The effect of misclassification on sample size: one-sample proportion test

PharmD Péter Hársfalvi[1] and Jenő Reiczigel[1]

[1]University of Veterinary Medicine Budapest

June 20, 2023

Abstract

Potential misclassification of a binary outcome measure is often ignored in study design, causing considerable loss of power, and threatening the quality of research. Although there exist studies taking misclassification into account in data analysis, we argue that it should be accounted for already in sample size calculation. We illustrate this by comparing sample sizes needed with and without misclassification in case of the binomial test. Our sample size procedure, implemented as an R function, calculates exact power, and accounts for non-monotonicity of power as a function of sample size, and for potential drop-out or lack of data in the study. The necessary sample size is computed from the null proportion p_0, the assumed true proportion p_a, and the probabilities of correct classification, sensitivity (*Se)* and specificity (*Sp)*. Our results show that misclassification may drastically affect the necessary sample size. For $p_0<0.5$, the effect of specificity is stronger than that of sensitivity, whereas for $p_0>.5$ it is the other way round. Effects are strongest when p_0 is near 0 or 1, especially for one-sided tests with p_a located farther from 0.5 than the null value p_0. For example, even with $Se = Sp = 99\%$, $p_0 = 0.01$, and left-sided alternative, sample size is more than fourfold of that without misclassification (3-fold if $p_0=0.02$; 1.4-fold if $p_0=0.05$).

Title:

The effect of misclassification on sample size: one-sample proportion test

Author information:

Corresponding author: Péter Hársfalvi, PharmD

Affiliation(s): 1. University of Veterinary Medicine Budapest, Department of Biostatistics, Budapest, Hungary 2. BiTrial Clinical Research, Budapest, Hungary**E-mail address:***harsfalvipeter@gmail.com***ORCID no.:** 0000-0002-6048-5017**Author:** Prof. Dr. Jenő Reiczigel**Affiliation(s):** 1. University of Veterinary Medicine Budapest, Department of Biostatistics, Budapest, Hungary**E-mail address:***Reiczigel.Jeno@univet.hu***ORCID no.:** 0000-0003-4232-6386

Abstract

Potential misclassification of a binary outcome measure is often ignored in study design, causing considerable loss of power, and threatening the quality of research. Although there exist studies taking misclassification into account in data analysis, we argue that it should be accounted for already in sample size calculation. We illustrate this by comparing sample sizes needed with and without misclassification in case of the binomial test. Our sample size procedure, implemented as an R function, calculates exact power, and accounts for non-monotonicity of power as a function of sample size, and for potential drop-out or lack of data in the study. The necessary sample size is computed from the null proportion p_0, the assumed true proportionp_a , and the probabilities of correct classification, sensitivity (*Se)* and specificity (*Sp)* . Our results show that misclassification may drastically affect the necessary sample size. For $p_0<0.5$, the effect of specificity is stronger than that of sensitivity, whereas for$p_0>.5$ it is the other way round. Effects are strongest when p

$_0$ is near 0 or 1, especially for one-sided tests with p_a located farther from 0.5 than the null value p_0. For example, even with $Se = Sp = 99\%$,$p_0 = 0.01$, and left-sided alternative, sample size is more than fourfold of that without misclassification (3-fold ifp_0=0.02; 1.4-fold ifp_0=0.05).

Keywords

binomial test, sample size, misclassification, diagnostic test, power

Highlights

- Potential misclassification of a binary outcome must be considered in study design
- Power of binomial test is non-monotonic: larger samples may lead to smaller power
- Misclassification rates of 1-2% may imply 3-4-fold increase in sample size
- Increase of sample size is greatest if probability of outcome is near 0 or 1

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics statement – Not applicable.

Acknowledgements

The second author was partly supported by the National Research Development and Innovation Fund of Hungary (K143622).

