Assessment of Required Sample Sizes for Estimating Proportions

Methods Article

Abstract

When estimating a population proportion p within margin of error m, a preliminary sample of size n is taken to produce a preliminary sample proportion y/n, which is then used to determine the required sample size $(y/n)(1-y/n)(z/m)^2$, where z is the critical value for a given level of confidence. The population is assumed to be infinite, so these Bernoulli(p) observations are mutually independent. Upon taking a new sample based on the required sample size, the coverage probabilities on p are determined exactly for various values of m, n, p, and z, using a commonly-used formula for a confidence interval on p. The coverage probabilities tend to be somewhat smaller that their nominal values, and tend to be a lot smaller when np or $n(1-p)$ is small, which would result in anti-conservative confidence intervals. As a more minor conclusion, since the given margin of error m is not relative to the population proportion p, then the required sample size is larger for values of p nearest to 0.5. The mean and standard deviation of the required sample size are also computed exactly to provide prospective, regarding just how large or how small these required sample sizes need to be.

Keywords: Bernoulli distribution, Binomial distribution, sample size determination, confidence interval.

2020 Mathematics Subject Classification: 62F10

1 Introduction

Determining the appropriate sample size is an important component of research design, as it directly influences the credibility and utility of study outcomes; it impacts both a study's ability to find meaningful effects and the accuracy of the estimates derived from the data. If the sample size is insufficient, valid conclusions about the data often cannot be made. This introduction synthesizes insights from a range of sources, focusing on the methodologies for determining sample sizes for means and proportions.

These approaches predicate the need for the evaluation of the common formula for sample size determination using the coverage probabilities we are evaluating. Accurately determining sample size is crucial in probability as it significantly impacts the statistical power of a study. Foundational principles guide researchers in the statistical community, underscoring the theoretical underpinnings necessary for understanding sample size calculations and their implications for research outcomes using a commonly-used equation [1, 2, 3]. The importance of sample size determination is studied across various study designs and its impact on the validity of research findings [4]. This work provides a practical perspective on the challenges and considerations involved in sample size estimation, offering valuable guidance for researchers.

The complexities associated with calculating sample size for two proportions is addressed, highlighting how the choice of formula and software can significantly influence the calculated sample sizes [5]. This shows the variability and potential inconsistencies in sample size determination practices.

Sample size estimation for health and social science research is explored, presenting principles and considerations tailored to different study designs [6]. This review contributes to a deeper understanding of the factors influencing sample size decisions and their implications for research in these fields.

Methodological insights into sample size calculation for comparing proportions and estimating intraclass correlation coefficients, respectively, are provided [7, 8]. These contributions highlight the mathematical and statistical considerations essential for accurate sample size determination in specific statistical analyses. An approach to determine the optimal sample size for clinical trials, accounting for the population size, is proposed [9]. This approach emphasizes the significance of incorporating broader population characteristics into sample size calculations, providing a nuanced perspective that improves upon conventional methodologies.

The foundational aspects of calculating sample sizes in clinical research is discussed [10]. This article articulates the basic principles underlying sample size calculations, such as the importance of specifying the margin of error, confidence level, and the expected effect size. A method for estimating population proportions is presented, highlighting the potential for significant improvements in efficiency compared to traditional estimators [11]. This article shows the importance of statistical techniques in sample size calculation. Having a sufficient sample size is needed when showing the prevalence of smoking, heart disease, diabetes, and other matters related to health [12]. Showing differences between users and nonusers of electronic cigarettes regarding heart rate, blood pressure, and oral temperature also requires a sufficient sample size [13].

Our paper seeks to bridge the gap between the theoretical and practical aspects of sample size determination by evaluating a commonly-used formula for sample size through the generation of coverage probabilities. This approach contrasts with the existing literature which predominantly focuses on deriving sample sizes for specific situations or using software tools for estimation. We are examining the performance of this particular formula under different scenarios. By considering values producing extra low coverage probabilities, such as 94.0% or lower when the nominal level is 95%, we seek to uncover potential limitations and biases associated with using this formula in practical settings and applied research practices.

