Print ISSN: 2319-2003 | Online ISSN: 2279-0780

IJBCP International Journal of Basic & Clinical Pharmacology

doi: 10.5455/2319-2003.ijbcp000912

Educational Forum

Small size sampling

Rakesh R. Pathak[a,]*

[a]Department of Pharmacology, C.U. Shah Medical College, Surendranagar 363001, India

Received: 6 August 2012
Accepted: 13 August 2012

***Correspondence to:**
Dr. Rakesh R. Pathak,
Email: rr_pathak@yahoo.com

ABSTRACT

Based on the law of large numbers which is derived from probability theory, we tend to increase the sample size to the maximum. Central limit theorem is another inference from the same probability theory which approves largest possible number as sample size for better validity of measuring central tendencies like mean and median. Sometimes increase in sample-size turns only into negligible betterment or there is no increase at all in statistical relevance due to strong dependence or systematic error. If we can afford a little larger sample, statistically power of 0.90 being taken as acceptable with medium Cohen's d (<0.5) and for that we can take a sample size of 175 very safely and considering problem of attrition 200 samples would suffice.

Keywords: Systematic errors, Distribution, Sample size, Probability

We tend to increase the sample size to the maximum - based on the law of large numbers which is derived from probability theory. Central limit theorem is another inference from the same probability theory which approves largest possible number as sample size for better validity of measuring central tendencies like mean and median.

But it may not work every time. Sometimes increase in sample-size turns only into negligible betterment or there is no increase at all in statistical relevance due to strong dependence or systematic error. For example, due to systematic error, distance measured by radar will be systematically overestimated if the slight slowing down of the waves in air is not accounted for.

The same systematic error is seen when suspending support also shakes with each oscillation of pendulum and unaccounted as error in time measurements. Otherwise too, a sample size doesn't increase necessarily with population size.

For example, a margin of error ±5% and 95% confidence interval and nearly equal probability of getting or not getting an outcome is set (giving largest sample size), the sample size of 375 would suffice for population sizes of 15,000 and for 30,000 population the sample size is just 380. By a sample size of 385 with similar other parameters, we can safely deal a population of any size whatsoever.

Thus to take maximal benefit of the increasing sample size, please make sure that systematic errors of measurement or observation don't creep in. The second source of error that limits utility of larger samples is the dependence (i.e. two variables to be compared are not totally independent of each other). For example we know that genetically inherited trait of tallness can't make the heights of father and son independent.

Otherwise too, if we consider the correlation coefficient between the heights of fathers and their sons over all adult males, and compare it to the same correlation coefficient calculated when the fathers are selected to be between 165 cm and 170 cm in height, the correlation will be weaker in the latter case. In such a case just increasing the sample size is not remedial but instead we need the broadening of inclusion criteria.

For another example, we know that obesity and diabetes are not independent and if studying the comorbidities of diabetes obesity is taken as independent factor, the study design would be faulty, which no size of sample can rectify. Another such example is analgesic and healing effect of a drug – as stress impairs healing, a centrally acting painkiller can be easily dubbed as a good healing agent as well.

Coming to our current concern, except in some industry or government sponsored project (or those by international organizations like WHO), we as students of masters or Ph. D degrees have much constraint of time,

money and man power. In such cases, small size sampling with validity is mostly a must.

According to Mead's resource equation E = N – B – T (all notations explained ahead; where E should be between 10-20). If the values < 10, the sample is too small and if >20, too large. For example, suppose we plan a study using laboratory animals with four treatment groups having eight animals per group.

Making 32 animals in total (*N*=31; degree of freedom is one less), without any further stratification (*B*=0), and for 4 treatment group *T*=3 (degree of freedom is one less), then *E* (degree of freedom of error) would equal to 28. Thus the value is above the cutoff of 20, indicating that sample size may be a bit too large, and six animals per group may suffice.[1] But many statisticians think E = 12 or 15 as better lower limit instead of 10, upper limit remaining the same to avoid wastage of resources.

There can be case of binomial distribution in our data. For example, we may be interested to know the number of people >65 years in a given populations (say for planning a budget for senior citizen subsidy). In such sampling, anyone can be either yes (> 65 years) or no (< 65 years)" and the distribution would be called binomial. In such sampling $4\sqrt{(0.25/n)} = W$ and we can find thereby that $n = 4/W^2 = 1/B^2$ where *B* is the error bound on the estimate.

The estimate is usually given as *within* ± *B*. So, for *B* = ±10% [in the equation $4\sqrt{(0.25/n)}$, total range of error = 20% = 0.2 calculated = W] one requires *n* = 100 (±10% error is permitted in budgeting). Similarly for *B* = ±5% one needs *n* = 400, for *B* = ± 3% the requirement approximates to *n* = 1000, while for *B* = ±1% a sample size of *n* = 10000 is required.

When confidence interval is defined large (± 10 mm Hg for maintenance antihypertensive therapy), considering 95% confidence level enough - for a population of 5000 (catchment area of a hospital), the sample size would be <100 (between 93 and 94, to be exact).[2]

The resource websites[2-4] quoted above also give an excellent dealing on the terminology and process along with online estimate of our required sample size. The last one gives a printable tabulation that can be saved offline for future use - but maximal error calculated is 5%.

For a more strict confidence interval (± 5% variation in BP to decide probability of brain hemorrhage) with same sample size of <100, we would require a case in which response distribution in earlier calculation is 85% or 15% (for example, either 85% people have faced hemorrhage by variation of ± 5% in BP or just 15% people have had it).

With an estimates of 2.5% Indian population having G6PD deficiency[5], for 5% estimated error and population >200 (right from a study within college campus up to worldwide survey), a sample size of 40 is more than enough and by increasing sample size to 65 we can have > 99% confidence interval. If we can increase the sample size to 101, we can manage estimated error of ±4%.

Otherwise, the sample size $n = 16\sigma^2/W^2$. For example, if we are interested in estimating the amount by which a drug lowers a subject's blood pressure with a confidence interval that is six units (mm Hg) wide, and we know that the standard deviation of blood pressure in the population is 15, then the required sample size is 100. The same sample size of 100 can also work if we target for confidence interval that is 5 units (mm Hg) wide, and we know that the standard deviation of blood pressure in another population is 10.

If we can afford a little larger sample, statistically power of 0.90 being taken as acceptable with medium Cohen's d (<0.5) and for that we can take a sample size of 175 very safely and considering problem of attrition (sample lost or rejected while study) 200 samples would suffice.[6]

REFERENCES

1. Kenny David A. The two group design. In: Kenny David A, eds. Statistics for the social and behavioral sciences. Boston: Little, Brown; 1987:215.
2. Sample size calculator. Available at http://www.surveysystem.com/sscalc.htm. Accessed 2 August 2012.
3. Sample size calculator. Available at http://www.raosoft.com/samplesize.html. Accessed 2 August 2012.
4. Sample size table. Available at http://research-advisors.com/tools/SampleSize.htm. Accessed 2 August 2012.
5. Glader BE. Glucose-6-phosphate dehydrogenase deficiency and related disorders of hexose monophosphate shunt and glutathione metabolism. In: Wintrobe's Clinical Hematology, 10th ed, Lee GR, Foerster J, Lukens J, et al. (Eds), Baltimore, Williams & Wilkins; 1999:1178.
6. Mead R. The design of experiments. Cambridge, New York: Cambridge University Press; 1988:620.