MSC Classification Code :

62P10

1. Introduction

One-sample inference for binary data is one of the most common task in epidemiology and medical statistics (Bland, 2015). One-sided testing is used more often than two-sided one, among others when assessing freedom from disease or approving diagnostic tests, and in industrial quality control, evaluation of medical devices, and in clinical trials of rare diseases (Cameron & Baldock, 1998; Cheng & Zhen, 2021; Feld et al., 2015; Khan, Sarker, & Hackshaw, 2012; Lu, Li, & Xu, 2020). The left-sided alternative (H_0: $p = p_0$ against$H_a : p < p_0$) is applied for example in proving freedom from disease, while the right-sided one (H_0: $p = p_0$ against$H_a : p > p_0$) is used if one wants to prove that a particular exposure increases the probability of getting a disease. There exist several different tests, exact as well as asymptotic, for all testing scenarios. Sample size calculation is available for each method, it needs the prescribed alpha and power, the null proportion p_0, and the assumed true proportion p_a for which the prescribed power should be reached (Chow, Shao, & Wang, 2008; Suresh & Chandrashekara, 2012).

In many cases, the outcome may be wrongly classified. When the outcome is disease status and a diagnostic test is applied, the two usual measures of test quality are sensitivity, the proportion of correct diagnosis given the subject has the disease, and specificity, the proportion of correct diagnosis given the subject does not have the disease (Yerushalmy, 1947). Usually, a diagnostic test has less than 100% sensitivity and specificity which must be accounted for in both the design and analysis of a study. There are analysis methods accounting for misclassification (Lang & Reiczigel, 2014; Reiczigel, Földi, & Ózsvári, 2010; Hársfalvi & Reiczigel, 2023) but the sample size needed for the same power is higher than it would be without misclassification. Thus, ignoring the possibility of misclassification in the sample size calculation may result in an underpowered, inconclusive study, causing considerable financial loss and raising ethical concerns.

Most books on sample size calculation do not mention misclassification at all. Others have a short note declaring this as a problem advised to account for but none of them have clear instructions for researchers. (Chow et al., 2008; Julious, 2009; Kieser, 2020; Ryan, 2013). Intuition may suggest that if probability of misclassification is as low as a few percent in both directions, the increase in sample size is ignorable, but this is not true. To show this, we develop sample size calculation for the one-sample proportion test under

misclassification and investigate how the necessary sample size depends on the sensitivity, specificity, and effect size.

2. Methods

In this study we focus on the exact binomial test (a.k.a. the Clopper and Pearson test) (Chow et al., 2008; Clopper & Pearson, 1934). For a short description of the test, let X denote a variable from a binomial $(n,\ p\ )$ distribution and x its observed value. Let$n$ be the sample size and$b_{n,p}\ (x\ ) = P\ (X\ = x\ )$ and$B_{n,p}\ (x\ ) = P\ (X\ < x\ )$ denote the probability mass function and the cumulative distribution function of the binomial distribution with parameters p and$n$. Assume that based on the observation of x we test for$H_{\ 0}$: $p\ = p_{\ 0}$ against$H_a : p\ $ [?] $p_{\ 0}$(two-tailed test) or$H_a : p\ < p_{\ 0}$or$H_a : p\ > p_{\ 0}$(one-tailed tests). The alpha level critical region of the test, calculated by inverting the Clopper-Pearson exact confidence interval, is as follows:

$\{x\ :\ \sum_{i=0}^{x} b_{n,p}\,(i)\ \leq \alpha\ \}$*.(left-tailed test)*$\{x\ :\ \sum_{i=x}^{n} b_{n,p}\,(i)\ \leq \alpha\ \}$*(right-tailed test)*$\{x:\ \sum_{i=0}^{x} b_{n,p}\,(i)\ \leq \alpha/2\ \} \cup \{x : \sum_{i=x}^{n} b_{n,p}\,(i)\ \leq \alpha/2\}$ *(two-tailed test)*

If the outcome is subject to misclassification with known sensitivity and specificity, the so-called Rogan and Gladen formula can be applied to calculate the true proportion from the observed one (Rogan & Gladen, 1978). The formula for this adjustment looks like

$p_{adj} = (p_{obs}\ +\ Sp - 1)\ /\ (\ Se\ +\ Sp\ -\ 1)$

where Se and Sp denote the sensitivity and specificity of the diagnostic test and p_{adj} andp_{obs} denote the adjusted and observed proportions.