2 Method for Determining Coverage Probabilities

A common question in research is what sample size, N, is required when estimating a population proportion or Bernoulli probability, p, for a given value of m, the margin of error, and nominal level of confidence, often set to 95%. The required conservative sample size is

$$N \;=\; 0.25\,(z/m)^2, \tag{2.1}$$

where z is the standard normal critical value and is 1.96 for 95% confidence. Note that a confidence interval on p is often defined to be

$$\widehat{p} \pm z \sqrt{\widehat{p}\,(1-\widehat{p}\,)\,/\,N}\,, \tag{2.2}$$

where $\widehat{p}$ is the sample proportion of successes based on N independent Bernoulli observations.

If a small preliminary sample of size n produces y Bernoulli success and $(n-y)$ Bernoulli failures, then a preliminary estimate of p is $p^* = y/n$. When p is near 0 or 1, then an approach more efficient than using Equation 2.1 is setting the required sample size to $N_y = \lceil\, p^*(1-p^*)(z/m)^2 \,\rceil$, which depends on y, where $\lceil \eta \rceil$ is the ceiling function which rounds any number η upward to its nearest integer. In the extremely rare situation where $y = 0$ or $y = n$, we redefine the required sample size of the new sample to be $N_y = 1$ rather than $N_y = 0$. Therefore,

$$N_y = \max\left\{ \left\lceil \frac{y}{n}\left(1-\frac{y}{n}\right)\left(\frac{z}{m}\right)^2 \right\rceil \,,\, 1 \right\}\,. \tag{2.3}$$

Once the required sample size, N_y, is determined for the new sample, then the final estimate of p is simply $\widehat{p} = x/N_y$, where x is the number of Bernoulli successes from the new sample of size N_y. The required sample of N_y Bernoulli observations for the new sample does not include the n Bernoulli observations from the preliminary sample, and all Bernoulli observations are assumed to be independently sampled with common mean, p.

Conditional on y, the conditional coverage probability using a new sample is the weighted average of $1(\mid x/N_y - p \mid < m)$, where the weights are the Binomial probabilities evaluated at x for a sample of size N_y and the given probability p. The indicator function $1(A)$ is defined to be one if the event A is true and zero otherwise. Therefore, conditional on y, the conditional coverage probability is

$$\sum_{x=0}^{N_y} 1\left(\left| \frac{x}{N_y} - p \right| < m \right) \binom{N_y}{x} p^x\,(1-p)^{N_y-x}\,, \tag{2.4}$$

The values of y, the number of successes in the preliminary sample, are weighted according to the Binomial probabilities evaluated at y for a preliminary sample of size n and the given probability p. Thus, for given values of m, p, z, and preliminary sample size n, the unconditional coverage probability is

$$\sum_{y=0}^{n} \left[\sum_{x=0}^{N_y} 1\left(\left| \frac{x}{N_y} - p \right| < m \right) \binom{N_y}{x} p^x\,(1-p)^{N_y-x} \right] \binom{n}{y} p^y\,(1-p)^{n-y}\,, \tag{2.5}$$

where the values of N_y are defined by Equation 2.3.

Thus, the unconditional mean of the required sample size N_y is

$$\begin{aligned} E(N_y) &= \sum_{y=0}^{n} N_y \binom{n}{y} p^y\,(1-p)^{n-y} \\ &= \sum_{y=0}^{n} \max\left\{ \left\lceil \frac{y}{n}\left(1-\frac{y}{n}\right)\left(\frac{z}{m}\right)^2 \right\rceil \,,\, 1 \right\} \binom{n}{y} p^y\,(1-p)^{n-y}\,, \end{aligned} \tag{2.6}$$

and the unconditional second population moment of N_y is

$$\begin{aligned} E(N_y^2) &= \sum_{y=0}^{n} N_y^2 \binom{n}{y} p^y\,(1-p)^{n-y} \\ &= \sum_{y=0}^{n} \max\left\{ \left\lceil \frac{y}{n}\left(1-\frac{y}{n}\right)\left(\frac{z}{m}\right)^2 \right\rceil^2 \,,\, 1 \right\} \binom{n}{y} p^y\,(1-p)^{n-y}\,, \end{aligned} \tag{2.7}$$

