.) with adverse consequences in ethical, economic and health terms. In the study conducted by Duquia et al., when the authors contend that the use of sunscreen was different according to sex, the p value presented (<0.001) indicates that the probability of not observing such difference in the target population is less that 0.1% (confidence level >99.9%).[3]

Although the type II or Beta error is less harmful, it should also be avoided, since if a study contends that a given exposure/intervention does not affect the outcome, when this effect actually exists in the target population, the consequence may be that a new medication with better therapeutic effects is not administered or that some aspects related to the etiology of the damage are not considered. This is the reason why the value of the type II error is usually fixed at a maximum value of 20% (or 0.20). In publications, this value tends to be mentioned as the power of the study, which is the ability of the test to detect a difference, when in fact it exists in the target population (usually fixed at 80%, as a result of the 1-Beta calculation).

SAMPLE CALCULATION FOR STUDIES THAT AIM AT TESTING THE ASSOCIATION BETWEEN A RISK/PROTECTIVE FACTOR AND AN OUTCOME, EVALUATED DICHOTOMOUSLY

In cases where the exposure variables are dichotomous (intervention/control, man/woman, rich/poor etc.) and so is the outcome (negative/positive outcome, to use sunscreen or not), the required parameters to calculate sample size are those described in chart 3. According to the previously mentioned example, it would be interesting to know whether sex, skin color, schooling level and income are associated with the use of sunscreen at work, while doing sports and at the beach. Thus, when the four exposure variables are crossed with the three outcomes, there would be 12 different questions to be answered and consequently an equal number of sample size calculations to be performed. Using the information in the article by Duquia et al.[3] for the prevalence of exposures and outcomes, a simulation of sample size calculations was used for each one of these situations (Chart 4).

Estimates show that studies with more power or that intend to find a difference of a lower magnitude in the frequency of the outcome (in this case, the prevalence rates) between exposed and non-exposed groups require larger sample sizes. For these reasons, in sample size calculations, an effect measure between 1.5 and 2.0 (for risk factors) or between 0.50 and 0.75 (for protective factors), and an 80% power are frequently used.

Considering the values in each column of chart 3, we may conclude also that, when the non-exposed/exposed relationship moves away from one (similar proportions of exposed and non-exposed individuals in the sample), the sample size increases. For this reason, intervention studies usually work with the same proportion of individuals in the intervention and control groups. Upon analysis of the values on each line, it can be concluded that there is an inverse relationship between the prevalence of the outcome and the required sample size.

Based on these estimates, assuming that the authors intended to test all of these associations, it would be necessary to choose the largest estimated sample size (2,630 subjects). In case the required sample size is larger than the target population, the investigators may decide to perform a multicenter study, lengthen the period for data collection, modify the research question or face the possibility of not having sufficient power to draw valid conclusions.

Additional aspects need to be considered in the previous estimates to arrive at the final sample size, which may include the possibility of refusals and/or losses in the study (an additional 10-15%), the need for adjustments for confounding factors (an additional 10-20%, applicable to observational studies), the possibility of effect modification (which implies an analysis of subgroups and the need to duplicate or triplicate the sample size), as well as the existence of design effects (multiplication of sample size by 1.5 to 2.0) in case of cluster sampling.

SAMPLE CALCULATIONS FOR STUDIES THAT AIM AT TESTING THE ASSOCIATION BETWEEN A DICHOTOMOUS EXPOSURE AND A NUMERICAL OUTCOME

Suppose that the investigators intend to evaluate whether the daily quantity of sunscreen used (in grams), the time of daily exposure to sunlight (in minutes) or a laboratory parameter (such as vitamin D levels) differ according to the socio-demographic variables mentioned. In all of these cases, the outcomes are numerical variables (discrete or continuous)[1], and the objective is to answer whether the mean outcome in the exposed/intervention group is different from the non-exposed/control group.