Reiczigel et al (Reiczigel et al., 2010) showed that applying the Rogan & Gladen formula to the endpoints of a confidence interval for the sample proportion results in a valid confidence interval for the true proportion. Furthermore, the adjustment preserves exactness of the CI. These properties of the CIs have similar implications on testing.

We carried out the investigations setting the alpha error rate to 5% and power to 80%. For some selected null proportions$p_{\ 0}$ in $H_{\ 0}$ and assumed true proportions p_a (Table 1) we determined the necessary sample size n by exact power calculation. For each$n$ we calculated the power so that we determined the alpha-level critical region C of the test and calculated the probability of$C$ assuming a binomial distribution with$p = p_a$.

We calculated sample sizes for sensitivity and specificity values 1, 0.99, 0.98, 0.95, 0.90. As we suspected that increase in necessary sample size may differ for the two one-tailed tests (even for the two-tailed test depending on whether p_a is located left or right from $p_{\ 0}$), we investigated each one separately. Thus, we set up two p_a for each $p_{\ 0}$: one left and the other right from$p_{\ 0}$ (see Table 1). These were selected so that the sample size in case of no misclassification takes a few hundreds. We did not include $p_{\ 0}$ values above 0.5 because results for $p_{\ 0}$>0.5 are mirror-images of those for $p_{\ 0}$< 0.5. For example, power of test for $p_{\ 0}$=0.9 withp_a =0.96, Se =0.99, and Sp =.95 is same as that for $p_{\ 0}$=0.1 withp_a =0.04, Se =0.95, and Sp =.99.

It is known that the power of the binomial test does not depend monotonically on sample size but displays a saw-tooth pattern (Chernick & Liu, 2002), thus, it may occur that for some n the power is above 80% but for a greater sample size it falls again under 80%. An example of this is shown in Figure 1.

Hosted file

`image1.emf` available at https://authorea.com/users/630849/articles/650461-the-effect-of-misclassification-on-sample-size-one-sample-proportion-test

Figure 1. Power of the binomial test is not a monotonic function of sample size ($p_{\ 0}$ = 0.5, p_a = 0.4, alternative = "left-sided", Se = Sp = 1)

Unfortunately, the actual sample size of a study, despite the hardest efforts, may differ from the planned one, and non-monotonicity of power invalidates the simplest method of handling this "to play safe, add 10% to

the calculated sample size", which works well for continuous outcomes. Even though the saw-tooth pattern of the power function is well known, it is easy to find clinical trials still using this simple but risky method of handling potential drop-out patients (clinicaltrials.gov, NCT01693614 and NCT02844582). To avoid this trap, some authors recommend choosing the smallest n so that for each$m$ [?] n the power is at least 80% (Chernick & Liu, 2002). However, if it can be ensured that the drop-out rate remains under a certain limit λ , say, under 5%, a smaller sample size than that is sufficient. Therefore, we propose a sample size procedure that searches for the minimal sample size n so that even in case of some drop-out, not exceeding the specified proportion λ , the power reaches the prescribed value. That is, for each sample size$m$ from $(n - \lambda \cdot n)$ to n the power reaches the prescribed value, say 80%.

For the analyses, we prepared an R function that carries out the sample size calculation in case of known sensitivity and specificity of the diagnostic test used for the prevalence estimation. The function can compute sample size for five tests: Clopper-Pearson exact test, Wald-test, Wilson's score test, Agresti-Coull-test, and Blaker's exact test. It has an additional argument to specify the highest proportion of data loss λ (due to drop-out or other reasons), which still must not result in power less than the prescribed value. The function returns the minimal sample size n so that prescribed power is reached for each sample size from $(n - \lambda \cdot n)$ to n . The function is available at GitHub:*https://github.com/Ragnar0ss/*.

We calculated sample sizes assuming a drop-out rate of λ =0.15, that is, power remains at least 80% up to 15% drop-out.