based on Equation 2.3. Therefore, the unconditional standard deviation of N_y is

$$\sigma_{N_y} = \sqrt{E(N_y^2) \; - \; [E(N_y)]^2} \; . \tag{2.8}$$

The R-code used to produce the coverage probabilities as defined by Equation 2.5, along with the unconditional mean and standard deviation of the required sample size N_y as defined by Equations 2.6 and 2.8, is shown below. This R-code, therefore, produced all of the results in the tables below, and the required sample size is abbreviated as N. Hence, these results are based on exact calculation, not simulation.

```
coverage <- function( n=100, p=0.5, m=0.01, nom.prob=0.95 ) {
# INPUT
# 'n' is the preliminary sample size.
# 'p' is the true probability of success.
# 'm' is the desired margin of error.
# 'nom.prob' is the nominal probability.
# OUTPUT
# 'coverage.prob' is the true coverage probability.
# 'mean.N' is the average required sample size.
# 'sd.N' is the standard deviation of the required sample size.
z <- qnorm( (nom.prob+1)/2 )
coverage.prob <- 0 ; mean.N <- 0 ; mean.N.squared <- 0
for (y in 0:n) {
    N <- max( ceiling( y*(n-y)*(z/n/m)^2 ), 1 ) ; x <- 0:N
    coverage.prob <- coverage.prob + sum( ( abs(x/N-p) <= m ) *
      dbinom( x, N, p) ) * dbinom(y,n,p)
    mean.N <- mean.N + N * dbinom(y,n,p)
    mean.N.squared <- mean.N.squared + N^2 * dbinom(y,n,p) }
sd.N <- sqrt( mean.N.squared - mean.N^2)
return( list( coverage.prob=coverage.prob, mean.N=mean.N, sd.N=sd.N ) ) }
```

Values selected for margin of error are $m = 0.01$, $m = 0.02$, and $m = 0.03$, corresponding to Tables 1, 2, and 3, respectively. Nominal probabilities are set to 90%, 95%, and 99%. Preliminary sample sizes are $n = 25$, 50, 75, and 100. The population proportion was set to $p = 0.05$, 0.1, 0.2, 0.3, 0.4, and 0.5, noting that selecting values of p above 0.5 would be redundant due to symmetry.

3 Results and Discussion

Every coverage probability in Tables 1, 2, and 3 fails to achieve the nominal probability of 90%, 95%, or 99%, although in many cases the difference between the coverage and nominal probabilities is evident only in the third significant digit. When $\min\{np, \; n(1-p)\}$, the mean number preliminary Bernoulli success or failures, is small, the coverage probability can be substantially lower than the nominal probability, as shown in yellow in the tables. This difference in probabilities is only slightly affected by the margin of error being $m = 0.01$, 0.02, or 0.03. Therefore, Tables 1, 2, and 3 produce fairly similar results. For large values of $\min\{np, \; n(1-p)\}$, the coverage probability tends to be extremely close to the nominal probability, as shown in pink.

The mean of the N_y, the required sample size, is larger for the values of p closest to 0.5, as anticipated, and the preliminary sample sizes n have little impact on the mean of N_y. However, the standard deviation of N_y decreases for the larger values of n, as anticipated. The larger values of the margin of error obviously produce smaller values of the mean of N_y.