In this case, the first three parameters from chart 4 (alpha error, power of the study and relationship between non-exposed/exposed groups) are required, and the conclusions about their influences on the final sample size are also applicable. In addition to defining the expected outcome means in each group or the expected mean difference between non-exposed/exposed groups (usually at least 15% of the mean value in non-exposed group), they also need to define the standard deviation value for each group. There is a direct relationship between the standard deviation value and the sample size, the reason why in case of asymmetric variables the sample size would be overestimated. In such cases, the option may be to estimate sample sizes based on specific calculations for asymmetric variables, or the investigators may choose to use a percentage of the median value (for example, 25%) as a substitute for the standard deviation.

SAMPLE SIZE CALCULATIONS FOR OTHER TYPES OF STUDY

There are also specific calculations for some other quantitative studies, such as those aiming to assess correlations (exposure and outcome are numerical variables), time until the event (death, cure, relapse etc.) or the validity of diagnostic tests, but they are not described in this article, given that they were discussed elsewhere.[5]

Chart 3: Description of different parameters to be considered in the calculation of sample size for a study aiming at estimating the frequency of health ouctomes, behaviors or conditions

Parameter	Description	Remark
Type I or Alpha error	It is the probability of rejecting H0, when H0 is false in the target population. Usually fixed as 5%.	It is expressed by the p value. It is usually 5% ($p<0.05$). For sample size calculation, the confidence level may be adopted (usually 95%), calculated as 1-Alpha. The smaller the Alpha error (greater confidence level), the larger will be the sample size.
Statistical Power (1-Beta)	It is the ability of the test to detect a difference in the sample, when it exists in the target population. A value between 80%-90% is usually used.	Calculated as 1-Beta. The greater the power, the larger the required sample size will be.
Relationship between non-exposed/exposed groups in the sample	It indicates the existing relationship between non-exposed and exposed groups in the sample.	For observational studies, the data are usually obtained from the scientific literature. In intervention studies, the value 1:1 is frequently adopted, indicating that half of the individuals will receive the intervention and the other half will be the control or comparison group. Some intervention studies may use a larger number of controls than of individuals receiving the intervention. The more distant this ratio is from one, the larger will be the required sample size.
Prevalence* of outcome in the non-exposed group** (percentage of positive among the non-exposed)	Proportion of individuals with the disease (outcome) among those non-exposed to the risk factor (or that are part of the control group).	Data usually obtained from the literature. When this information is not available but there is information on general prevalence/incidence in the population, this value may be used in sample size calculation (values attributed to the control group in intervention studies) or estimated based on the following formula: PONE=pO/(pNE+(pE*PR)) where pO = prevalence of outcome; pNE = percentage of non-exposed; pE = percentage of exposed; PR = prevalence* ratio (usually a value between 1.5 and 2.0).
Expected prevalence* ratio	Relationship between the prevalence* of disease in the exposed (intervention) group and the prevalence* of disease in the non-exposed group, indicating how many times it is expected that the prevalence* will be higher (or lower) in the exposed compared to non-exposed group. Usually, a value between 1.50 and 2.00 is used (exposure as risk factor) or between 0.50 and 0.75 (protective factor).	It is the value that the investigators intend to find as HA, with the corresponding H0 equal to one (similar prevalence* of the outcome in both exposed and non-exposed groups). For the sample size estimates, the expected outcome prevalence* may be used for the non-exposed group, or the expected difference in the prevalence* between the exposed and the non-exposed groups. For intervention studies, the clinical relevance of this value should be considered. The smaller the prevalence rate (the smaller the expected difference between the groups), the larger the required sample size.
Type of statistical test	The test may be one-tailed or two-tailed, depending on the type of the HA.	Two-tailed tests require larger sample sizes

* It may be prevalence, incidence or risk, according to type of study; ** Non-exposed or control group; Ho - null hypothesis; Ha - alternative hypothesis

Chart 4: Sample size calculation to estimate the frequency (prevalence) of sunscreen use in the population, considering different scenarios but keeping the significance level (95%) and the design effect (1.0) constant