Table 1. Null and assumed true probabilities used in the study. Left- and right-sided tests were evaluated separately.

p_0	.01	.02	.03	.05	.10	.20	.30	.50
p_{aL}	.0005	.001	.003	.01	.04	.12	.20	.40
p_{aR}	.04	.07	.09	.12	.18	.32	.42	.62

3. Results and discussion

We found that even small misclassification probabilities may result in considerable increase of sample size necessary to reach the prescribed power. Table 2 illustrate dependence of sample size onp_0, *Se* , and *Sp* . Full details of the results are given in the supplementary material.

Hosted file

`image2.emf` available at https://authorea.com/users/630849/articles/650461-the-effect-of-misclassification-on-sample-size-one-sample-proportion-test

Figure 2. Dependence of the necessary sample size on sensitivity and specificity for p_0=0.05, with left- and right-sided alternative. Sensitivity is coded by letters (a-100%, b-95%, c-90%, d-85%, e-80%)

For p_0<0.5 specificity has a stronger effect on the sample size. It is strongest if $p_0$is close to 0 and decreases towards 0.5. Effect of specificity is stronger for left-sided than for right-sided alternative (Figure 2). Effect of sensitivity is about the same in the whole range ofp_0, for left- as well as right-sided alternative.

Due to symmetry, for p_0>0.5 it is the other way round, influence of sensitivity is stronger than that of specificity. The latter does not depend much onp_0, nor on the alternative, whereas the former has the most dramatic effects for p_0 near 1, with right-sided alternative.

Se / Sp	**p0 (alternative = left-sided)**	**p0 (alternative = left-sided)**	**p0 (alter**
	0.01	**0.02**	**0.03**
1	352	176	185
0.99	1440	520	312

Se / Sp	**p0 (alternative = left-sided)**	**p0 (alternative = left-sided)**	**p0 (alter**
0.98	2376	762	436
0.95	5437	1526	840
Se / Sp	**p0 (alternative = two-sided with pa < p0)**	**p0 (alternative = two-sided with pa < p0)**	**p0 (alter**
	0.01	**0.02**	**0.03**
1	433	216	217
0.99	1830	623	386
0.98	2971	922	562
0.95	6846	1903	1051

Table 2. - Sample sizes for all alternatives in scenarios with equal sensitivity and specificity

Although the most dramatic effects of misclassification were observed when p_0 was near 0 or 1 (more than fourfold increase in sample size for p_0=0.01 with left-sided alternative and $Se = Sp = 99\%$), even in the best case, that is when p_0=0.5, the increase in necessary sample size was 22% with $Se = Sp = 95\%$, and 9% with $Se = Sp = 98\%$.

For the two-sided alternatives, results differ depending on whether the assumed true proportion p_a is smaller or greater than p_0. Increase of necessary sample size is greater than that for the respective one-sided alternative illustrated by the extreme sample size increase of more than 6400 individuals (from an initial 433) in case of the two-sided alternative withp_a<p_0 whenp_0=0.01 with $Se = Sp = 95\%$.

Our results showed that ignoring even small misclassification probabilities may result in considerable power loss. Here we studied the exact binomial test in detail, but results are similar for other tests for the binomial proportion. Our R function enables sample size calculation for any test given an R function for the test is available. Presumably similar tendencies could be observed in the comparison of two or more binomial samples, which will be investigated later.

Although not presented in the article, we performed the same calculations for several asymptotic methods as well (Agresti-Coull, Wald and Wilson) resulting in similar amount of sample size increase.

4. Conclusion

Potential misclassification must be considered in sample size calculation for the one-sample binomial test. Even if sensitivity and specificity are high (98-99%), necessary sample sizes may be much higher than without misclassification.

5. References

Agresti, A., & Coull, B. A. (1998). Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician, 52* (2), 119-126. doi:10.1080/00031305.1998.10480550

Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics, 28* , 783-798. doi:10.2307/3315916

Cameron, A. R., & Baldock, F. C. (1998). Two-stage sampling in surveys to substantiate freedom from disease. *Preventive Veterinary Medicine, 34* (1), 19-30. doi:10.1016/s0167-5877(97)00073-1

Carlsson, S., Hammar, N., Hakala, P., Kaprio, J., Marniemi, J., & Rönnemaa, T. (2003). Assessment of alcohol consumption by mailed questionnaire in epidemiological studies: evaluation of misclassification using a dietary history interview and biochemical markers. *European journal of epidemiology, 18* (6), 493-501. doi:10.1023/a:1024694816036.