4 Conclusions

A preliminary sample of independent Bernoulli random variables may be taken to determine the required sample size for estimating the Bernoulli mean, p, within a given margin of error for a fixed level of confidence. When the mean number of successes or failures of this Bernoulli random variable in the preliminary sample is no more than 5, then the coverage probabilities tend to be *severely* smaller than their nominal levels, leading to severely anti-conservative confidence intervals. When the mean number of success or failures is between 5 and 15, the coverage probabilities tend to be only *somewhat* smaller. When the mean number of success or failures is larger than 15, the coverage probabilities tend to be extremely close and only *slightly* smaller than their nominal levels. Therefore, in the preliminary sample, taking a preliminary sample size to ensure that np and $n(1-p)$ both exceed 15 is crucial when trusting the commonly-used equation 2.3 for determining the required sample size. These conclusions herein are limited to situations where equation 2.3 is used to determine the required sample size. For example, future research might involve replicating these studies for finite population sizes, in which case equation 2.3 would be replaced by an equation which includes a finite population correction [14]. The mean and standard deviation of the required sample sizes N_y are listed in the table mainly for perspective of the reader, but do show that they are only slightly impacted by the preliminary sample size n for a given margin of error m and nominal probability.

References

[1] NAVIDI, W., MONK, B. (2022). *Essential Statistics,* 3rd edition. New York: *McGraw Hill,* pp. 334–335.

[2] HOGG, R., TANIS, E., ZIMMERMAN, D. (2023). *Probability and Statistical Inference,* 10th edition. *Pearson,* section 7.4.

[3] SULLIVAN, III, M. (2022). *Fundamentals of Statistics: Informed Decisions Using Data,* 6th edition. *Pearson,* pp. 404–405.

[4] NAING, N. N. (2003). Determination of sample size. *The Malaysian Journal of Medical Sciences,* **10**(2), 84–86. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3561892/

[5] BELL, M. L., TEIXEIRA-PINTO, A., MCKENZIE, J. E., OLIVIER, J. (2014). A myriad of methods: Calculated sample size for two proportions was dependent on the choice of sample size formula and software. *Journal of Clinical Epidemiology,* **67**(5), 601–605. https://doi.org/https://doi.org/10.1016/j.jclinepi.2013.10.008/

[6] BOLARINWA, O. A. (2020). Sample size estimation for health and social science researchers: The principles and considerations for different study designs. *Nigerian Postgraduate Medical Journal,* **27**(2), p. 67. https://doi.org/10.4103/npmj.npmj_19_20/

[7] WANG, H., CHOW, S. (2007). Sample size calculation for comparing proportions. *Wiley Encyclopedia of Clinical Trials,* pp. 1–11. https://doi.org/10.1002/9780471462422.eoct005/

[8] ZOU, G. Y. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in Medicine,* **31**(29), 3972–3981. https://doi.org/10.1002/sim.5466/

[9] STALLARD, N., MILLER, F., DAY, S., HEE, S. W., MADAN, J., ZOHAR, S., POSCH, M. (2016). Determination of the optimal sample size for a clinical trial accounting for the population size. *Biometrical Journal,* **59**(4), 609–625. https://doi.org/10.1002/bimj.201500228/

[10] DAS, S., MITRA, K., MANDAL, M. (2016). Sample size calculation: Basic principles. *Indian Journal of Anaesthesia,* **60**(9), p. 652. https://doi.org/10.4103/0019-5049.190621/

[11] ZAMANZADE, E., WANG, X. (2017). Estimation of population proportion for judgment post-stratification. *Computational Statistics and Data Analysis,* **112**, 257–269. https://doi.org/10.1016/j.csda.2017.03.016/

[12] NAING, L., NORDIN, R. B., RAHMAN, H. A., NAING, Y. T. (2022). Sample size calculation for prevalence studies using Scalex and ScalaR calculators. *BMC Medical Research Methodology,* **22**(209). https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01694-7/

[13] MCCLELLAND, M. L., SESOKO, C. S., MACDONALD, D. A., DAVIS, L. M., MCCLELLAND, S. C. (2021). The immediate physiological effects of e-cigarette use and exposure to secondhand e-cigarette vapor. *Respiratory Care,* **66**(6), 943–950. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10027181/

[14] SCHEAFFER, R. L., MENDENHALL, III, W., OTT, R. L., GEROW, K. G. (2012). *Elementary Survey Sampling,* 7th edition. *Cengage,* section 4.5.