Exposure		Outcome Prevalence					
		Sunscreen use at work p=13.7% p=13.7%		Sunscreen use in sports p=30.2%		Sunscreen use at the beach p=60.8%	
		Expected PR 1.50	Expected PR 2.00	Expected PR 1.50	Expected PR 2.00	Expected PR 1.50	Expected PR 2.00
Sex:	**Power**	**PONE: 10.7%**		**PONE: 23.6%**	**PONE: 47.5%**		
Female: 56%(E)	**80%**	n=1298	n=388	n=487	n=134	n=136	n=28
Male:44%(NE)	**90%**	n=1738	n=519	n=652	n=179	n=181	n=38
r: 0.79							
Skin Color:	**Power**	**PONE: 9.7%**		**PONE: 21.4%**	**PDNE: 43.1%**		
White: 82%(E)	**80%**	n=2630	n=822	n=970	n=276	n=275	n=49
Other: 18%(NE)	**90%**	n=3520	n=1100	n=1299	n=370	n=368	n=66
r: 0.22							
Schooling:	**Power**	**PONE: 12.2%**		**PONE: 26.8%**	**PONE: 54.0%**		
0-4 years: 25%(E)	**80%**	n=1340	n=366	n=488	n=131	n=138	ND
>4 anos: 75%(NE)	**90%**	n=1795	n=490	n=654	n=175	n=184	ND
r: 3.00							
Per capita income:	**Power**	**PONE: 11.0%**		**PONE: 24.2%**	**PONE: 48.6%**		
≤133: 50%(E)	**80%**	n=1228	n=360	n=458	n=124	n=128	n=28
>133: 50%(NE)	**90%**	n=1644	n=480	n=612	n=166	n=170	n=36
r: 1.00							

E=exposed group; NE=non-exposed group; r=NE/E relationship; PONE=prevalence of outcome in the non-exposed group (percentage of positives in non-exposed group), estimated based on formula from chart 3, considering an PR of 1.50; PR=prevalence ratio/incidence or expected relative risk; n= minimum necessary sample size; ND=value could not be determined, as prevalence of outcome in the exposed would be above 100%, according to specified parameters.

CONCLUSION

Sample size calculation is always an essential step during the planning of scientific studies. An insufficient or small sample size may not be able to demonstrate the desired difference, or estimate the frequency of the event of interest with acceptable precision. A very large sample may add to the complexity of the study, and its associated costs, rendering it unfeasible. Both situations are ethically unacceptable and should be avoided by the investigator. ❑

REFERENCES

1. Duquia RP, Bastos JL, Bonamigo RR, González-Chica DA, Martínez-Mesa J. Presenting data in tables and charts. An Bras Dermatol. 2014;89:280-5.
2. OpenEpi.com [Internet]. Dean AG, Sullivan KM, Soe MM. OpenEpi: Open Source Epidemiologic Statistics for Public Health, Version. [updated 2013 Apr 6; cited 2014 Mar 22]. Available from: www.OpenEpi.com
3. Duquia RP, Baptista Menezes AM, Reichert FF, de Almeida HL Jr. Prevalence and associated factors with sunscreen use in Southern Brazil: A population-based study. J Am Acad Dermatol. 2007;57:73-80.
4. Porta M, editor. A dictionary of epidemiology. 5th. ed. New York: Oxford University Press; 2008.
5. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J. 2003;20:453-8.

Mailing address:
Jeovany Martínez-Mesa
Latin American Cooperative Oncology Group
Av. Ipiranga 6681 TECNOPUC Prédio 99 A Sala 806
90619-900 - Porto Alegre - RS
Brazil
E-mail: jeovanymm@gmail.com

Como citar este artigo: Martínez-Mesa J, González-Chica DA, Bastos JL, Bonamigo RR, Duquia RP. Sample size: how many participants do I need in my research? An Bras Dermatol. 2014;89(4):609-15.