Cheng, C., & Zhen, B. (2021). Binomial sampling plans for validation and quality control in blood product manufacturing. *Transfusion* . doi:10.1111/trf.16436

Chernick, M., & Liu, C. (2002). The Saw-Toothed Behavior of Power Versus Sample Size and Software Solutions. *American Statistician - AMER STATIST, 56* , 149-155. doi:10.1198/000313002317572835

Chow, S.-C., Shao, J., & Wang, H. (2008). *Sample size calculations in clinical research, second edition* (Second Edition ed.): Taylor & Francis Group.

Clopper, C. J., & Pearson, E. S. (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of Binomial. *Biometrika, 26* , 404-413. doi:10.1093/biomet/26.4.404

England, L. Bland, M. (2015). *An introduction to medical statistics* . Oxford university press.

Hársfalvi, P., & Reiczigel, J. (2023). Profile likelihood confidence interval for the prevalence assessed by an imperfect diagnostic test.*Preventive Veterinary Medicine* , *214* , 105886.

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques.*Public Health Reports (1896-1970)* , 1432–1449.

J., Grauman, A., Qian, C., Wilkins, D. G., Schisterman, E. F., Yu, K. F., & Levine, R. J. (2007). Misclassification of maternal smoking status and its effects on an epidemiologic study of pregnancy outcomes.*Nicotine & Tobacco Research, 9* (10), 1005-1013. doi:10.1080/14622200701491255

Feld, J. J., Jacobson, I. M., Hézode, C., Asselah, T., Ruane, P. J., Gruener, N., . . . Chan, H. L. (2015). Sofosbuvir and velpatasvir for HCV genotype 1, 2, 4, 5, and 6 infection. *New England Journal of Medicine, 373* (27), 2599-2607. doi:10.1056/NEJMoa1512610

Julious, S. A. (2009). *Sample Sizes for Clinical Trials* .

Khan, I., Sarker, S.-J., & Hackshaw, A. K. (2012). Smaller sample sizes for phase II trials based on exact tests with actual error rates by trading-off their nominal levels of significance and power.*British journal of cancer, 107* , 1801-1809. doi:10.1038/bjc.2012.444

Kieser, M. (2020). *Methods and applications of sample size calculation and Recalculation in clinical trials* : Springer Nature.

Lang, Z., & Reiczigel, J. (2014). Confidence limits for prevalence of disease adjusted for estimated sensitivity and specificity.*Preventive Veterinary Medicine, 113* , 13–22. doi:10.1016/j.prevetmed.2013.09.015

Lu, N., Li, H., & Xu, Y.-L. (2020). Use of Weighted Performance Goals in Prospective Single-Arm Clinical Studies Designed to Assess the Safety and Effectiveness of Medical Devices. *Statistics in Biopharmaceutical Research* , 1-4. doi:10.1080/19466315.2020.1799853

Reiczigel, J., Földi, J., & Ózsvári, L. (2010). Exact confidence limits for prevalence of a disease with an imperfect diagnostic test.*Epidemiology and infection, 138* , 1674-1678. doi:10.1017/S0950268810000385

Rogan, W., & Gladen, B. (1978). Estimating Prevalence From Results of A Screening-test. *American journal of epidemiology, 107* , 71-76. doi:10.1093/oxfordjournals.aje.a112510

Ryan, T. P. (2013). *Sample size determination and power* : John Wiley & Sons.

Suresh, K., & Chandrashekara, S. (2012). Sample Size estimation and Power analysis for Clinical research studies. *Journal of human reproductive sciences, 5* , 7-13. doi:10.4103/0974-1208.97779

Vollset, S. (1993). Confidence intervals for a binomial proportion.*Statistics in medicine, 12 9* , 809-824. doi:10.1002/sim.4780120902