Table 1: Coverage Probabilities for Margin of Error of 1%

$m = 0.01$		Nominal Probab. is 0.9			Nominal Probab. is 0.95			Nominal Probab. is 0.99		
p	n	Coverage Probab.	Mean of N_y	SD of N_y	Coverage Probab.	Mean of N_y	SD of N_y	Coverage Probab.	Mean of N_y	SD of N_y
0.05	25	0.6629	1234	1021	0.6884	1753	1450	0.7148	3026	2505
0.05	50	0.7991	1260	736	0.8494	1789	1046	0.8965	3089	1806
0.05	75	0.8472	1268	605	0.8973	1801	859	0.9498	3110	1484
0.05	100	0.8662	1273	526	0.9157	1807	746	0.9653	3121	1289
0.1	25	0.8114	2338	1254	0.8593	3320	1780	0.9032	5733	3075
0.1	50	0.8670	2387	903	0.9201	3389	1281	0.9687	5853	2213
0.1	75	0.8805	2403	741	0.9332	3412	1052	0.9792	5892	1818
0.1	100	0.8847	2411	644	0.9372	3423	914	0.9828	5912	1579
0.2	25	0.8729	4156	1269	0.9254	5901	1803	0.9736	10192	3113
0.2	50	0.8884	4243	908	0.9403	6024	1289	0.9847	10404	2227
0.2	75	0.8925	4272	744	0.9440	6065	1057	0.9868	10475	1825
0.2	100	0.8944	4286	646	0.9457	6085	917	0.9878	10510	1584
0.3	25	0.8865	5455	1003	0.9386	7745	1424	0.9842	13376	2459
0.3	50	0.8943	5569	706	0.9450	7906	1002	0.9877	13655	1730
0.3	75	0.8962	5606	575	0.9469	7960	816	0.9886	13748	1410
0.3	100	0.8977	5625	498	0.9479	7987	706	0.9890	13794	1220
0.4	25	0.8907	6234	623	0.9437	8851	885	0.9873	15287	1529
0.4	50	0.8954	6364	410	0.9471	9036	582	0.9889	15606	1005
0.4	75	0.8973	6407	326	0.9482	9097	462	0.9893	15712	798
0.4	100	0.8979	6429	278	0.9485	9128	395	0.9895	15765	682
0.5	25	0.8923	6494	375	0.9449	9220	532	0.9881	15924	919
0.5	50	0.8963	6629	189	0.9476	9412	269	0.9892	16256	465
0.5	75	0.8982	6674	127	0.9488	9476	180	0.9895	16367	311
0.5	100	0.8981	6697	95	0.9487	9508	135	0.9896	16422	233

Coverage probabilities smaller than 86%, 92%, and 97% regarding nominal probabilities of 90%, 95%, and 99%, respectively, are in yellow.

Coverage probabilities within 0.5% of the nominal probabilities are in pink.

Table 2: Coverage Probabilities for Margin of Error of 2%

$m = 0.02$		Nominal Probab. is 0.9			Nominal Probab. is 0.95			Nominal Probab. is 0.99		
p	n	Coverage Probab.	Mean of N_y	SD of N_y	Coverage Probab.	Mean of N_y	SD of N_y	Coverage Probab.	Mean of N_y	SD of N_y
0.05	25	0.6702	309	255	0.6850	438	362	0.7141	757	626
0.05	50	0.8229	315	184	0.8518	448	261	0.8961	773	451
0.05	75	0.8550	317	151	0.8950	450	215	0.9506	778	371
0.05	100	0.8716	318	131	0.9154	452	187	0.9665	780	322
0.1	25	0.8153	585	314	0.8574	830	445	0.9045	1433	769
0.1	50	0.8693	597	226	0.9184	848	320	0.9697	1463	553
0.1	75	0.8828	601	185	0.9320	853	263	0.9790	1473	454
0.1	100	0.8842	603	161	0.9369	856	229	0.9830	1478	395
0.2	25	0.8769	1040	317	0.9268	1476	451	0.9739	2548	778
0.2	50	0.8912	1061	227	0.9404	1506	322	0.9847	2601	557
0.2	75	0.8936	1068	186	0.9439	1517	264	0.9867	2619	456
0.2	100	0.8968	1072	161	0.9456	1522	229	0.9878	2628	396
0.3	25	0.8865	1364	251	0.9384	1937	356	0.9843	3344	615
0.3	50	0.8933	1392	176	0.9448	1977	250	0.9878	3414	433
0.3	75	0.8961	1402	144	0.9468	1990	204	0.9887	3437	353
0.3	100	0.8967	1407	124	0.9474	1997	177	0.9890	3449	305
0.4	25	0.8916	1559	156	0.9438	2213	221	0.9874	3822	382
0.4	50	0.8968	1591	103	0.9474	2259	146	0.9890	3902	251
0.4	75	0.8983	1602	81	0.9483	2275	116	0.9894	3928	200
0.4	100	0.8985	1608	70	0.9488	2282	99	0.9895	3942	170
0.5	25	0.8921	1624	94	0.9432	2305	133	0.9882	3982	230
0.5	50	0.8965	1658	47	0.9463	2353	67	0.9892	4064	116
0.5	75	0.8984	1669	32	0.9478	2369	45	0.9895	4092	78
0.5	100	0.8993	1675	24	0.9482	2377	34	0.9896	4106	58

Coverage probabilities smaller than 86%, 92%, and 97% regarding nominal probabilities of 90%, 95%, and 99%, respectively, are in yellow.

Coverage probabilities within 0.5% of the nominal probabilities are in pink.

Table 3: Coverage Probabilities for Margin of Error of 3%

$m = 0.03$		Nominal Probab. is 0.9			Nominal Probab. is 0.95			Nominal Probab. is 0.99		
p	n	Coverage Probab.	Mean of N_y	SD of N_y	Coverage Probab.	Mean of N_y	SD of N_y	Coverage Probab.	Mean of N_y	SD of N_y
0.05	25	0.6628	138	113	0.6936	195	161	0.7147	337	278
0.05	50	0.7890	140	82	0.8562	199	116	0.8996	344	201
0.05	75	0.8665	142	67	0.8892	200	96	0.9513	346	165
0.05	100	0.8740	142	58	0.9269	201	83	0.9669	347	143
0.1	25	0.8072	260	139	0.8552	369	198	0.9026	637	341
0.1	50	0.8721	266	100	0.9216	377	142	0.9674	651	246
0.1	75	0.8814	268	82	0.9320	380	117	0.9784	655	202
0.1	100	0.8910	268	72	0.9388	381	102	0.9824	657	175
0.2	25	0.8770	462	141	0.9272	656	200	0.9734	1133	346
0.2	50	0.8916	472	101	0.9406	670	143	0.9848	1157	248
0.2	75	0.8958	475	83	0.9446	674	117	0.9869	1164	203
0.2	100	0.8964	477	72	0.9460	676	102	0.9880	1168	176
0.3	25	0.8883	607	111	0.9400	861	158	0.9844	1487	273
0.3	50	0.8951	619	78	0.9462	879	111	0.9878	1518	192
0.3	75	0.8972	623	64	0.9473	885	91	0.9885	1528	157
0.3	100	0.8978	626	55	0.9483	888	78	0.9889	1533	135
0.4	25	0.8895	693	69	0.9431	984	98	0.9875	1699	170
0.4	50	0.8947	708	45	0.9469	1004	65	0.9889	1734	112
0.4	75	0.8969	712	36	0.9482	1011	51	0.9894	1746	89
0.4	100	0.8970	715	31	0.9482	1015	44	0.9896	1752	76
0.5	25	0.8938	722	42	0.9439	1025	59	0.9883	1770	102
0.5	50	0.8958	737	21	0.9479	1046	30	0.9891	1807	52
0.5	75	0.8971	742	14	0.9477	1053	20	0.9895	1819	34
0.5	100	0.8964	745	11	0.9485	1057	15	0.9895	1825	26

Coverage probabilities smaller than 86%, 92%, and 97% regarding nominal probabilities of 90%, 95%, and 99%, respectively, are in yellow.

Coverage probabilities within 0.5% of the nominal probabilities are in pink.